

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>

IELM, Hong Kong UST, September 2011

Research Partners

- ▶ **Students:**

Aldor*, Baron*, Carmeli, Feldman*, Garnett*, Gurvich*, Huang, Khudiakov*, Maman*, Marmor*, Reich, Rosenshmidt*, Shaikhet*, Senderovic, Tseytlin*, Yom-Tov*, Zaied, Zeltyn*, Zychlinski, Zohar*, Zviran, ...

- ▶ **Theory:**

Armony, Atar, Gurvich, Jelenkovic, Kaspi, Massey, Momcilovic, Reiman, Shimkin, Stolyar, Wasserkrug, Whitt, Zeltyn, ...

- ▶ **Industry:**

IBM Research (OCR: Carmeli, Vortman, Wasserkrug, Zeltyn), Rambam Hospital, Hapoalim Bank, Mizrahi Bank, Pelephone Cellular, ...

- ▶ **Technion SEE Center / Laboratory:**

Feigin; Trofimov, Nadjharov, Gavako, Kutsyy; Liberman, Koren, Rom, Plonsky; Research Assistants, ...

- ▶ **Empirical/Statistical Analysis:**

Brown, Gans, Zhao; Shen; Ritov, Goldberg; Allon, Ata, Bassamboo; Gurvich, Huang, Liberman; Armony, Marmor, Tseytlin, Yom-Tov; Zeltyn, Nardi, ...

History, Resources (Downloadable)

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

The Case for Service Science / Engineering

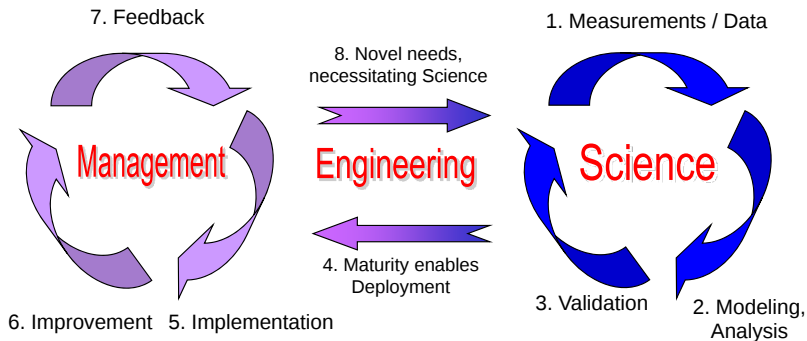
- ▶ **Service Science / Engineering** (vs. Management) are emerging **Academic Disciplines**. For example, universities (world-wide), IBM (SSME, a la Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ...
- ▶ Models that explain **fundamental phenomena**, which are **common** across applications:
 - **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.

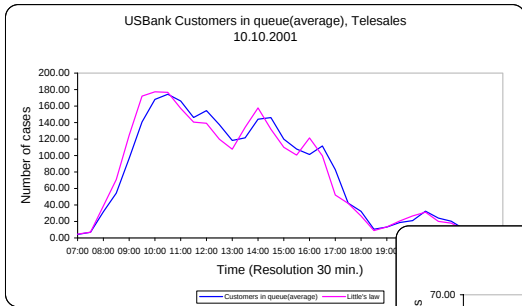
Starting with **Data**, expand to:



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

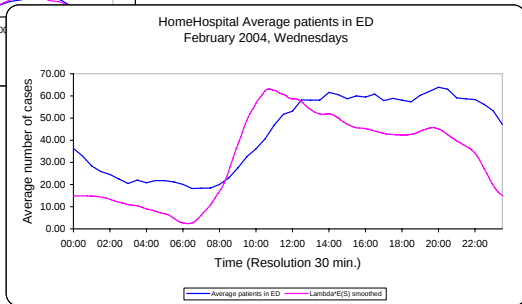
Little's Law: Call Center & Emergency Department

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



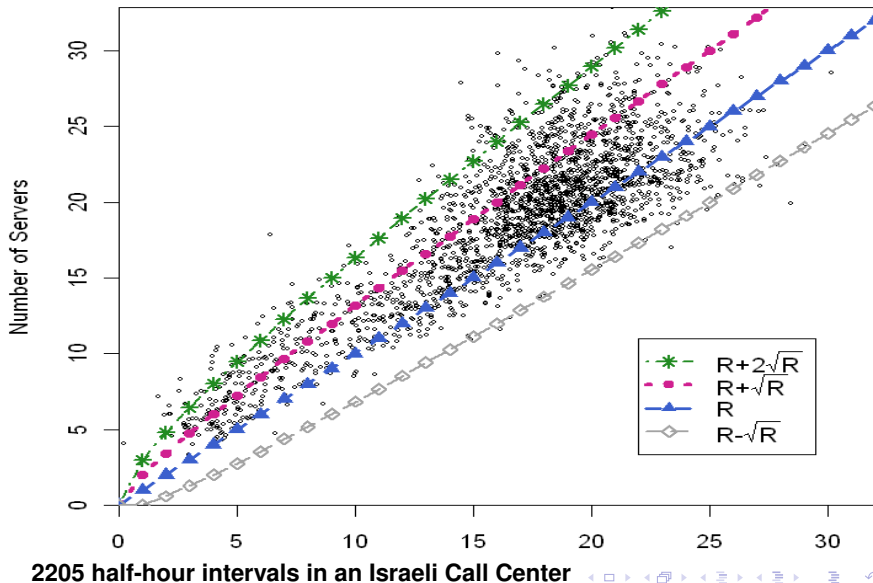
⇒ **Time-Varying Little's Law**

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



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

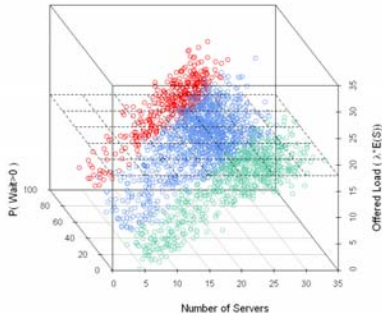
IL Telecom; June-September, 2004; w/ Nardi, Plonski, Zeltyn



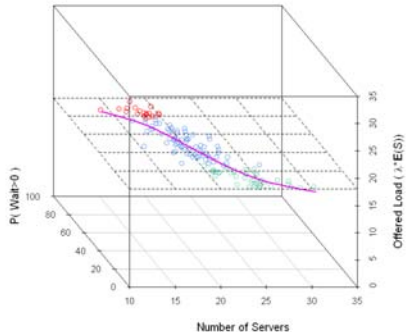
QED Call Center: Performance

Large Israeli Bank

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

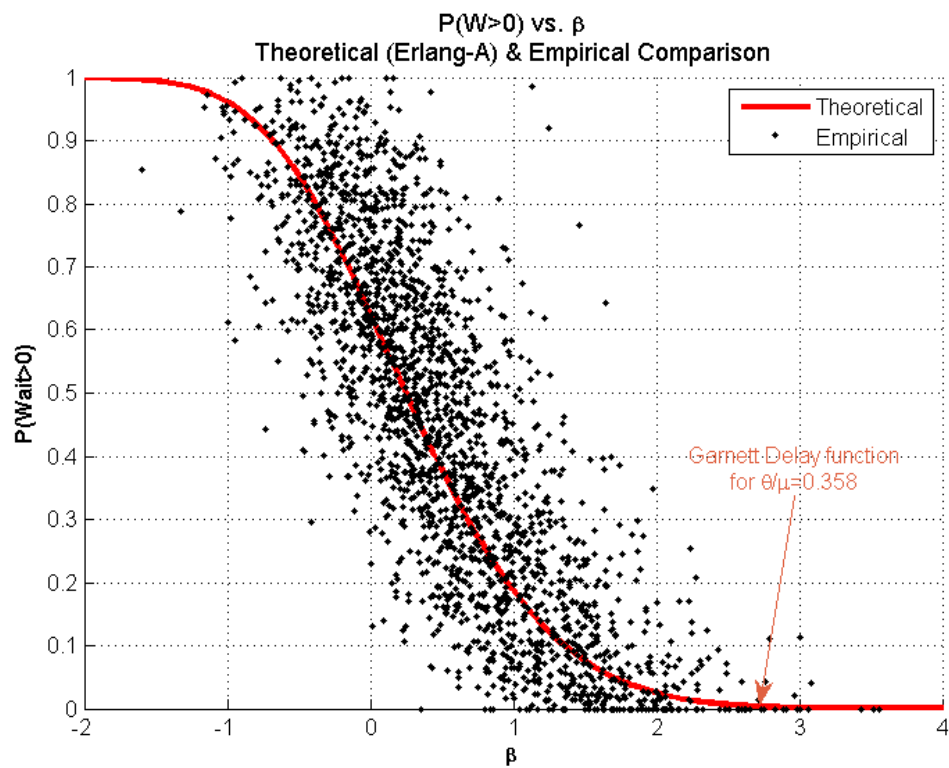


R-Slice: $P\{W_q > 0\}$ vs. N

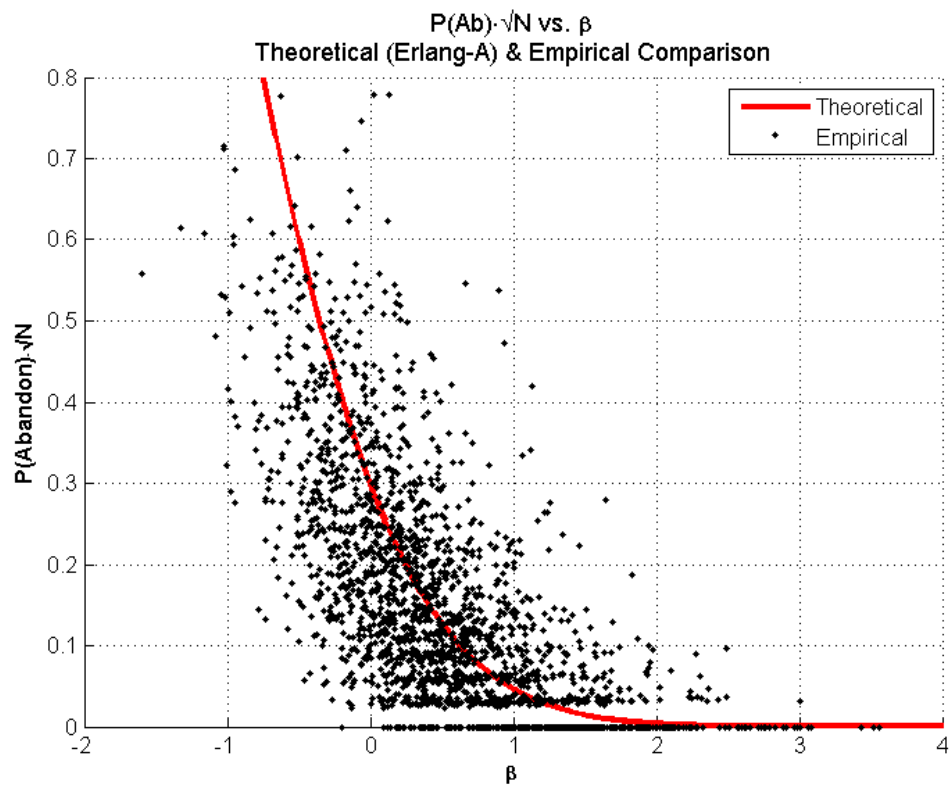


3 Operational Regimes:

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



IT Telecom Technical University of Sofia, week days



IT Telecom Technical University of Sofia, week days

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

Erlang-A μ fixed	Conventional scaling			MS scaling				NDS scaling		
	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 λ	$\frac{\mu}{1+\delta}$	$\mu - \frac{\beta}{\sqrt{n}}\mu$	$\frac{\mu}{1-\gamma}$	$\frac{n\mu}{1+\delta}$	$n\mu - \beta\mu\sqrt{n}$	$\frac{n\mu}{1-\gamma}$	$\frac{n\mu}{1-\gamma} - \beta\mu\sqrt{\frac{n}{(1-\gamma)^3}}$	$\frac{n\mu}{1+\delta}$	$n\mu - \beta\mu$	$\frac{n\mu}{1-\gamma}$
Number of servers	1			n				n		
Time-scale	n			1				n		
Abandonment rate	θ/n			θ				θ/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
$\mathbb{E}(Q)$	$\frac{\alpha}{\delta}$	$\sqrt{n} \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}]$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{1}{\sqrt{2\pi}} \frac{1+\delta}{\delta^2} \varrho^n \frac{1}{\sqrt{n}}$	$\sqrt{n} \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}] \alpha_2$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{n\mu}{\theta(1-\gamma)} (\gamma - \frac{\beta}{\sqrt{n(1-\gamma)}})$	$o(1)$	$n \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}]$	$\frac{n^2\mu\gamma}{\theta(1-\gamma)}$
$\mathbb{P}(Ab)$	$\frac{1}{n} \frac{1+\delta}{\delta} \frac{\theta}{\mu} \alpha_1$	$\frac{1}{\sqrt{n}} \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}]$	γ	$\frac{1}{\sqrt{2\pi}} \frac{\theta}{\delta^2} \varrho^n \frac{1}{n^{3/2}}$	$\frac{1}{\sqrt{n}} \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}] \alpha_2$	γ	$\gamma - \frac{\beta\sqrt{1-\gamma}}{\sqrt{n}}$	$o(\frac{1}{n^2})$	$\frac{1}{n} \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}]$	γ
$\mathbb{P}(W_q > 0)$	$\alpha_1 \in (0, 1)$	≈ 1		$\frac{1}{\sqrt{2\pi}} \frac{1+\delta}{\delta} \varrho^n \frac{1}{\sqrt{n}} \approx 0$	$\alpha_2 \in (0, 1)$	≈ 1	≈ 1	≈ 0	≈ 1	
$\mathbb{P}(W_q > T)$	$\alpha_1 e^{-\frac{1}{1+\delta}\mu T}$	$1 + O(\frac{1}{\sqrt{n}})$	$1 + O(\frac{1}{n})$	≈ 0		$\tilde{G}(T) 1_{\{G(T) < \gamma\}}$	α_3 , if $G(T) = \gamma$	≈ 0	$\frac{\Phi(\hat{\beta} + \sqrt{\theta\mu}T)}{\Phi(\hat{\beta})}$	$1 + O(\frac{1}{n})$
Congestion $\frac{\mathbb{E}W_q}{\mathbb{E}S}$	$\alpha_1 \frac{1+\delta}{\delta}$	$\sqrt{n} \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}]$	$n\mu\gamma/\theta$	$\frac{1}{\sqrt{2\pi}} \frac{(1+\delta)^2}{\delta^2} \varrho^n \frac{1}{n^{5/2}}$	$\frac{1}{\sqrt{n}} \sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}] \alpha_2$	$\mu \int_0^{x^*} \tilde{G}(s) ds$	$\mu \int_0^{x^*} \tilde{G}(s) ds - \frac{\mu\beta\sqrt{1-\gamma}}{h_G(x^*)\sqrt{n}}$	$o(\frac{1}{n})$	$\sqrt{\frac{\mu}{\delta}} [h(\hat{\beta}) - \hat{\beta}]$	$n\mu\gamma/\theta$

• $\delta > 0, \gamma \in (0, 1)$ and $\beta \in (-\infty, \infty)$;

• QD: $\varrho = \frac{1}{1+\delta} e^{\frac{\delta}{1+\delta}} < 1$;

• ED (ED+QED): $G(x^*) = \gamma$;

• QED: $\alpha_2 = [1 + \sqrt{\frac{\theta}{\mu} \frac{h(\hat{\beta})}{n-\hat{\beta}}}]^{-1}$, here $\hat{\beta} = \beta \sqrt{\frac{\mu}{\theta}}$ and $h(x) = \frac{\phi(x)}{\Phi(x)}$;

• ED+QED: $\alpha_3 = \tilde{G}(T) \frac{\Phi(\hat{\beta} \sqrt{\frac{\mu}{\theta}} T)}{\Phi(\hat{\beta})}$;

• 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**:

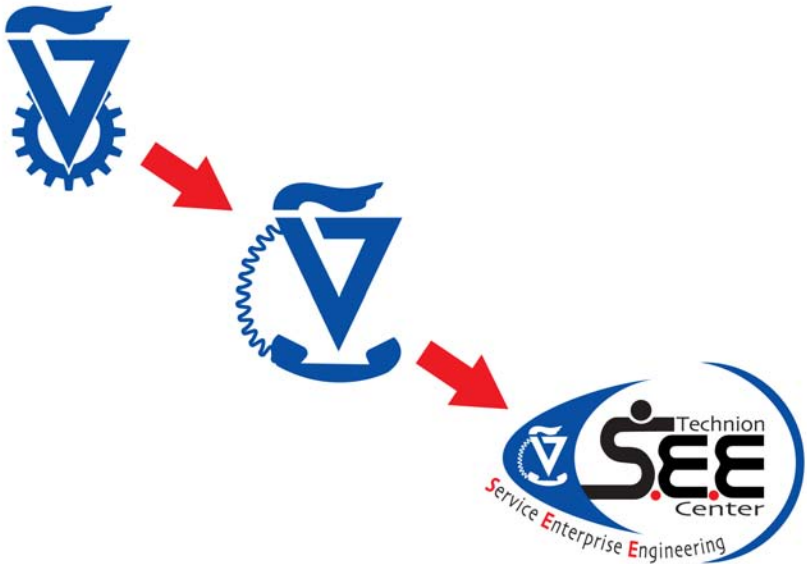
For each service transaction (during a phone-service in a call center, or a patient's visit in a hospital, or browsing in a website, or ...), its

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 = 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.**

SEESStat: 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

Center for Service Enterprise Engineering (SEE)
<http://ie.technion.ac.il/Labs/Serveng/>

SEESat 3.0 Tutorial



HKUST, Hong Kong, September 2011

Note: This tutorial is customized to HKUST.

To become a regular user of SEESat, please go to

<http://seeserver.iem.technion.ac.il/see-terminal/>

click on “Register” (left menu), and follow the registration procedure.

As a participant in the **HKUST** seminar/mini-course, you are able to **connect** (from your PC or laptop) to the **SEELab Server** at the Technion.

Once connected, you will be able to go through the self-taught **SEE Tutorial** that follows.

To start, you need a “User Name” and “Password”.
Use the Number allocated to you in the seminar/mini-course (from 1-20):

On Page 3, you are asked to provide the following LogIn information: Your User Name: visitor_1_20 Your Password: visitor_1_20	Log In Log In User Name: Password: Create a new account I forgot my password Log In
You will be needing the above also for the following step (Page 4):	Log On to Windows Log On to Windows Microsoft Windows Server 2003 R2 Standard Edition Copyright © 2005 Microsoft Corporation User name: Your User Name Password: Your Password OK Cancel Options >>

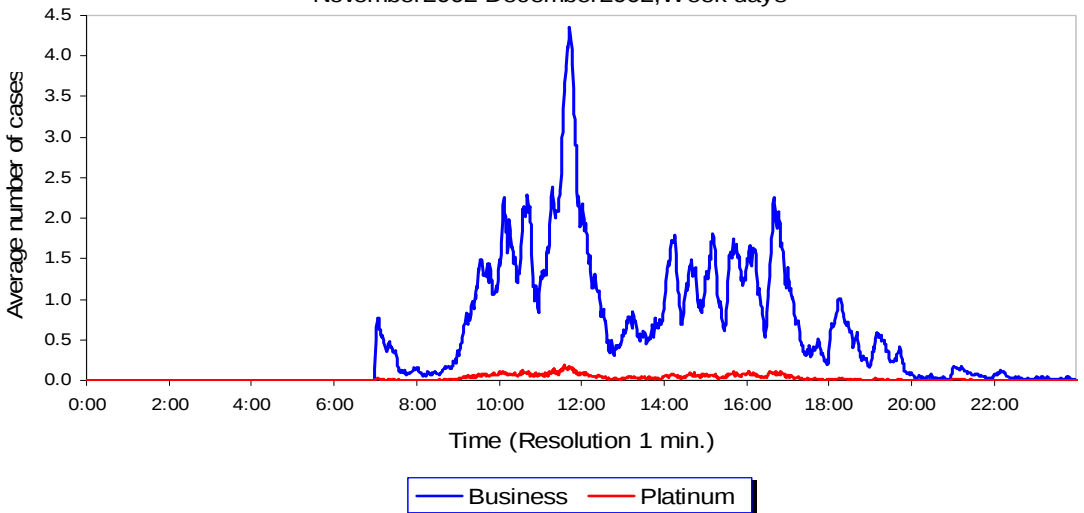
Introduction: SEEStat is a environment for Exploratory Data Analysis (EDA) in real-time. It enables users to easily conduct statistical and performance analyses of massive datasets; in particular, datasets representing operational histories of large service operations (e.g. call centers, hospitals, internet sites), as available through the SEELab server. SEEStat can also automatically create sophisticated reports in Microsoft Excel, which support research and teaching.

Both SEEStat and the SEELab Server were developed at the Faculty of Industrial Engineering and Management, Technion, Israel Institute of Technology. More information on the SEELab can be found at its homepage <http://ie.technion.ac.il/Labs/Serveng/>

SEESat 3.0 Tutorial

Introduction.....	2
Connecting to SEESat on the Technion SEELab Server.....	3
SEESat Tutorial.....	7
Part 1	7
<i>Example 1.1: Distributions</i>	<i>8</i>
<i>Example 1.2: Intraday time series</i>	<i>15</i>
<i>Example 1.3: Time series (Daily totals)</i>	<i>20</i>
Part 2	24
<i>Example 2.1: Distribution fitting</i>	<i>24</i>
<i>Example 2.2: Distribution mixture fitting.....</i>	<i>26</i>
<i>Example 2.3: Survival analysis with smoothing of hazard rates</i>	<i>28</i>
<i>Example 2.4: Smoothing of intraday time series</i>	<i>30</i>
Part 3	32
<i>Example 3.1: Queue regulated by a protocol & announcements</i>	<i>32</i>
<i>Example 3.2: Queue length & state-space collapse</i>	<i>34</i>
<i>Example 3.3: Change-of_Shifts phenomena</i>	<i>36</i>
<i>Example 3.4: Daily flow of calls.....</i>	<i>40</i>

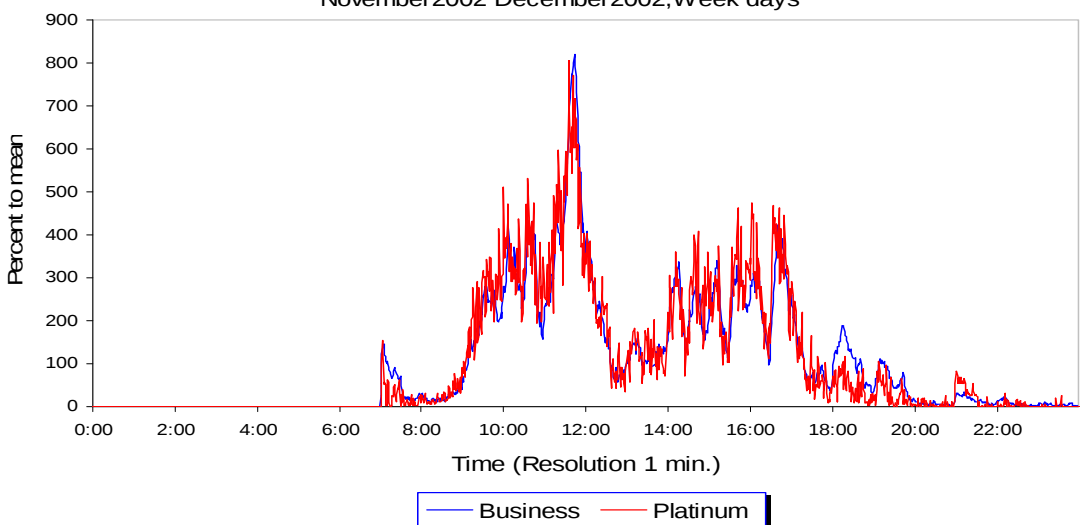
USBank Customers in queue(average)
 Total for May2002 June2002 July2002 August2002 September2002 October2002
 November2002 December2002, Week days



Platinum is a small-scale service. You will now normalize the chart in order to identify patterns.

Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.
 Open the **"Options"** tab and select **Percent to mean**. Click **"OK"**.

USBank Customers in queue(average)
 Total for May2002 June2002 July2002 August2002 September2002 October2002
 November2002 December2002, Week days



Note the essentially overlapping patterns of the queue lengths of the two customer types. (This phenomenon is predicted by asymptotic analysis of queues in heavy traffic, where it is referred to as State-Space-Collapse.)

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

Drill: Chemical MCE, Rambam Hospital, May 2010



Focus on **severely wounded** casualties (≈ 40 in drill)

Note: 20 observers support real-time control (helps validation)

Data Cleaning: MCE with RFID Support

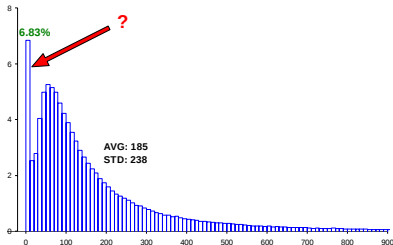
Data-base				Company report		comment
Asset id	order	Entry date	Exit date	Entry date	Exit date	
4	1	1:14:07 PM		1:14:00 PM		
6	1	12:02:02 PM	12:33:10 PM	12:02:00 PM	12:33:00 PM	
8	1	11:37:15 AM	12:40:17 PM	11:37:00 AM		exit is missing
10	1	12:23:32 PM	12:38:23 PM	12:23:00 PM		
12	1	12:12:47 PM	12:35:33 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) !

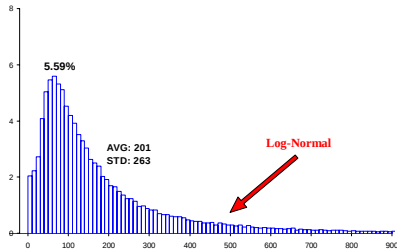
Beyond Averages: The Human Factor

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

January-October



November-December

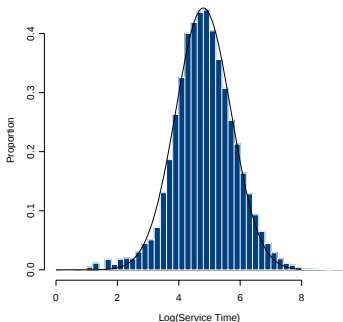


- ▶ **6.8% Short-Services:** Agents' "Abandon" (improve bonus, rest), (mis)lead by **incentives**
- ▶ **Distributions** must be measured (in **seconds** = **natural scale**)
- ▶ **LogNormal** service times common in call centers

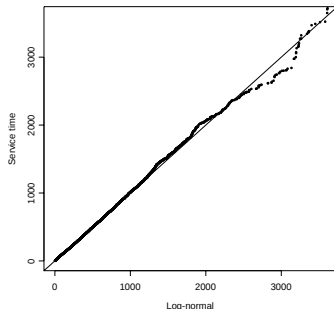
Validating LogNormality of Service-Duration

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:**

Statistics OR IE

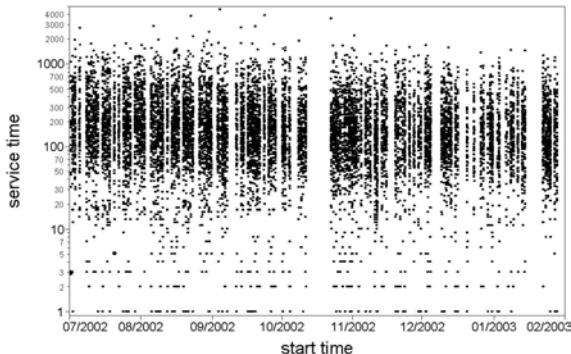


Individual Agents: Service-Duration, Variability

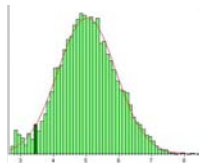
w/ Gans, Liu, Shen & Ye

Agent 14115

Service-Time Evolution: 6 month



Log(Service-Time)

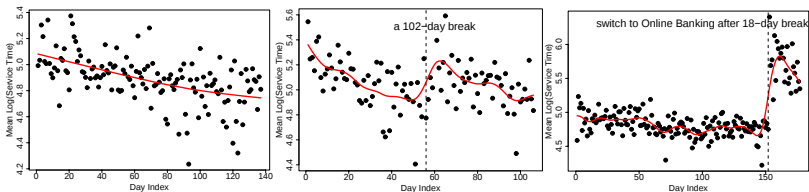


- ▶ **Learning**: Noticeable decreasing-trend in service-duration
- ▶ **LogNormal** Service-Duration, individually and collectively

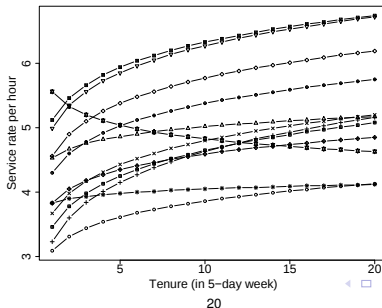
Individual Agents: Learning, Forgetting, Switching

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:

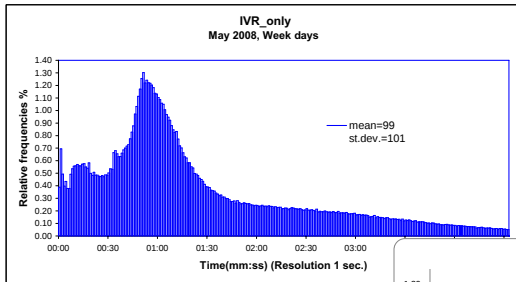
+One Second to Service-Time implies **+Millions** in costs, annually

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

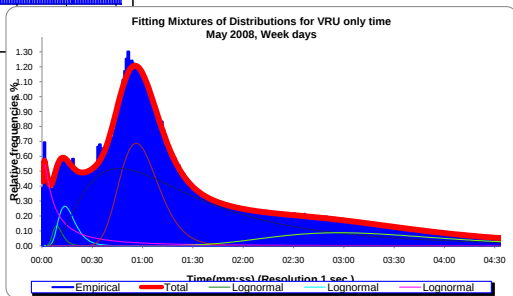
- ▶ **Service-Process Model**: Customer-Agent Interaction
 - ▶ **Work Design** (w/ **Khudiakov**)
eg. **Cross-Selling**: higher profit vs. longer (costlier) services;
Analysis yields (congestion-dependent) cross-selling protocols
 - ▶ **"Worker" Design** (w/ **Gans, Liu, Shen & Ye**)
eg. **Learning, 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

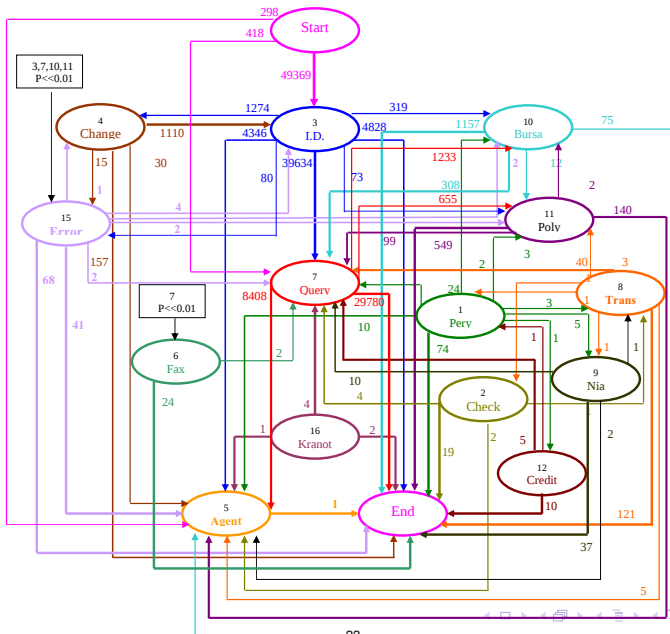
Israeli Bank: IVR/VRU Only, May 2008



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)

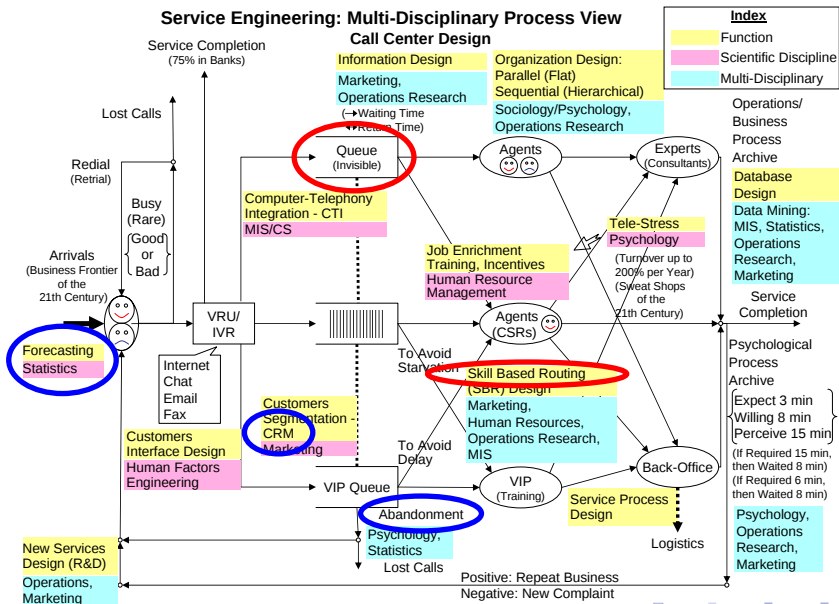
Healthcare - similar and unique challenges:

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

Call-Center Environment: Service Network



Call-Center Network: Gallery of Models



Skills-Based Routing in Call Centers

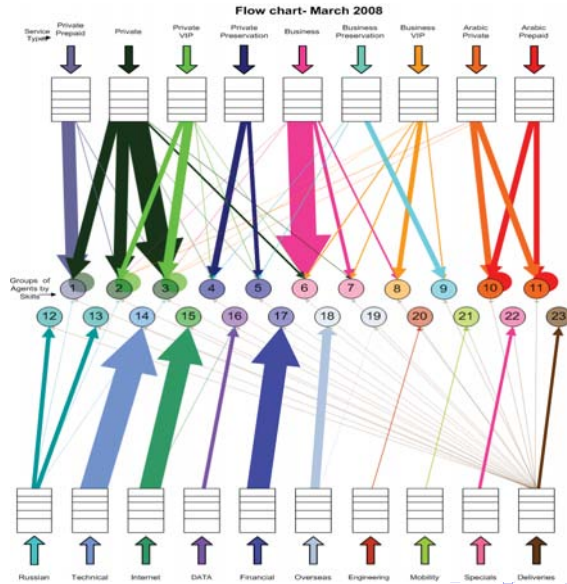
EDA and OR, with I. Gurvich and P. Liberman

Mktg. ⇒

OR ⇒

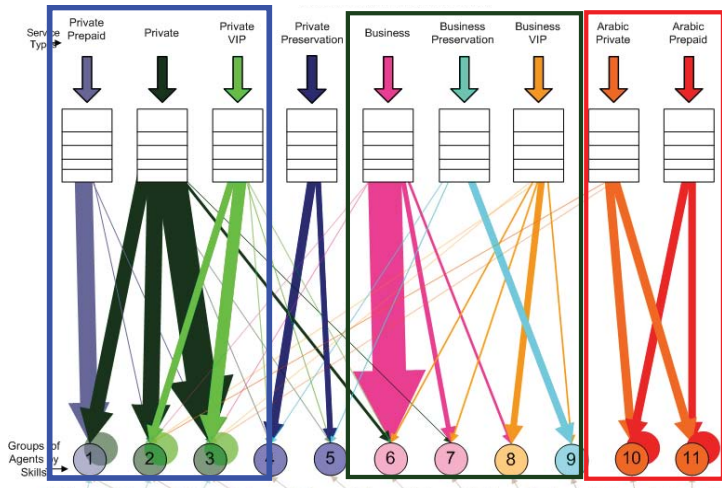
HRM ⇒

MIS ⇒



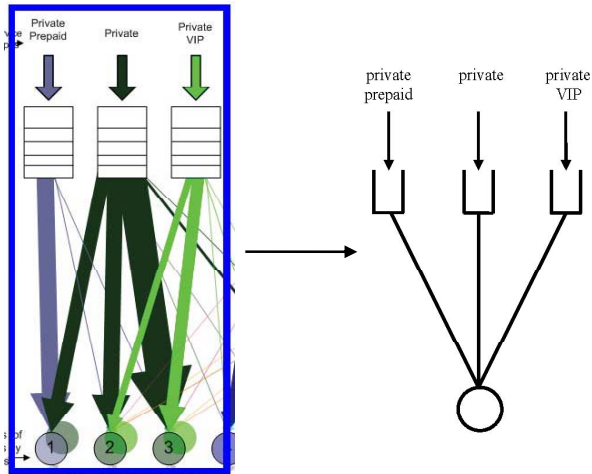
SBR Topologies: I; V, Reversed-V; N, X; W, M

Israeli Cellular, March 2008



SBR: Class-Dependent Services

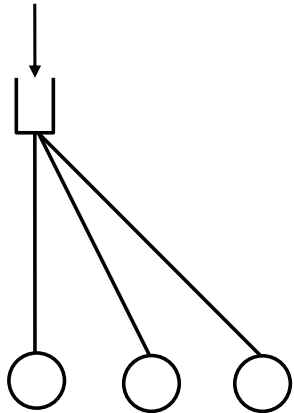
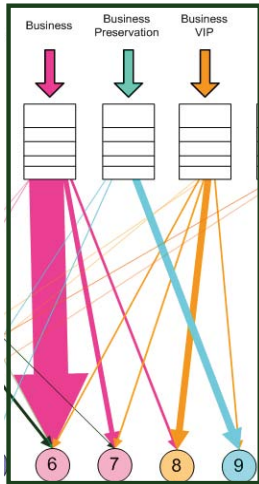
“Reduction” to V-Topology (Equivalent Brownian Control)



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

SBR: Pool-Dependent Services

“Reduction” to Reversed-V and I (Equivalent Brownian Control)

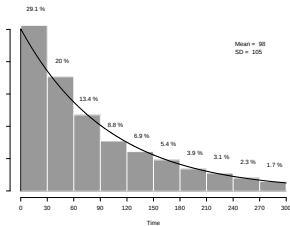


PhD's: **Tezcan**, Dai; **Shaikhet**, w/ Atar; **Gurvich**, Whitt

Waiting Times in a Call Center (Theory?)

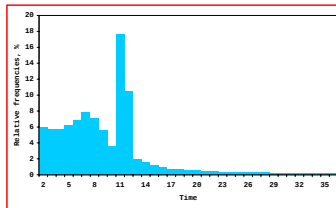
Exponential in Heavy-Traffic (min.)

Small Israeli Bank



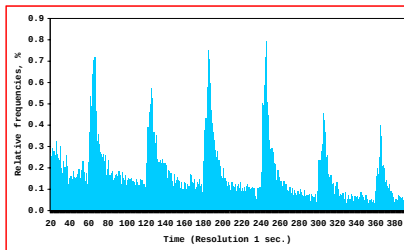
Routing via Thresholds (sec.)

Large U.S. Bank



Scheduling Priorities (sec) (later: Hospital LOS, hr.)

Medium Israeli Bank



ER / ED Environment: Service Network

Acute (Internal, Trauma)



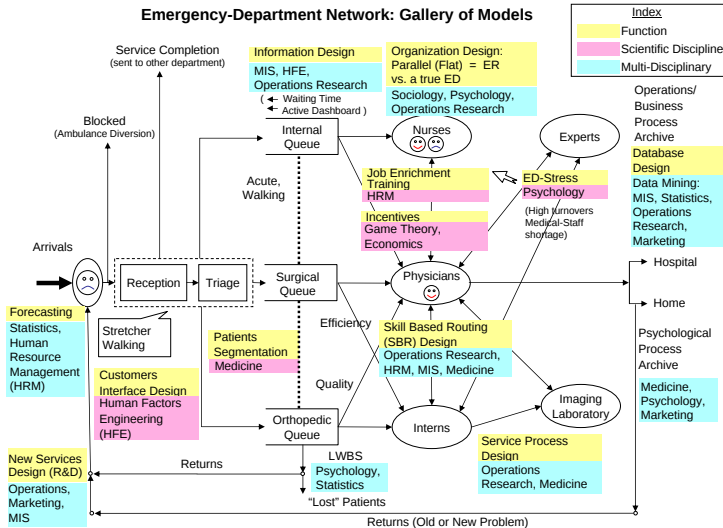
Walking



Multi-Trauma



Emergency-Department Network: Gallery of Models

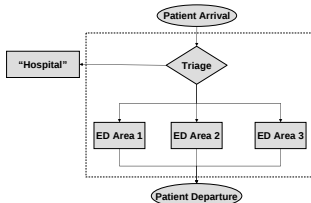


► **Forecasting**, Abandonment = **LWBS**, SBR $\approx$ **Flow Control**

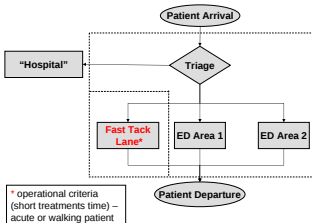
ED Design, with B. Golany, Y. Marmor, S. Israelit

Routing: **Triage (Clinical)**, **Fast-Track (Operational)**, ... (via DEA)

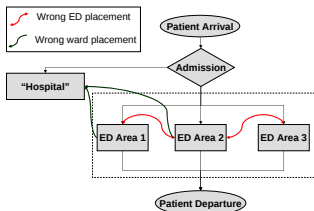
eg. Fast Track most suitable when elderly dominate



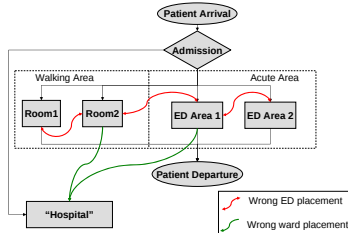
(a) Triage Model



(b) Fast-Track Model



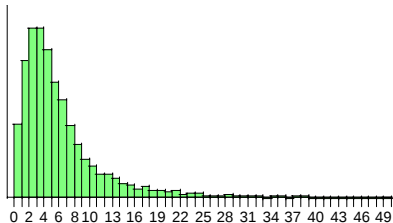
(c) Illness-based Model



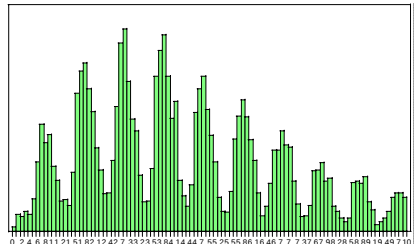
(d) Walking-Acute Model

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

Israeli Hospital, in Days: LN



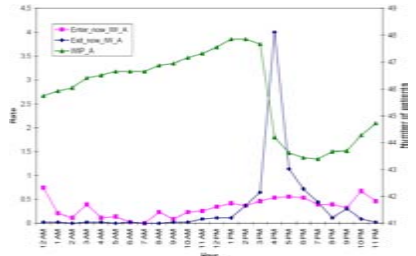
Israeli Hospital, in Hours: Mixture



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

Why Bother ?

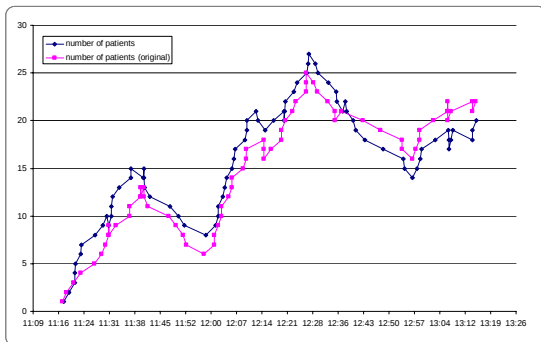
- ▶ Hourly Scale: Staffing,...
- ▶ Daily: Flow / Bed Control,...



Prerequisite II: Models (Fluid Q's)

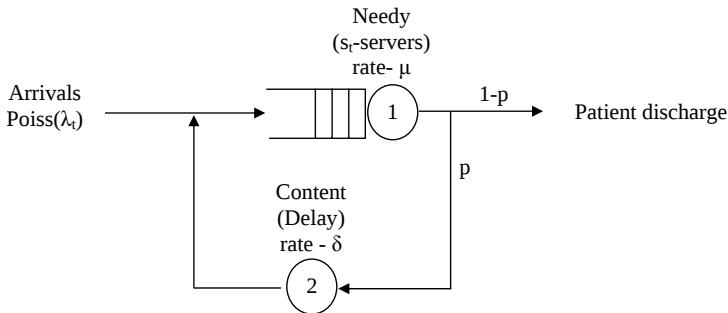
“Laws of Large Numbers” capture **Predictable** Variability
Deterministic Models: Scale Averages-out **Stochastic Individualism**

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



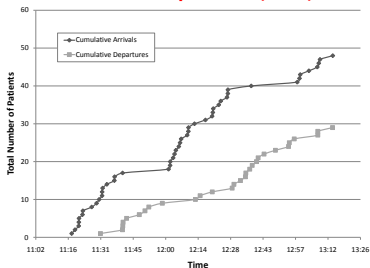
Erlang-R (IE: Repairman Problem 50's; CS: Central-Server 60's) =
2-station "Jackson" Network = (M/M/S, M/M/ ∞) :

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

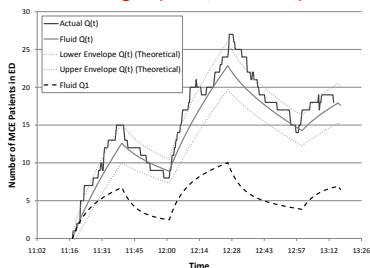
Erlang-R: Fitting a Simple Model to a Complex Reality

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

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

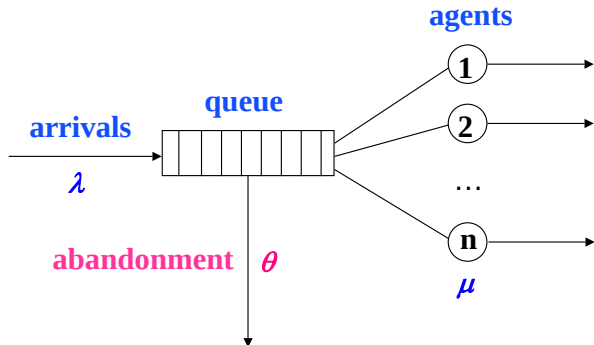
and only **9%** of the customers are served immediately upon arrival.

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:

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

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 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**

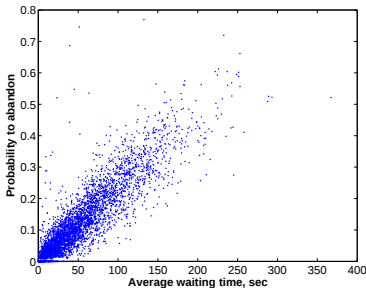
Estimating (Im)Patience: via $P\{Ab\} \propto E[W_q]$

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

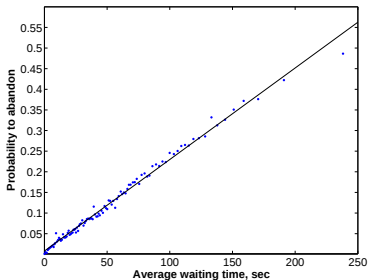
% Abandonment vs. Average Waiting-Time

Bank Anonymous (JASA): Yearly Data

Hourly Data



Aggregated

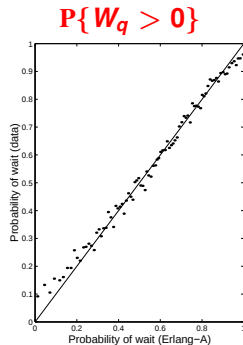
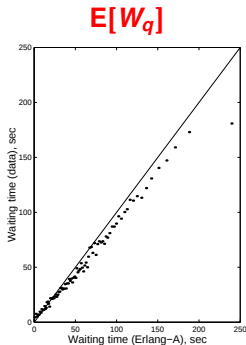
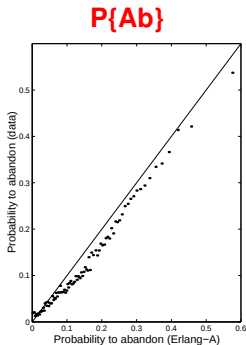


Graphs based on 4158 hour intervals.

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 (θ) estimated via $P\{Ab\} / E[W_q]$
- ▶ Graphs: **Hourly Performance vs. Erlang-A Predictions**, during 1 year (aggregating groups with 40 similar hours).



Erlang-A: Simple, but Not Too Simple

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:

- ▶ **Efficiency-Driven (ED)** regime: Fluid models (deterministic)
- ▶ **Quality- and Efficiency-Driven (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 Leeuwen, B. Zhang, B. Zwart**.

Operational Regimes: Conceptual Framework

R : Offered Load

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

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

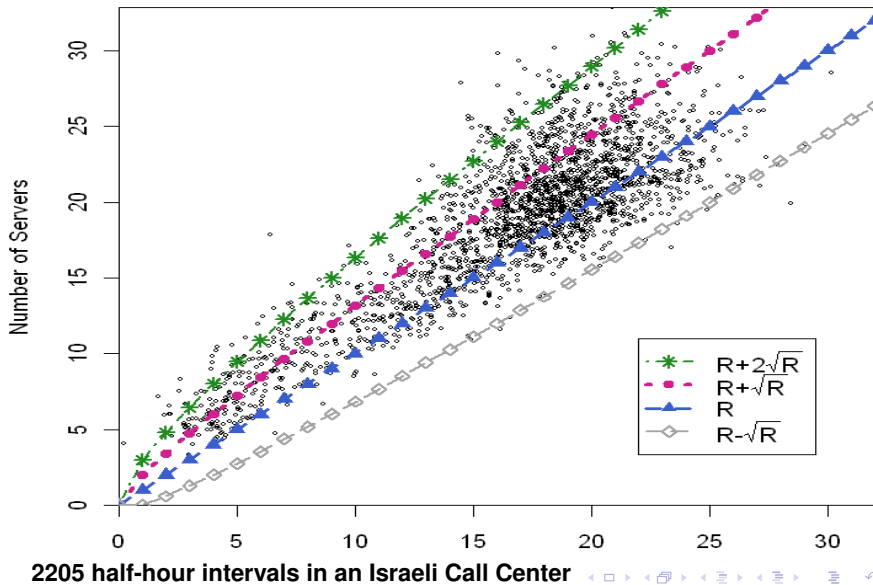
N = #Agents **?** **Intuition**, as R or N increase unilaterally.

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

Erlang-A	Conventional scaling			MS scaling				NDS scaling		
	Sub	Critical	Super	QD	QED	ED	ED+QED	Sub	Critical	Super
Offered load per server	$\frac{1}{1+\delta} < 1$	$1 - \frac{\beta}{\sqrt{n}} \approx 1$	$\frac{1}{1-\gamma} > 1$	$\frac{1}{1+\delta}$	$1 - \frac{\beta}{\sqrt{n}}$	$\frac{1}{1-\gamma}$	$\frac{1}{1-\gamma} - \beta\sqrt{\frac{1}{n(1-\gamma)^2}}$	$\frac{1}{1+\delta}$	$1 - \frac{\beta}{n}$	$\frac{1}{1-\gamma}$
Arrival rate λ	$\frac{\mu}{1+\delta}$	$\mu - \frac{\beta}{\sqrt{n}}\mu$	$\frac{\mu}{1-\gamma}$	$\frac{n\mu}{1+\delta}$	$n\mu - \beta\mu\sqrt{n}$	$\frac{n\mu}{1-\gamma}$	$\frac{n\mu}{1-\gamma} - \beta\mu\sqrt{\frac{n}{(1-\gamma)^2}}$	$\frac{n\mu}{1+\delta}$	$n\mu - \beta\mu$	$\frac{n\mu}{1-\gamma}$
Number of servers	1			n				n		
Time-scale	n			1				n		
Abandonment rate	θ/n			θ				θ/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
$\mathbb{E}(Q)$	$\frac{\lambda}{\mu}$	$\sqrt{n}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\mu}\varrho^n\frac{1}{\sqrt{n}}$	$\sqrt{n}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{n\mu}{\theta(1-\gamma)}(\gamma - \frac{\beta}{\sqrt{n(1-\gamma)}})$	$o(1)$	$n\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$	$\frac{n^2\mu\gamma}{\theta(1-\gamma)}$
$\mathbb{P}(Ab)$	$\frac{1}{n}\frac{1+\delta}{\delta}\frac{\theta}{\mu}\alpha_1$	$\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$	γ	$\frac{1}{\sqrt{2\pi}}\frac{\theta}{\mu}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$	$\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$	γ	$\gamma - \frac{\beta\sqrt{1-\gamma}}{\sqrt{n}}$	$o(\frac{1}{n^2})$	$\frac{1}{n}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$	γ
$\mathbb{P}(W_q > 0)$	$\alpha_1 \in (0, 1)$	≈ 1		$\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\delta}\varrho^n\frac{1}{\sqrt{n}} \approx 0$	$\alpha_2 \in (0, 1)$	≈ 1	≈ 1	≈ 0	≈ 1	
$\mathbb{P}(W_q > T)$	$\alpha_1 e^{-\frac{\delta}{1+\delta}\mu T}$	$1 + O(\frac{1}{\sqrt{n}})$	$1 + O(\frac{1}{n})$	≈ 0		$\tilde{G}(T)1_{\{G(T) < \gamma\}}$	α_3 , if $G(T) = \gamma$	≈ 0	$\frac{\tilde{\Phi}(\hat{\beta} + \sqrt{\theta\mu}T)}{\tilde{\Phi}(\hat{\beta})}$	$1 + O(\frac{1}{n})$
Congestion $\frac{\mathbb{E}W_q}{\mathbb{E}S}$	$\alpha_1\frac{1}{1+\delta}$	$\sqrt{n}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$	$n\mu\gamma/\theta$	$\frac{1}{\sqrt{2\pi}}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$	$\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]\alpha_2$	$\mu \int_0^x \tilde{G}(s)ds$	$\mu \int_0^x \tilde{G}(s)ds - \frac{\mu\beta\sqrt{1-\gamma}}{h_G(x)\sqrt{n}}$	$o(\frac{1}{n})$	$\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta}) - \hat{\beta}]$	$n\mu\gamma/\theta$

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

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



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, ...
Then the following points of view are **equivalent**, as $N \uparrow \infty$:

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

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

Erlang-A: QED Approximations (Examples)

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

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:

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

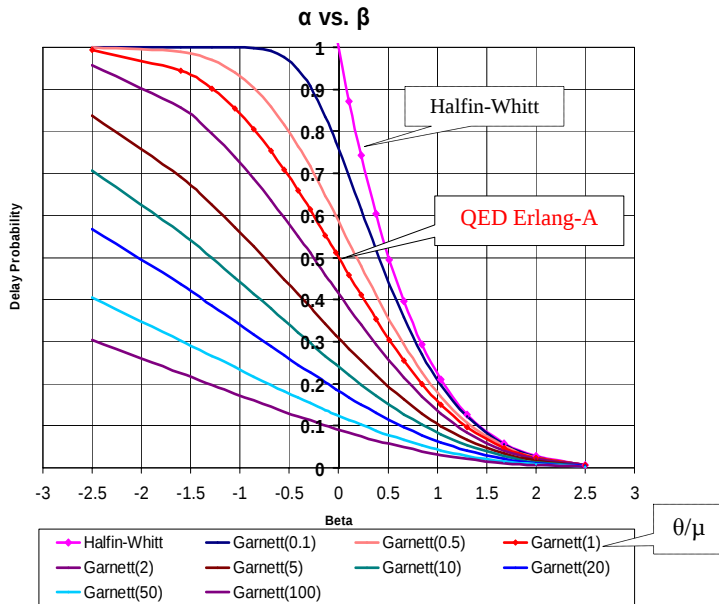
► Probability to Abandon:

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

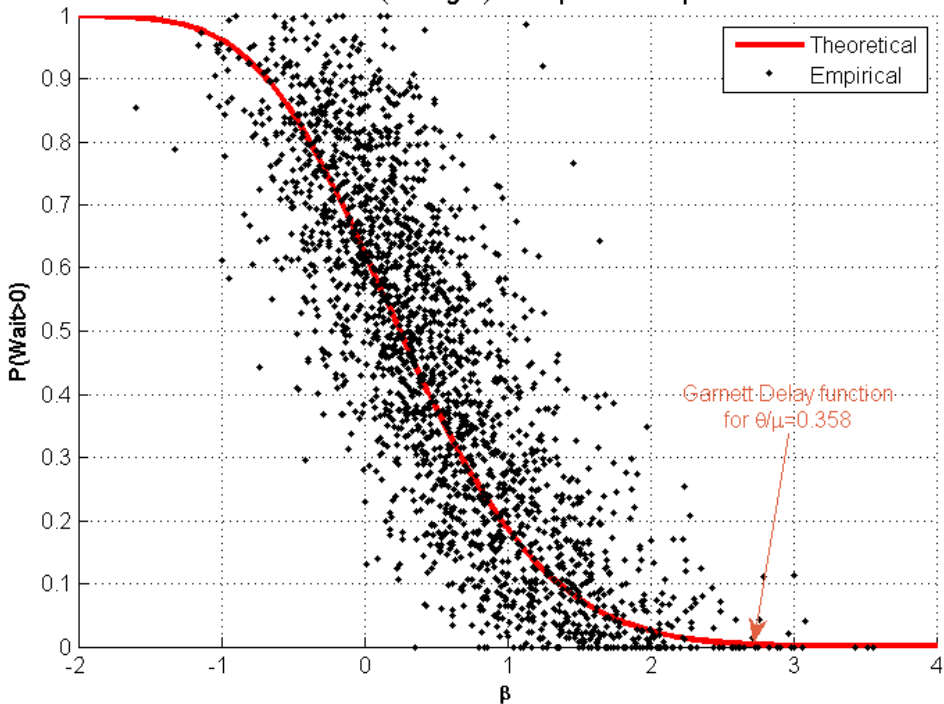
► $P\{\text{Ab}\} \propto E[W_q]$, both order $\frac{1}{\sqrt{n}}$:

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

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



$P(W>0)$ vs. β
Theoretical (Erlang-A) & Empirical Comparison

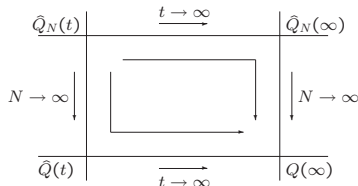


Process Limits (Queueing, Waiting)

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

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

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 Leeuwen
- ▶ Refined Asymptotics (Erlang A/B/C), 2008-2011: Janssen, van Leeuwen, 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), $G/Ph/n$, $\dots$, $G/GI/n+GI$, Measure-Valued Diffusions
 - ▶ **Dimensioning** (Staffing): $M/M/n$, $\dots$, time-varying Q's, V- and Reversed-V, $\dots$
 - ▶ **Control**: V-network, Reversed-V, $\dots$, SBRNets

Back to “Why does Erlang-A Work?”

Theoretical (Partial) Answer:

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

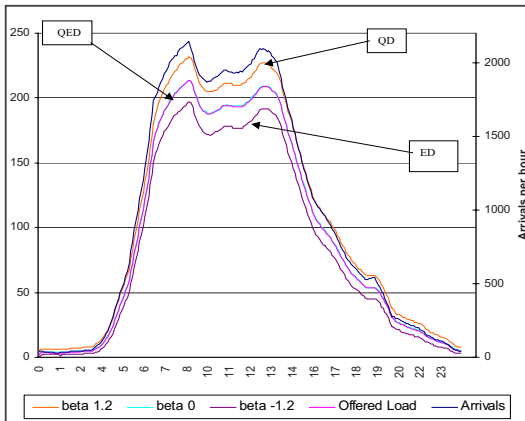
- ▶ **Over-Dispersed Arrivals:** $R + \beta R^c$, c -Staffing ($c \geq 1/2$).
- ▶ **General Patience:** Behavior at the origin matters most (only).
- ▶ **General Services:** Empirical insensitivity beyond the mean.
- ▶ **Heterogeneous Customers / Servers:** State-Collapse.
- ▶ **Time-Varying Arrivals:**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

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

Square-Root Staffing: $N_t = R_t + \beta\sqrt{R_t}$, $-\infty < \beta < \infty$

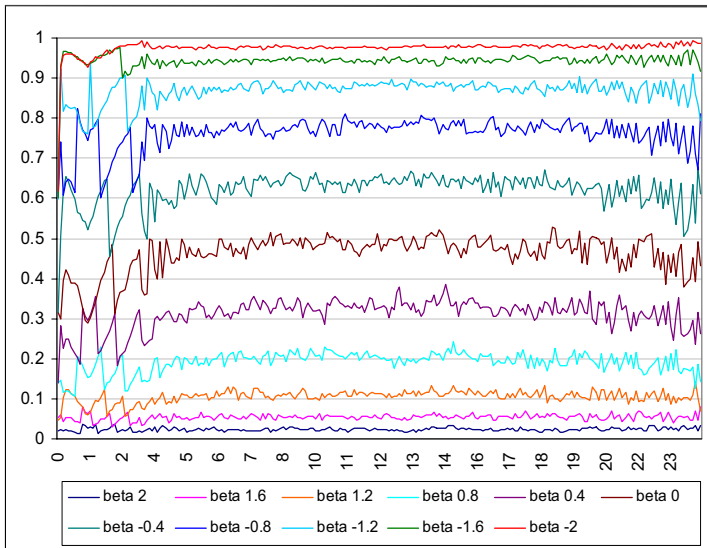
What is R_t , the **Offered-Load** at time t ? ($R_t \neq \lambda_t \times E[S]$)

Arrivals, Offered-Load and Staffing



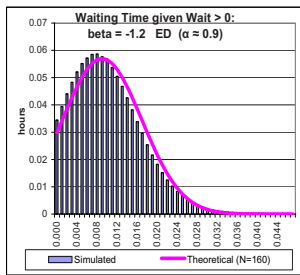
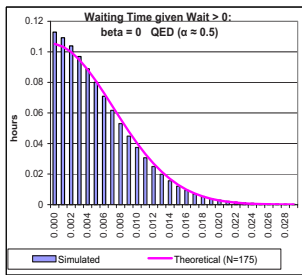
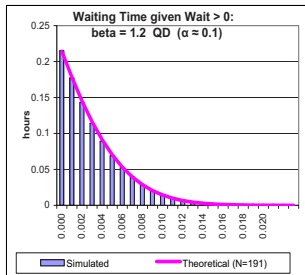
Time-Stable Performance of Time-Varying Systems

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

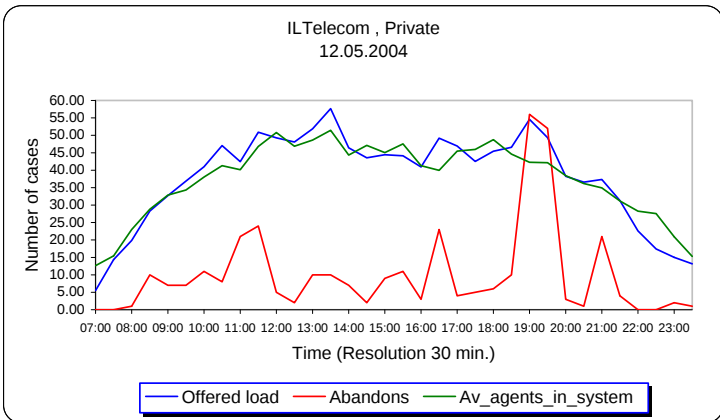
Four (all useful) representations, capturing “**workload before t**”:

$$\begin{aligned} R(t) = E[L(t)] &= \int_{-\infty}^t \lambda(u) \cdot P(S > t - u) du = E[A(t) - A(t - S)] = \\ &= E\left[\int_{t-S}^t \lambda(u) du\right] = E[\lambda(t - S_e)] \cdot E[S] \approx \dots \end{aligned}$$

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

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

- ▶ **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: $L(\cdot) = \text{Least number of servers that guarantees no delay.}$
- ▶ **Offered-Load** Function $R(\cdot) = E[L(\cdot)]$ ($M_t/G/N_t^? + G \leftrightarrow M_t/G/\infty$).



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 $\Rightarrow$ **emotional** load dominates
- ▶ $\Rightarrow$
 - ▶ Offered **Operational** Load
 - ▶ Offered **Cognitive** Load
 - ▶ Offered **Emotional** Load
 - ▶ $\Rightarrow$ **Fair** Division of Load (Routing) between 2 Maternity Wards:
One attending to complications before birth, the other to complications after birth, and both share normal birth

The Technion SEE Center / Laboratory

Data-Based Service Science / Engineering

