QED Q's:

Quality- and Efficiency-Driven Queues

with a focus on Call/Contact Centers

Avishai Mandelbaum

Technion, Haifa, Israel

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

Research Partners

- Empirical/Statistical Analysis:
  Brown, Gans, Ritov, Sakov, Shen, Zeltyn, Zhao
- Students:
  Garnett, Aldor, Khudiakov, Feldman, Rosenshmidt, Maman, Yom-Tov, Marmor, Tseytlin
- Colleagues:
  Armony, Atar, Gurvich, Massey, Momcilovic, Shaikhet, Whitt
- Technion SEE Lab:
  Feigin, Trofimov, Nadjharov, Gavako, Liberman; RA's

Contents

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

Background Material (Downloadable)

- Technion's "**Service-Engineering" Course** ($\geq$ 1995): http://ie.technion.ac.il/serveng
- Google Scholar - search <Call Centers>:
  - Gans (U.S.A.), Koole (Europe), and M. (Israel): "Telephone Call Centers: Tutorial, Review and Research Prospects." MSOM, 2003.
  - Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." JASA, 2005.

Main Messages

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

Note: Simple rooted in **deep** analysis.

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

Supported by the **SEE** Lab.

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

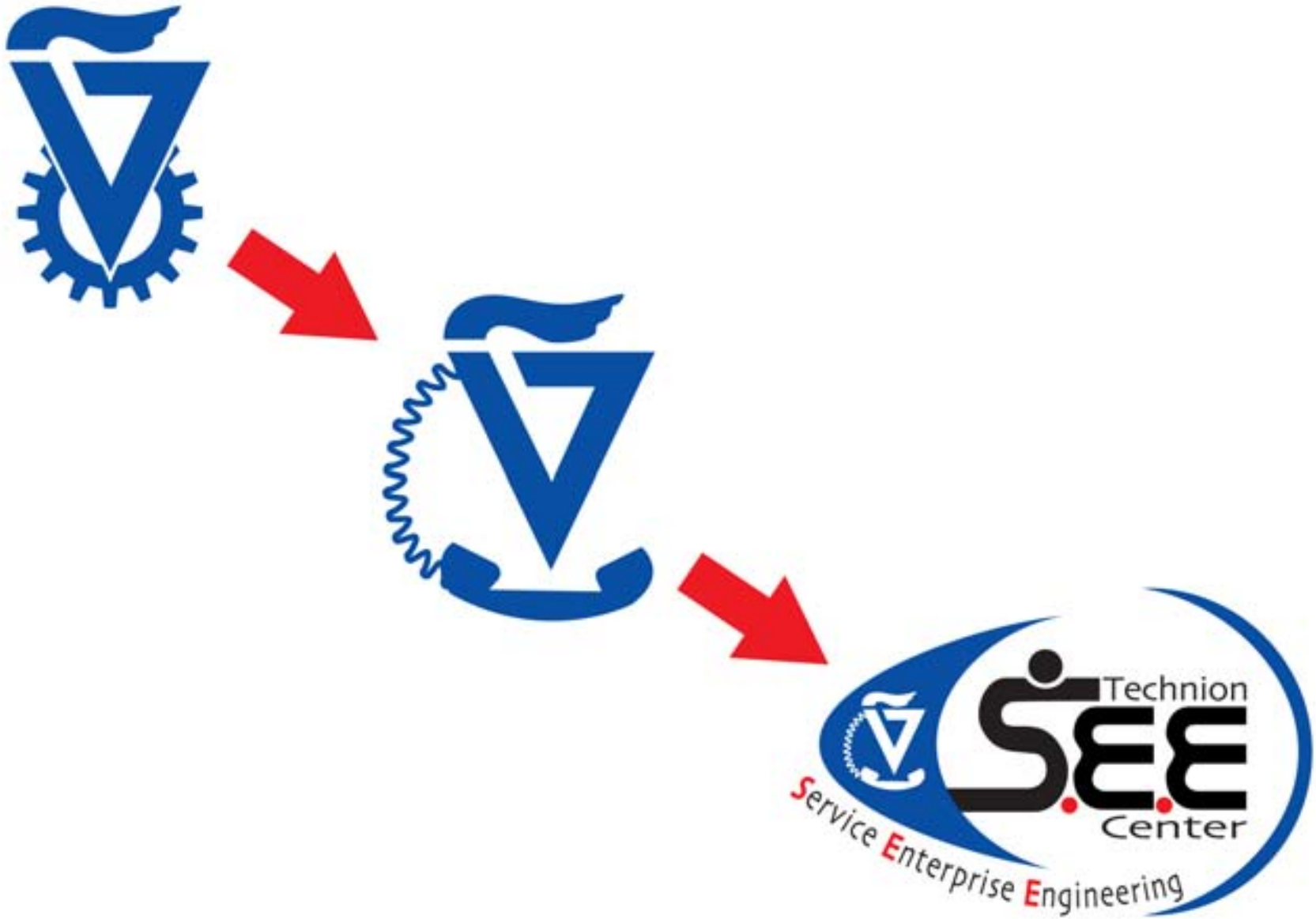
4. Ancestors & **Practitioners** often knew/apply the **"right answer"**: simply did/do not have our tools/desire/need to prove it so.

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

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

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

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), Zhao.
- **industry**:
  - US Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
  - IL Cellular: **3.5 years, 110M / 25M calls, 800 agents; ongoing**.
  - IL Bank: **16 months, ongoing**.

Project Goal: Design and Implement 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)**.

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

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

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

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

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

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

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

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

Prerequisite: Data

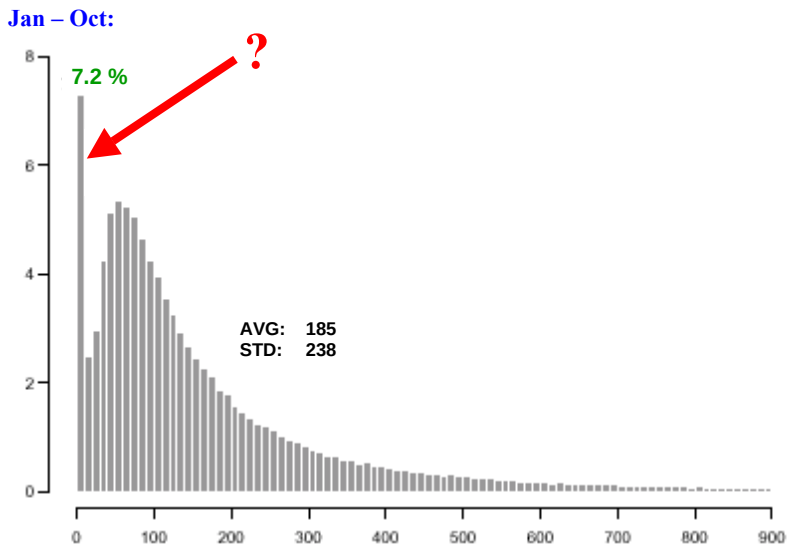
Averages Prevalent.

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

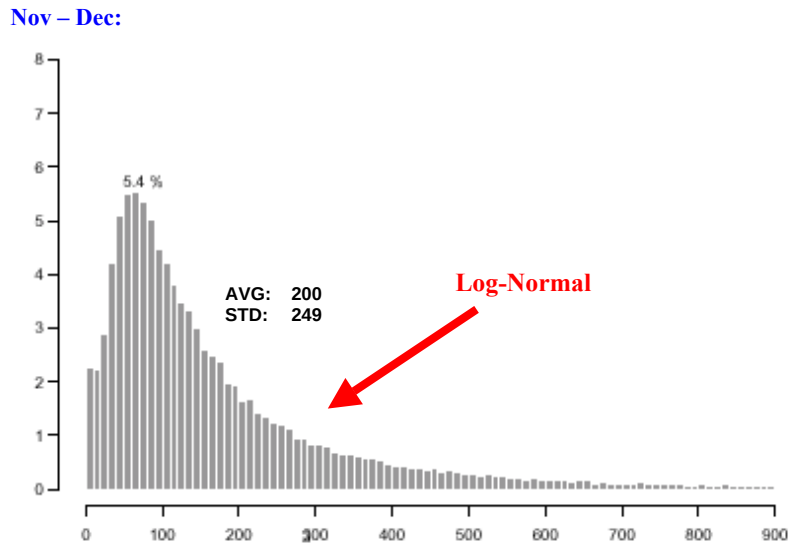
Beyond Averages (+ The Human Factor)

Histogram of Service Times in an Israeli Call Center

January-October



November-December



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

Present Focus: Call Centers

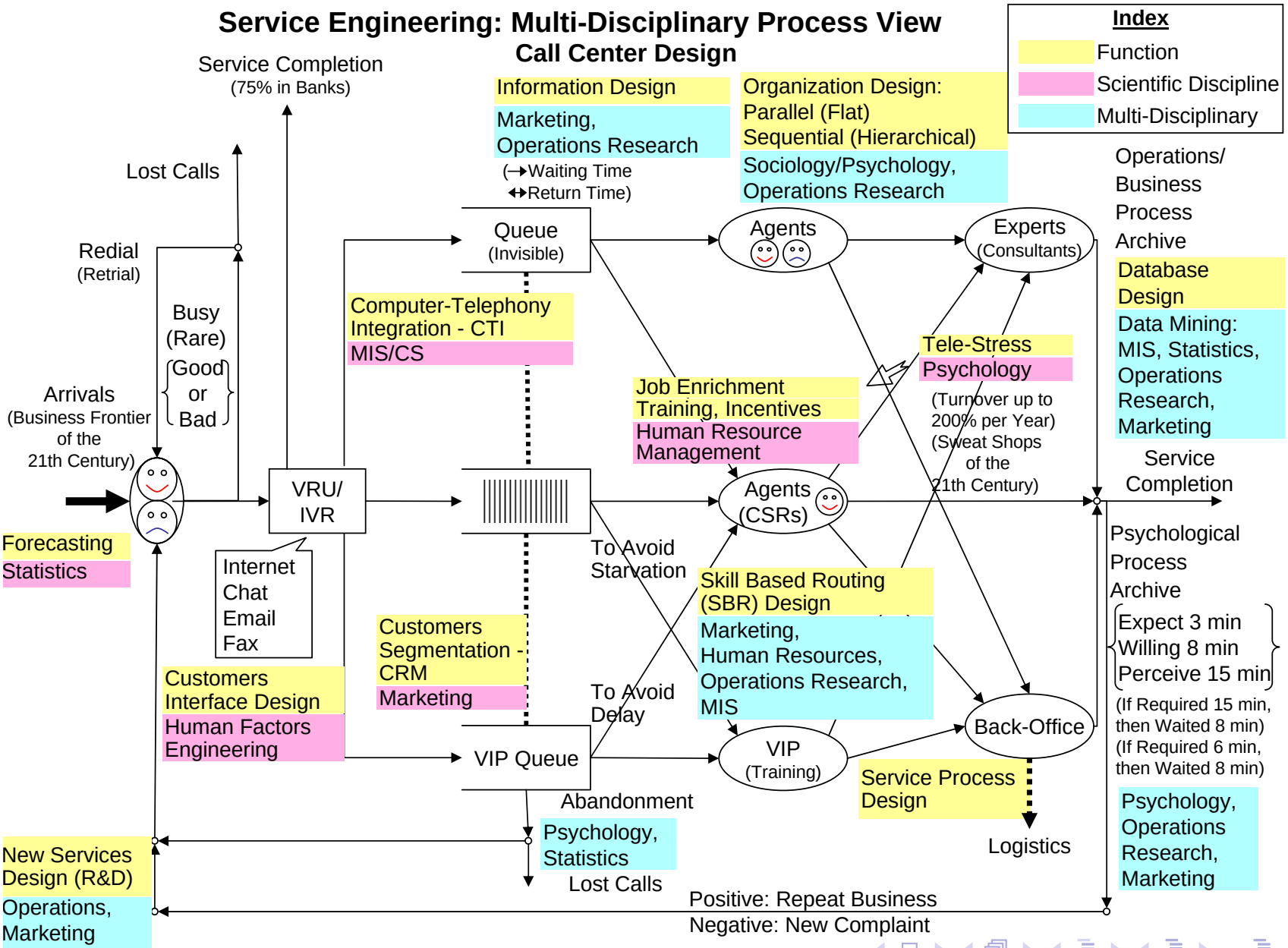
U.S. Statistics (Relevant Elsewhere)

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

Call-Center Environment: Service Network

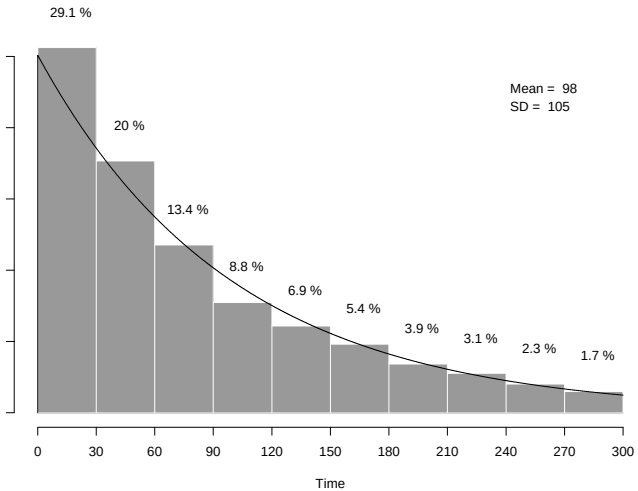
Call-Centers: “Sweat-Shops of the 21st Century"

Call-Center Network: Gallery of Models

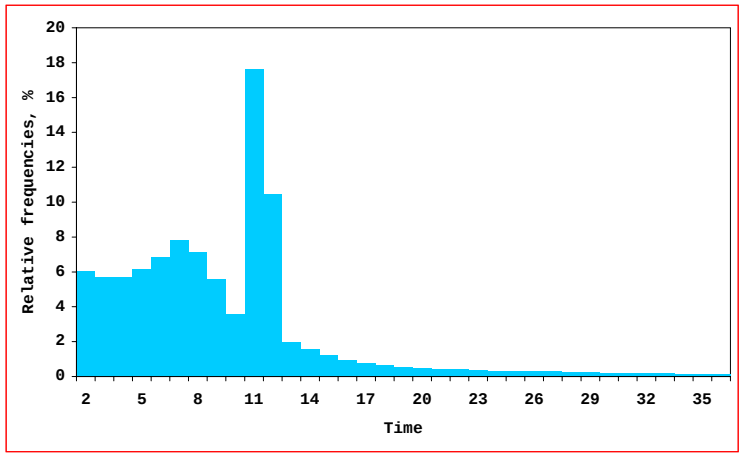


Beyond Averages: Waiting Times in a Call Center

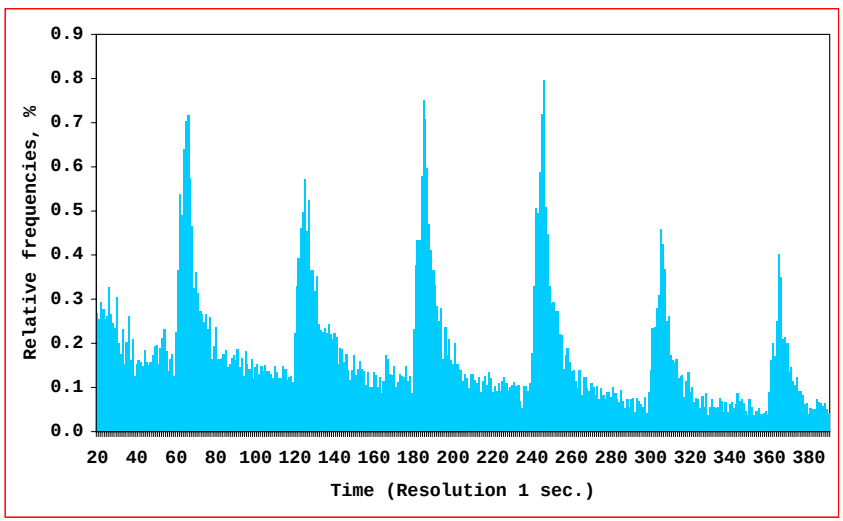
Small Israeli Bank



Large U.S. Bank



Medium Israeli Bank



The “Anatomy of Waiting" for Service

Common Experience:

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

An attempt at “Modeling the Experience":

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

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

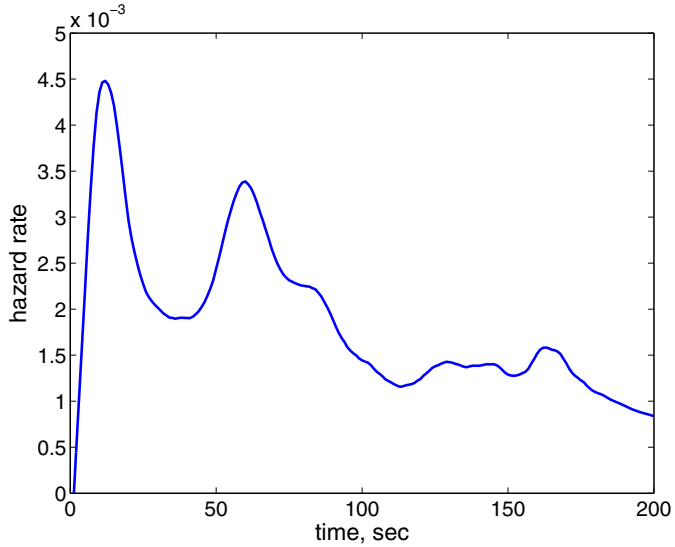
Then left with (τ, V).

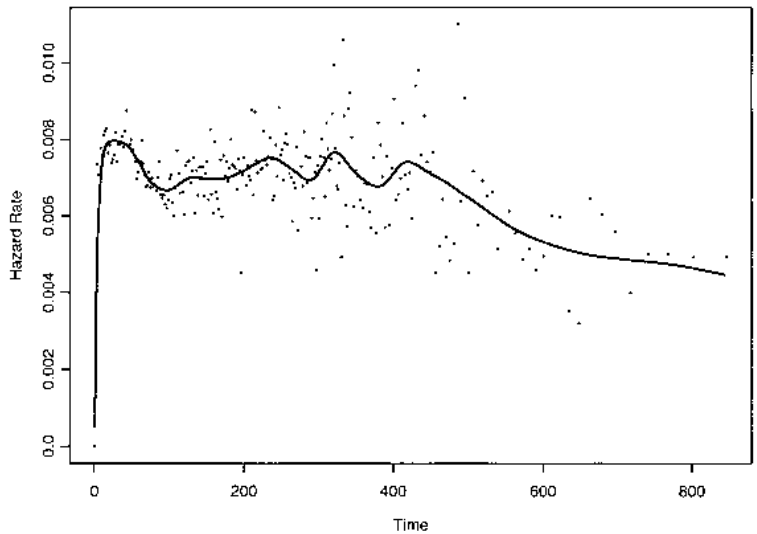
Call Center Data: Hazard Rates (Un-Censored)

(Im)Patience Time τ

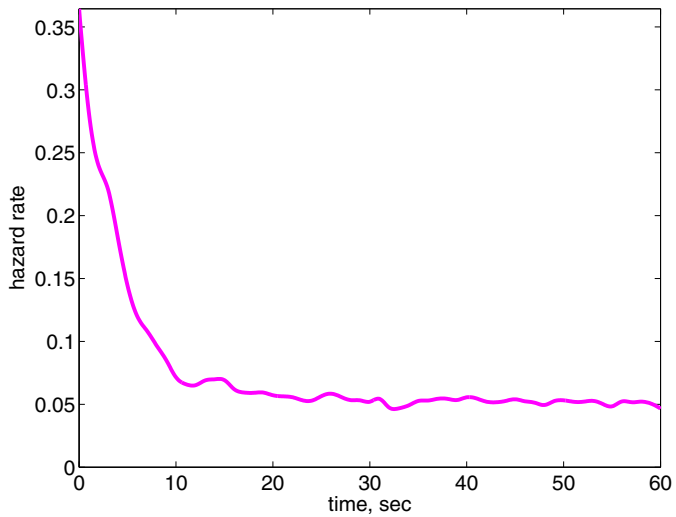
Required/Offered Wait V

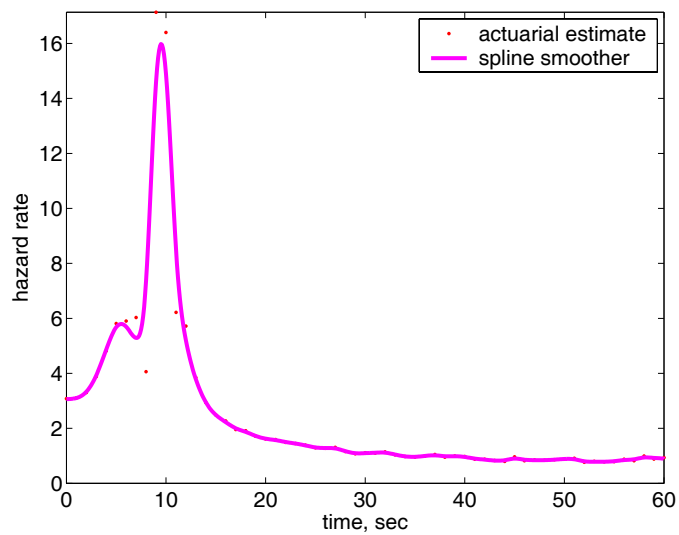
Israel





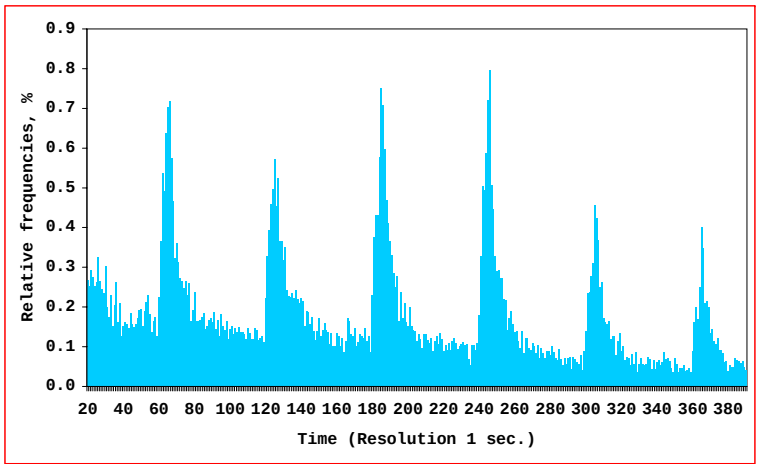
U.S.





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

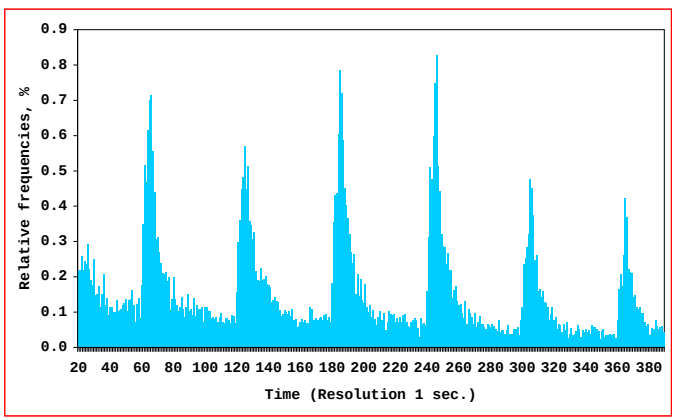
“Waiting-Times" Puzzle at a Large Israeli Bank



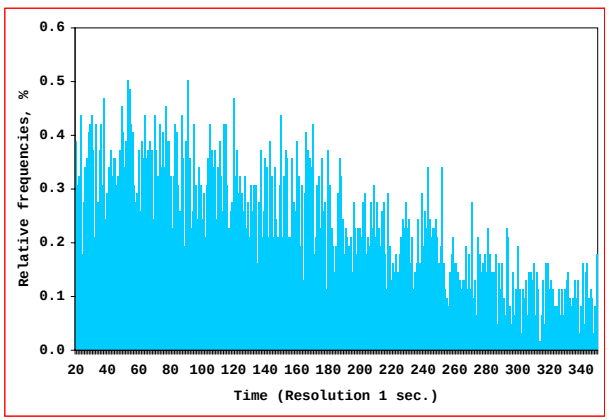
Peaks Every 60 Seconds. **Why?**

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

Served Customers

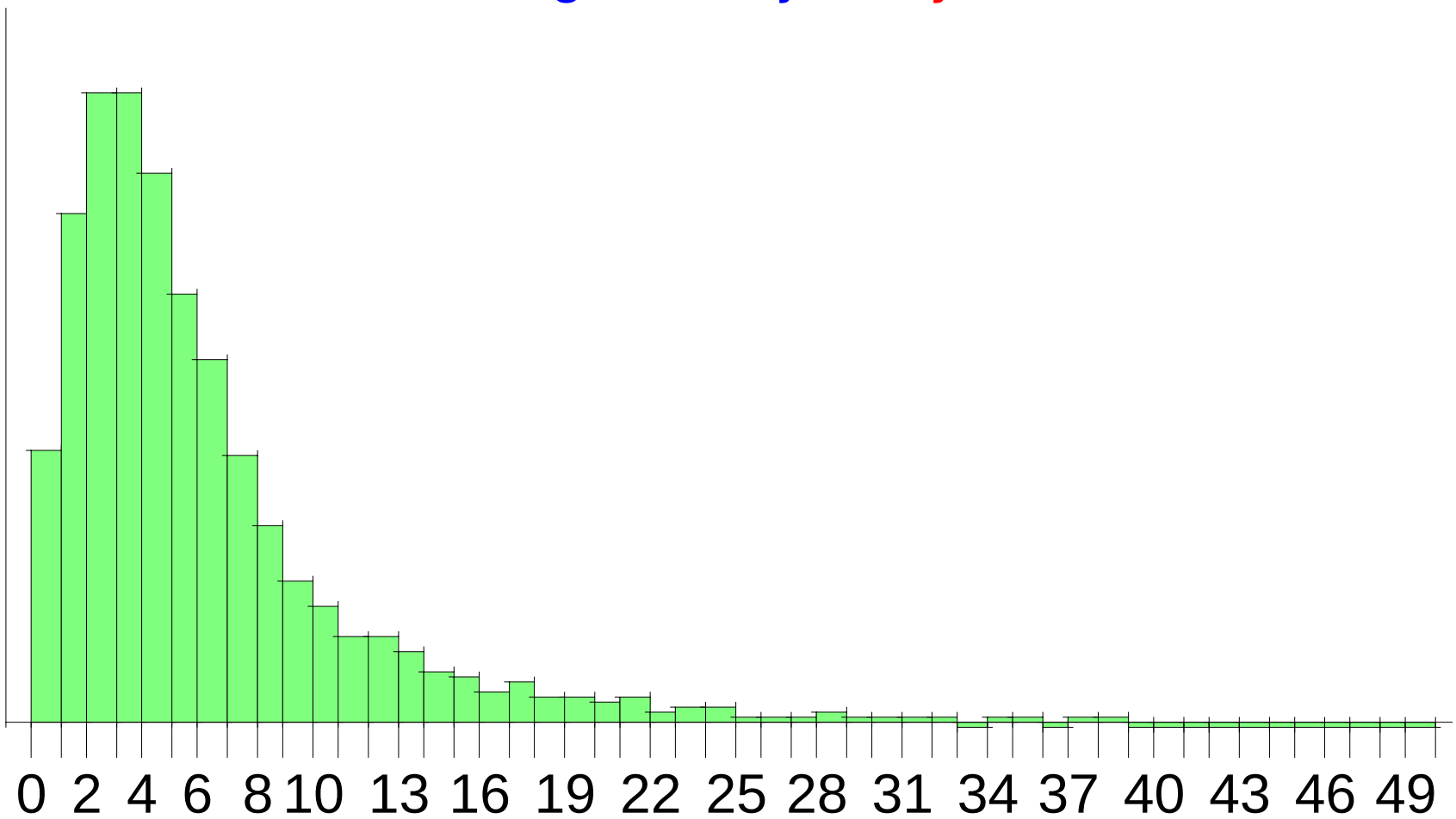


Abandoning Customers



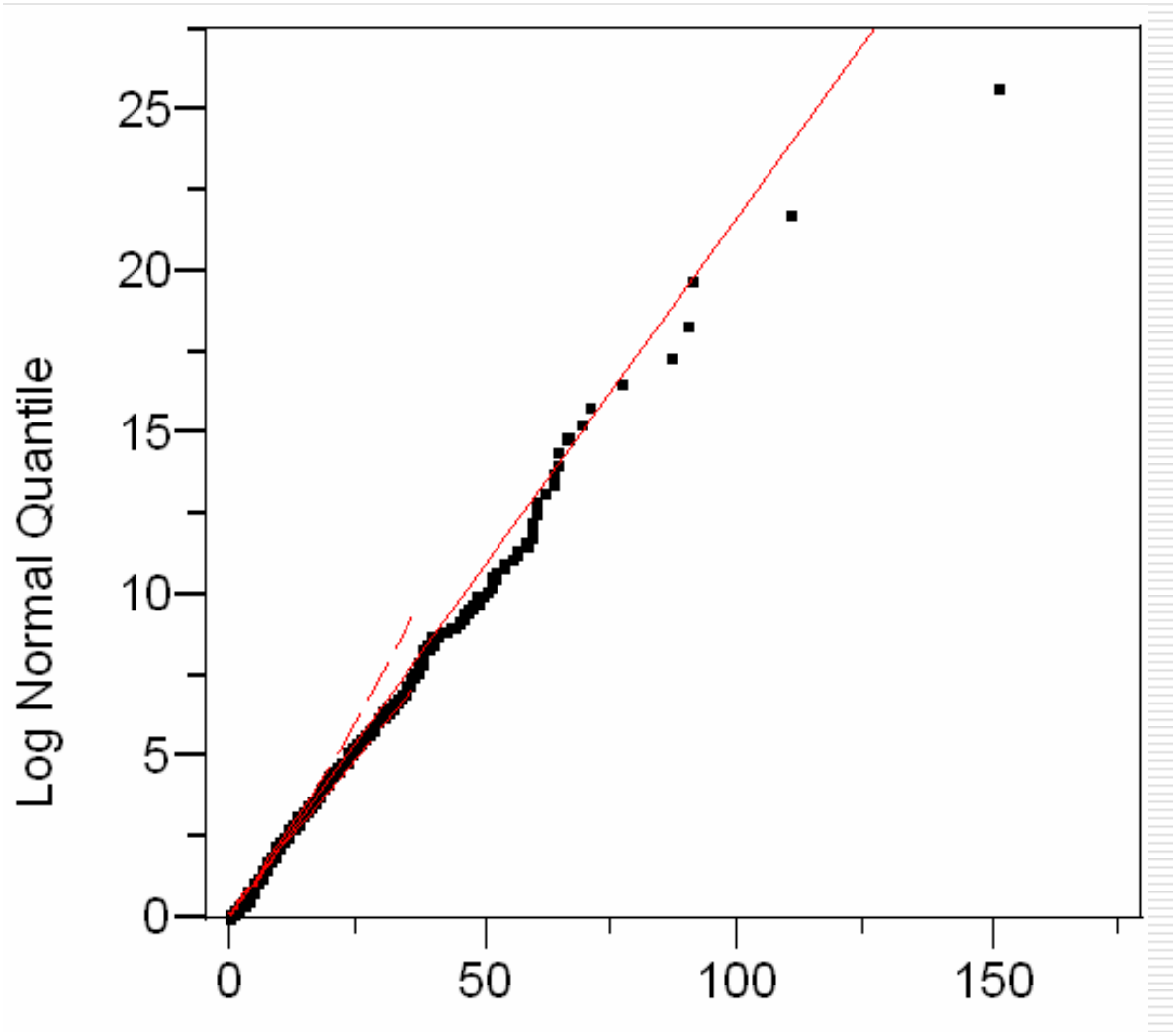
LOS at a Large Israeli Hospital

Length-of-Stay in Days



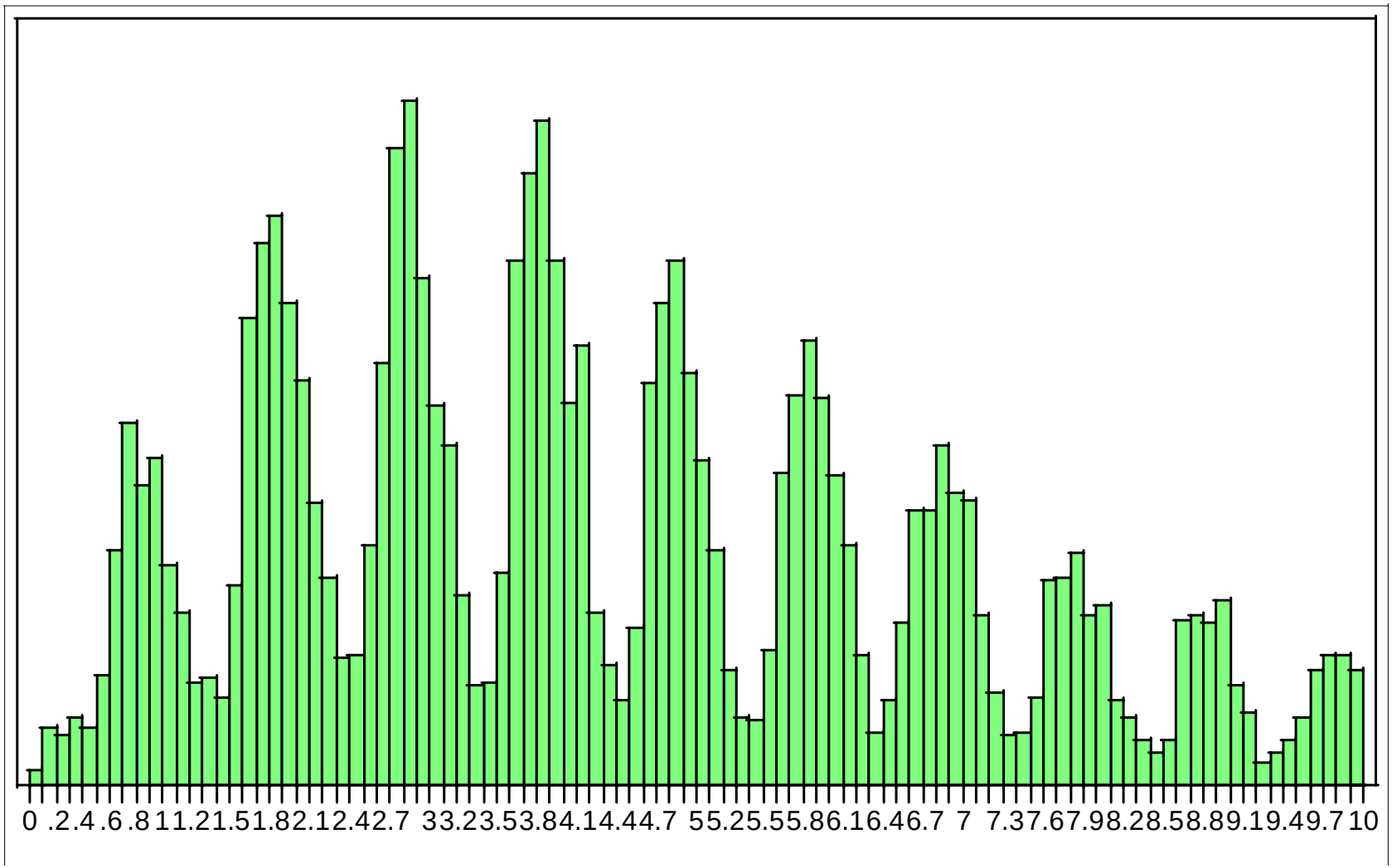
LOS at a Large Israeli Hospital

Length-of-Stay in Days: LogNormal QQ Plot



LOS Puzzle at a Large Israeli Hospital

Length-of-Stay in Hours (0 to 10 days): LN = Normal Mixture



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

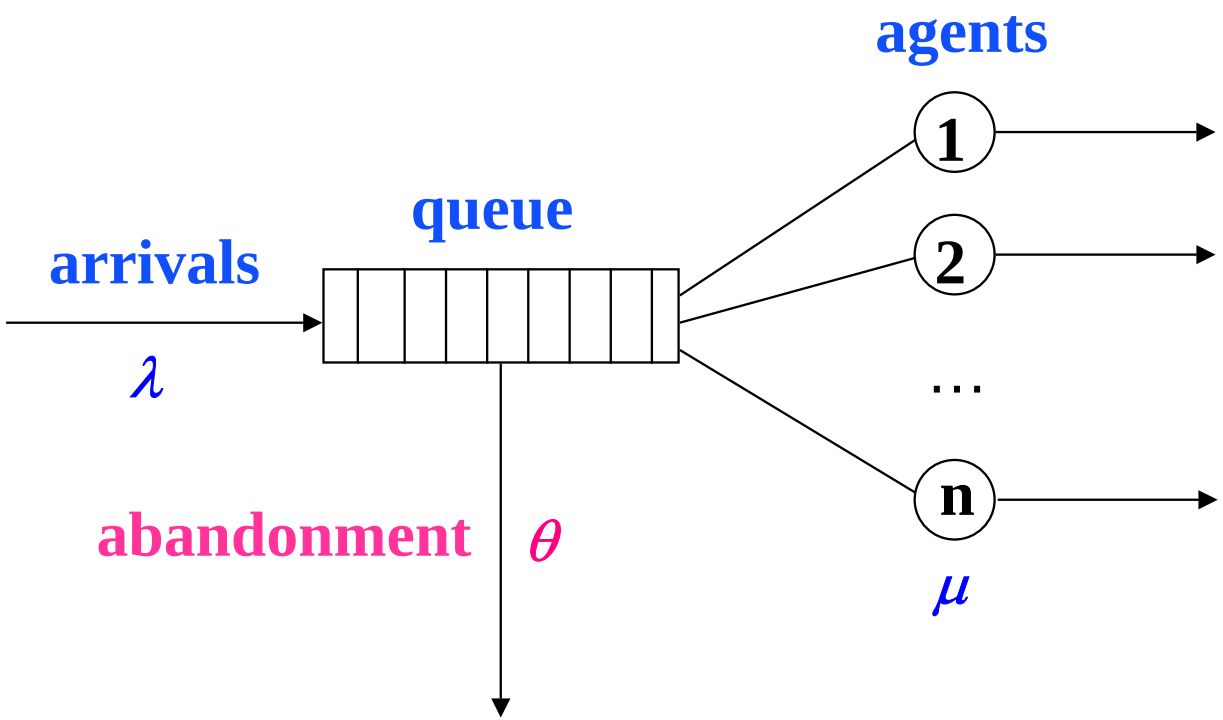
Modeling: τ = **input** to the model, V = **output**.

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

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

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/n – Number of **Service-Agents**.

Testing the Erlang-A Primitives

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

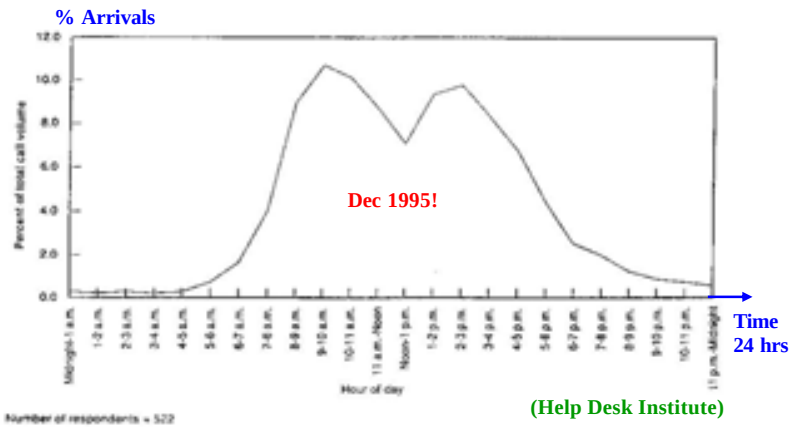
- Primitives independent?
- Customers / Servers Heterogeneous?
- Service discipline FCFS?
- . . . ?

Validation: Support? Refute?

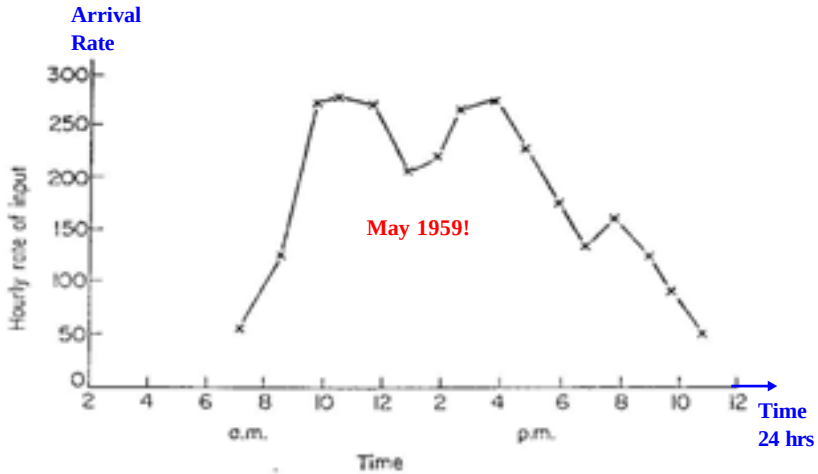
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers

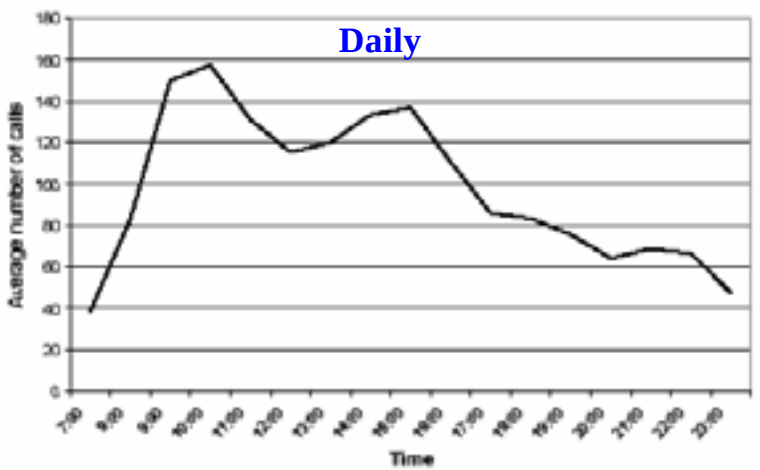
Dec. **1995** (U.S. 700 Helpdesks)



May **1959** (England)

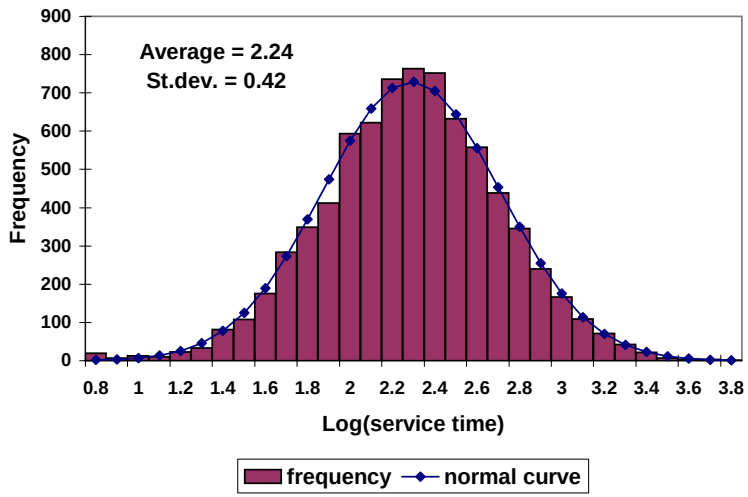


November **1999** (Israel)

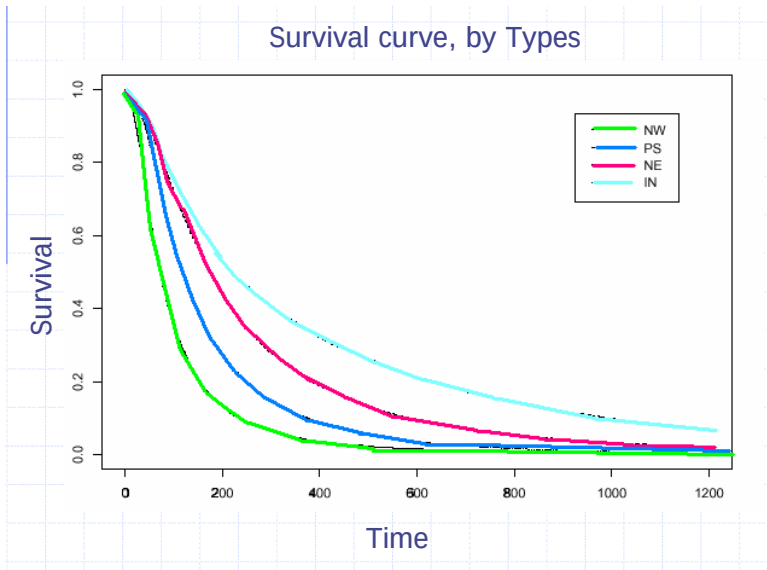


Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram



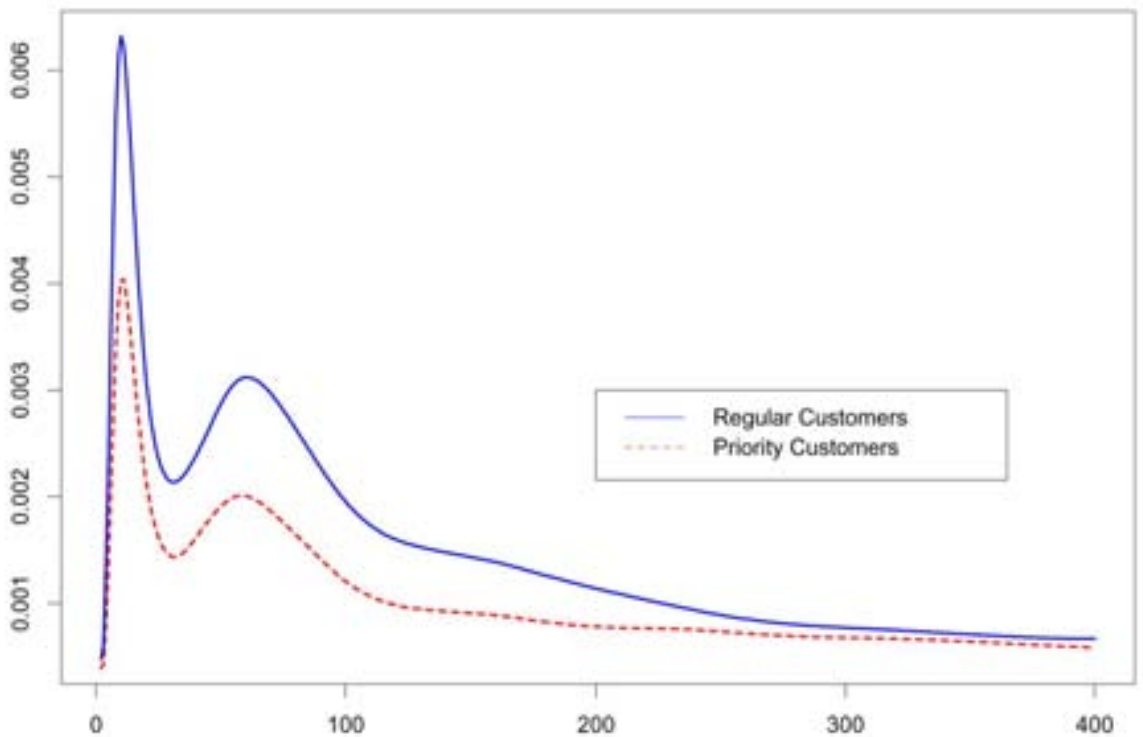
Survival-Functions by Service-Class



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

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

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

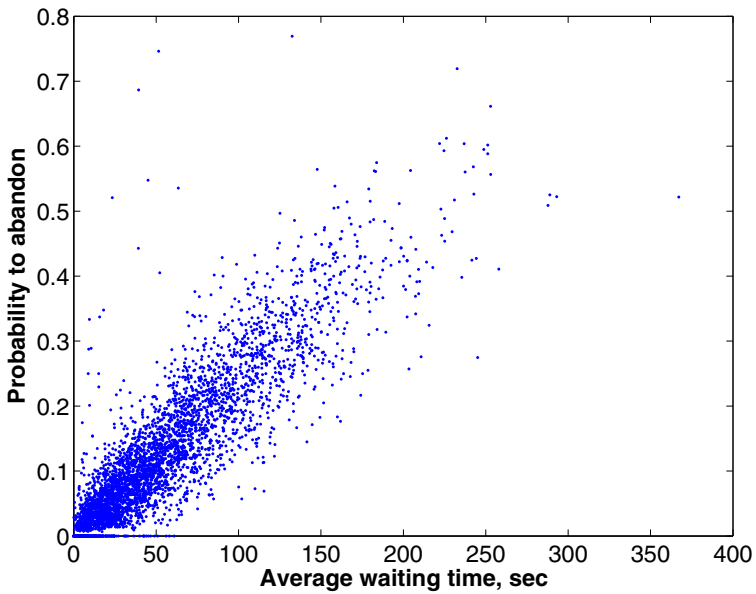


Estimating (Im)Patience: via P{Ab} $\propto$ E[W_q]

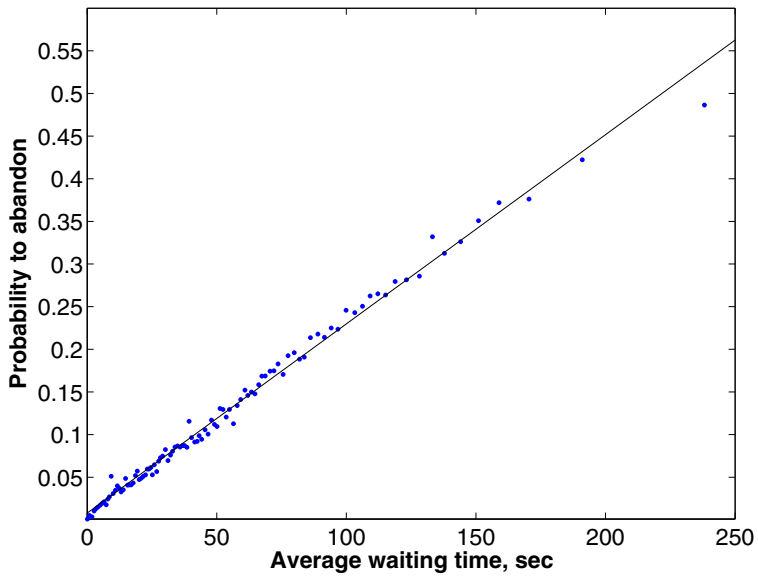
Assume **Exp(θ)** (im)patience. Then, $\mathrm{P\{Ab\}} = \theta \cdot \mathrm{E}[W_q]$.

Israeli Bank: Yearly Data

Hourly Data



Aggregated



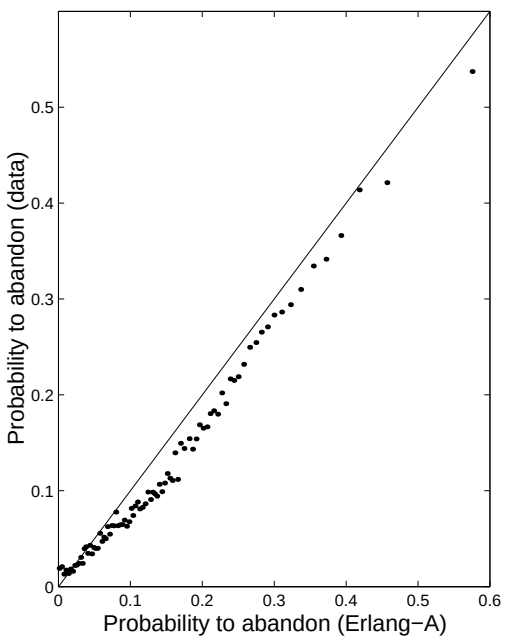
Graphs based on 4158 hour intervals.

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

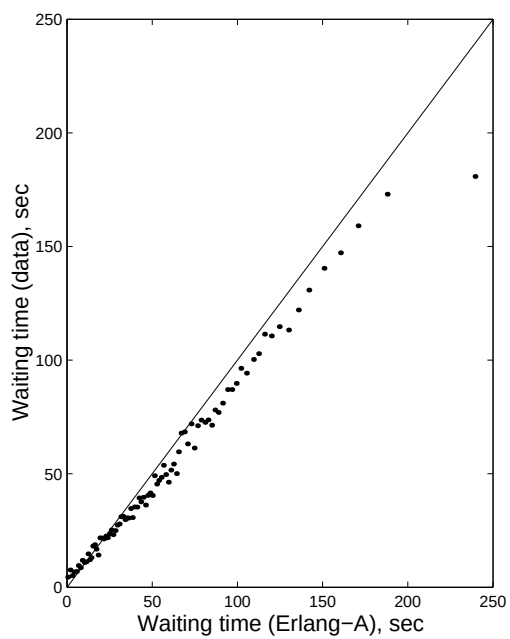
Erlang-A: Fitting a Simple Model to a Complex Reality

- **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).

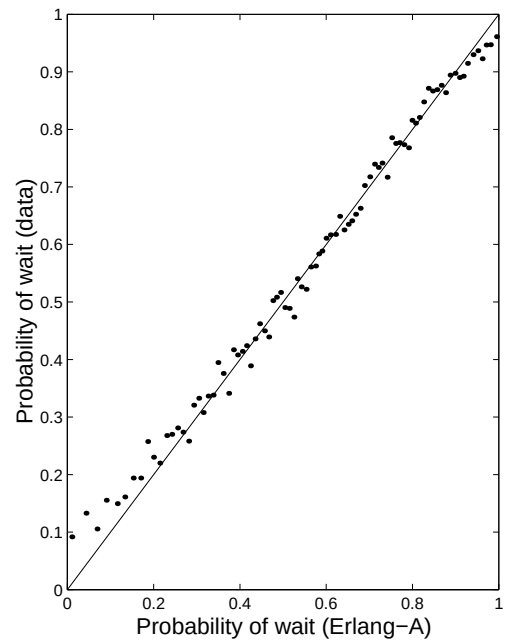
P{Ab}



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 , . . .

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

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

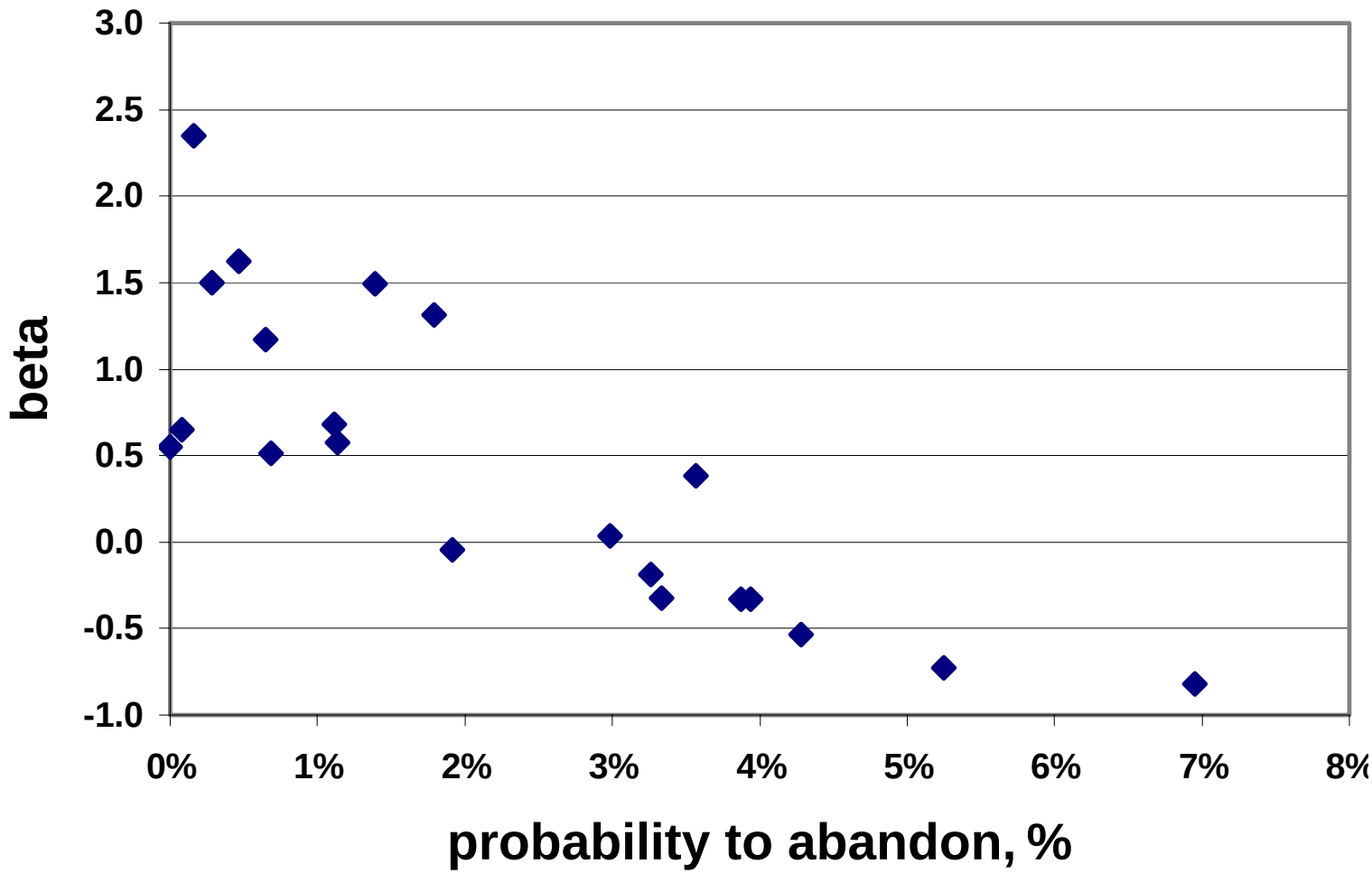
- Erlang 1913/24, Halfin & Whitt 1981
- %Delayed between 25% and 75%
- Wait one-order below service-time (sec vs. min); 1-5% Abandon.

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

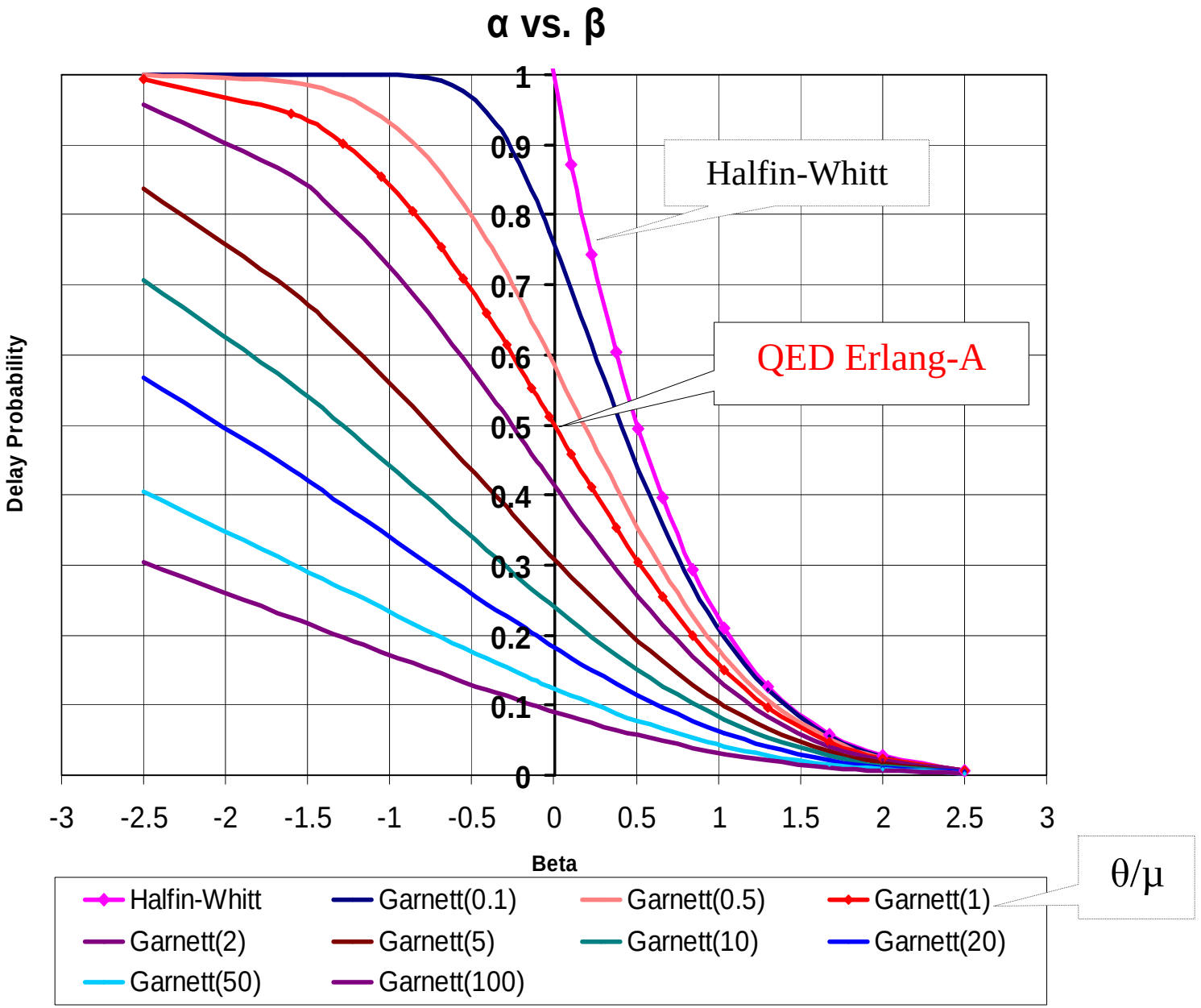
Consider a sequence of M/M/**N**+G models, **N=1,2,3,…**

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

- **QED** $\%\{\text{Wait} > 0\} \approx \alpha, \qquad 0 < \alpha < 1;$
- **Customers** $\%\{\text{Abandon}\} \approx \frac{\gamma}{\sqrt{N}}, \qquad 0 < \gamma;$
- Agents $\text{OCC} \approx 1 - \frac{\beta + \gamma}{\sqrt{N}} \qquad -\infty < \beta < \infty;$
- **Managers** $N \approx R + \beta\sqrt{R}$, $R = \lambda \times \text{E(S)}$ not small;

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

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



QED Approximations (Zeltyn, M. '06)

G – patience distribution,

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

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

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

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

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

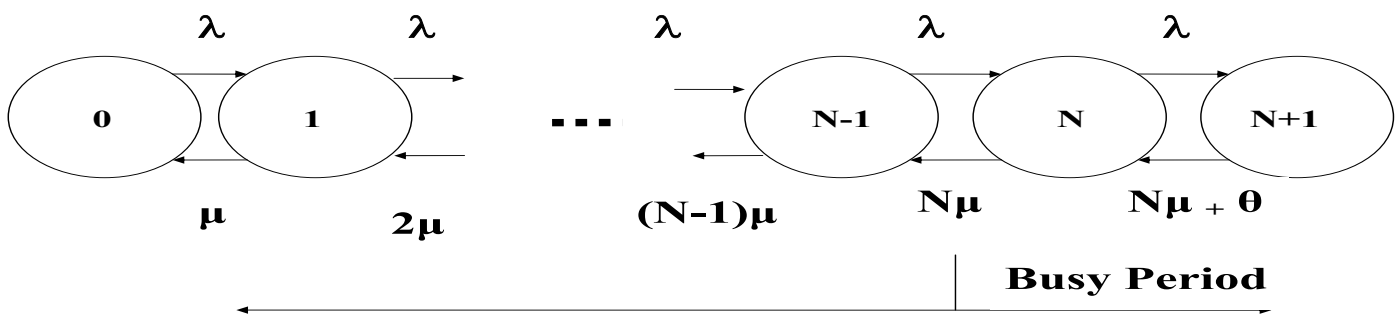
Here

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

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

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

QED Intuition via Excursions: Busy/Idle Periods



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

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

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

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

QED Intuition via Excursions: Asymptotics

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

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

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

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

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

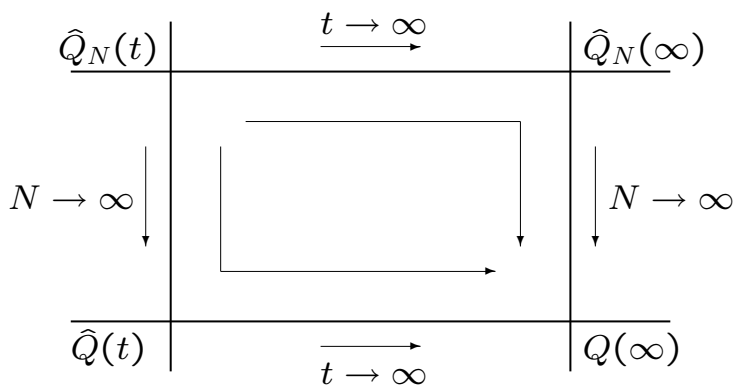
Process Limits (Queueing, Waiting)

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

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

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

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



Approximating (Virtual) Waiting Time

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

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 $\leq 10\%\mathrm{E}[\tau]$	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).

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.

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

Finally,

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

M/M/n+G Performance Measures

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

$$\mathrm{P}\{V > t\} = \frac{\lambda J(t)}{\mathcal{E} + \lambda J}, \quad (\textbf{Baccelli} \ \& \ \textbf{Hebuterne}, 1981)$$

$$\mathrm{P}\{W_q > 0\} = \frac{\lambda J}{\mathcal{E} + \lambda J} \cdot \bar{G}(0)\,,$$

$$\mathrm{P}\{\mathrm{Ab}\} = \frac{1 + (\lambda - n\mu)J}{\mathcal{E} + \lambda J}\,,$$

$$\mathrm{E}[V] = \frac{\lambda J_1}{\mathcal{E} + \lambda J}\,,$$

$$\mathrm{E}[W_q] = \frac{\lambda J_H}{\mathcal{E} + \lambda J}\,,$$

$$\mathrm{E}[Q] = \frac{\lambda^2 J_H}{\mathcal{E} + \lambda J}\,,$$

$$\mathrm{E}[W_q \mid \mathrm{Ab}] = \frac{J + \lambda J_H - n\mu J_1}{(\lambda - n\mu)J + 1}\,,$$

$$\mathrm{P}\{W_q > t\} = \frac{\lambda \bar{G}(t) J(t)}{\mathcal{E} + \lambda J}\,,$$

$$\mathrm{E}[W_q \mid W_q > t] = \frac{J_H(t) - (H(t) - t\bar{G}(t)) \cdot J(t)}{\bar{G}(t) J(t)}\,,$$

$$\mathrm{P}\{\mathrm{Ab} \mid W_q > t\} = \frac{\lambda - n\mu - G(t)}{\lambda \bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda \bar{G}(t) J(t)}\,.$$

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

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

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^*:

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

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

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

Asymptotic distribution of wait:

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

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:

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

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:

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

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

Back to "Why does Erlang-A Work?"

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

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

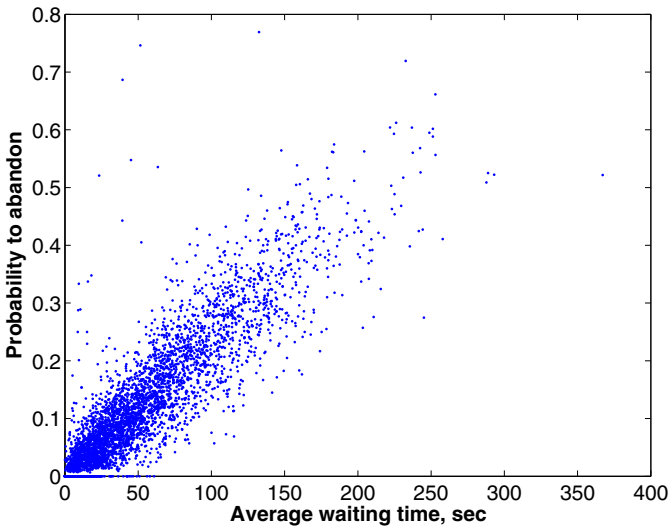
Practically: Why do (stochastic-ignorant) Call Centers work?

"The right answer for the wrong reason"

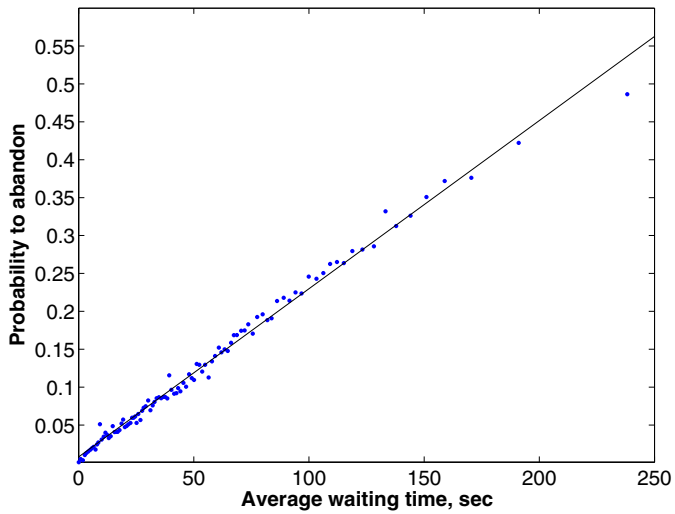
General Patience: Fitting Erlang-A

Israeli Bank: Yearly Data

Hourly Data



Aggregated



Theory:

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

Recipe:

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

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 $\lambda = \mathrm{E}\lambda(t - S_e)$.
     ($S_e \stackrel{d}{=}$ residual-service: congestion-lag behind peak-load.)
2. Square-Root Staffing: N_t
   - Let $R_t = \mathrm{E}\lambda(t - S_e) \times \mathrm{E}S$ be the **Offered-Load** at time t
     (R_t = Number-in-system in a corresponding $M_t/G/\infty$.)
   - Staff $N_t = R_t + \beta\sqrt{R_t}$.

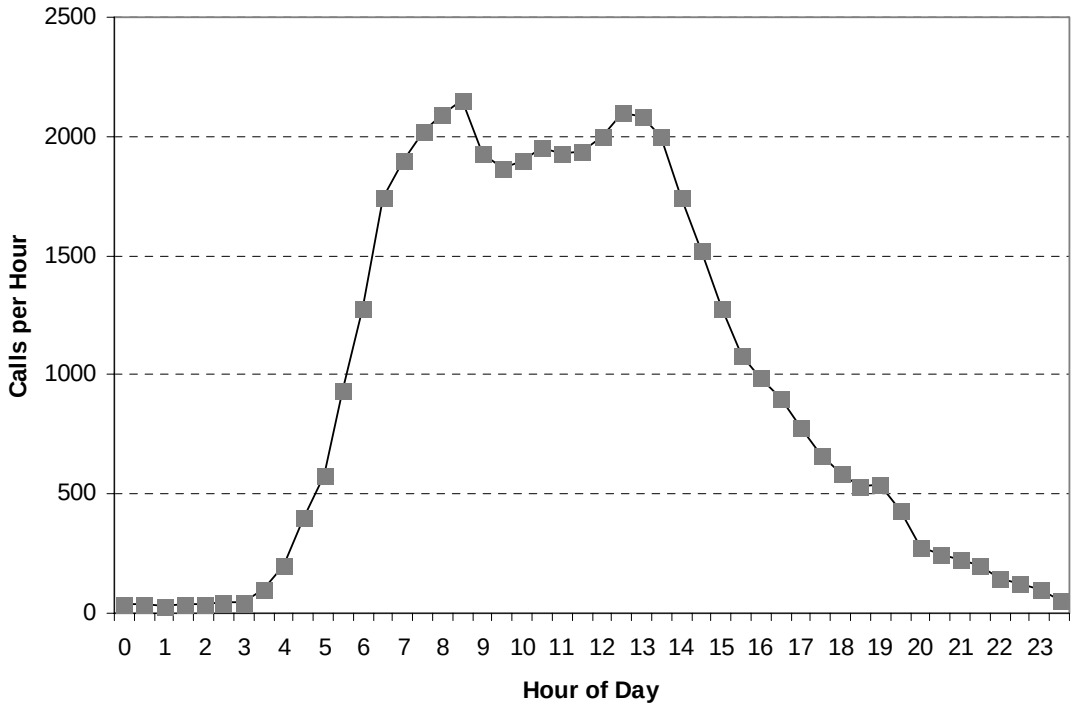
Serendipity: **Time-stable** performance, supported by **ISA** = Iterative Staffing Algorithm, and QED diffusion limits ($M_t/M/N + M, \ \mu = \theta$).

Example: "Real" Call Center

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

Time-Varying (two-hump) arrival functions common

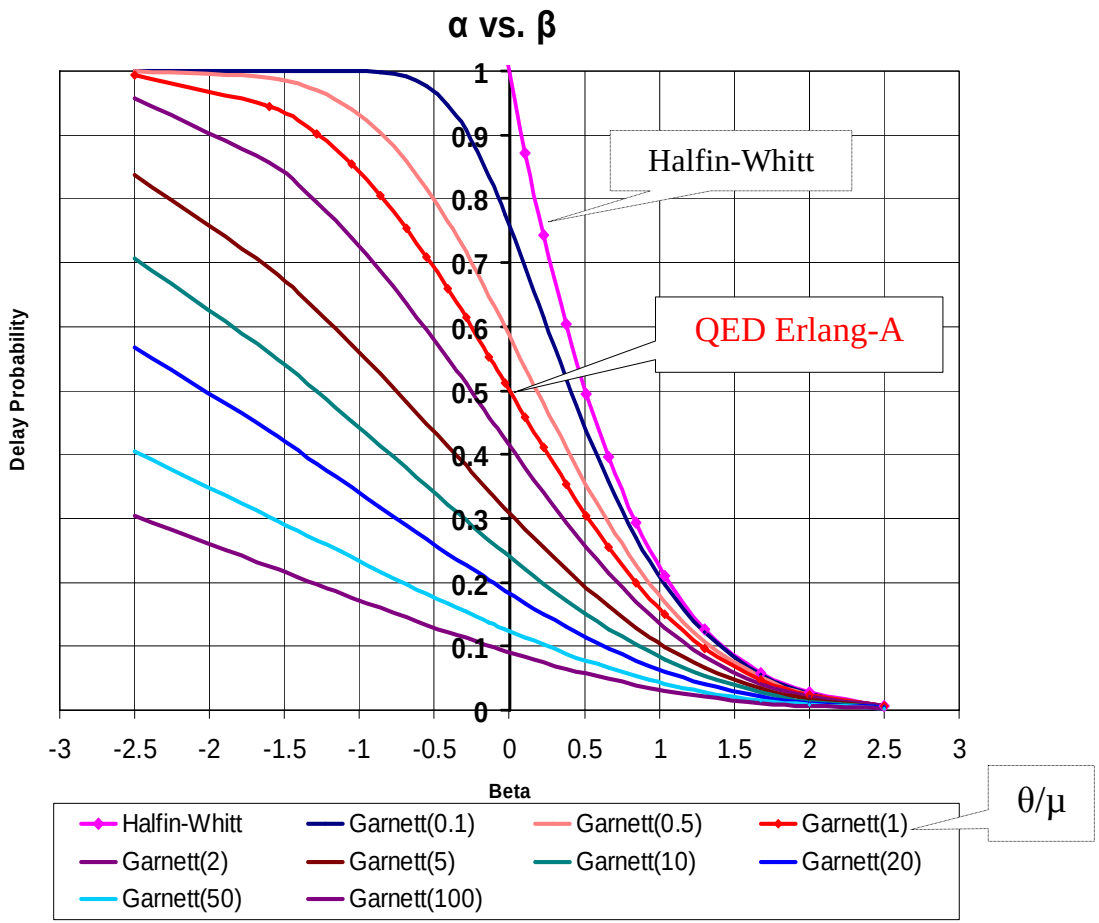
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)



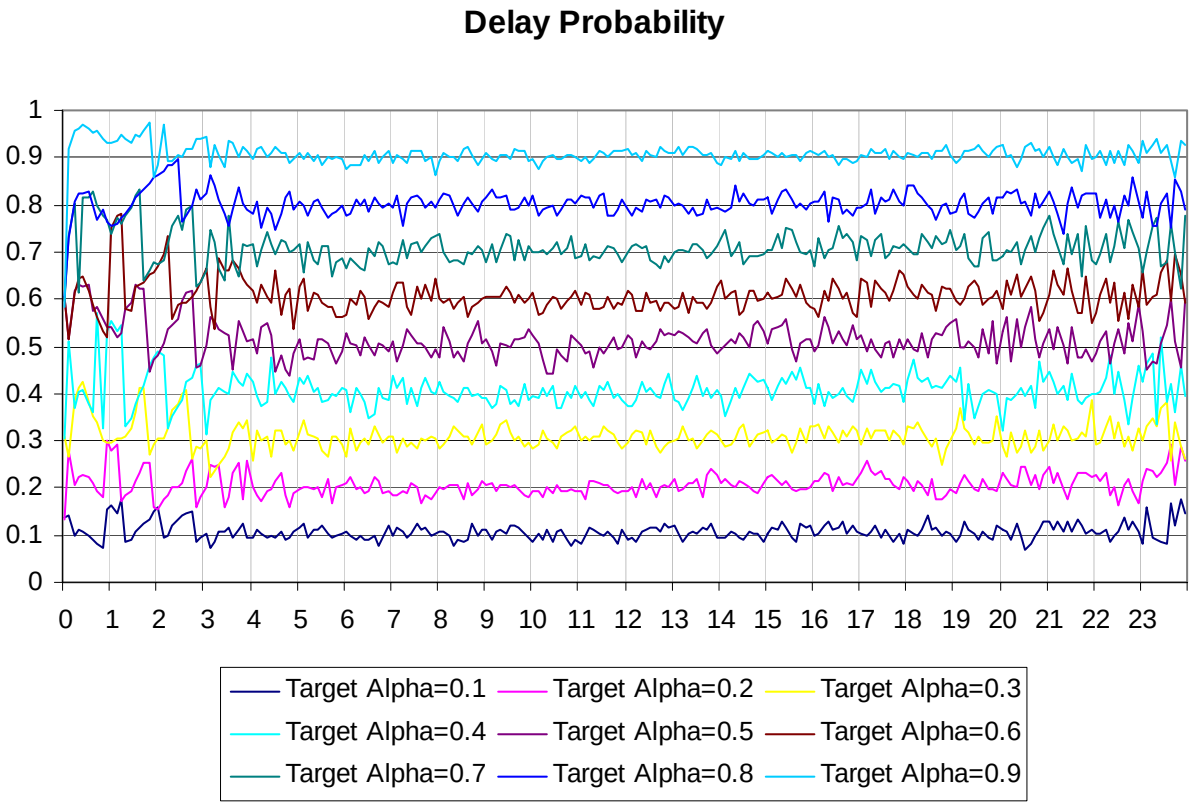
Assume: Service and abandonment times are both

Exponential, with mean 0.1 (6 min.)

HW/GMR Delay Functions

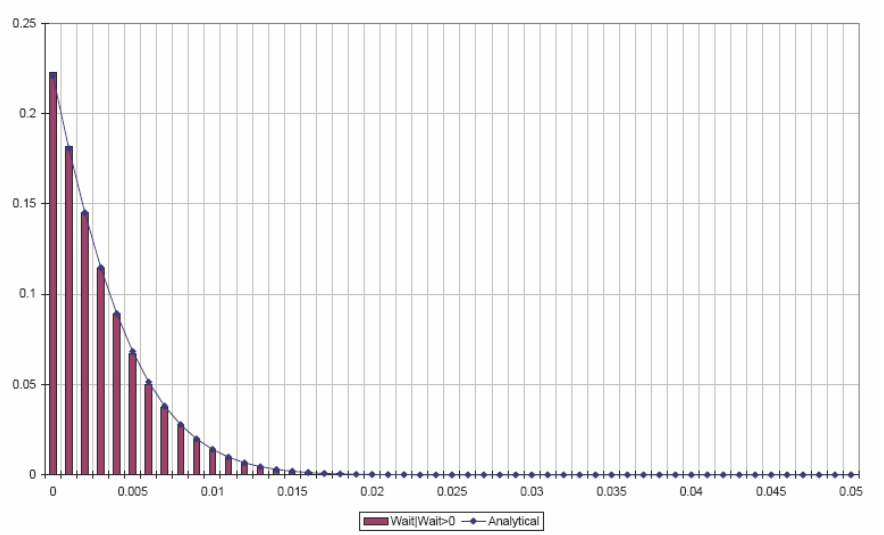


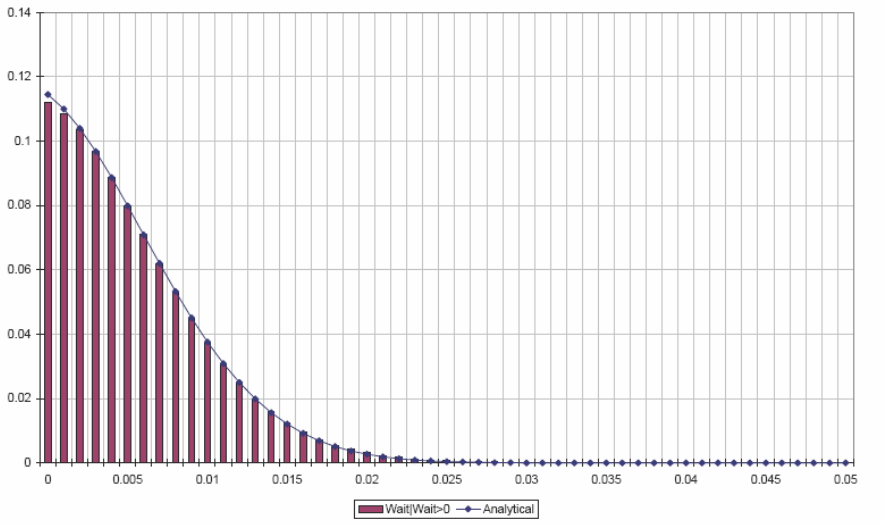
Delay Probability α

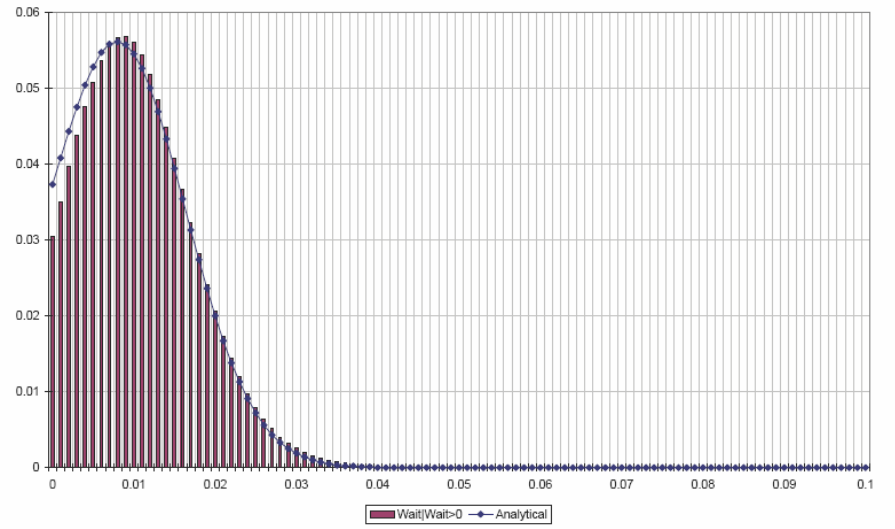


Real Call Center: Empirical waiting time, given positive wait

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







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

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ **?** θ

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

$\mu = \theta$: $L_t \overset{d}{=} \text{Poisson}(R_t) \overset{d}{\approx} N(R_t, R_t)$, 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 $L_t \approx R_t + Z\sqrt{R_t}$, $Z \overset{d}{=} N(0,1)$

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

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

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

Indeed, but in fact TIME-STABLE PERFORMANCE

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

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

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?**

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** $\boldsymbol{t}$ with a **stationary** threshold-controlled $\boldsymbol{M^J/G/N_t + G}$, in which $\lambda_j = \mathrm{E}\lambda_j(\boldsymbol{t - S_e})$.

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