# Data-Based Science for Service Engineering and Management

## or: Empirical Adventures in Call-Centers and Hospitals

Avi Mandelbaum

Technion, Haifa, Israel

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



## Research Partners

- **Students**: **Aldor*, Baron*, Carmeli, Feldman*, Garnett*, Gurvich*, Huang, Khudiakov*, Maman*, Marmor*, Reich, Rosenshmidt*, Shaikhet*, Senderovic, Tseytlin*, Yom-Tov*, Yuviler, Zaied, Zeltyn*, Zychlinski, Zohar*, Zviran*, . . .**
- **Theory**: **Armony, Atar, Gurvich, Jelenkovic, Kaspi, Massey, Momcilovic, Reiman, Shimkin, Stolyar, Wasserkrug, Whitt, Zeltyn, . . .**
- **Industry**: **Mizrahi Bank (A. Cohen, U. Yonissi), Rambam Hospital (R. Beyar, S. Israelit, S. Tzafrir), IBM Research (OCR Project), Hapoalim Bank (G. Maklef, T. Shlasky), Pelephone Cellular, . . .**
- **Technion SEE Center / Laboratory**: **Feigin; Trofimov, Nadjharov, Gavako, Kutsy; Liberman, Koren, Plonsky, Senderovic**; Research Assistants, . . .
- **Empirical/Statistical Analysis**: **Brown, Gans, Zhao; Shen; Ritov, Goldberg; Gurvich, Huang, Liberman; Armony, Marmor, Tseytlin, Yom-Tov; Zeltyn, Nardi, Gorfine, . . .**



## History, Resources (Downloadable)

- Math. + C.S. + Stat. + O.R. + Mgt. ⇒ **IE** ($\geq$ 1990)
- **Teaching**: **"Service-Engineering" Course** ($\geq$ 1995): http://ie.technion.ac.il/serveng - **website** http://ie.technion.ac.il/serveng/References/teaching_**paper**.pdf
- **Call-Centers Research** ($\geq$ 2000) e.g. <**Call Centers**> in Google-Scholar
- **Healthcare Research** ($\geq$ 2005) e.g. **OCR Project**: IBM + Rambam Hospital + Technion
- **The Technion SEE Center** ($\geq$ 2007)



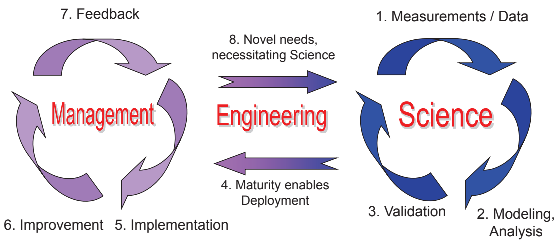
## The Case for Service Science / Engineering

- **Service Science** / **Engineering** (vs. Management) are emerging **Academic Disciplines**. For example, universities (world-wide), IBM (SSME, a là Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ...
- Models that explain fundamental phenomena, which are **common** across applications:
  - **Call Centers**
  - **Hospitals**
  - **Transportation**
  - Justice, Fast Food, Police, Internet, . . .
- Simple models at the Service of Complex Realities (Human) Note: Simple yet rooted in **deep analysis**.
- Mostly **What Can Be Done** vs. **How To**



## Title: Expands the Scientific Paradigm

Physics, Biology, . . . : Measure, Model, Experiment, Validate, Refine. **Human-complexity** triggered above in Transportation, Economics. Starting with **Data**, expand to:

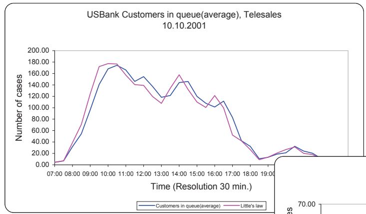


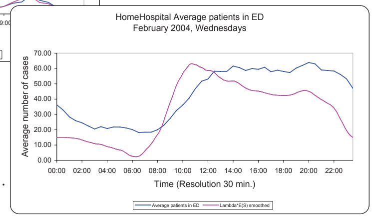

e.g. Validate, refute or discover **congestion laws** (Little, PASTA, SSC, ?, ?,...), in call centers and hospitals



## Little's Law: Call Center & Emergency Department

**Time-Gap**: **# in System** lags behind **Piecewise-Little** ($L = \lambda \times W$)



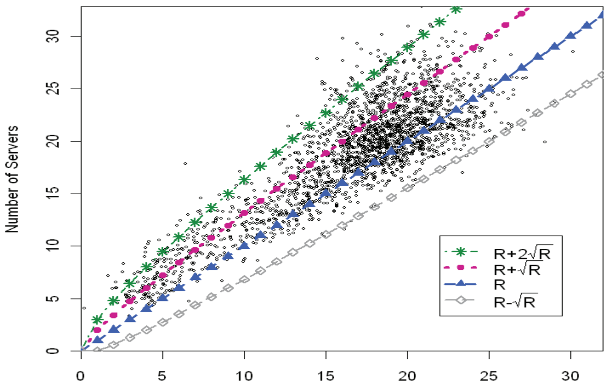

⇒ **Time-Varying Little's Law**

- **Berstemas & Mourtzinou;**
- **Fralix, Riano, Serfozo;** . . .



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

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



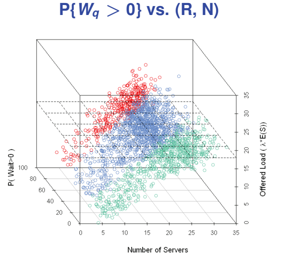

**2205 half-hour intervals in an Israeli Call Center**



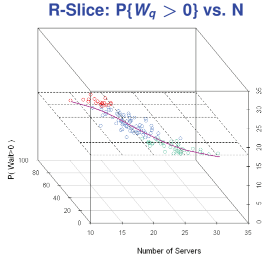
## QED Call Center: Performance

**Large Israeli Bank**

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




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




**3 Operational Regimes**:

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



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

| Erlang-A | Conventional scaling | | | MS scaling | | | | NDS scaling | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ fixed | Sub | Critical | Super | QD | QED | ED | ED+QED | Sub | Critical | Super |
| Offered load per server | $\frac{1}{1+\delta} < 1$ | $1 - \frac{\beta}{\sqrt{n}} \approx 1$ | $\frac{1}{1-\gamma} > 1$ | $\frac{1}{1+\delta}$ | $1 - \frac{\beta}{\sqrt{n}}$ | $\frac{1}{1-\gamma}$ | $\frac{1}{1-\gamma} - \beta\sqrt{\frac{1}{n(1-\gamma)^3}}$ | $\frac{1}{1+\delta}$ | $1 - \frac{\beta}{n}$ | $\frac{1}{1-\gamma}$ |
| Arrival rate $\lambda$ | $\frac{\mu}{1+\delta}$ | $\mu - \frac{\beta}{\sqrt{n}}\mu$ | $\frac{\mu}{1-\gamma}$ | $\frac{n\mu}{1+\delta}$ | $n\mu - \beta\mu\sqrt{n}$ | $\frac{n\mu}{1-\gamma}$ | $\frac{n\mu}{1-\gamma} - \beta\mu\sqrt{\frac{n}{(1-\gamma)^3}}$ | $\frac{n\mu}{1+\delta}$ | $n\mu - \beta\mu$ | $\frac{n\mu}{1-\gamma}$ |
| Number of servers | 1 | | | $n$ | | | | $n$ | | |
| Time-scale | $n$ | | | 1 | | | | $n$ | | |
| Abandonment rate | $\theta/n$ | | | $\theta$ | | | | $\theta/n$ | | |
| Staffing level | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu}(1+\frac{\beta}{\sqrt{n}})$ | $\frac{\lambda}{\mu}(1-\gamma)$ | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}$ | $\frac{\lambda}{\mu}(1-\gamma)$ | $\frac{\lambda}{\mu}(1-\gamma) + \beta\sqrt{\frac{\lambda}{\mu}}$ | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu} + \beta$ | $\frac{\lambda}{\mu}(1-\gamma)$ |
| Utilization | $\frac{1}{1+\delta}$ | $1 - \sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{\sqrt{n}}$ | 1 | $\frac{1}{1+\delta}$ | $1 - \sqrt{\frac{\theta}{\mu}}\frac{(1-\alpha_2)\hat{\beta}+\alpha_2 h(\hat{\beta})}{\sqrt{n}}$ | 1 | 1 | $\frac{1}{1+\delta}$ | $1 - \sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{n}$ | 1 |
| $\mathbb{E}(Q)$ | $\frac{\alpha_1}{\delta}$ | $\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\delta^2}\varrho^n\frac{1}{\sqrt{n}}$ | $\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $\frac{n\mu}{\theta(1-\gamma)}(\gamma - \frac{\beta}{\sqrt{n(1-\gamma)}})$ | $o(1)$ | $n\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $\frac{n^2\mu\gamma}{\theta(1-\gamma)}$ |
| $\mathbb{P}(Ab)$ | $\frac{1}{n}\frac{1+\delta}{\delta}\frac{\theta}{\mu}\alpha_1$ | $\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$ | $\gamma$ | $\frac{1}{\sqrt{2\pi}}\frac{\theta}{\mu}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$ | $\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$ | $\gamma$ | $\gamma - \frac{\beta\sqrt{1-\gamma}}{\sqrt{n}}$ | $o(\frac{1}{n^2})$ | $\frac{1}{n}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$ | $\gamma$ |
| $\mathbb{P}(W_q > 0)$ | $\alpha_1 \in (0,1)$ | $\approx 1$ | | $\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\delta}\varrho^n\frac{1}{\sqrt{n}} \approx 0$ | $\alpha_2 \in (0,1)$ | $\approx 1$ | $\approx 1$ | $\approx 0$ | $\approx 1$ | |
| $\mathbb{P}(W_q > T)$ | $\alpha_1 e^{-\frac{\delta}{1+\delta}\mu t}$ | $1 + O(\frac{1}{\sqrt{n}})$ | $1 + O(\frac{1}{n})$ | $\approx 0$ | | $\bar{G}(T)1_{\{G(T)<\gamma\}}$ | $\alpha_3$, if $G(T) = \gamma$ | $\approx 0$ | $\frac{\bar{\Phi}(\hat{\beta}+\sqrt{\theta\mu}T)}{\bar{\Phi}(\hat{\beta})}$ | $1 + O(\frac{1}{n})$ |
| Congestion $\frac{\mathbb{E}W_q}{\mathbb{E}S}$ | $\alpha_1\frac{1+\delta}{\delta}$ | $\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $n\mu\gamma/\theta$ | $\frac{1}{\sqrt{2\pi}}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$ | $\frac{1}{\sqrt{n}}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$ | $\mu\int_0^{x^*}\bar{G}(s)ds$ | $\mu\int_0^{x^*}\bar{G}(s)ds - \frac{\mu\beta\sqrt{1-\gamma}}{h_G(x^*)\sqrt{n}}$ | $o(\frac{1}{n})$ | $\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $n\mu\gamma/\theta$ |

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

## Prerequisite I: Data

**Averages Prevalent** (and could be useful / interesting).
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 visit in a hospital, or browsing in a website, or . . .), 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, IVR/VRU)
- **Hospitals** (Emergency Departments, . . .)
- Expanding:
  - Hospitals, via **RFID**
  - Operational + Financial + Contents (Marketing, Clinical)
  - Internet, Chat (multi-media)



## Pause for a Commercial: The Technion SEE Center

Technion SEE Center
Service Enterprise Engineering



## Technion SEE = Service Enterprise Engineering

**SEELab**: **Data-repositories for research and teaching**

- For example:
  - Bank Anonymous: **1 years, 350K calls by 15 agents** - in 2000. **Brown, Gans, Sakov, Shen, Zeltyn, Zhao** (JASA), paved the way for:
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**.
  - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**.
  - Israeli Hospital: **4 years, 1000 beds; 8 ED's- Sinreich's data**.

**SEEStat**: Environment for graphical **EDA** in real-time

- **Universal Design, Internet Access, Real-Time Response**.

**SEEServer**: **Free for academic use**
Register, then access (presently) U.S. Bank and Bank Anonymous.
**Visitor**: run mstsc, seeserver.iem.technion.ac.il ; Self-Tutorial



## Tutorial Cover; State-Space Collapse from Tutorial

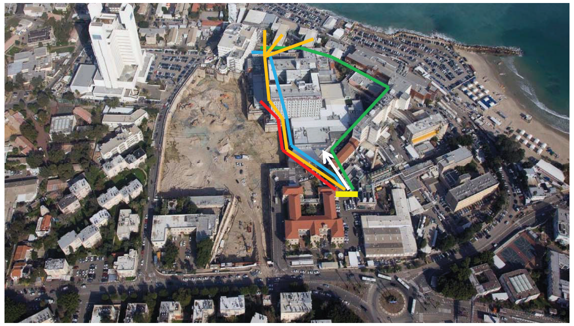
4 overheads:

- Cover (make sure relevant to the lecture (e.g. APS, HKUST)
- Page 2 (again, make sure relevant to the lecture)
- Contents (with Stat-Space Collapse yellowed)
- The page with State-Space Collapse.



## eg. RFID-Based Data: Mass Casualty Event (MCE)

**Drill: Chemical MCE, Rambam Hospital, May 2010**



Focus on **severely wounded** casualties ($\approx$ 40 in drill)
**Note**: 20 observers support real-time control (helps validation)



## Data Cleaning: MCE with RFID Support

| Data-base | | | | Company report | | comment |
|---|---|---|---|---|---|---|
| **Asset id** | order | **Entry** date | **Exit** date | Entry date | Exit date | |
| 4 | 1 | 1:14:07 PM | | 1:14:00 PM | | |
| 6 | 1 | 12:02:02 PM | 12:33:10 PM | 12:02:00 PM | 12:33:00 PM | |
| 8 | 1 | 11:37:15 AM | 12:40:17 PM | 11:37:00 AM | | **exit is missing** |
| 10 | 1 | 12:23:32 PM | 12:38:23 PM | 12:23:00 PM | | |
| 12 | 1 | 12:12:47 PM | 12:35:33 PM | | 12:35:00 PM | **entry is missing** |
| 15 | 1 | 1:07:15 PM | | 1:07:00 PM | | |
| 16 | 1 | 11:18:19 AM | 11:31:04 AM | 11:18:00 AM | 11:31:00 AM | |
| 17 | 1 | 1:03:31 PM | | 1:03:00 PM | | |
| 18 | 1 | 1:07:54 PM | | 1:07:00 PM | | |
| 19 | 1 | 12:01:58 PM | | 12:01:00 PM | | |
| 20 | 1 | 11:37:21 AM | 12:57:02 PM | 11:37:00 AM | 12:57:00 PM | |
| 21 | 1 | 12:01:16 PM | 12:37:16 PM | 12:01:00 PM | | |
| 22 | 1 | 12:04:31 PM | 12:20:40 PM | | | first customer is missing |
| 22 | 2 | 12:27:37 PM | | 12:27:00 PM | | |
| 25 | 1 | 12:27:35 PM | 1:07:28 PM | 12:27:00 PM | 1:07:00 PM | |
| 27 | 1 | 12:06:53 PM | | 12:06:00 PM | | |
| 28 | 1 | 11:21:34 AM | 11:41:06 AM | 11:41:00 AM | 11:53:00 AM | **exit time instead of entry time** |
| 29 | 1 | 12:21:06 PM | 12:54:29 PM | 12:21:00 PM | 12:54:00 PM | |
| 31 | 1 | 11:40:54 AM | 12:30:16 PM | 11:40:00 AM | 12:30:00 PM | |
| 31 | 2 | 12:37:57 PM | 12:54:51 PM | 12:37:00 PM | 12:54:00 PM | |
| 32 | 1 | 11:27:11 AM | 12:15:17 PM | 11:27:00 AM | 12:15:00 PM | |
| 33 | 1 | 12:05:50 PM | 12:13:12 PM | 12:05:00 PM | 12:15:00 PM | wrong exit time |
| 35 | 1 | 11:31:48 AM | 11:40:50 AM | 11:31:00 AM | 11:40:00 AM | |
| 36 | 1 | 12:06:23 PM | 12:29:30 PM | 12:06:00 PM | 12:29:00 PM | |
| 37 | 1 | 11:31:50 AM | 11:48:18 AM | 11:31:00 AM | 11:48:00 AM | |
| 37 | 2 | 12:59:21 PM | | 12:59:00 PM | | |

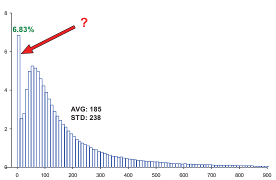
Imagine **"Cleaning" 60,000+ customers per day** (call centers) !



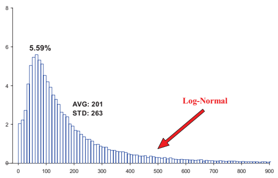
## Beyond Averages: The Human Factor

**Histogram of Service-Time in a (Small Israeli) Bank, 1999**

**January-October**




**November-December**



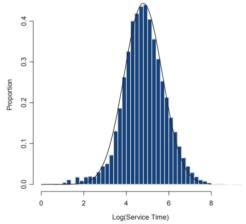

- **6.8% Short-Services:** Agents' "Abandon" (improve bonus, rest), (mis)lead by **incentives**
- **Distributions** must be measured (in **seconds** = **natural scale**)
- **LogNormal** service times common in call centers



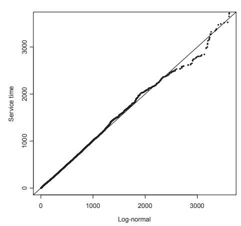
## Validating LogNormality of Service-Duration

**Israeli Call Center, Nov-Dec, 1999**

**Log(Service Times)**




**LogNormal QQPlot**

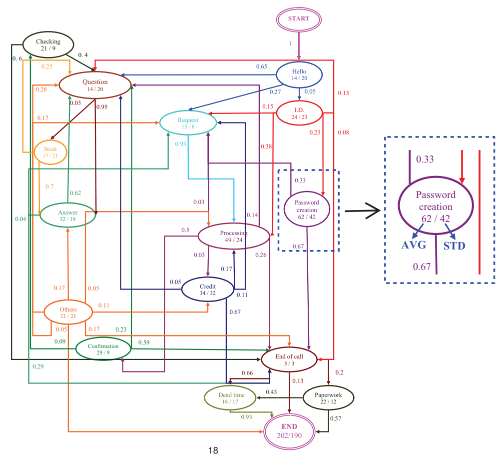



- **Practically Important**: (mean, std)(log) characterization
- **Theoretically Intriguing**: Why LogNormal ? Naturally multiplicative but, in fact, also **Infinitely-Divisible** (Generalized Gamma-Convolutions)
- Simple-model of a complex-reality? The **Service Process:**



## (Telephone) Service-Process = "Phase-Type" Model

**Retail Service (Israeli Bank)**



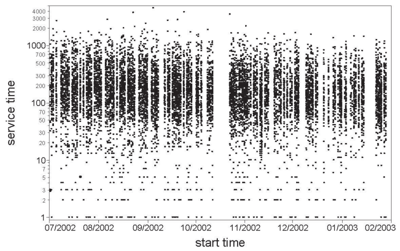

**Statistics OR IE**

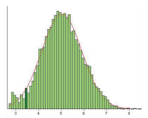
## Individual Agents: Service-Duration, Variability

w/ **Gans, Liu, Shen & Ye**

**Agent 14115**

**Service-Time Evolution: 6 month** — **Log(Service-Time)**



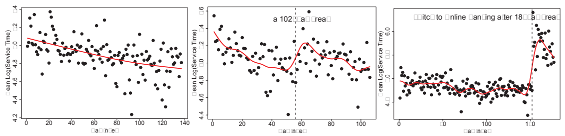



- **Learning**: Noticeable decreasing-trend in service-duration
- **LogNormal** Service-Duration, individually and collectively

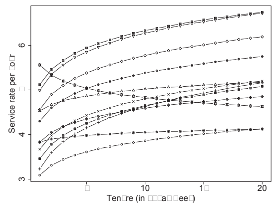


## Individual Agents: Learning, Forgetting, Switching

**Daily-Average Log(Service-Time), over 6 months**
**Agents 14115, 14128, 14136**



**Weakly Learning-Curves for 12 Homogeneous(?) Agents**





## Why Bother?

In large call centers:
**+One Second** to Service-Time implies **+Millions** in costs, annually

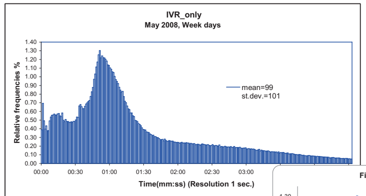
⇒ **Time and "Motion" Studies** (**Classical IE** with New-age IT)

- **Service-Process Model**: Customer-Agent Interaction
  - **Work Design** (w/ **Khudiakov**)
    eg. **Cross-Selling**: higher profit vs. longer (costlier) services;
    Analysis yields (congestion-dependent) cross-selling protocols
  - **"Worker" Design** (w/ **Gans, Liu, Shen & Ye**)
    eg. **Learning, Forgetting**, . . . : Staffing & individual-performance prediction, in a heterogenous environment
- **IVR-Process Model**: Customer-Machine Interaction
  **75% bank-services**, poor design, yet scarce research;
  Same approach, automatic (easier) data (**w/ Yuviler**)

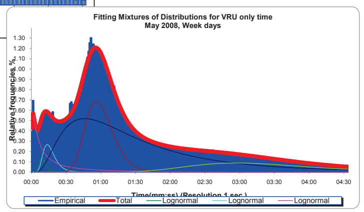


## IVR-Time: Histograms

**Israeli Bank: IVR/VRU Only, May 2008**

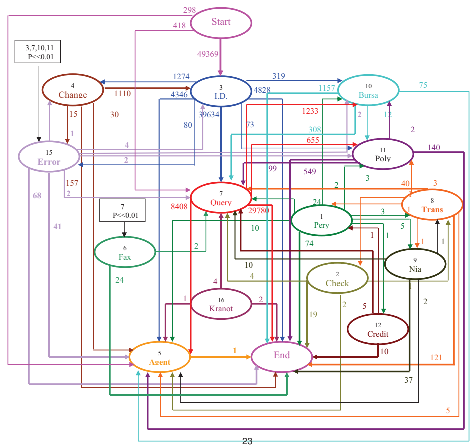



**Mixture: 7 LogNormals**






## IVR-Process: "Phase-Type" Model






## Started with Call Centers, Expanded to Hospitals

**Call Centers - U.S. (Netherlands) Stat.**

- $200 – $300 billion annual expenditures (0.5)
- 100,000 – 200,000 call centers (1500-2000)
- "Window" into the company, for better or worse
- Over 3 million agents = **2% – 4% workforce** (100K)

**Healthcare** - similar and unique challenges:

- Cost-figures far more staggering
- Risks much higher
- ED (initial focus) = hospital-window
- Over 3 million nurses



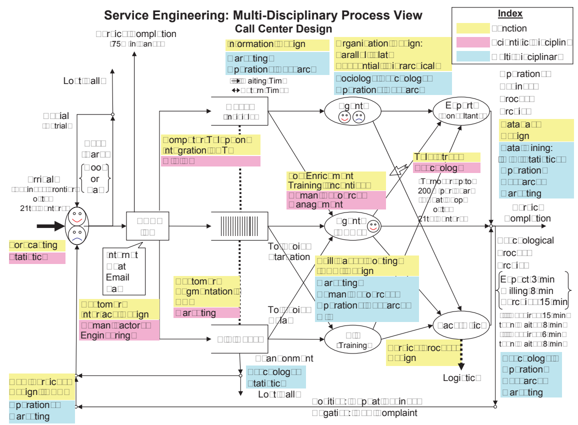
## Call-Center Environment: Service Network



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



## Call-Center Network: Gallery of Models



## Call-Center Network: Gallery of Models

Add marks of topics to focus on



## Skills-Based Routing in Call Centers

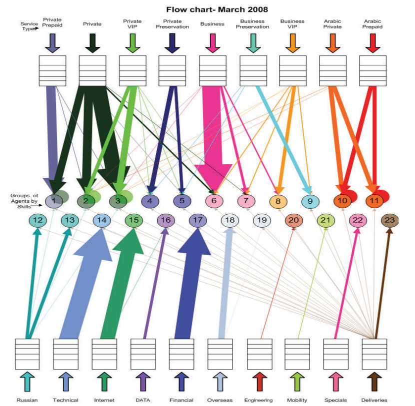
**EDA and OR**, with **I. Gurvich and P. Liberman**

Mktg. ⇒

OR ⇒

HRM ⇒

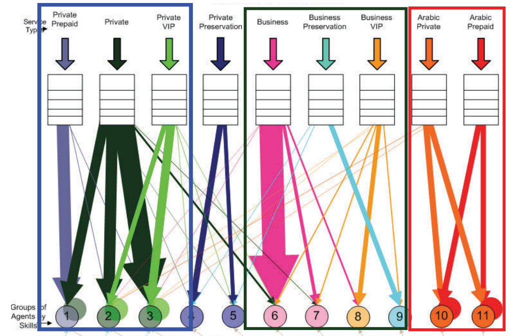
MIS ⇒






## SBR Topologies: I; V, Reversed-V; N, X; W, M

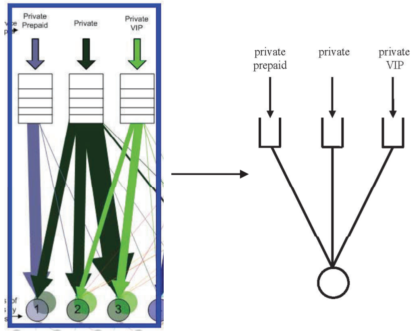
### Israeli Cellular, March 2008






## SBR: Class-Dependent Services

**"Reduction" to V-Topology** (**Equivalent Brownian Control**)

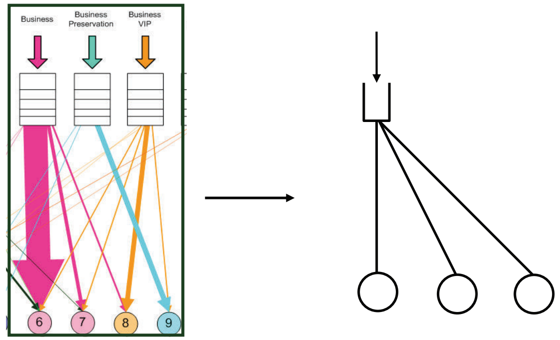



PhD's: **Tezcan**, **Dai**; **Shaikhet**, **w/ Atar**; **Gurvich**, **Whitt**



## SBR: Pool-Dependent Services

**"Reduction" to Reversed-V and I** (**Equivalent Brownian Control**)

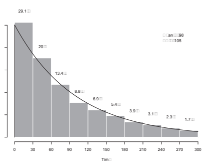



PhD's: **Tezcan**, **Dai**; **Shaikhet**, **w/ Atar**; **Gurvich**, **Whitt**

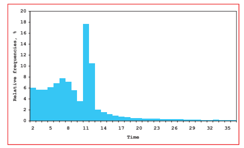


## Waiting Times in a Call Center (Theory?)

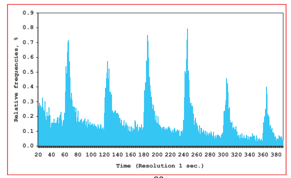
**Exponential in Heavy-Traffic (min.)**
Small Israeli Bank



**Routing via Thresholds (sec.)**
Large U.S. Bank



**Scheduling Priorities (sec) (later: Hospital LOS, hr.)**
Medium Israeli Bank






## ER / ED Environment: Service Network

**Acute (Internal, Trauma)**

**Walking**

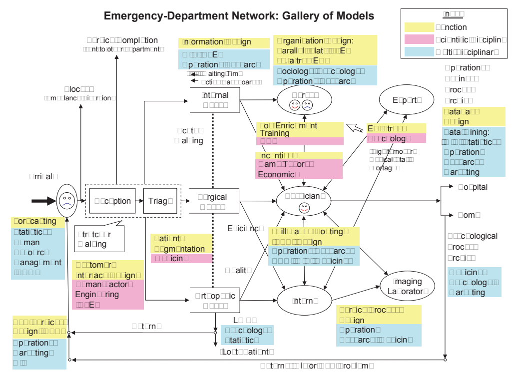
**Multi-Trauma**



## Queueing in a "Good" Beijing Hospital, at 6am




## Emergency-Department Network: Gallery of Models




▶ **Forecasting**, Abandonment = **LWBS**, SBR ≈ **Flow Control**

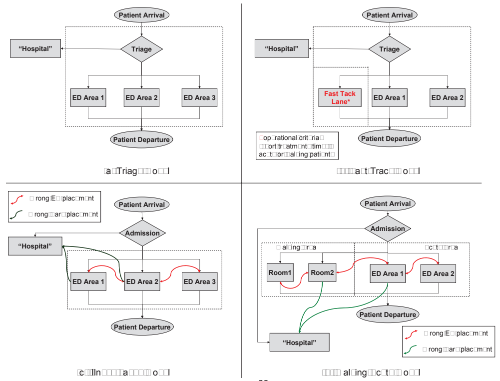
## Emergency-Department Network: Gallery of Models

Add ED-to-IW routing



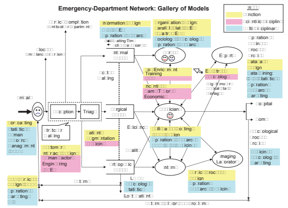
## ED Design, with B. Golany, Y. Marmor, S. Israelit

**Routing**: **Triage (Clinical), Fast-Track (Operational),** . . . (via DEA)
**eg. Fast Track most suitable when elderly dominate**





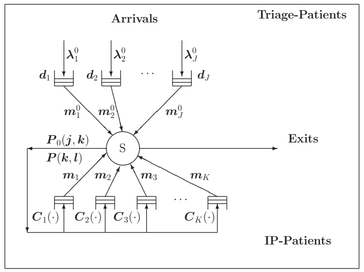

## Emergency-Department Network: Flow Control




- **Queueing-Science**, w/ **Armony, Marmor, Tseytlin, Yom-Tov**
- **Fair ED-to-IW Routing** (Patients vs. Staff), w/ **Momcilovic, Tseytlin**
- **Triage vs. In-Process / Release** in EDs, w/ **Carmeli, Huang, Shimkin**
- **Workload and Offered-Load** in **Fork-Join Networks**, w/ **Kaspi, Zaeid**
- **Synchronization Control** of Fork-Join Networks, w/ **Atar, Zviran**
- **Staffing Time-Varying Q's with Re-Entrant Customers**, w/ **Yom-Tov**



## ED Patient Flow: The Physicians View




- **Goal**: Adhere to **Triage-Constraints**, then **process/release In-Process** Patients
- **Model** = Multi-class Q with Feedback: Min. convex **congestion costs** of IP-Patients, s.t. **deadline constraints** on Triage-Patients.
- **Solution**: In **conventional** heavy-traffic, **asymptotic least-cost** s.t. **asymptotic compliance**, via threshold (**w/ B. Carmeli, J. Huang, S. Israelit, N. Shimkin**; as in Plambeck, Harrison, Kumar, who applied admission control).



## Operational Fairness

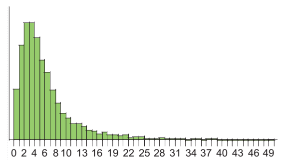
1. **"Punishing" fast wards in ED-to-IW Routing**:
   - Parallel IWs: similar clinically , differ operationally
   - Problem: Short Length-of-Stay goes hand in hand with high **bed-occupancy, bed-turnover**, yet clinically apt: **unfair!**
   - Solution: Both nurses and managers content, w/ **P. Momcilovic and Y. Tseytlin** (3 time-scales: hour, day, week; "compare" with call-centers SBR)
2. **Balancing Load across Maternity Wards**:
   - 2 Maternity Wards: 1 = **pre-birth**, 2 = **post-birth** complications
   - Problem: Nurses think the **"others-work-less"**: **unfair!**
   - Goal: Balance workload, mostly via normal births
   - Challenge: Workload is **Operational, Cognitive, Emotional**
     - **Operational**: Work content of a task, in time-units
     - **Emotional**: e.g. Mother and fetus-in-stress, suddenly fetus dies

⇒ Need **help**: **A. Rafaeli** & students (**Psychology**) - Ongoing

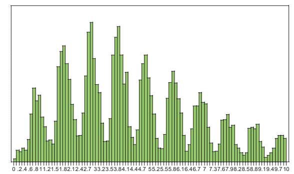


## LogNormal & Beyond: Length-of-Stay in a Hospital

**Israeli Hospital, in Days: LN**

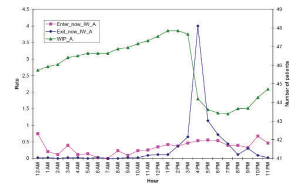


**Israeli Hospital, in Hours: Mixture**



**Explanation**: Patients released around **3pm** (1pm in Singapore)

**Why Bother ?**
- Hourly Scale: Staffing,. . .
- Daily: Flow / Bed Control,. . .

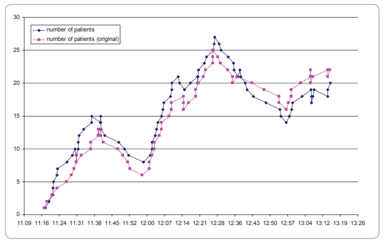




## Prerequisite II: Models (Fluid Q's)

**"Laws of Large Numbers"** capture **Predictable** Variability
**Deterministic** Models: Scale Averages-out **Stochastic Individualism**

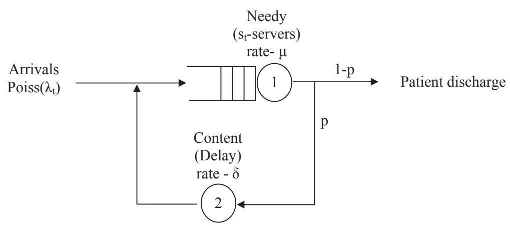
# **Severely-Wounded Patients, 11:00-13:00 (Censored LOS)**




- Paths of doctors, nurses, patients (100+, **1 sec.** resolution) eg. (could) Help predict "**What if** 150+ casualties severely wounded ?"
- **Transient** Q's:
  - Control of **Mass Casualty Events** (w/ **I. Cohen, N. Zychlinski**)
  - **Chemical MCE** = **Needy-Content Cycles** (w/ **G. Yom-Tov**)



## The Basic Service-Network Model: Erlang-R




**Erlang-R** (IE: Repairman Problem 50's; CS: Central-Server 60's) =
**2-station "Jackson" Network** = (M/M/S, M/M/$\infty$) :

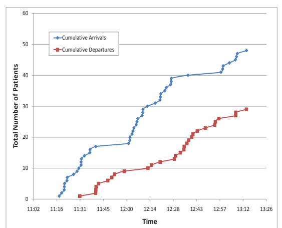
- $\lambda(t)$ – **Time-Varying** Arrival rate
- $S(\cdot)$ – Number of **Servers** (Nurses / Physicians).
- $\mu$ – **Service** rate ($E$[Service] $= \frac{1}{\mu}$)
- $p$ – **ReEntrant** (Feedback) fraction
- $\delta$ – **Content-to-Needy** rate ($E$[Content] $= \frac{1}{\delta}$)



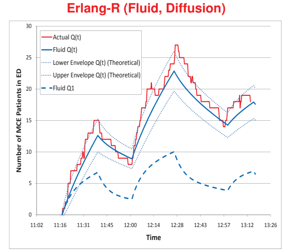
## Erlang-R: Fitting a Simple Model to a Complex Reality

### Chemical MCE Drill (Israel, May 2010)

**Arrivals & Departures (RFID)**




**Erlang-R (Fluid, Diffusion)**




- **Recurrent/Repeated** services in MCE Events: eg. Injection every 15 minutes
- **Fluid (Sample-path)** Modeling, via Functional Strong Laws of Large Numbers
- **Stochastic** Modeling, via Functional Central Limit Theorems
  - ED in **MCE**: Confidence-interval, usefully narrow for **Control**
  - ED in **normal** (**time-varying**) conditions: Personnel **Staffing**

## Prerequisite II: Models (Diffusion/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** (single-server queue): **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.

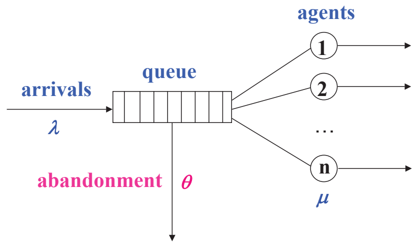
**Yet**, **heavily-loaded** queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

- **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 ) **?** **QED**



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




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

- $\lambda$ – **Arrival** rate (Poisson)
- $\mu$ – **Service** rate (Exponential; $E[S] = \frac{1}{\mu}$)
- $\theta$ – **Patience** rate (Exponential, $E[\text{Patience}] = \frac{1}{\theta}$)
- $n$ – Number of **Servers** (Agents).



## Testing the Erlang-A Primitives

- **Arrivals**: Poisson?
- **Service-durations**: Exponential?
- **(Im)Patience**: Exponential?
- Primitives independent (eg. Impatience and Service-Durations)?
- Customers / Servers Homogeneous?
- Service discipline FCFS?
- . . . ?

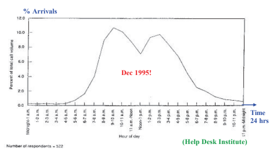
**Validation**: Support? Refute?



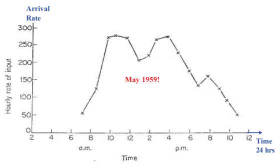
## Arrivals to Service

### Arrival-Rates to Three Call Centers

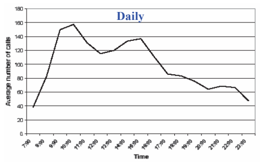
Dec. **1995** (U.S. 700 Helpdesks)




May **1959** (England)




November **1999** (Israel)



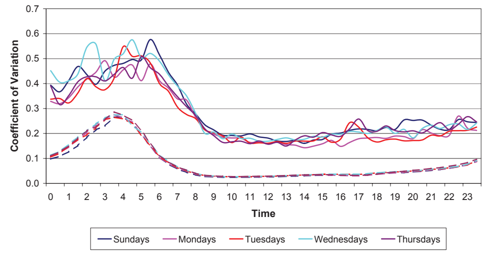

**Random Arrivals** "must be" (Axiomatically)
**Time-Inhomogeneous Poisson**



## Arrivals to Service: only Poisson-Relatives

**Arrival-Counts: Coefficient-of-Variation (CV)**, per 30 min.

**Israeli-Bank Call-Center, 263 regular days (4/2007 - 3/2008)**



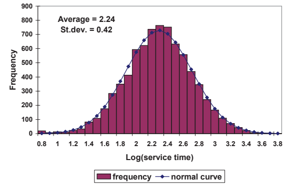

- **Poisson CV** (Dashed Line) = $1/\sqrt{\text{mean arrival-rate}}$
- Poisson CV's $\ll$ **Sampled CV's** (Solid) $\Rightarrow$ **Over-Dispersion**

$\Rightarrow$ **Modeling** (Poisson-Mixture) of and **Staffing** ( $> \sqrt{\cdot}$ ) against **Time-Varying Over-Dispersed** Arrivals (w/ **S. Maman & S. Zeltyn**)

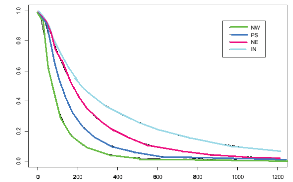


## Service Durations: LogNormal Prevalent

**Israeli Bank Log-Histogram**




**Service-Classes Survival-Functions**



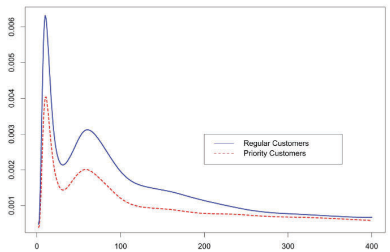

- **New** Customers: **2** min (NW);
- **Regulars**: **3** min (PS);
- **Stock**: **4.5** min (NE);
- Tech-Support: **6.5** min (IN).

- Service Durations are **LogNormal (LN)** and **Heterogeneous**



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

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

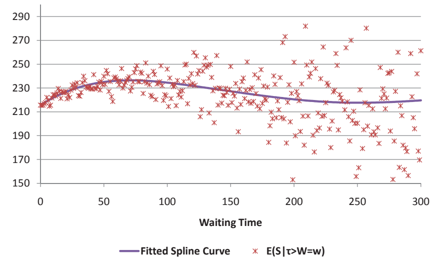



- **VIP** Customers are **more Patient** (Needy)
- **Peaks** of abandonment at times of **Announcements**
- Challenges: **Un-Censoring, Dependence (vs. KM), Smoothing** - requires **Call-by-Call Data**



## Dependent Primitives: Service- vs. Waiting-Time

**Average Service-Time as a function of Waiting-Time**
**U.S. Bank, Retail, Weedays, January-June, 2006**




$\Rightarrow$ Focus on **( Patience, Service-Time ) jointly** , w/ **Reich and Ritov**.
$E[S \mid \text{Patience} = w]$, $w \geq 0$: **Service-Time of the Unserved**.



## Erlang-A: Practical Relevance?

**Experience:**

- Arrival process **not pure Poisson** (time-varying, $\sigma^2$ 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 associated with long service; with **w/ M. Reich and Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment; dependently; **P. Khudiakov, M. Gorfine, P. Feigin**)
- · · · , and more.

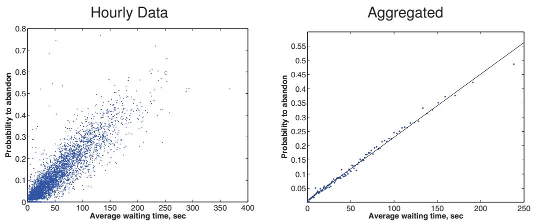
Question: **Is Erlang-A Relevant?**

**YES !** **Fitting a Simple Model to a Complex Reality**, both **Theoretically** and **Practically**

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

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

**% Abandonment vs. Average Waiting-Time**
**Bank Anonymous (JASA): Yearly Data**



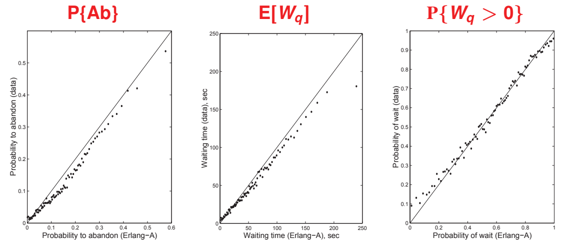

Graphs based on 4158 hour intervals.

Estimate of mean (im)patience: 250/0.55 sec. $\approx$ **7.5 minutes**.



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

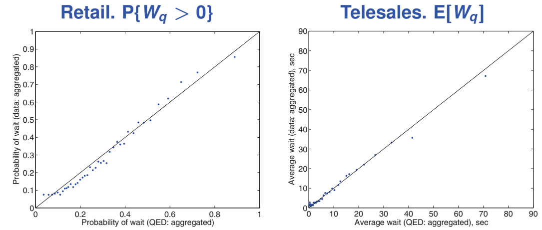
- **Bank Anonymous** Small Israeli Call-Center
- (Im)Patience ($\theta$) estimated via P{Ab} / E[$W_q$]
- Graphs: **Hourly Performance vs. Erlang-A Predictions**, during 1 year (aggregating groups with 40 similar hours).






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

**Large U.S. Bank**




**Partial success** – in **some** cases Erlang-A **does not work** well (Networking, SBR).

Ongoing **Validation** Project, **w/ Y. Nardi, O. Plonsky, S. Zeltyn**



## Erlang-A: Simple, but Not Too Simple

**Practical** (Data-Based) questions, started in **Brown et al. (JASA)**:

1. Fitting Erlang-A (**Validation**, **w/ Nardi, Plonsky, Zeltyn**).
2. Why does it practically work? justify **robustness**.
3. When does it fail? chart **boundaries**.
4. Generate needs for **new theory**.

**Theoretical Framework**: **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** servers:

- **Practically Important**: Relevant to **Healthcare** (First: F. de Véricourt and O. Jennings; w/ **G. Yom-Tov; Y. Marmor, S. Zeltyn; H. Kaspi, I. Zaeid**)
- **Theoretically Justifiable**: Gap-Analysis by **A. Janssen, J. van Leeuwaarden, B. Zhang, B. Zwart**.



## Operational Regimes: Conceptual Framework

***R*: Offered Load**

Def. $\boldsymbol{R}$ = Arrival-rate $\times$ Average-Service-Time = $\frac{\lambda}{\mu}$

eg. $\boldsymbol{R}$ = 25 calls/min. $\times$ 4 min./call = **100**

$\boldsymbol{N}$ = #Agents **?** **Intuition**, as $\boldsymbol{R}$ or $\boldsymbol{N}$ increase unilaterally.

**QD Regime:** $\boldsymbol{N \approx R + \delta R}$ , $\quad \boldsymbol{0.1 < \delta < 0.25}$ (eg. $\boldsymbol{N = 115}$)

- Framework developed in **O. Garnett**'s MSc thesis
- Rigorously: $(\boldsymbol{N - R})/\boldsymbol{R} \to \delta$, as $\boldsymbol{N}, \lambda \uparrow \infty$, with $\mu$ fixed.
- Performance: Delays are rare events

**ED Regime:** $\boldsymbol{N \approx R - \gamma R}$ , $\quad \boldsymbol{0.1 < \gamma < 0.25}$ (eg. $\boldsymbol{N = 90}$)

- Essentially **all** customers are delayed
- Wait same order as service-time; $\gamma$% Abandon (10-25%).

**QED Regime:** $\boldsymbol{N \approx R + \beta\sqrt{R}}$, $\quad \boldsymbol{-1 < \beta < +1}$ (eg. $\boldsymbol{N = 100}$)

- Erlang 1913-24, **Halfin & Whitt** 1981 (for Erlang-C)
- %Delayed between 25% and 75%
- E[Wait] $\propto \frac{1}{\sqrt{N}} \times$ E[Service] (**sec vs. min**); 1-5% Abandon.



## Operational Regimes: Rules-of-Thumb, w/ S. Zeltyn

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

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

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

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

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

**WFM**: How to determine specific staffing level $\boldsymbol{N}$ ? e.g. $\beta$.



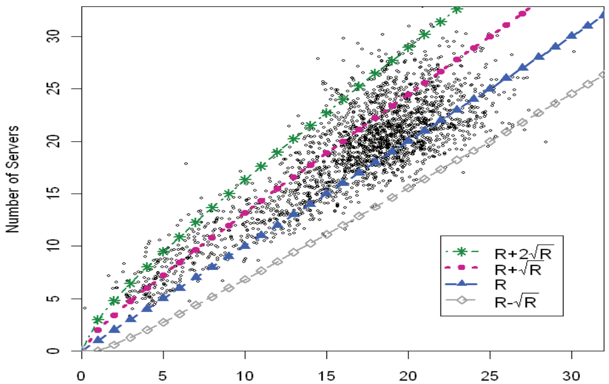
## Operational Regimes: Scaling, Performance, w/ I. Gurvich & J. Huang

| Erlang-A | Conventional scaling | | | MS scaling | | | | NDS scaling | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ fixed | Sub | Critical | Super | QD | QED | ED | ED+QED | Sub | Critical | Super |
| Offered load per server | $\frac{1}{1+\delta} < 1$ | $1 - \frac{\beta}{\sqrt{n}} \approx 1$ | $\frac{1}{1-\gamma} > 1$ | $\frac{1}{1+\delta}$ | $1 - \frac{\beta}{\sqrt{n}}$ | $\frac{1}{1-\gamma}$ | $\frac{1}{1-\gamma} - \beta\sqrt{\frac{1}{n(1-\gamma)^3}}$ | $\frac{1}{1+\delta}$ | $1 - \frac{\beta}{n}$ | $\frac{1}{1-\gamma}$ |
| Arrival rate $\lambda$ | $\frac{\mu}{1+\delta}$ | $\mu - \frac{\beta}{\sqrt{n}}\mu$ | $\frac{\mu}{1-\gamma}$ | $\frac{n\mu}{1+\delta}$ | $n\mu - \beta\mu\sqrt{n}$ | $\frac{n\mu}{1-\gamma}$ | $\frac{n\mu}{1-\gamma} - \beta\mu\sqrt{\frac{n}{(1-\gamma)^3}}$ | $\frac{n\mu}{1+\delta}$ | $n\mu - \beta\mu$ | $\frac{n\mu}{1-\gamma}$ |
| Number of servers | 1 | | | $n$ | | | | $n$ | | |
| Time-scale | $n$ | | | 1 | | | | $n$ | | |
| Abandonment rate | $\theta/n$ | | | $\theta$ | | | | $\theta/n$ | | |
| Staffing level | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu}(1+\frac{\beta}{\sqrt{n}})$ | $\frac{\lambda}{\mu}(1-\gamma)$ | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}$ | $\frac{\lambda}{\mu}(1-\gamma)$ | $\frac{\lambda}{\mu}(1-\gamma) + \beta\sqrt{\frac{\lambda}{\mu}}$ | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu} + \beta$ | $\frac{\lambda}{\mu}(1-\gamma)$ |
| Utilization | $\frac{1}{1+\delta}$ | $1 - \sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{\sqrt{n}}$ | 1 | $\frac{1}{1+\delta}$ | $1 - \sqrt{\frac{\theta}{\mu}}\frac{(1-\alpha_2)\hat{\beta}+\alpha_2 h(\hat{\beta})}{\sqrt{n}}$ | 1 | 1 | $\frac{1}{1+\delta}$ | $1 - \sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{n}$ | 1 |
| $\mathbb{E}(Q)$ | $\frac{\alpha_1}{\delta}$ | $\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\delta^2}\varrho^n\frac{1}{\sqrt{n}}$ | $\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $\frac{n\mu}{\theta(1-\gamma)}(\gamma - \frac{\beta}{\sqrt{n(1-\gamma)}})$ | $o(1)$ | $n\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $\frac{n^2\mu\gamma}{\theta(1-\gamma)}$ |
| $\mathbb{P}(Ab)$ | $\frac{1}{n}\frac{1+\delta}{\delta}\frac{\theta}{\mu}\alpha_1$ | $\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$ | $\gamma$ | $\frac{1}{\sqrt{2\pi}}\frac{\theta}{\mu}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$ | $\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$ | $\gamma$ | $\gamma - \frac{\beta\sqrt{1-\gamma}}{\sqrt{n}}$ | $o(\frac{1}{n^2})$ | $\frac{1}{n}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$ | $\gamma$ |
| $\mathbb{P}(W_q > 0)$ | $\alpha_1 \in (0,1)$ | $\approx 1$ | | $\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\delta}\varrho^n\frac{1}{\sqrt{n}} \approx 0$ | $\alpha_2 \in (0,1)$ | $\approx 1$ | $\approx 1$ | $\approx 0$ | $\approx 1$ | |
| $\mathbb{P}(W_q > T)$ | $\alpha_1 e^{-\frac{\delta}{1+\delta}\mu t}$ | $1 + O(\frac{1}{\sqrt{n}})$ | $1 + O(\frac{1}{n})$ | $\approx 0$ | | $\bar{G}(T)1_{\{G(T)<\gamma\}}$ | $\alpha_3$, if $G(T) = \gamma$ | $\approx 0$ | $\frac{\bar{\Phi}(\hat{\beta}+\sqrt{\theta\mu}T)}{\bar{\Phi}(\hat{\beta})}$ | $1 + O(\frac{1}{n})$ |
| Congestion $\frac{\mathbb{E}W_q}{\mathbb{E}S}$ | $\alpha_1\frac{1+\delta}{\delta}$ | $\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $n\mu\gamma/\theta$ | $\frac{1}{\sqrt{2\pi}}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$ | $\frac{1}{\sqrt{n}}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$ | $\mu\int_0^{x^*}\bar{G}(s)ds$ | $\mu\int_0^{x^*}\bar{G}(s)ds - \frac{\mu\beta\sqrt{1-\gamma}}{h_G(x^*)\sqrt{n}}$ | $o(\frac{1}{n})$ | $\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta}) - \hat{\beta}]$ | $n\mu\gamma/\theta$ |



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

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

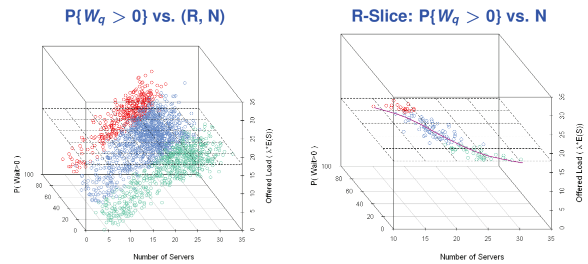



**2205 half-hour intervals in an Israeli Call Center**



## QED Call Center: Performance

**Large Israeli Bank**




**3 Operational Regimes**:

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

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

Consider a sequence of **steady-state** M/M/$N$ + G queues, $N = 1, 2, 3, \ldots$
Then the following points of view are **equivalent**, as $N \uparrow \infty$:

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

- **QED performance**: **Laplace Method** (asymptotics of integrals).
- **Parameters**: Arrivals and Staffing - $\beta$, Services - $\mu$, (Im)Patience - $g(0)$ = **patience density at the origin**.



## Erlang-A: QED Approximations (Examples)

Assume **Offered Load** $R$ not small ($\lambda \to \infty$).

Let $\hat{\beta} = \beta\sqrt{\frac{\mu}{\theta}}$, $h(\cdot) = \frac{\phi(\cdot)}{1 - \Phi(\cdot)}$ = hazard rate of $\mathcal{N}(0,1)$.

- **Delay Probability:**

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

- **Probability to Abandon:**

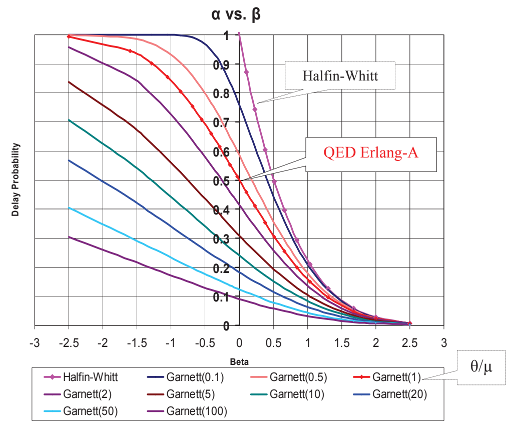
$$\text{P}\{\text{Ab}|W_q > 0\} \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\theta}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right].$$

- $P\{Ab\} \propto \text{E}[W_q]$ , both order $\frac{1}{\sqrt{n}}$ :

$$\frac{\text{P}\{\text{Ab}\}}{\text{E}[W_q]} = \theta.$$



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






## QED Intuition: Why $P\{W_q > 0\} \in (0,1)$ ?

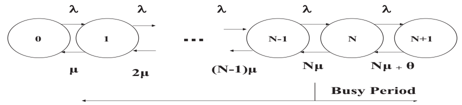
1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $n \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$.
   - Fix $n$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $n$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $n \approx R + \beta\sqrt{R}$.
2. **Erlang-A** (M/M/n+M), with parameters $\lambda, \mu, \theta; n$, in which $\mu = \theta$: (Im)Patience and Service-times are equally distributed.
   - Steady-state: $L(M/M/n+M) \stackrel{d}{=} L(M/M/\infty) \stackrel{d}{=} Poisson(R)$, with $R = \lambda/\mu$ (Offered-Load)
   - Poisson$(R) \stackrel{d}{\approx} R + Z\sqrt{R}$, with $Z \stackrel{d}{=} N(0,1)$.
   - $P\{W_q(M/M/n+M) > 0\} \stackrel{PASTA}{=} P\{L(M/M/n+M) \geq n\} \stackrel{\mu=\theta}{=}$

     $P\{L(M/M/\infty) \geq n\} \approx P\{R + Z\sqrt{R} \geq n\} =$

     $P\{Z \geq (n - R)/\sqrt{R}\} \stackrel{\sqrt{\cdot}\ staffing}{\approx} P\{Z \geq \beta\} = 1 - \Phi(\beta)$.
3. QED **Excursions**



## QED Intuition via Excursions: Busy-Idle Cycles




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

Let $T_{N,N-1}$ = E[Busy Period] down-crossing $N \downarrow N-1$

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

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



## QED Intuition via Excursions: Asymptotics

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

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

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

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

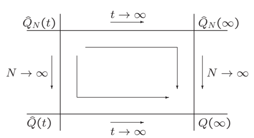


## Process Limits (Queueing, Waiting)

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




Approximating (Virtual) Waiting Time

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



## QED Erlang-X (Markovian Q's: Performance Analysis)

- Pre-History, 1914: **Erlang** (Erlang-B = M/M/n/n, Erlang-C = M/M/n)
- Pre-History, 1974: Jagerman (Erlang-B)
- History Milestone, 1981: **Halfin-Whitt** (Erlang-C, GI/M/n)
- Erlang-A (M/M/N+M), 2002: w/ **Garnett** & Reiman
- Erlang-A with General (Im)Patience (M/M/N+G), 2005: w/ Zeltyn
- Erlang-C (ED+QED), 2009: w/ Zeltyn
- Erlang-B with Retrial, 2010: Avram, Janssen, van Leeuwaarden
- Refined Asymptotics (Erlang A/B/C), 2008-2011: Janssen, van Leeuwaarden, Zhang, Zwart
- NDS Erlang-C/A, 2009: Atar
- Production Q's, 2011: Reed & Zhang
- Universal Erlang-R, ongoing: w/ Gurvich & Huang
- Queueing Networks:
  - (Semi-)Closed: Nurse Staffing (Jennings & de Vericourt), CCs with IVR (w/ Khudiakov), Erlang-R (w/ Yom-Tov)
  - CCs with Abandonment and Retrials: w. Massey, Reiman, Rider, Stolyar
  - Markovian Service Networks: w/ Massey & Reiman
- Leaving out:
  - **Non-Exponential Service Times**: M/D/n (Erlang-D), G/Ph/n, $\cdots$, G/GI/n+GI, Measure-Valued Diffusions
  - **Dimensioning** (Staffing): M/M/n, $\cdots$, time-varying Q's, V- and Reversed-V, $\cdots$
  - **Control**: V-network, Reversed-V, $\cdots$, SBRNets



## Back to "Why does Erlang-A Work?"

**Theoretical** (Partial) Answer:

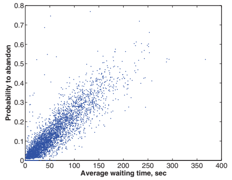
$M_t^{?,J}/G^*/N_t + G \stackrel{d}{\approx} (M/M/N+M)_t$ , $t \geq 0$.

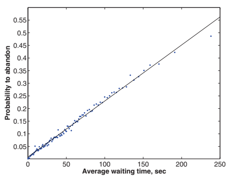
- **Over-Dispersed Arrivals**: $R + \beta R^c$, $c$-Staffing ($c \geq 1/2$).
- **General Patience**: Behavior at the origin matters most (only).
- **General Services**: Empirical insensitivity beyond the mean.
- **Heterogeneous Customers / Servers**: State-Collapse.
- **Time-Varying Arrivals**: Modified Offered-Load approximations.
- **Dependent Building-Blocks**: eg. When (Im)Patience and Service-Times correlated (positively):
  - Predict performance with $E[S\,|\,\text{Served}]$.
  - Calculate offered-load with $E[S] = E[S\,|\,\text{Wait} = 0]$.
  - Note: staffing $\leftarrow$ service-times $\leftarrow$ waiting / abandonment $\leftarrow$ staffing

## "Why does Erlang-A Work?" General Patience

### Israeli Bank: Yearly Data

Hourly Data | Aggregated




**Theory:**

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

**M/M/N+G:** $\mathrm{P}\{Ab\} \approx g(0) \cdot \mathrm{E}[W_q]$.
$g(0)$ = Patience-density at origin

**Recipe:**

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

**References** on $g(0)$:

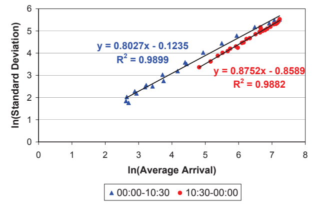
- Stationary M/M/N+GI, w/ **Zeltyn**
- Process G/GI/N+GI: w/ **Momcilovic; Dai & He**;

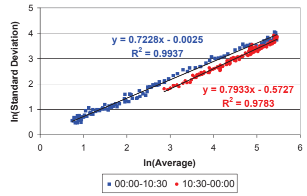


## "Why does Erlang-A Work?" Over-Dispersion

### ln(*STD*) vs. ln(*AVG*) (Israeli Bank, 4/2007-3/2008)

Tue-Wed, 30 min resolution | Tue-Wed, 5 min resolution




Significant linear relations (w/ **Aldor & Feigin**; then w/ **Maman & Zeltyn**):

$$\ln(\mathrm{STD}) = c \cdot \ln(\mathrm{AVG}) + a$$

(Poisson: STD $=$ AVG$^{1/2}$, hence $c = 1/2, a = 0$.)



## Over-Dispersion: Random Arrival-Rates

**Linear relation** between ln(STD) and ln(AVG) gives rise to:

**Poisson-Mixture** (Doubly-Poisson, Cox) model for Arrivals:
**Poisson($\Lambda$)** with **Random-Rate** of the form

$$\Lambda = \lambda + \lambda^c \cdot X, \quad c \le 1\,;$$

- $c$ determines magnitude of over-dispersion ($\lambda^c$)
  $c = 1$, proportional to $\lambda$; $c \le 1/2$, Poisson-level;
  - In **Call Centers**: $c \approx 0.75 - 0.85$ (significant over-dispersion).
  - In **Emergency Departments**, $c \approx 0.5$ (Poisson).
- $X$ random-variable with $E[X] = 0$ ($E[\Lambda] = \lambda$), capturing the magnitude of **stochastic deviation** from mean arrival-rate: under conventional Gamma prior ($\lambda$ large), $X$ can be taken Normal with std. derived from the intercept.

**QED-c** Regime: Erlang-A, with Poisson($\Lambda$) arrivals, amenable to asymptotic analysis (with **S. Maman & S. Zeltyn**)



## Over-Dispersion: The QED-c Regime

**QED-c Staffing**: Under offered-load $R = \lambda \cdot E[S]$,

$$N = R + \beta \cdot R^c\,, \quad 0.5 < c < 1$$

**Performance measures** (M/M/N + G):

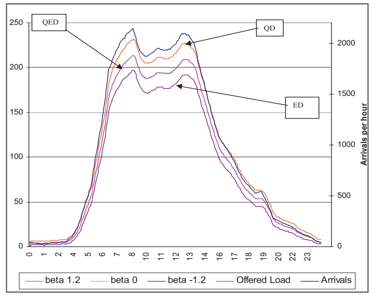
- Delay probability: $P\{W_q > 0\} \sim 1 - G(\beta)$
- Abandonment probability: $P\{Ab\} \sim \dfrac{E[X-\beta]_+}{n^{1-c}}$
- Average offered wait: $E[V] \sim \dfrac{E[X-\beta]_+}{n^{1-c} \cdot g_0}$
- Average actual wait: $E_{\Lambda,N}[W] \sim E_{\Lambda,N}[V]$



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

**Square-Root Staffing:** $N_t = R_t + \beta\sqrt{R_t}\,, \quad -\infty < \beta < \infty$
What is $R_t$, the **Offered-Load** at time $t$? ( $R_t \ne \lambda_t \times \mathrm{E[S]}$ )

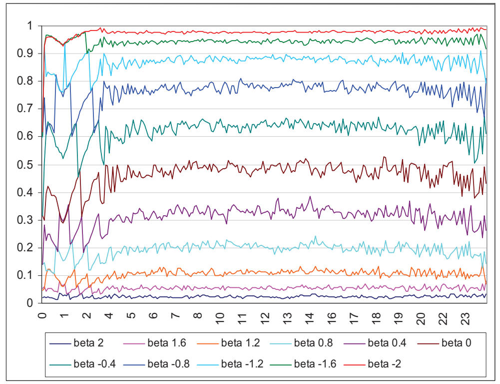
### Arrivals, Offered-Load and Staffing






## Time-Stable Performance of Time-Varying Systems

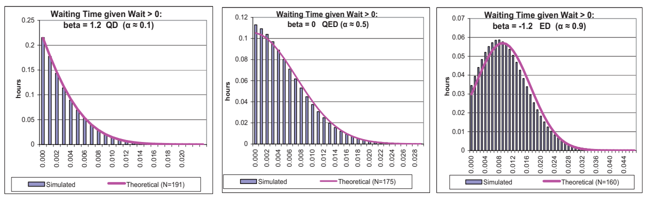
**Delay Probability** = As in the **Stationary Erlang-A** (Garnett)






## Time-Stable Performance of Time-Varying Systems

### Waiting Time, Given Waiting: Empirical vs. Theoretical Distribution




- **Empirical**: Simulate **time-varying** $M_t/M/N_t + M$ ($\lambda_t, N_t = R_t + \beta\sqrt{R_t}$)
- **Theoretical**: Naturally-corresponding **stationary** Erlang-A, with QED $\beta$-staffing (some **Averaging** Principle?)
- **Generalizes** up to a single-station within a complex network (eg. Doctors in an Emergency Department).



## What is the Offered-Load $R(t)$?

- Offered-Load Process: $L(\cdot)$ = **Least** number of **servers** that guarantees **no delay**.
- **Offered-Load** Function $R(t) = E[L(t)]$, $t \ge 0$.
  Think $M_t/G/N_t^? + G$ vs. $M_t/G/\infty$: **Ample-Servers**.

Four (all useful) representations, capturing **"workload before t"**:

$$R(t) = E[L(t)] = \int_{-\infty}^{t} \lambda(u) \cdot P(S > t-u)du = E\Big[A(t) - A(t-S)\Big] =$$
$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t - S_e)] \cdot E[S] \approx \ldots.$$

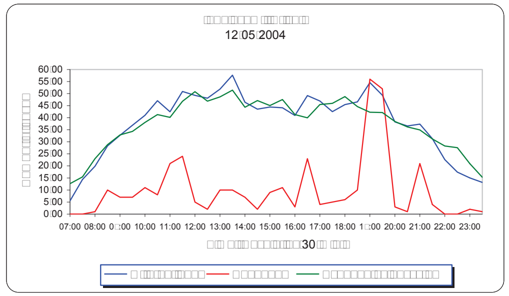
- $\{A(t),\ t \ge 0\}$ Arrival-Process, rate $\lambda(\cdot)$;
- $S$ ($S_e$) generic Service-Time (Residual Service-Time).
- Relating $L, \lambda, S$ ("$W$"): **Time-Varying Little's Formula.**
  **Stationary models**: $\lambda(t) \equiv \lambda$ then $R(t) \equiv \lambda \times E[S]$.

**QED-c**: $N_t = R_t + \beta R_t^c$, $1/2 \le c < 1$; ($c = 1$ separate analysis).



## The Offered-Load $R(t), t \ge 0$

- **Backbone** of time-varying staffing:
  - Practically **robust**: up to a station within a complex network (ED).
  - Theoretically **challenging**: only Erlang-A with $\mu = \theta$ tractable.
- Process: $L(\cdot)$ = **Least** number of **servers** that guarantees **no delay**.
- **Offered-Load** Function $R(\cdot) = E[L(\cdot)]$ ($M_t/G/N_t^? + G \leftrightarrow M_t/G/\infty$).

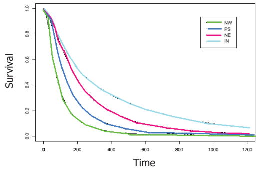


## Estimating / Predicting the Offered-Load

Must account for **"service times of abandoning customers"**.

- Prevalent Assumption: Services and (Im)Patience independent.
- But recall Patient VIPs: Willing to wait more for longer services.

**Survival Functions by Type (Small Israeli Bank)**



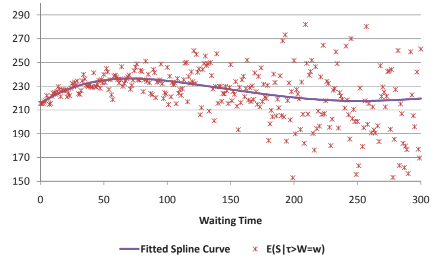

**Service times** stochastic order: $S_{New} \overset{st}{<} S_{Reg} \overset{st}{<} S_{VIP}$

**Patience times** stochastic order: $\tau_{New} \overset{st}{<} \tau_{Reg} \overset{st}{<} \tau_{VIP}$



## Dependent Primitives: Service- vs. Waiting-Time

**Average Service-Time as a function of Waiting-Time**
**U.S. Bank, Retail, Weedays, January-June, 2006**




⇒ Focus on **( Patience, Service-Time ) jointly** , w/ **Reich and Ritov**.
$E[S \,|\, \text{Patience} = w]$, $w \geq 0$: **Service-Time of the Unserved**.



## (Imputing) Service-Times of Abandoning Customers

**w/ M. Reich, Y. Ritov**:

1. **Estimate** $g(w) = E[S \,|\, \text{Patience} > \text{Wait} = w]$, $w \geq 0$:
   Mean service time of those **served after waiting exactly** $w$ units of time (via non-linear regression: $S_i = g(W_i) + \varepsilon_i$);
2. **Calculate**
   $$E[S \,|\, \text{Patience} = w] = g(w) - \frac{g'(w)}{h_\tau(w)} ;$$
   $h_\tau(w)$ = hazard-rate of (im)patience (via un-censoring);
3. **Offered-load** calculations: Impute $E[S \,|\, \text{Patience} = w]$ (or the conditional distribution).

**Challenges**: Stable and accurate inference of $g, g', h_\tau$.



## Extending the Notion of the "Offered-Load"

- **Business** (Banking Call-Center): Offered **Revenues**
- **Healthcare** (Maternity Wards): Fetus in stress
  - 2 patients (Mother + Child) = high **operational** and **cognitive** load
  - Fetus dies ⇒ **emotional** load dominates
- ⇒
  - Offered **Operational** Load
  - Offered **Cognitive** Load
  - Offered **Emotional** Load
  - ⇒ **Fair** Division of Load (Routing) between 2 Maternity Wards: One attending to complications <u>before</u> birth, the other to complications <u>after</u> birth, and both share normal birth



## The Technion SEE Center / Laboratory

**Data-Based Service Science / Engineering**

Technion S.E.E Center
Service Enterprise Engineering