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

SP&**Applications**, August 2007

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

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

Contents

- **Data-Based** Introduction (Service Engineering, Call Centers)
  - “Production" of Health, Justice, Banking-Services, Tele-Services
  - **Simple Models at the Service of Complex Realities**

Contents

- **Data-Based** Introduction (Service Engineering, Call Centers)
  - “Production" of Health, Justice, Banking-Services, Tele-Services
  - **Simple Models at the Service of Complex Realities**
- The Basic Call-Center Model: **Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**

Contents

- **Data-Based** Introduction (Service Engineering, Call Centers)
  - “Production" of Health, Justice, Banking-Services, Tele-Services
  - **Simple Models at the Service of Complex Realities**
- The Basic Call-Center Model: **Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?**
  Framework - Asymptotic Regimes: **QED, ED, ED+QED**

Contents

- **Data-Based** Introduction (Service Engineering, Call Centers)
  - “Production" of Health, Justice, Banking-Services, Tele-Services
  - **Simple Models at the Service of Complex Realities**
- The Basic Call-Center Model: **Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?**
  Framework - Asymptotic Regimes: **QED, ED, ED+QED**
  - General Patience: M/M/N+**G**
  - General Services: M/**G**/N+G
  - Time-Varying Arrivals: M_t/G/N_t+G
  - Heterogenous Customers: M^J/M/N+M
  - Current Research: M_t^J/G/N_t+G

Contents

- **Data-Based** Introduction (Service Engineering, Call Centers)
  - “Production" of Health, Justice, Banking-Services, Tele-Services
  - **Simple Models at the Service of Complex Realities**
- The Basic Call-Center Model: **Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?**
  Framework - Asymptotic Regimes: **QED, ED, ED+QED**
  - General Patience: M/M/N+**G**
  - General Services: M/**G**/N+G
  - Time-Varying Arrivals: M_t/G/N_t+G
  - Heterogenous Customers: M^J/M/N+M
  - Current Research: M_t^J/G/N_t+G
- Explain **Practice**: “Right Answers for the Wrong Reasons"

Contents

- **Data-Based** Introduction (Service Engineering, Call Centers)
  - "Production" of Health, Justice, Banking-Services, Tele-Services
  - **Simple Models at the Service of Complex Realities**
- The Basic Call-Center Model: **Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?**
  Framework - Asymptotic Regimes: **QED, ED, ED+QED**
  - General Patience: M/M/N+**G**
  - General Services: M/**G**/N+G
  - Time-Varying Arrivals: $\boldsymbol{M_t}$/G/$\boldsymbol{N_t}$+G
  - Heterogenous Customers: $\boldsymbol{M^J}$/M/N+M
  - Current Research: $\boldsymbol{M_t^J}$/G/$\boldsymbol{N_t}$+G
- Explain **Practice**: "Right Answers for the Wrong Reasons"
- The Technion SEE Center / Laboratory: **DataMOCCA**

Main Messages

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

Main Messages

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

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

Main Messages

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

Note: Useful must be Simple; Simple often 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.
Initiated with Wharton, developed at Technion, available for adoption.

Main Messages

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

Note: Useful must be Simple; Simple often 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.
Initiated with Wharton, developed at Technion, available for adoption.

3. Back to the **Basic-Research Paradigm** (Physics, Biology, . . .):
Measure, Model, Experiment, Validate, Refine, etc.

Main Messages

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

Note: Useful must be Simple; Simple often 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.
Initiated with Wharton, developed at Technion, available for adoption.

3. Back to the **Basic-Research Paradigm** (Physics, Biology, . . .):
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 (1915), Palm (1945)**,..., thoughtful managers.

Main Messages

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

Note: Useful must be Simple; Simple often 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.
Initiated with Wharton, developed at Technion, available for adoption.

3. Back to the **Basic-Research Paradigm** (Physics, Biology, . . .):
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 (1915), Palm (1945)**,..., thoughtful managers.

5. Scientifically-based design principles and tools (software),
that support the balance of service **quality**, process **efficiency** and
business **profitability**, from the (often-conflicting) views of
customers, servers, managers: **Service Engineering**.

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.

Present Focus: Call Centers, but Expanding

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

Present Focus: Call Centers, but Expanding

U.S. Statistics (Relevant Elsewhere)

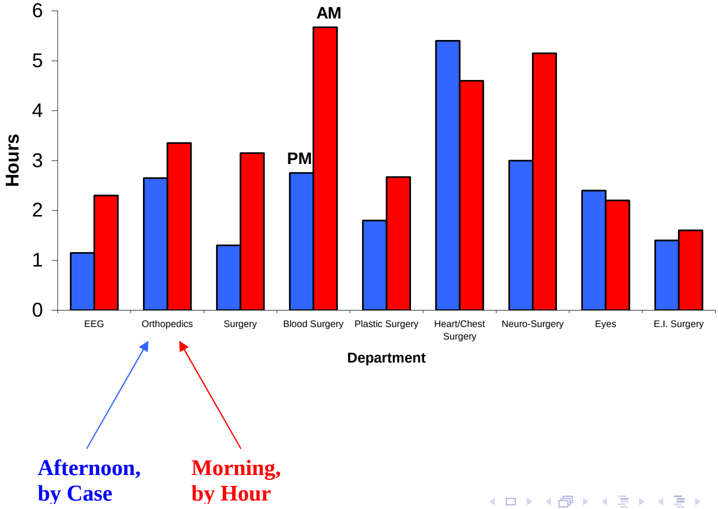
- Over 60% of annual business volume via the telephone
- 100,000 – 200,000 call centers
- 3 – 6 million employees (**2% – 4% workforce**)
- 1000's agents in a "single" call center = 70 % costs.
- 20% annual growth rate
- \$200 – \$300 billion annual expenditures

Expanding, eg. **Healthcare**:

- **Similar** Challenges: Scarce transactional data, natural queueing-network view, human-operations interface (**7% LWBS**), nurse-staffing (over 3 millions), . . .
- **Unique** Challenges: Risk, economies vs. dis-economies of scale, synchronization gaps, . . . ,

The Human Factor, or
Even “Doctors" Can Manage

Operations Time - Morning (by Hour) vs. Afternoon (by Case):



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.

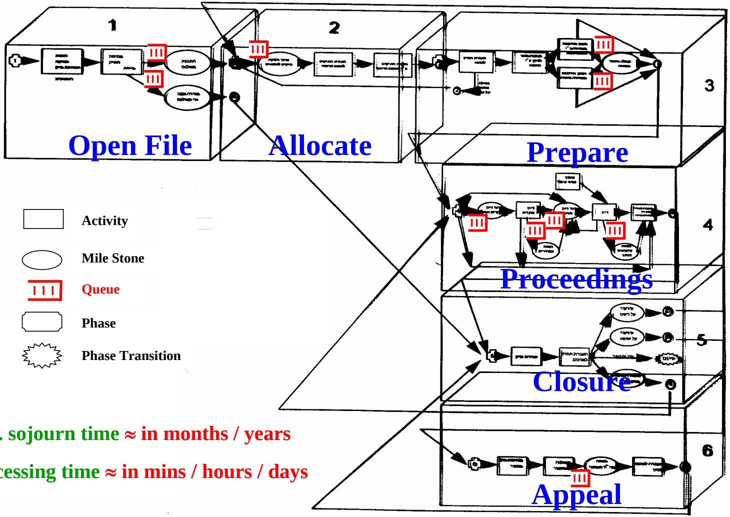
But I need data at the level of the **Individual Transaction**: For each service transaction (during a phone-service in a call center, or a patient's stay in a hospital), its **operational history** = time-stamps of events.

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

“Production of Justice" (Administrative) Network

The Labor-Court Process in Haifa, Israel

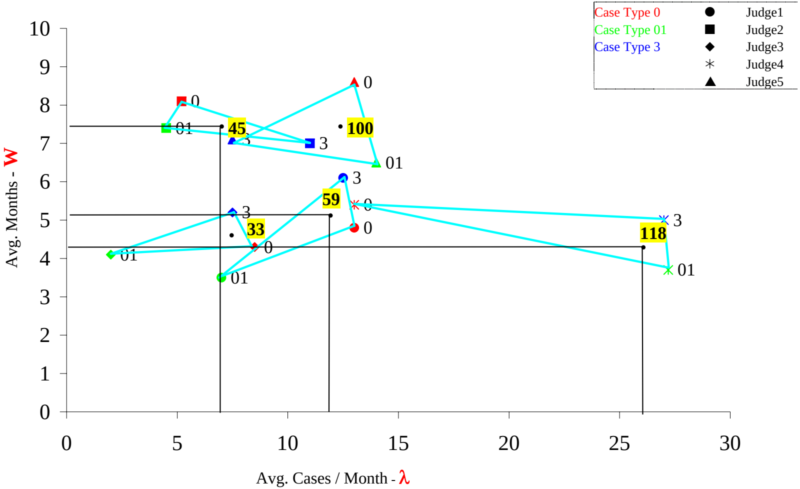


Avg. sojourn time ≈ in months / years

Processing time ≈ in mins / hours / days

Little's Law in Court (Creative Averaging)

Judges: The Best/Worst (Operational) Performer



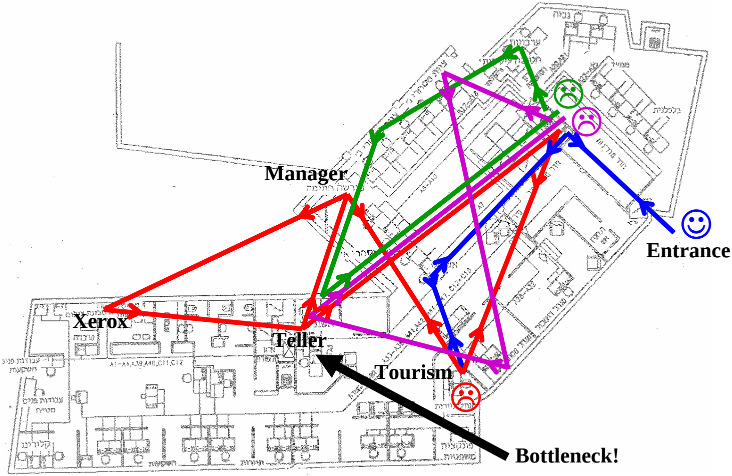
Measurements: Face-to-Face Services

23 Bar-Code Readers at a Bank Branch

Bank – 2nd Floor Measurements

"Face-to-Face Services" Network

Bank Branch = Queueing Network



Transition Probabilities of a Jackson Network

Transition Frequencies Between Units in The Private and Business Sections:

		Private Banking				Business				
	To Unit / From Unit	Bankers	Authorized Personal	Compens - - ations	Tellers	Tellers	Overdrafts	Authorized Personal	Full Service	Exit
Private Banking	Bankers		1%	1%	4%	4%	0%	0%	0%	90%
	Authorized Personal	12%		5%	4%	6%	0%	0%	0%	73%
	Compensations	7%	4%		18%	6%	0%	0%	1%	64%
	Tellers	6%	0%	1%		1%	0%	0%	0%	90%
Services	Tellers	1%	0%	0%	0%		1%	0%	2%	94%
	Overdrafts	2%	0%	1%	1%	19%		5%	8%	64%
	Authorized Personal	2%	1%	0%	1%	11%	5%		11%	69%
	Full Service	1%	0%	0%	0%	8%	1%	2%		88%
	Entrance	13%	0%	3%	10%	58%	2%	0%	14%	0%

Legend:

0%-5%	5%-10%	10%-15%	>15%

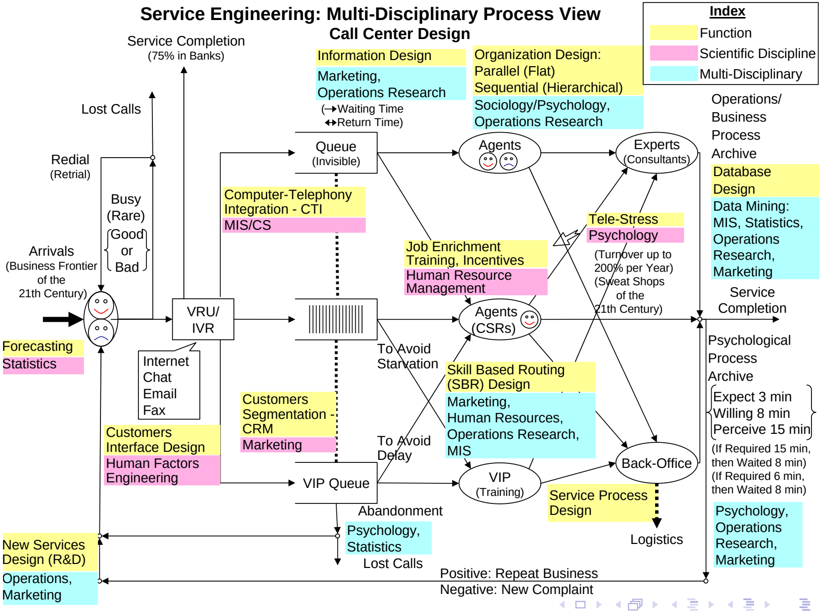
Dominant Paths - Business:

Unit Parameter	Station 1 Tourism	Station 2 Teller	Total Dominant Path
Service Time	12.7	4.8	17.5
Waiting Time	8.2	6.9	15.1
Total Time	20.9	11.7	32.6
Service Index	**0.61**	**0.41**	**0.53**

Call-Center Environment: Service Network

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

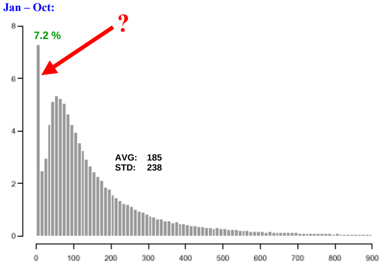
Call-Center Network: Gallery of Models



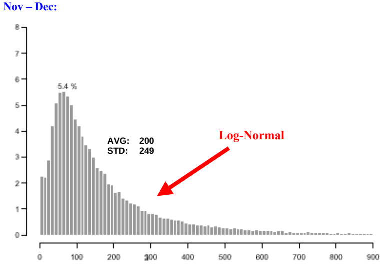
Beyond Averages: Service Times in a Call Center

Histogram of Service Times in an Israeli Call Center

January-October



November-December

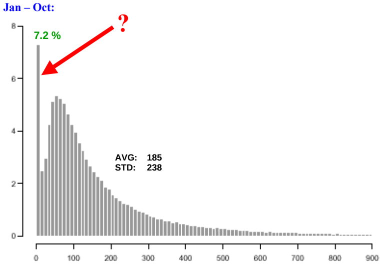


- **7.2% Short-Services:**

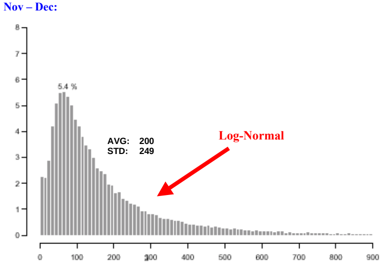
Beyond Averages: Service Times in a Call Center

Histogram of Service Times in an Israeli Call Center

January-October



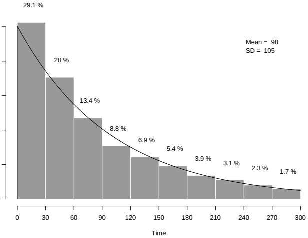
November-December



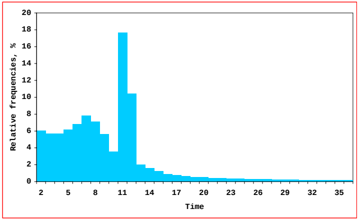
- **7.2% Short-Services:** Agents' "Abandon" (improve bonus, rest)
- **Distributions**, not only Averages, must be measured.
- **Lognormal** service times prevalent in call centers

Beyond Averages: Waiting Times in a Call Center

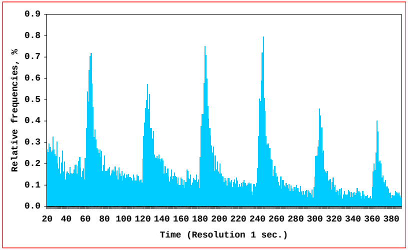
Small Israeli Bank



Large U.S. Bank



Medium Israeli Bank



The “Phases of Waiting" for Service

Common Experience:

- Expected to wait 5 minutes, Required to 10
- Felt like 20, Actually waited 10 (hence Willing $\geq$ 10)

The "Phases of Waiting" for Service

Common Experience:

- Expected to wait 5 minutes, Required to 10
- Felt like 20, Actually waited 10 (hence Willing $\geq$ 10)

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 "Phases of Waiting" for Service

Common Experience:

- Expected to wait 5 minutes, Required to 10
- Felt like 20, Actually waited 10 (hence Willing $\geq$ 10)

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

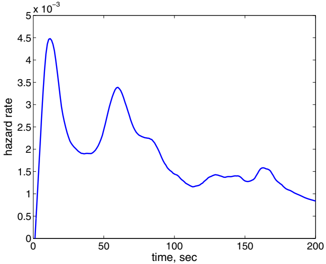
"Rational" customers $\Rightarrow$ Perceived = Actual.

Then left with (τ, V).

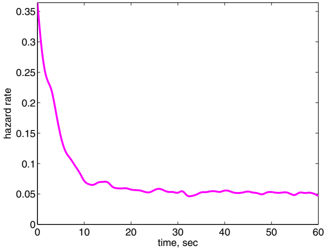
Call Center Data: Hazard Rates (Un-Censored)

(Im)Patience Time τ

Israel



U.S.

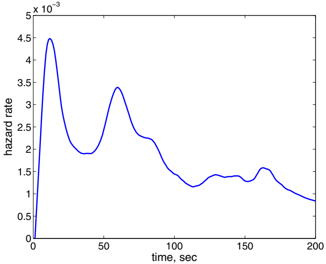


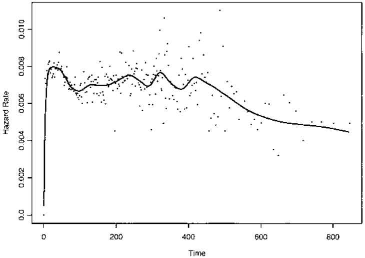
Call Center Data: Hazard Rates (Un-Censored)

(Im)Patience Time τ

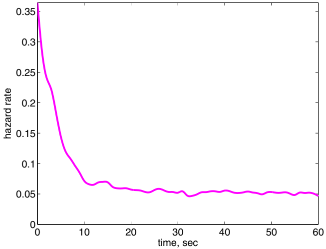
Required/Offered Wait V

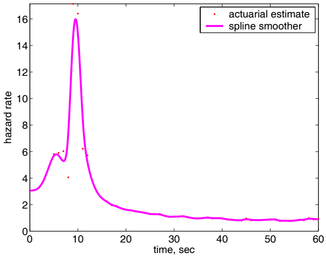
Israel





U.S.

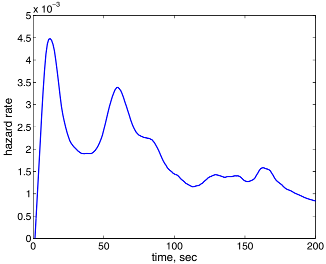


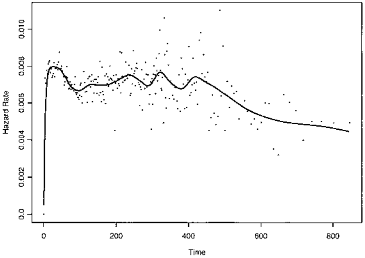


Call Center Data: Hazard Rates (Un-Censored)

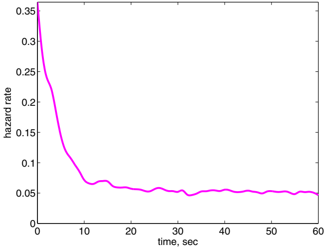
(Im)Patience Time τ — Required/Offered Wait V

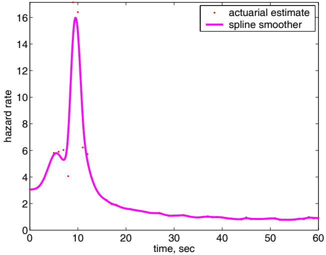
Israel





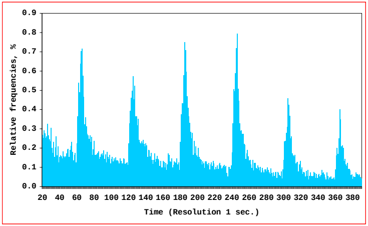
U.S.





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

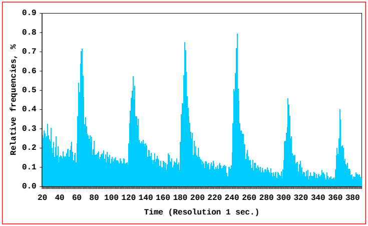
A “Waiting-Times" Puzzle at a Large Israeli Bank



Peaks Every 60 Seconds. Why?

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

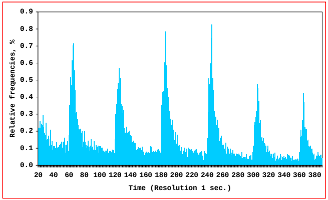
A "Waiting-Times" Puzzle at a Large Israeli Bank

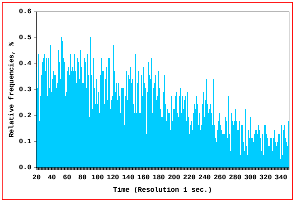


Peaks Every 60 Seconds. Why?

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

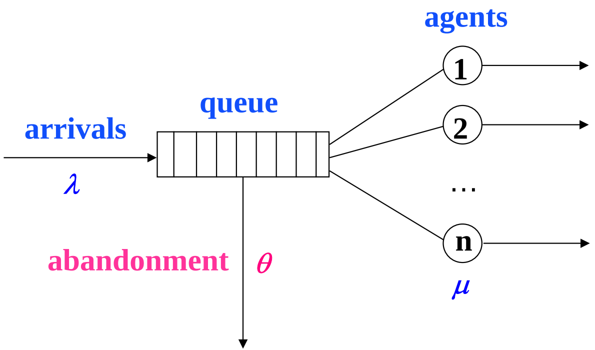
Modeling: τ = **input** to model, V = **output**.

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?** (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?

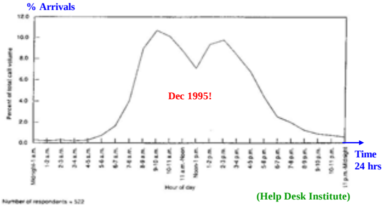
- Primitives independent?
- Customers / Servers Heterogeneous?
- Service discipline FCFS?
- . . . ?

Validation: Support? Refute?

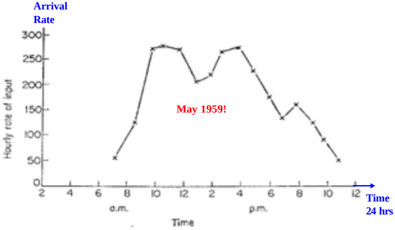
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers

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



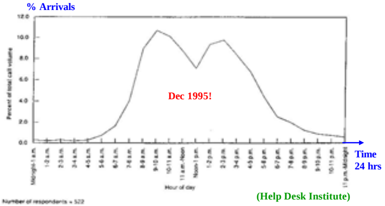
May **1959** (England)



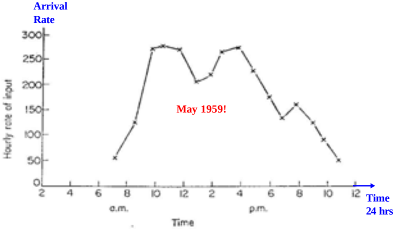
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers

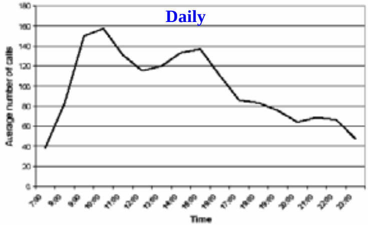
Dec. **1995** (U.S. 700 Helpdesks)



May **1959** (England)



November **1999** (Israel)

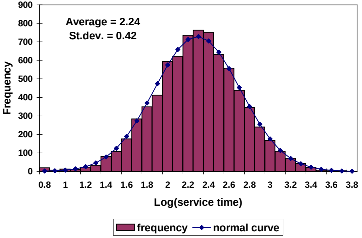


Observation:

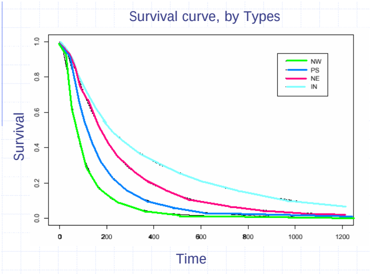
Peak Loads at **10:00 & 15:00**

Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram



Survival-Functions by Service-Class

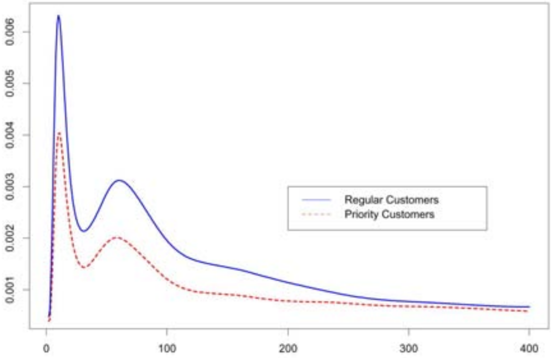


- **New** Customers: 2 min (NW);
- **Regulars**: 3 min (PS);
- **Stock**: 4.5 min (NE);
- Tech-Support: 6.5 min (IN).

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

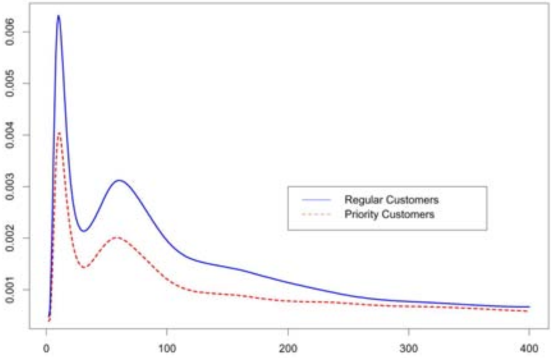
(Im)Patience while Waiting (Palm 1943-53)

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



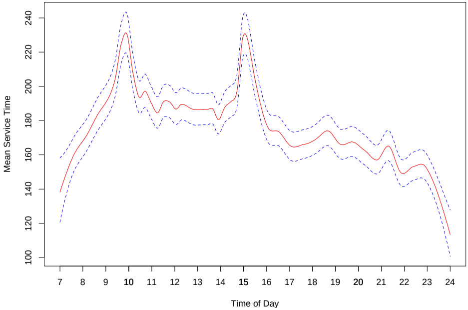
- **Peaks** of abandonment at times of **Announcements**
- **Call-by-Call Data (DataMOCCA)** required (& Un-Censoring).

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

A "Service-Time" Puzzle at a Small Israeli Bank

Inter-related Primitives

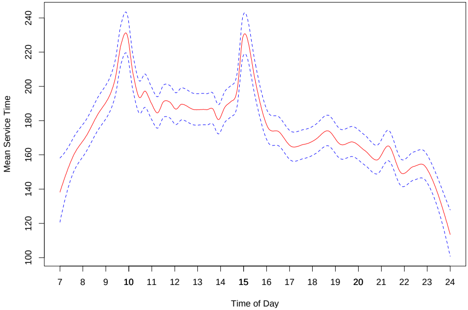
Average Service Time over the Day – Israeli Bank



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

A "Service-Time" Puzzle at a Small Israeli Bank
Inter-related Primitives

Average Service Time over the Day – Israeli Bank



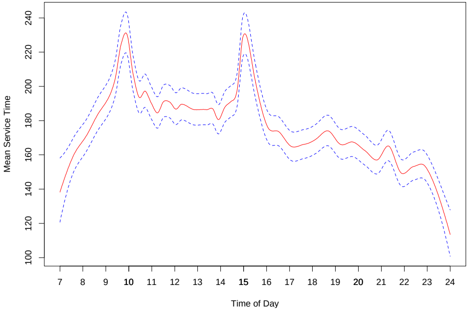
Prevalent: **Longest services** at **peak-loads** (10:00, 15:00). **Why?**

Explanations:

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

A “Service-Time" Puzzle at a Small 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).

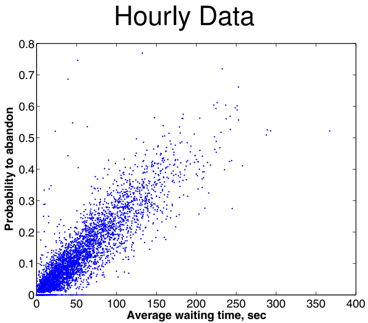
- Building Blocks need **not be independent** (eg. long wait possibly implies long service)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment)
- $\cdots$, and more.

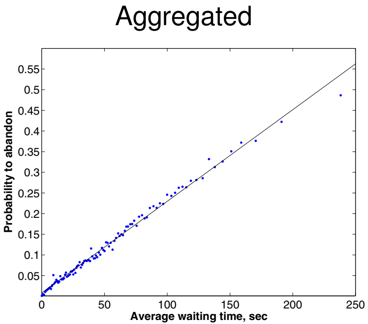
Question: **Is Erlang-A Practically Relevant?**

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

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

Israeli Bank: Yearly Data



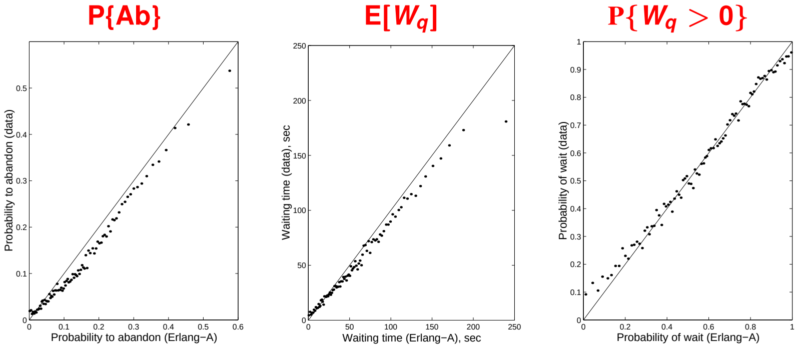


Graphs based on 4158 hour intervals.

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).



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** $R = \frac{\lambda}{\mu}$ $(= \lambda \times \mathrm{E}[S])$ 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: $N \approx R + \delta R$ $\quad [(N - R)/R \to \delta, \text{ as } N, \lambda \uparrow \infty]$

- 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: $N \approx R + \beta\sqrt{R}$

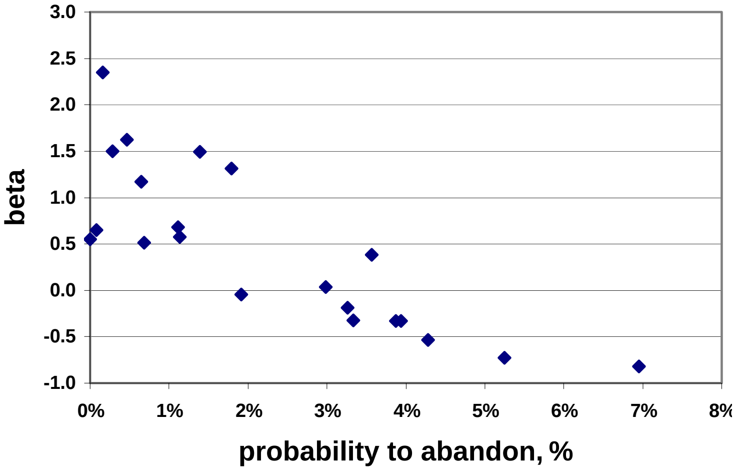
- Erlang 1924, Halfin & Whitt 1981
- %Delayed between 25% and 75%
- Wait one-order below service-time (sec vs. min); 1-5% Abandon.

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 $\quad (g_0 = \theta, \text{ if } \exp(\theta))$.

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

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

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

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

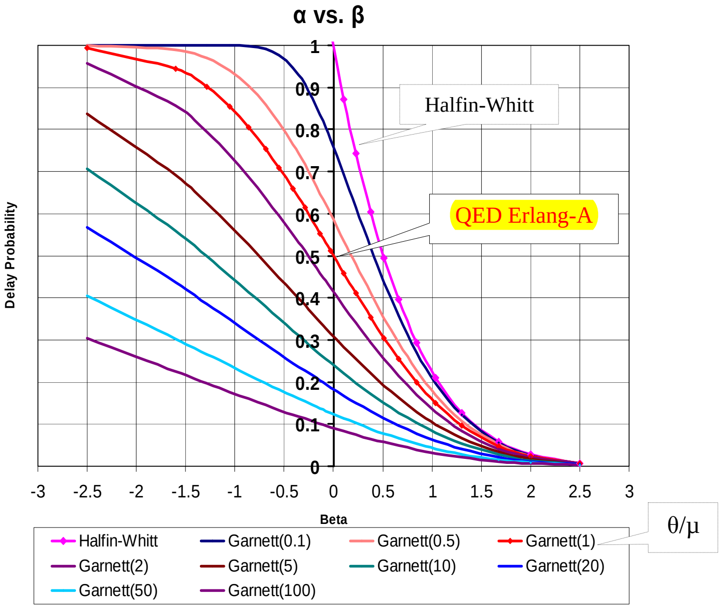
Here

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

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

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

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



Process Limits (Queueing, Waiting)

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

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

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

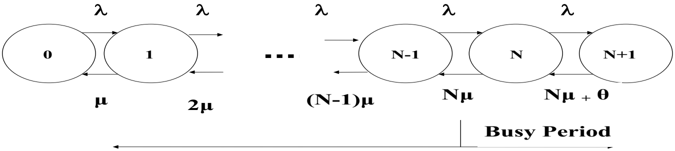
- $\widehat{Q} = \{\widehat{Q}(t), t \geq 0\}$: process defined by: $\widehat{Q}_N(t) \xrightarrow{d} \widehat{Q}(t)$.

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

Approximating (Virtual) Waiting Time

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

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

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 case: $\mu = \theta$:

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

and $P\{Wait > 0\} \approx 1/2$, since $\delta = \beta$.

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**:
   - **Constraint Satisfaction**: eg. Min. n s.t. QOS constraints.
   - **Cost / Profit Optimization**: eg. Min costs of Staffing + Congestion.

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**:
   - **Constraint Satisfaction**: eg. Min. n s.t. QOS constraints.
   - **Cost / Profit Optimization**: eg. Min costs of Staffing + Congestion.
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 $\leq 10\%\mathrm{E}[\tau]$	Loose $\geq 10\%\mathrm{E}[\tau]$	Tight $0 \leq T \leq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\%\mathrm{E}[\tau]$ $5\% \leq \alpha \leq 50\%$
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).

The ED Regime: M/M/n+G

ED – **E**fficiency-**D**riven.

Assume $G(x) = \gamma$ has a unique solution x^* and $g(x^*) > 0$.

Staffing: $\boldsymbol{n = R \cdot (1 - \gamma) + o(\sqrt{R})}$, $0 < \gamma < 1$.

The ED Regime: M/M/n+G

ED – **E**fficiency-**D**riven.

Assume $G(x) = \gamma$ has a unique solution x^* and $g(x^*) > 0$.

Staffing: $\boldsymbol{n = R \cdot (1-\gamma) + o(\sqrt{R})}$, $0 < \gamma < 1$.

Performance Measures

- $\mathrm{P}\{W_q = 0\}$ decreases exponentially in n.
- Probability to abandon converges to:

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

- Offered wait converges to x^*:

$$\mathrm{E}[V] \approx x^*, \qquad V \xrightarrow{p} x^*.$$

- Distribution G^* of $\min(x^*, \tau)$:

$$G^*(x) = \begin{cases} G(x)/\gamma, & x \leq x^* \\ 1, & x > x^* \end{cases}$$

  Asymptotic distribution of wait:

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

ED+QED Regime: Motivation

Min n, s.t.

$\mathrm{P}\{W_q > 0\} \le \alpha$ – use **QED** staffing.

$\mathrm{E}[W_q] \le T$ – use **ED** staffing.

What about $\boldsymbol{\mathrm{P}\{W_q > T\} \le \alpha}$, $T > 0$?
(Most prevalent SL constraint in call centers.)

ED+QED Regime: Motivation

Min n, s.t.

$\mathrm{P}\{W_q > 0\} \leq \alpha$ – use **QED** staffing.

$\mathrm{E}[W_q] \leq T$ – use **ED** staffing.

What about $\boldsymbol{\mathrm{P}\{W_q > T\} \leq \alpha}$, $T > 0$?
(Most prevalent SL constraint in call centers.)

ED approximation:

$$\mathrm{P}\{W_q > T\} \;\approx\; \begin{cases} \bar{G}(T)\,, & T < G^{-1}(\gamma)\,, \\ 0\,, & T > G^{-1}(\gamma)\,. \end{cases}$$

or (as a function of "staffing"):

$$\mathrm{P}\{W_q > T\} \;\approx\; \begin{cases} \bar{G}(T)\,, & \gamma > G(T)\,, \\ 0\,, & \gamma < G(T)\,, \end{cases}$$

Too crude to capture α exactly.

Solution: Refine around $\gamma = G(T)$,

$$\boldsymbol{n \;=\; (1 - G(T)) \cdot R + \beta\sqrt{R}}, \quad -\infty < \beta < \infty\,.$$

ED+QED Performance Measures

Theorem. The following statements are equivalent:

1. Staffing level: $\boldsymbol{n = (1-\gamma)R + \beta\sqrt{R} + o(\sqrt{R})\,;}$

2. Tail probability: $\mathrm{P}\{W_q > T\} = \alpha + o(1)\,;$

3. Probability to abandon:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \gamma - \frac{\beta}{\sqrt{R}} + o\left(\frac{1}{\sqrt{R}}\right)\;;$$

4. Average wait:

$$\mathrm{E}[W_q] = \int_0^T \bar{G}(u)du - \frac{\beta}{\sqrt{R}} \cdot \frac{1}{h_G(T)} + o\left(\frac{1}{\sqrt{R}}\right)\;.$$

ED+QED Performance Measures

Theorem. The following statements are equivalent:

1. Staffing level: $\boldsymbol{n = (1-\gamma)R + \beta\sqrt{R} + o(\sqrt{R})\,;}$

2. Tail probability: $\mathrm{P}\{W_q > T\} = \alpha + o(1)\,;$

3. Probability to abandon:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \gamma - \frac{\beta}{\sqrt{R}} + o\left(\frac{1}{\sqrt{R}}\right)\;;$$

4. Average wait:

$$\mathrm{E}[W_q] = \int_0^T \bar{G}(u)du - \frac{\beta}{\sqrt{R}} \cdot \frac{1}{h_G(T)} + o\left(\frac{1}{\sqrt{R}}\right)\;.$$

Here $0 < \alpha < \bar{G}(T)$, $\gamma = G(T)$,
$h_G(T)$ = patience hazard-rate at T and

$$\beta \;=\; \bar{\Phi}^{-1}\left(\frac{\alpha}{\bar{G}(T)}\right) \cdot \sqrt{\frac{g(T)}{\mu}}\,.$$

Corollary. Approximation for the tail probability:

$$\mathbf{P}\{\boldsymbol{W_q > T}\} \;\sim\; \boldsymbol{\bar{G}(T) \cdot \bar{\Phi}\left(\beta\sqrt{\frac{\mu}{g(T)}}\right)}\;.$$

Note: If $\alpha \geq \bar{G}(T)$ then $\boldsymbol{n = 0}$ satisfies
$\mathrm{P}\{W_q > T\} \leq \alpha$.

M/M/n+G Performance Measures: Building Blocks

$$H(x) \triangleq \int_0^x \bar{G}(u)du \,,$$

where $\bar{\boldsymbol{G}}(\cdot) = \mathbf{1} - \boldsymbol{G}(\cdot)$, the **survival-function** of patience.

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

Finally,

$$\mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1} \frac{1}{j!} \left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!} \left(\frac{\lambda}{\mu}\right)^{n-1}} = \int_0^\infty e^{-t} \left(1 + \frac{t\mu}{\lambda}\right)^{n-1} dt \,.$$

M/M/n+G Performance Measures

P{Ab} = probability to abandon, W_q = waiting time, V = offered wait, Q = queue length.

$$
\begin{aligned}
\mathrm{P}\{V > t\} &= \frac{\lambda J(t)}{\mathcal{E} + \lambda J}\,, \quad (\textbf{Baccelli} \,\&\, \textbf{Hebuterne}, 1981) \\
\mathrm{P}\{W_q > 0\} &= \frac{\lambda J}{\mathcal{E} + \lambda J} \cdot \bar{G}(0)\,, \\
\mathrm{P}\{\mathrm{Ab}\} &= \frac{1 + (\lambda - n\mu)J}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[V] &= \frac{\lambda J_1}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[W_q] &= \frac{\lambda J_H}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[Q] &= \frac{\lambda^2 J_H}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[W_q \mid \mathrm{Ab}] &= \frac{J + \lambda J_H - n\mu J_1}{(\lambda - n\mu)J + 1}\,, \\
\mathrm{P}\{W_q > t\} &= \frac{\lambda \bar{G}(t) J(t)}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[W_q \mid W_q > t] &= \frac{J_H(t) - (H(t) - t\bar{G}(t)) \cdot J(t)}{\bar{G}(t) J(t)}\,, \\
\mathrm{P}\{\mathrm{Ab} \mid W_q > t\} &= \frac{\lambda - n\mu - G(t)}{\lambda \bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda \bar{G}(t) J(t)}\,.
\end{aligned}
$$

M/M/n+G: Laplace Method

Asymptotic calculation of integrals:

1. Show that the integral (mass) is concentrated near a **certain point**.
2. Use **Taylor expansion** to approximate integrand near this point.

Apply to **Building Blocks** and **Performance Measures** above.

Examples:

QED regime: $\boldsymbol{n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}).}$

$$J \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h_G(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right),$$

where

$$\hat{\beta} \;\triangleq\; \beta\sqrt{\frac{\mu}{g_0}}.$$

ED+QED regime: $\boldsymbol{n = \bar{G}(T) \cdot \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}).}$

$$J \;\sim\; \exp\{\lambda H(T) - n\mu T\} \cdot \exp\left\{\frac{\beta^2\mu}{2g(T)}\right\} \cdot \sqrt{\frac{2\pi}{\lambda g(T)}}.$$

Back to "Why does Erlang-A Work?"

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

- **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?"

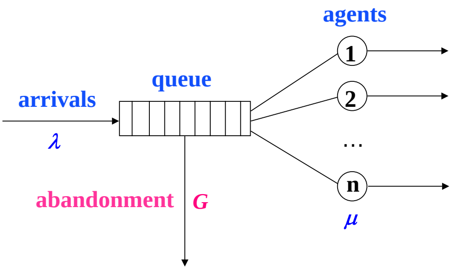
Theoretical Answer: $M_t^J/G/N_t + G \;\approx\; (M/M/N + M)_t, \quad t \geq 0.$

- **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"

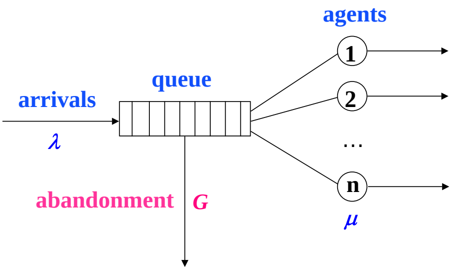
"Why does Erlang-A Work?" General Patience



(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/$\boldsymbol{n}$+G

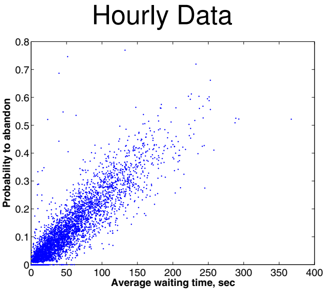
Exact analysis in steady-state (Baccelli & Hebuterne, 1981): solve Kolmogorov's PDE's (semi-Markov) for the offered-wait $\boldsymbol{V}$.

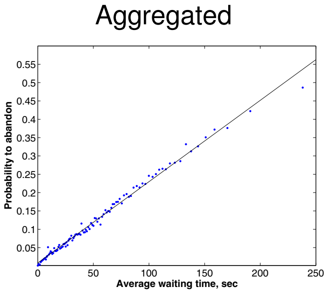
QED analysis (w/ Zeltyn, 2006): $\boldsymbol{n} \approx \boldsymbol{R} + \beta\sqrt{\boldsymbol{R}}$.

- Assume (Im)Patience density $\boldsymbol{g}(\mathbf{0}) > \mathbf{0}$.
- $\boldsymbol{V}$ asymptotics ($\lambda \uparrow \infty$): Laplace Method.
- **QED Approximations: Use Erlang-A**, with $\theta \leftrightarrow g(0)$.

General Patience: Fitting Erlang-A

Israeli Bank: Yearly Data





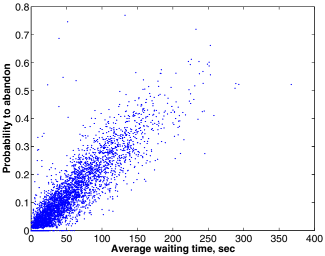
Theory:

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

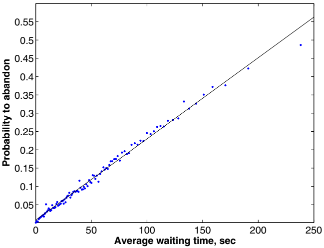
General Patience: Fitting Erlang-A

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].

Recipe:

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

Why Does Erlang-A Work? General Services

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

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

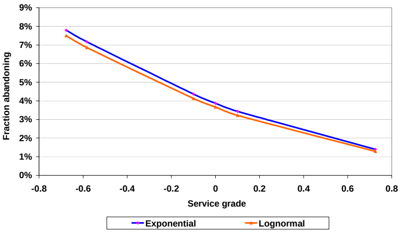
Established: M/M/N+**G** $\approx$ M/M/N+M $(\theta = g(0))$.

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

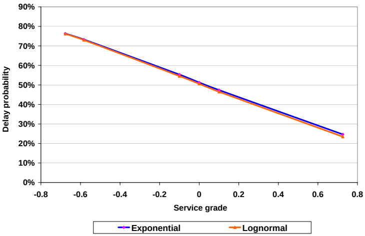
Numerical Experiments: Whitt (2004), Rosenshmidt (2006) demonstrate a **useful fit** for typical call-center parameters.

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

Fraction Abandoning



Delay Probability



Why Does Erlang-A Work? General Services

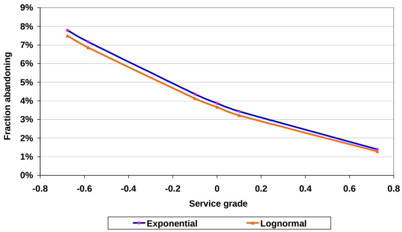
Established: M/M/N+**G** $\approx$ M/M/N+M ($\theta = g(0)$).

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

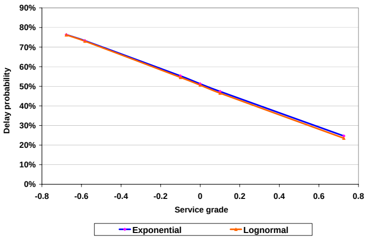
Numerical Experiments: Whitt (2004), Rosenshmidt (2006) demonstrate a **useful fit** for typical call-center parameters.

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

Fraction Abandoning



Delay Probability



QED G-Services: G/GI/N (Reed, 2007), G/D_K/N+G (w/ Momčilović), ...

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

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

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 $\quad (\theta = g(0))$.

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: $\boldsymbol{\lambda}$
   - 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.)

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: $\boldsymbol{\lambda}$
   - Consider $M_t/G/N_t + G$ with arrival rate $\boldsymbol{\lambda(t), t \geq 0}$.
   - Approximate its **time-varying** performance **at time** $\boldsymbol{t}$ 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 $(\theta = g(0))$.

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: 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 $(M_t/M/N + M, \ \mu = \theta)$.

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $M_t \,/\, M \,/\, N_t + \mathrm{M}$

Parameters $\lambda(t)$ μ **?** θ

? $N_t = R_t + \beta\sqrt{R_t}$

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $\quad M_t / M / N_t + \mathrm{M}$

Parameters $\quad \lambda(t) \quad \mu \quad$ **?** $\quad \theta$

? $\quad N_t = R_t + \beta\sqrt{R_t}$

$\mu = \theta$: $\quad L_t \overset{d}{=} \text{Poisson}(R_t) \overset{d}{\approx} N(R_t, R_t)$, since $M_t / M / \infty$

$$R_t = E\lambda(t - S) \cdot E(S) = E \int_{t-S}^{t} \lambda(u) du$$

offered load

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $\quad M_t / M / N_t + M$

Parameters $\quad \lambda(t) \quad \mu \quad$ **?** $\quad \theta$

? $\qquad N_t = R_t + \beta\sqrt{R_t}$

$$\mu = \theta: \quad L_t \overset{d}{=} \text{Poisson}(R_t) \overset{d}{\approx} N(R_t, R_t), \text{ since } M_t / M / \infty$$

$$R_t = E\lambda(t - S) \cdot E(S) = E \int_{t-S}^{t} \lambda(u)du \qquad \text{offered load}$$

Given $\quad L_t \approx R_t + Z\sqrt{R_t} \quad , \quad Z \overset{d}{=} N(0,1)$

choose $\quad N_t = R_t + \beta\sqrt{R_t}$

$$\Rightarrow \quad \alpha = P(W_t > 0) \underset{\text{PASTA}}{\approx} P(L_t \geq N_t) = P(Z \geq \beta) = 1 - \phi(\beta)$$

$$\Rightarrow \quad \beta = \phi^{-1}(1 - \alpha) \qquad \text{time-stable} \quad \alpha \equiv P(W_t > 0) \ ?$$

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $\quad M_t / M / N_t + \mathrm{M}$

Parameters $\quad \lambda(t) \quad \mu \quad$ **?** $\quad \theta$

? $\quad \mathrm{N_t = R_t + \beta\sqrt{R_t}}$

$\mu = \theta$: $\quad \mathrm{L_t \overset{d}{=} Poisson(R_t) \overset{d}{\approx} N(R_t, R_t)}$, since $M_t / M / \infty$

$$R_t = E\lambda(t - S) \cdot E(S) = E \int_{t-S}^{t} \lambda(u)du$$ offered load

Given $\quad \mathrm{L_t \approx R_t + Z\sqrt{R_t}} \quad , \quad \mathrm{Z \overset{d}{=} N(0,1)}$

choose $\quad \mathrm{N_t = R_t + \beta\sqrt{R_t}}$

$\Rightarrow \quad \alpha = \mathrm{P(W_t > 0) \underset{PASTA}{\approx} P(L_t \geq N_t) = P(Z \geq \beta) = 1 - \phi(\beta)}$

$\Rightarrow \quad \beta = \phi^{-1}(1 - \alpha) \qquad$ time-stable $\quad \alpha \equiv \mathrm{P(W_t > 0)}$?

Indeed, but in fact $\quad$ TIME-STABLE PERFORMANCE

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $\quad M_t / M / N_t + \mathrm{M}$

Parameters $\quad \lambda(t) \quad \mu \quad$ **?** $\quad \theta$

? $\quad \boxed{N_t = R_t + \beta\sqrt{R_t}}$

$\mu = \theta: \quad L_t \overset{d}{=} \text{Poisson}(R_t) \overset{d}{\approx} N(R_t, R_t)$, since $M_t / M / \infty$

$$R_t = E\lambda(t-S) \cdot E(S) = E \int_{t-S}^{t} \lambda(u)du$$ offered load

Given $\quad L_t \approx R_t + Z\sqrt{R_t} \quad , \quad Z \overset{d}{=} N(0,1)$

choose $\quad N_t = R_t + \beta\sqrt{R_t}$

$\Rightarrow \quad \alpha = P(W_t > 0) \underset{\text{PASTA}}{\approx} P(L_t \geq N_t) = P(Z \geq \beta) = 1 - \phi(\beta)$

$\Rightarrow \quad \boxed{\beta = \phi^{-1}(1-\alpha)} \quad$ time-stable $\quad \alpha \equiv P(W_t > 0)$?

Indeed, but in fact $\quad$ TIME-STABLE PERFORMANCE

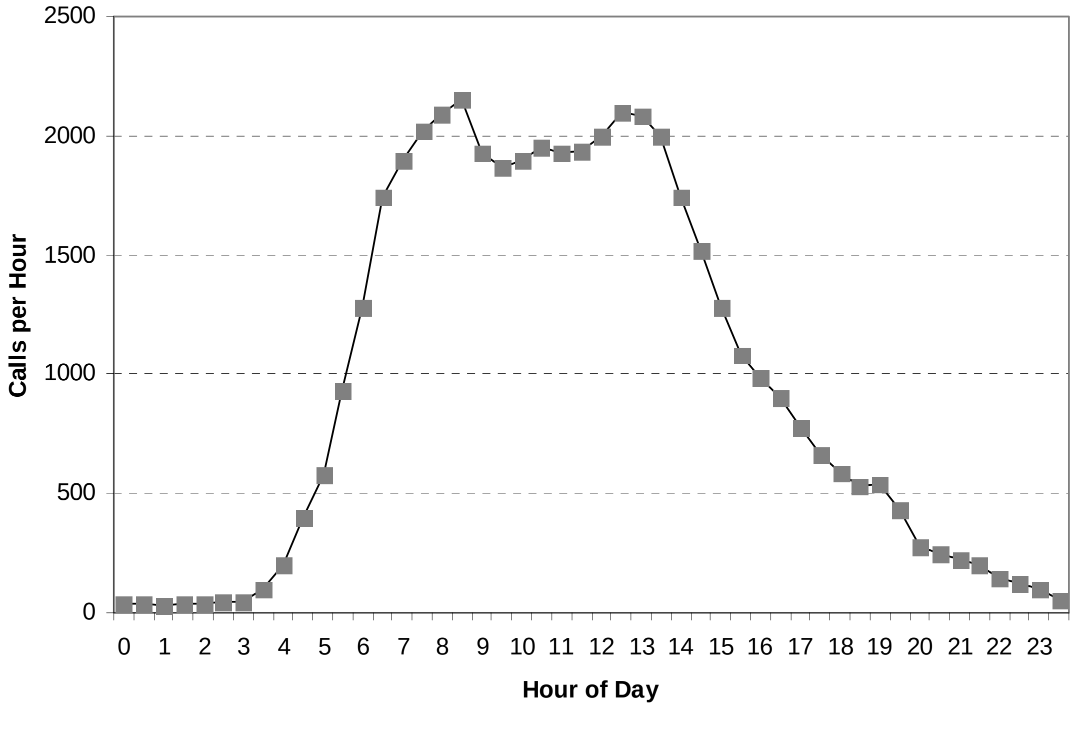
($\mu \neq \theta$, or generally : Iterative Simulation-Based Algorithm)

Example: "Real" Call Center

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

Time-Varying (two-hump) arrival functions common

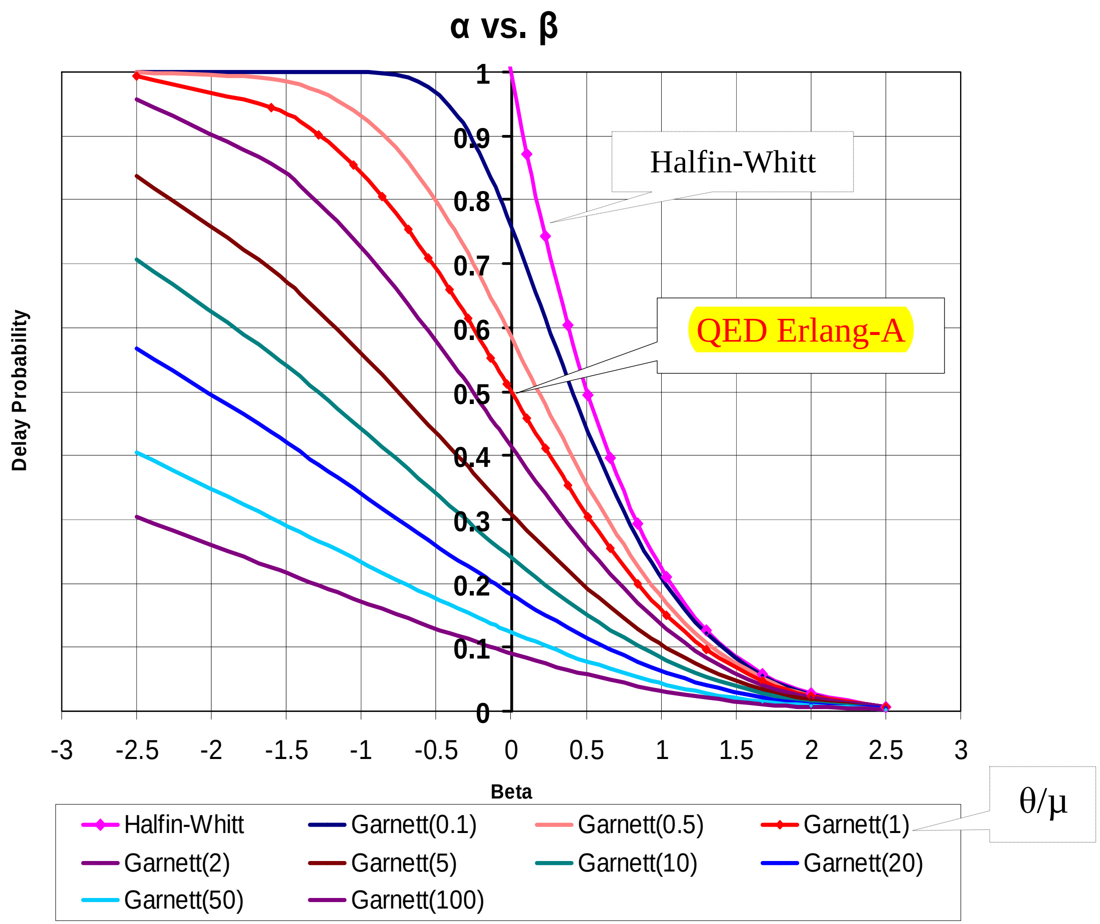
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)



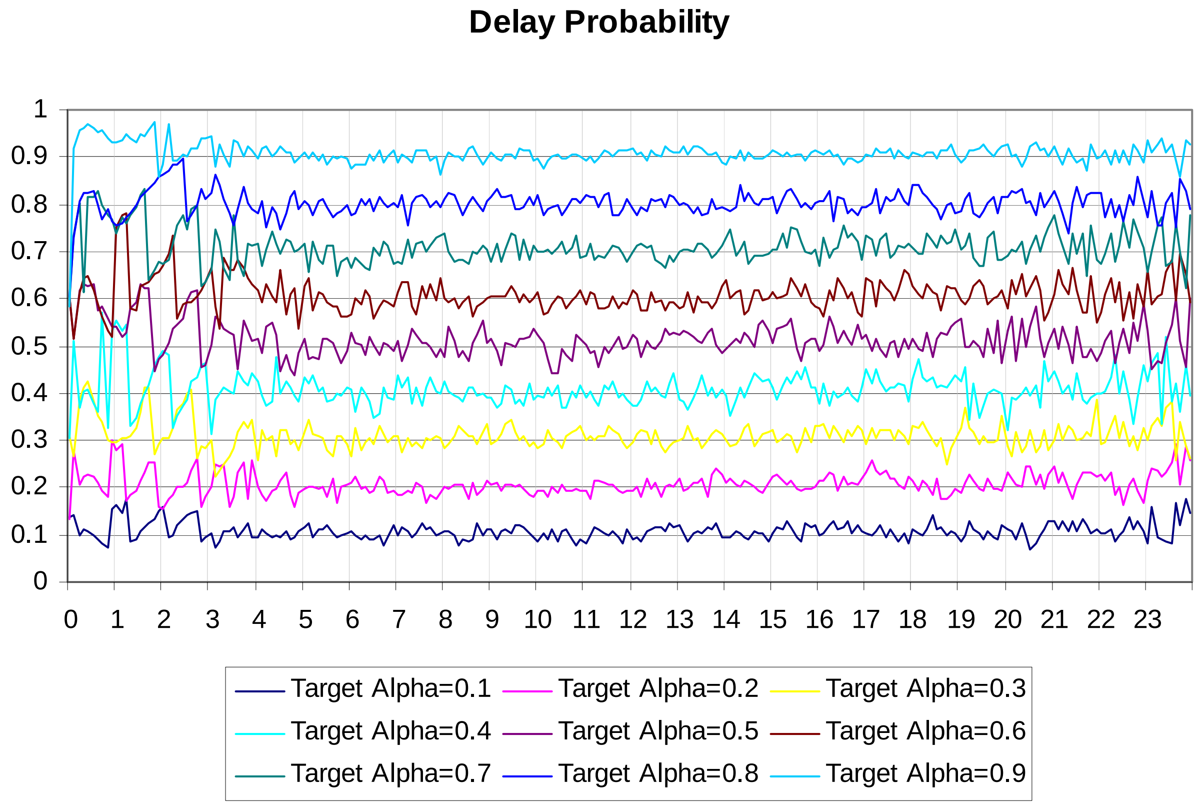
Assume: Service and abandonment times are both

Exponential, with mean 0.1 (6 min.)

HW/GMR Delay Functions

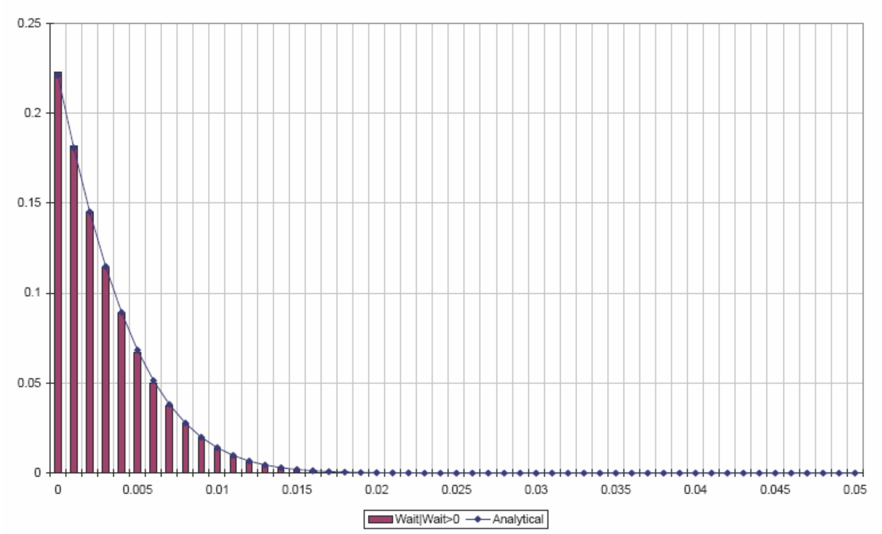


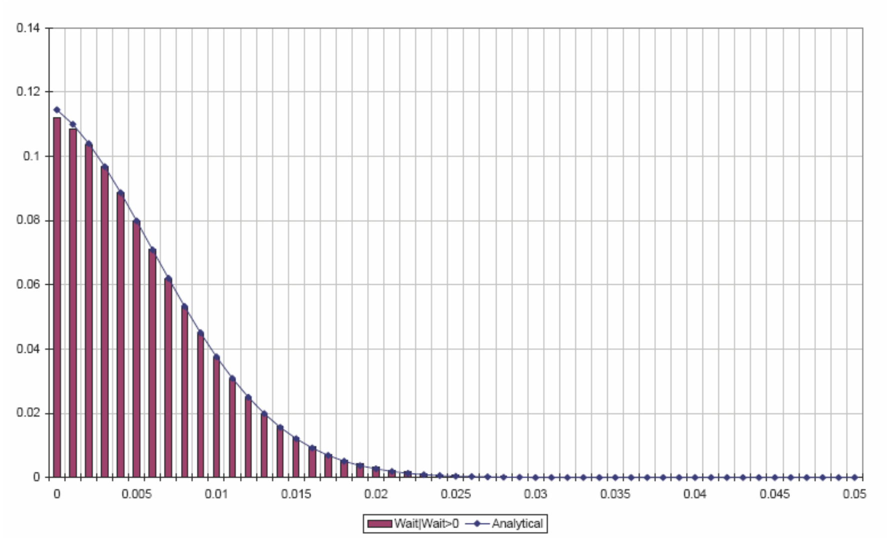
Delay Probability α

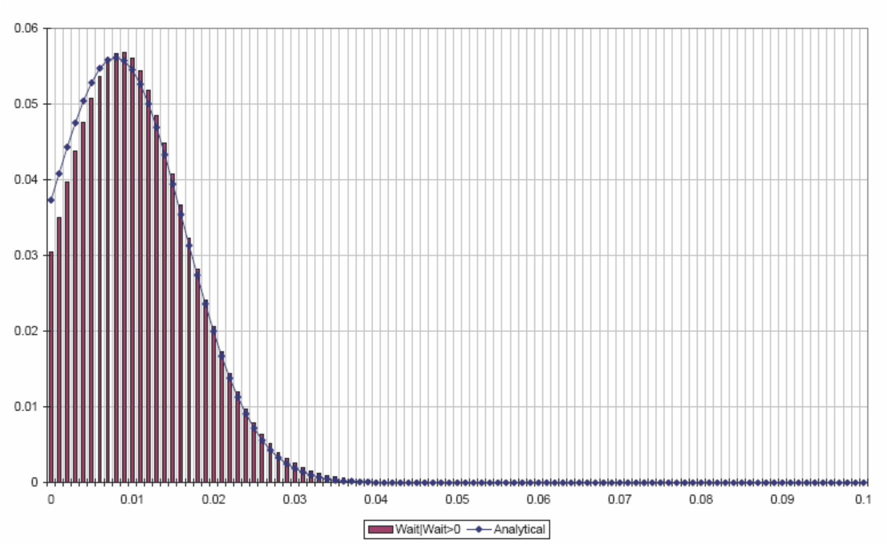


Real Call Center: Empirical waiting time, given positive wait

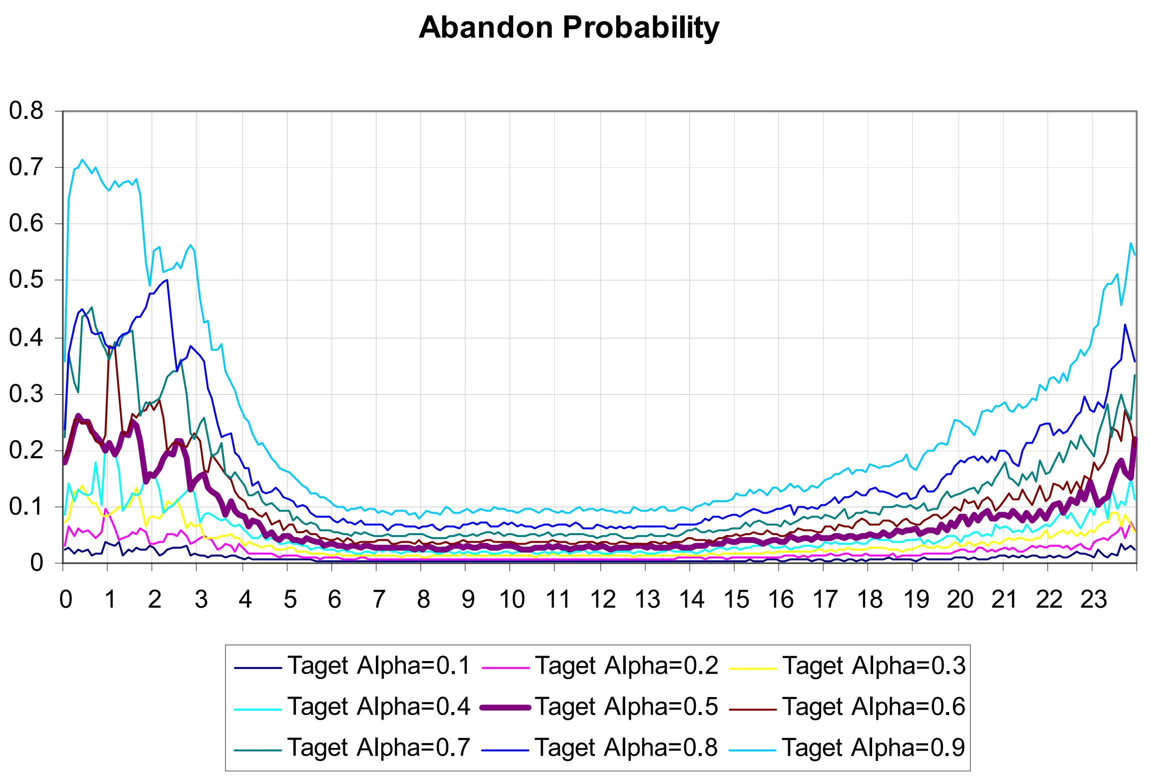
(1) α=0.1 (*QD*) **(2)** α=0.5 (*QED*) **(3)** α=0.9 (*ED*)

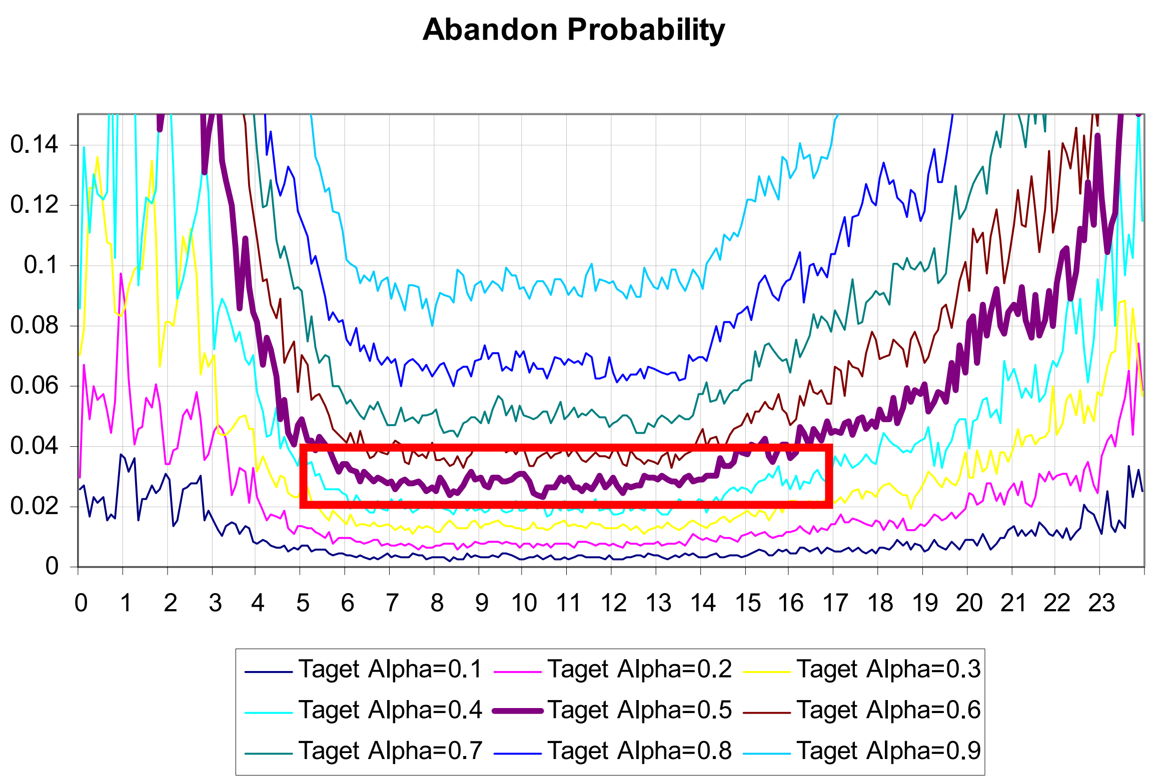






Abandon Probability





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

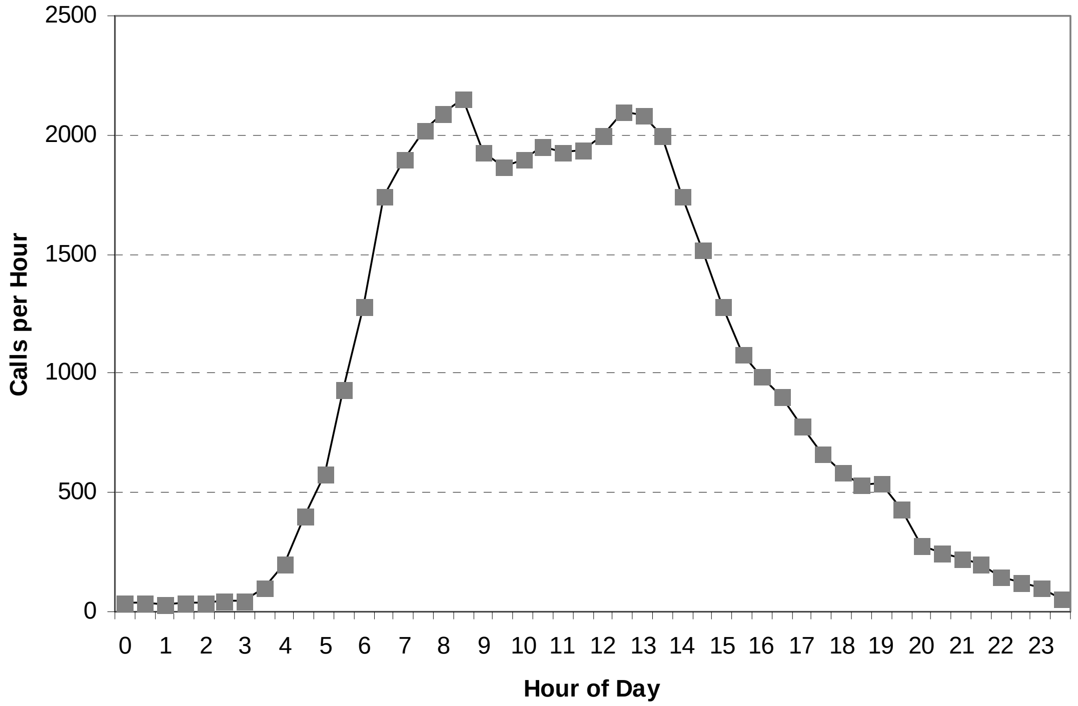
and $\lambda(t) \approx$ stable over service-durations

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

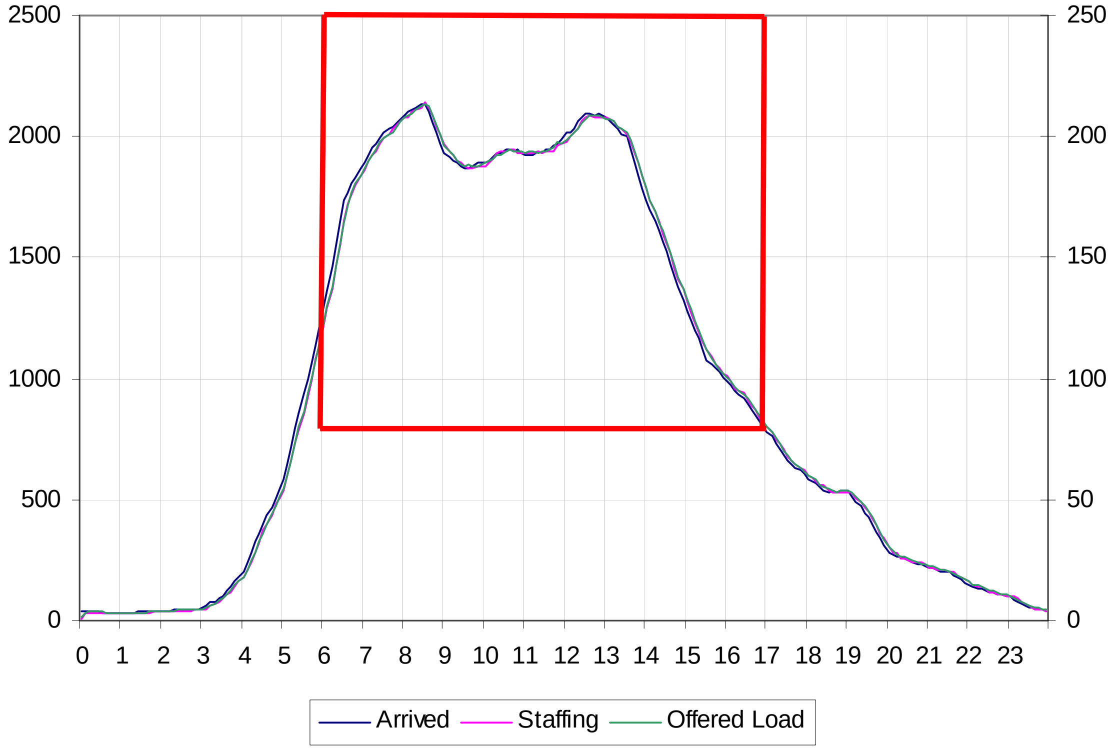
When Optimal ? for moderately-patient customers:

1. Satisfization $\Leftrightarrow$ At least 50% to be serve immediately

2. Optimization $\Leftrightarrow$ Customer-Time = 2 x Agent-Salary



QED Staffing (β=0 iff α=0.5)



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 $\boldsymbol{J}$.
- **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 $\boldsymbol{\lambda_j = \mathrm{E}\lambda_j(t - S_e)}$.

Why Does Erlang-A Work? Multi-Class Customers

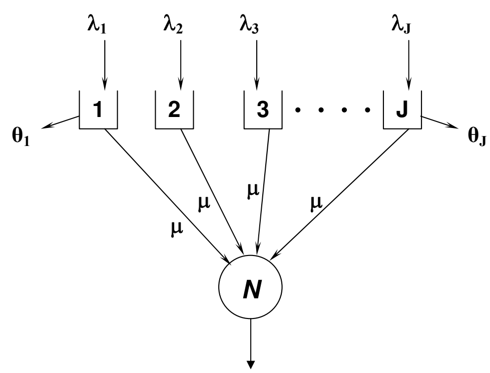
Now: $M_t^J/G/N_t + G \approx (M^J/G/N + G)_t$ (well staffed & controlled).
Service Levels: Class 1 = **VIP** , . . ., Class J = **best-effort**.

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 $\boldsymbol{J}$.
- **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).

The V-Design



- J customer classes: **Arrivals** - Poisson(λ_j), **Patience** - Exp(θ_j).
- N **iid** servers: service durations Exp(μ).
- Delay probabilities $\alpha_1 < \alpha_2 \ldots < \alpha_J$

Objective: **Min** N, subject to **service level differentiation:**

$$P\{W_j > 0\} \leq \alpha_j, \; j = 1, \ldots, J.$$

Proposed solution:

- **Static priorities** $1 > 2 > \ldots$ **with thresholds** $\mathrm{S}_1 > \mathrm{S}_2 > \ldots$
  i.e. a class-j customer served if it is of the present highest-priority and the number of idle servers is S_j or more.
- Performance analysis in steady-state **with no abandonments** (Schaack & Larson 1986).

The V-Design in the QED regime

Gurvich, Armony and Mandelbaum ('06)

Consider a sequence indexed by $N = 1, 2, ...$

Thresholds: $0 = S_1^N \leq S_2^N \leq ... \leq S_J^N \leq N$

Assume: $S_J^N = o(\sqrt{N})$

Assume: $\lambda_j^N / \lambda^N \mu \to \rho_j > 0\,, \forall j$ (all classes non-negligible)

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

- QED: $\lim_{N\to\infty} P_N\{W_J > 0\} = \alpha_J,\quad 0 < \alpha_J < 1;$
- Customer: $\lim_{N\to\infty} \sqrt{N} P_N\{Ab_J\} = \Delta,\quad 0 < \Delta < \infty;$
- Server: $\lim_{N\to\infty} \sqrt{N}\,(1 - \rho_N) = \beta,\quad -\infty < \beta < \infty;$
- Manager: $N = R + \beta\sqrt{R} + o(\sqrt{R})\,,\quad R = \lambda/\mu$ large.

Here, α_J and Δ are determined by the QED $M/M/N + M$.

Erlang-A is all that is needed:

$$\limsup_{N\to\infty} P\{W_j^N > 0\} \leq \alpha_j, \text{ if } \liminf_{N\to\infty} S_{j+1}^N - S_j^N \geq \frac{\ln \alpha_j/\alpha_{j+1}}{\ln \sum_{i=1}^j \rho_i}$$

The V-Design in the QED regime

Armony, Gurvich and Mandelbaum ('06)

High priorities ($j < J$):

- $W_j | W_j > 0$ as in $M(\lambda_1)/M(N\mu)/1 + M(\theta_1)$.
- $W_j | W_j > 0 \stackrel{d}{=} \Theta\left(\frac{1}{N}\right)$, but
- $W_j \stackrel{d}{=} o\left(\frac{1}{N}\right)$ for $S_j = O(\log(N))$.

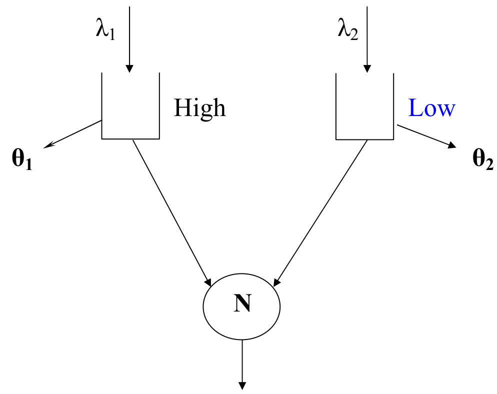
Low priority (W_J) unchanged as long as $S_J^N = o(\sqrt{N})$.

Comparison to static priorities:

- Static: All classes share same delay probability $P\{N$ servers busy$\}$
- Thresholds: $P\{W_j > 0\}$ differentiated through thresholds.

$o(\sqrt{N})$ thresholds guarantee QD waiting times for high-priority and, at the same time, are not hurting the low-priority (who are still QED-served).

What's Going On?



Thresholds: $0 = S_1^N \leq S_2^N \leq N$

1. **# High Priority Q_1^N:** "See" an $M(\lambda_1)/M(N\mu)/1 + M(\theta_1)$ queue in light traffic.

$$Q_1^N = o(\sqrt{N}).$$

Only Q_2^N and θ_2 play a role in asymptotics.

2. **# Busy Servers B^N:** Only high priority enter service when $B^N \geq N - S_2^N$. $(B^N - S_2^N)^+$ is in light traffic:

$$(B^N - S_2^N)^+ = o(\sqrt{N})$$

Conclusion:

$$\begin{aligned} B^N + Q_1^N + Q_2^N &\approx M/M/(N - S_2^N) + M(\theta_2) \\ &\approx M/M/N + M(\theta_2) \quad (S_2^N = o(\sqrt{N})) \end{aligned}$$

Asymptotic equivalence to $M/M/N + M$

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

Technion
S.E.E
Center
Service Enterprise Engineering

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**.

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: Mini version available on a DVD; working version **7GB tables**, or **20GB raw zipped**, for each call center – **ask for a DVD, or my mini-HD**.