# Data-Based Processing Networks

## Inference, Design & Control of Service Systems

Avi Mandelbaum

IE&M **SEE Laboratory**

Technion, Haifa, Israel

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

DSC/e Launch Symposium, December 2013

- **Lecture will be downloadable from my Technion website**

# Research Partners

- **Students**:
  **Aldor$^*$, Baron Yonit$^*$, Carmeli$^*$, Chen Hong$^*$, Cohen Izik$^*$, Feldman Zohar$^*$, Garnett$^*$, Gurvich$^*$, Khudiakov$^*$, Koren, Maman$^*$, Marmor$^*$, Reich$^*$, Rosenshmidt$^*$, Shaikhet$^*$, Senderovic$^*$, Tseytlin$^*$, Yom-Tov$^*$, Yuviler, Zaied$^*$, Zeltyn$^*$, Zychlinski$^*$, Zohar Eti$^*$, Zviran$^*$,** . . .
- **Theory**:
  **Armony, Atar, Azriel, Cohen Izik, Gurvich, Feigin, Huang Junfei, Jelenkovic, Kaspi, Massey, Momcilovic, Reiman, Shimkin, Stolyar, Trofimov, Wasserkrug, Whitt, Zeltyn, Zhang Jiheng, Zhang Hanqin** [1] **,** . . .
- **Empirical/Statistical Analysis**:
  **Brown, Gans, Shen Haipeng, Zhao Linda; Zeltyn; Ritov, Goldberg; Gurvich, Huang Junfei, Liberman; Liu Nan, Ye Han; Armony, Marmor, Tseytlin, Yom-Tov; Gorfine, Ghebali; Feigin, Azriel; Tezcan; Kim Song-Hee, Won Chul Cha; He Shuangchi, Sim Melvyn** . . .
- **Industry**:
  **Mizrahi Bank, Fleet Bank, Rambam Hospital, IBM Research, Hapoalim Bank, Pelephone Cellular, Samsung Hospital,** . . .
- **Technion SEE Center / Laboratory**:
  **Feigin; Trofimov, Nadjharov, Gavako, Kutsy; Senderovic; Liberman, Koren, Plonsky**; Research Assistants, . . .

[1] In many Western countries, there is a short list of popular "first names," but countless "last names." In China, it is just the reverse. The list of last names is short, and the number of first names is in the billions (from chinapage.com/biography/lastname.html).

# Research Goals

- Reality: Service Systems
  e.g. Hospitals, Call Centers, Websites, . . .
- Models = **ServNets**

  **Simple models at the service of complex realities**:

  **Q-Nets, Sim-Nets; F-Nets, D-Nets, . . .**

# Research Goals

- Reality: Service Systems
  e.g. Hospitals, Call Centers, Websites, . . .

- Models = **ServNets**

  **Simple models at the service of complex realities**:

  **Q-Nets, Sim-Nets; F-Nets, D-Nets, . . .**

- Research: **Data-based** **creation**, **analysis** and **validation**

# Research Goals

- Reality: Service Systems
  e.g. Hospitals, Call Centers, Websites, . . .

- Models = **ServNets**

  **Simple models at the service of complex realities**:

  **Q-Nets, Sim-Nets; F-Nets, D-Nets, . . .**

- Research: **Data-based creation**, **analysis** and **validation**

- Goals: (Reproducible) research & teaching (that impact practice)

⇒ Data & analysis of ServNets: accessible, useful

⇒ **Creation & Validation** of ServNets: Automatic, online
(Starting **w/ A. Gal, A. Senderovic, M. Weidlich)**

# Why be Optimistic?

- Research in **OR/QS/IE** + Technologies, at Technion / SEELab:
  - Need Process Mining to overcome "curse of simplicity": "my models too simple to be credible"
    – Validate my "simple" models against complex PM "realities"
    – Enrich them when desired (research) or needed (practice)
    – Create a Science for SEELab technologies/heuristics
- Research in **Process Mining** + Technologies, at TU/e / DSC/e

# Why be Optimistic?

- Research in **OR/QS/IE** + Technologies, at Technion / SEELab:
  - Need Process Mining to overcome "curse of simplicity": "my models too simple to be credible"
    – Validate my "simple" models against complex PM "realities"
    – Enrich them when desired (research) or needed (practice)
    – Create a Science for SEELab technologies/heuristics
- Research in **Process Mining** + Technologies, at TU/e / DSC/e
  - (Humbly submit that) OR/QS can help overcome "curse of dimensionality", through analysis of simple (parsimonious) yet valuable models

# Why be Optimistic?

- Research in **OR/QS/IE** + Technologies, at Technion / SEELab:
  - Need Process Mining to overcome "curse of simplicity": "my models too simple to be credible"
    – Validate my "simple" models against complex PM "realities"
    – Enrich them when desired (research) or needed (practice)
    – Create a Science for SEELab technologies/heuristics
- Research in **Process Mining** + Technologies, at TU/e / DSC/e
  - (Humbly submit that) OR/QS can help overcome "curse of dimensionality", through analysis of simple (parsimonious) yet valuable models
- **Data & Data-Based Models** = natural Meeting Ground, e.g.
  - PM: Create SimNet, QNet, FNet, DNet from hospital data
  - OR: Compare FNet against QNet – **Accuracy**
  - OR: Refine FNet with DNet – improve accuracy

# Why be Optimistic?

- Research in **OR/QS/IE** + Technologies, at Technion / SEELab:
  - Need Process Mining to overcome "curse of simplicity": "my models too simple to be credible"
    – Validate my "simple" models against complex PM "realities"
    – Enrich them when desired (research) or needed (practice)
    – Create a Science for SEELab technologies/heuristics
- Research in **Process Mining** + Technologies, at TU/e / DSC/e
  - (Humbly submit that) OR/QS can help overcome "curse of dimensionality", through analysis of simple (parsimonious) yet valuable models
- **Data & Data-Based Models** = natural Meeting Ground, e.g.
  - PM: Create SimNet, QNet, FNet, DNet from hospital data
  - OR: Compare FNet against QNet – **Accuracy**
  - OR: Refine FNet with DNet – improve accuracy
  - **OR+PM: Validate FNet+DNet against SimNet $\approx$ Reality** – **Value**
    Note: No need for QNets

# Contents

1. **2 simple models at the service of complex realities**
   - Emergency Department: **Time-Varying**
     - Mass Casualty Event: QNet and FNet (2 hours) - **performance**
     - Normal: QNet, DNet and SimNet (over 1 day) - **staffing**
   - Call Center: **Stationary**
     - Q-Net and D-Net (piecewise stationary) - **congestion laws**

2. Empirical adventures at the Technion IE&M **SEELab**:
   **Mining operational building blocks of ServNets**
   - Primitives
   - Structure
   - Protocols

# Applying (Queueing) Asymptotics

**There are by now numerous insightful asymptotic queueing models at our disposal, and many arise from, and create, deep beautiful theory:**

**Has it helped one approximate or simulate a service system more efficiently, estimate its parameter more accurately, teach it to our students more effectively, perhaps even manage the system better?**

# Applying (Queueing) Asymptotics

**There are by now numerous insightful asymptotic queueing models at our disposal, and many arise from, and create, deep beautiful theory:**

**Has it helped one approximate or simulate a service system more efficiently, estimate its parameter more accurately, teach it to our students more effectively, perhaps even manage the system better?**

***I am of the opinion that the answers to such questions have been too often negative, and that positive answers must have theory and applications nurture each other, which is good.***

***How to make this good happen? My approach has been to marry theory with data , which has been supported by (what I only recently came to realize is) process (Q-Net) mining: building-blocks, structure, protocols; laws of "nature".***

# Prevalent (Asymptotic) Approximations

System (Data)

Phenomenon

Model

SimNet Q-Net

# Prevalent (Asymptotic) Approximations

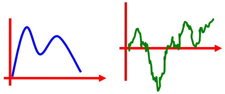
**System (Data)**

**Phenomenon**

**Accuracy**

**Model** ⟷ **Approximation**

**SimNet** **Q-Net** **F-Net** **D-Net**



**Data–Based** Prevalent (Asymptotic) Approximations **Models**

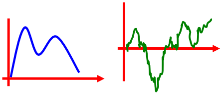
System (Data)

Phenomenon

Accuracy

Model ⟷ Approximation

SimNet Q-Net ⟷ F-Net D-Net **Models**



## Data–Based Prevalent (Asymptotic) Approximations Models

System (Data)

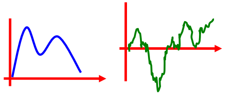
Phenomenon

Value

Accuracy

Model ⟷ Approximation

SimNet Q-Net ⟷ F-Net D-Net Models



## Data–Based Prevalent (Asymptotic) Approximations Models

System (Data)

Phenomenon

Value

Accuracy

Model ⟷ Approximation

SimNet Q-Net ⟷ F-Net D-Net Models

**System** = Coin Tossing, **Model** = Binomial ; de Moivre 1738

**Approx. / Models:** SLLN **(FNets)**, CLT **(DNets)** ; Laplace 1810

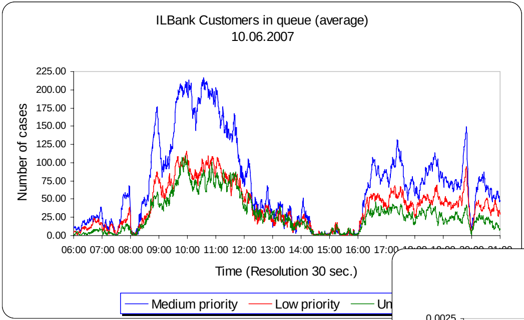
**Value:** Exceeds Value of originating stylized model

Normal, Brownian Motion ; Bachalier 1900

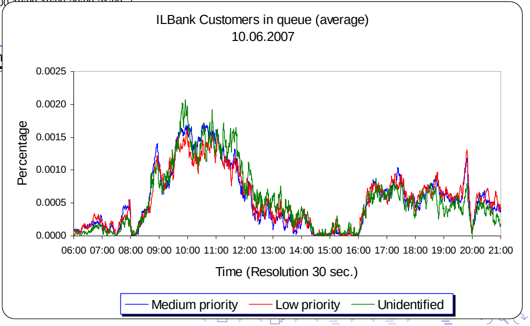
Poisson ; Poisson 1838

# Simple Models (Parsimony) Supported by Congestion Laws

## 3 Queue-Lengths at 30 sec. resolution (ILBank, 10/6/2007)



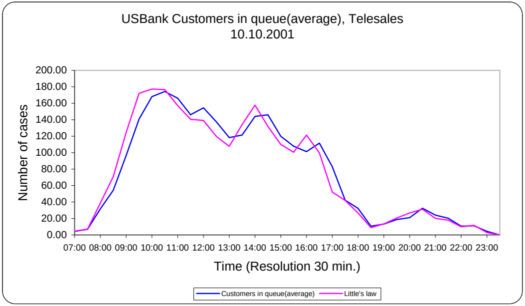

**1 Queue "Shape"**
**Deterministic Proportions**




- Area normalized to 100%
- **State-Space Collapse**

# Little's Law $L = \lambda \times W$, in a Time-Varying Environment

## Time-Gap: # in System lags behind Little / 30 min

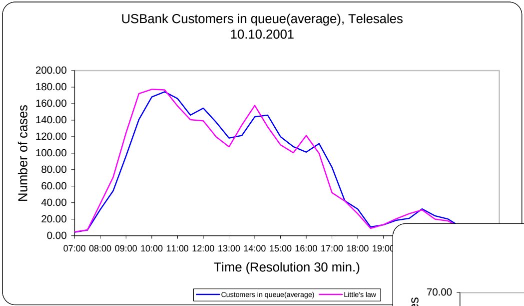



**Call Center**:

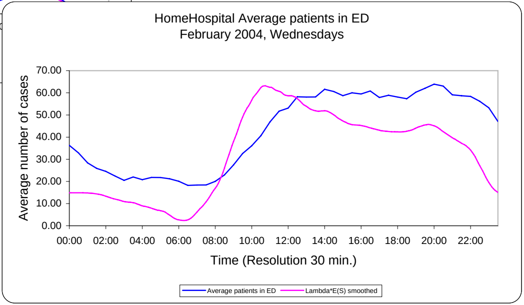
$\Rightarrow$ **Piecewise Steady-State**

# Little's Law $L = \lambda \times W$, in a Time-Varying Environment

## **Time-Gap**: **# in System** lags behind **Little / 30 min**




**Call Center**:
$\Rightarrow$ **Piecewise Steady-State**




**Emergency Dept**:
$\Rightarrow$ **Time-Varying Transient**

$$EL(t) = \tilde{\lambda}(t) \times EW,$$

$$\tilde{\lambda}(t) = E\lambda(t - W_e).$$

(Bertsimas, Mourtzinou;
Fralix, Riano, Serfozo)

# ER / ED Environment: Service Network

## Acute (Internal, Trauma)

## Walking

## Multi-Trauma

# ED-Environment in Israel

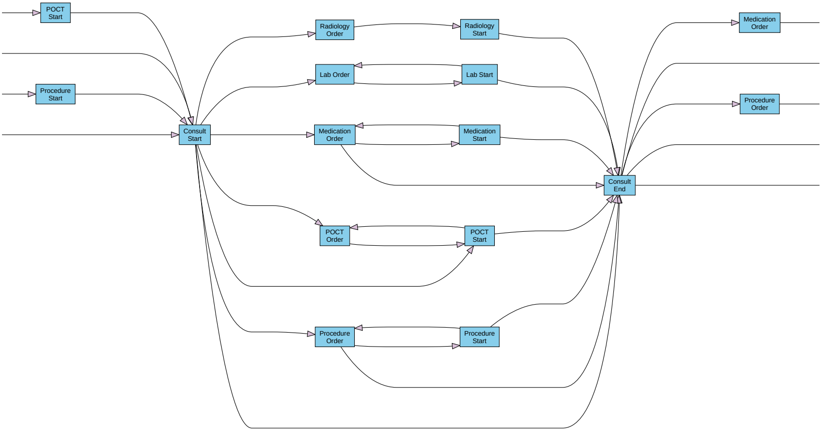
# Queueing in a “Good" Hospital

## Tong-ren Hospital at 6am, Beijing

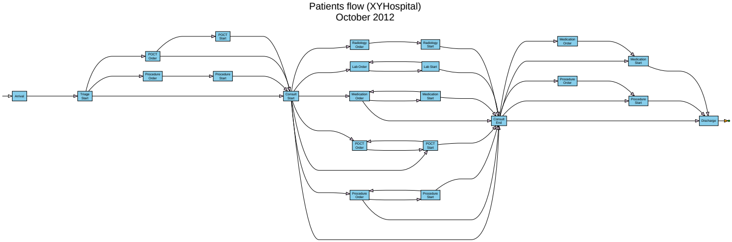
# Simple Models at the Service of Complex Realities: FNets

1. ED in Normal days (Time-Varying Periodic): Personnel Staffing (offline)
2. ED in **Mass Casualty Event** (**Transient**): Forecasting, Staffing (online)

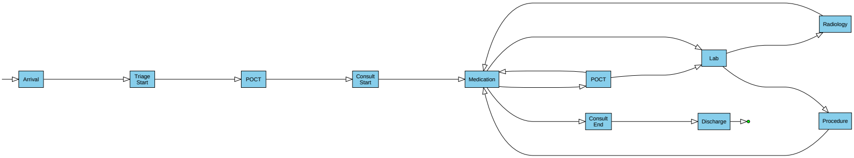
## Emergency Department in XYHospital, October 2012



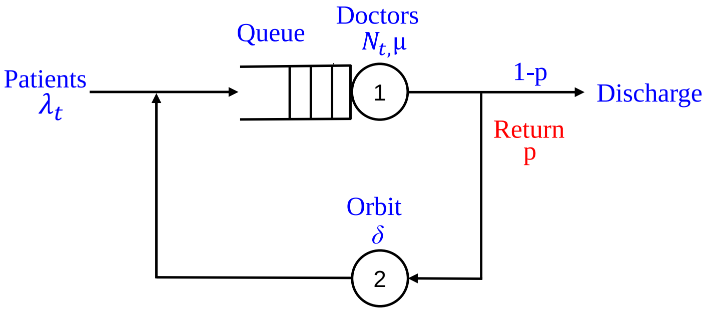
# Recurrent Service Process in the ED




Capture **Recurrent** nature of service process: **Multiple doctor visits**

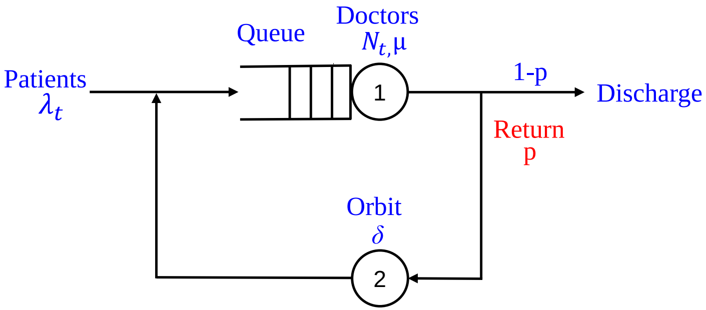


# The Basic Service-Network Model: Erlang-R




**w/ G. Yom-Tov**

# The Basic Service-Network Model: Erlang-R



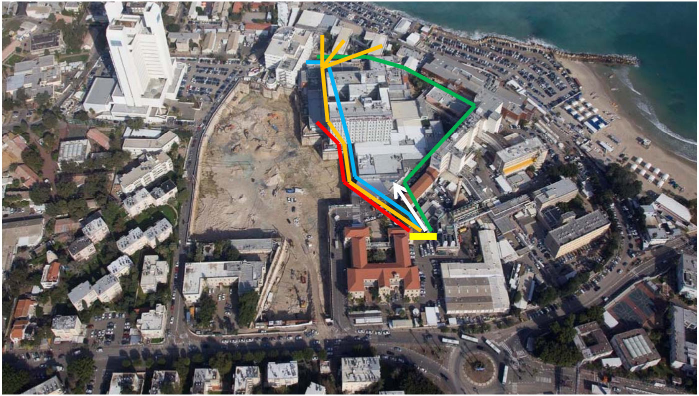

**w/ G. Yom-Tov**

**2-station "Jackson" Network** = (M/M/S, M/M/$\infty$) :

- $\lambda_t$ – **Time-Varying Arrival** rate
- $N_t$ – Number of **Servers** (Physicians, or Nurses)
- $\mu$ – **Service** rate ($E$[Service] $= \frac{1}{\mu}$)
- $p$ – **Return** (ReEntrant) fraction
- $\delta$ – **Orbit-to-Queue** rate ($E$[Delay] $= \frac{1}{\delta}$)

# RFID-Based Data in Mass Casualty Event (Drill)

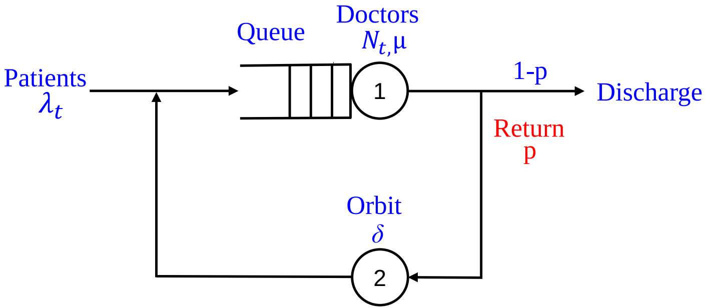
## Chemical MCE, Rambam Hospital (May 2010, 11:00-13:00)



**Fluid** Model:

- **Predictable** Variability ⇒ **Time-Varying**
- **Stochastic** Individualism averaged-out ⇒ **Deterministic**

# Fluid Model ↔ (Time-Varying) Erlang-R System




Functional Strong Law of Large Numbers, for a 2-station QNet. BUT

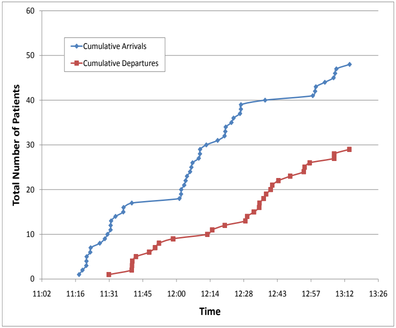
**FNet** = ODE: derived **directly** (no QNet), spreadsheet "solution"

$$\frac{d}{dt} q_t^1 = \lambda_t - \mu \cdot \left(q_t^1 \wedge N_t\right) + \delta \cdot q_t^2$$
$$\frac{d}{dt} q_t^2 = p \cdot \mu \cdot \left(q_t^1 \wedge N_t\right) - \delta \cdot q_t^2$$

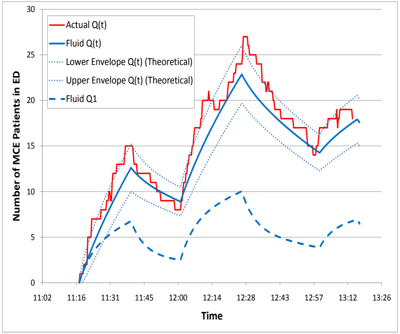
# Erlang-R Value: FNet vs. Data

## Chemical MCE Drill (Israel, May 2010, 11:00-13:00)

**Arrivals** & **Departures** (RFID)




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

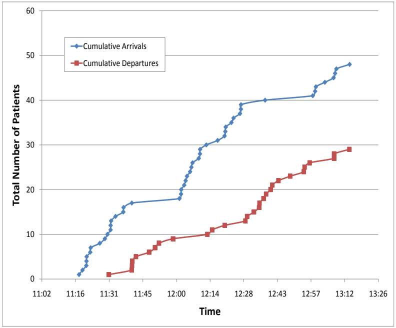



- **Recurrent/Repeated** services in Chemical MCE: injection every 15/30/60 min

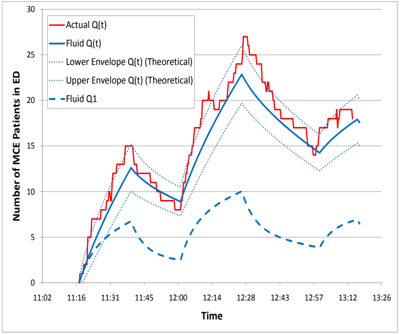
# Erlang-R Value: FNet vs. Data

## Chemical MCE Drill (Israel, May 2010, 11:00-13:00)

**Arrivals** & **Departures** (RFID)




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




- **Recurrent/Repeated** services in Chemical MCE: injection every 15/30/60 min
- **Fluid** = ODE
- **Diffusion** (confidence band), via F. Central Limit Theorem: Usefully narrow

# A Data-Based Framework, or “Erlang-R in the ED"

**System** = e.g. Emergency Department

- **QNet** = Erlang-R (time-varying 2-station Jackson; w/ Yom-Tov)
- **FNets** = 2-dim dynamical system (Massey & Whitt)
- **DNets** = 2-dim Markovian Service Net (w/ Massey and Reiman)
- **SimNet** = Customized ED-Simulator (Marmor & Sinreich)

# A Data-Based Framework, or “Erlang-R in the ED"

**System** = e.g. Emergency Department

- **QNet** = Erlang-R (time-varying 2-station Jackson; w/ Yom-Tov)
- **FNets** = 2-dim dynamical system (Massey & Whitt)
- **DNets** = 2-dim Markovian Service Net (w/ Massey and Reiman)
- **SimNet** = Customized ED-Simulator (Marmor & Sinreich)

**Framework**: Mining (all) ServNets from Data

- **MCE ED**: FNet $\Rightarrow$ Census, DNet = Confidence band
  Performance Analysis, Prediction
  Validated against Data

- **Normal ED**: FNet $\Rightarrow$ Physician offered-load $\Rightarrow$ $\sqrt{}$-Staffing
  Staffing to stabilize operational performance
  Validated against SimNet

# Call-Center Environment: Service Network

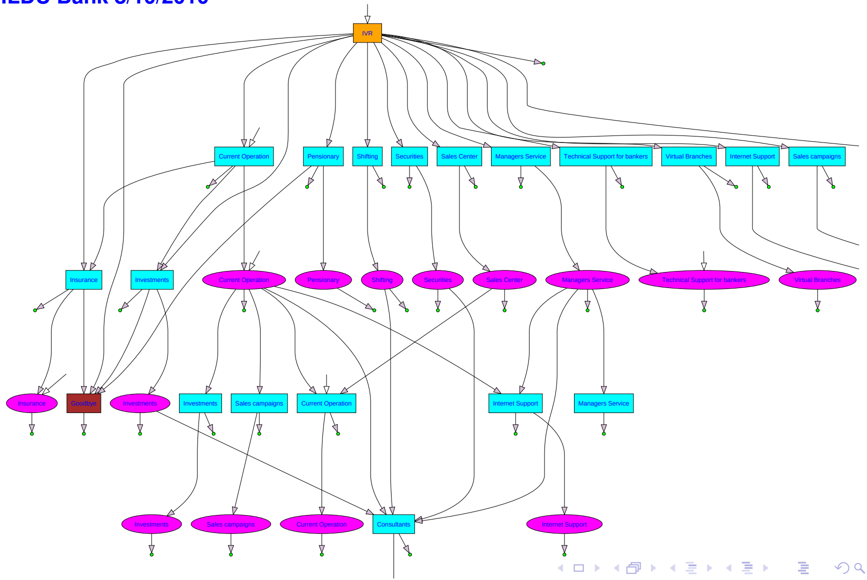
# Operational Focus

## Operational Measures:

- Surrogates for overall performance: Financial, Psychological; Clinical
- Easiest to quantify, measure, track online, react upon

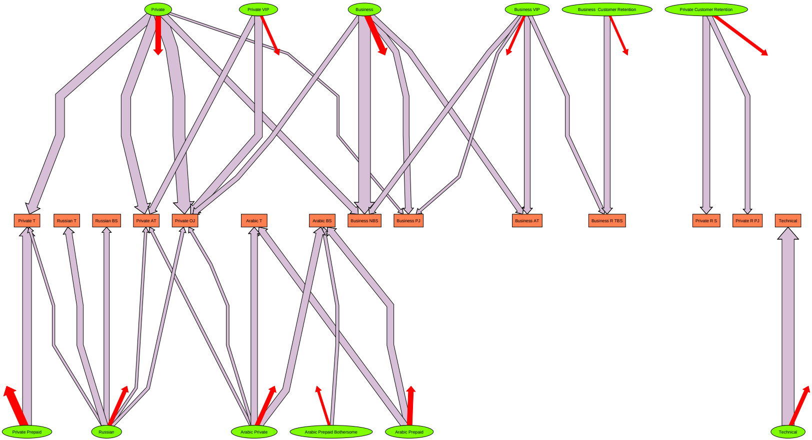
# Customer Flow in Call Centers

## ILDU Bank 8/10/2010



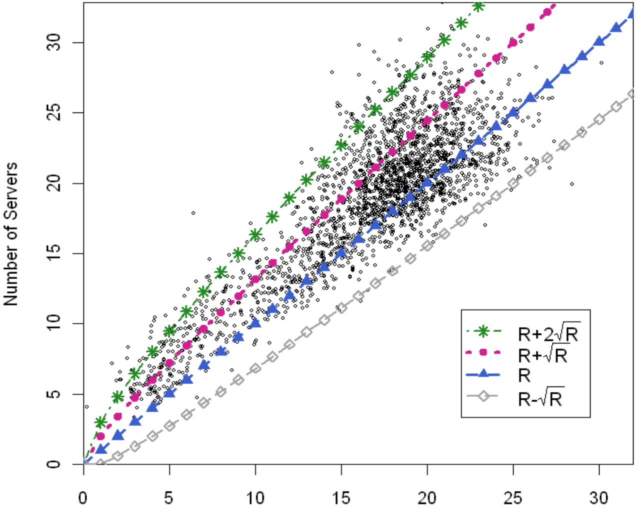
# Impatient Customers - Isolate or Aggregate

## ILTelecom 9/3/2008



# Beyond Fluid: $\#$Agents vs. Offered-Load ($N \approx R + \beta\sqrt{R}$)

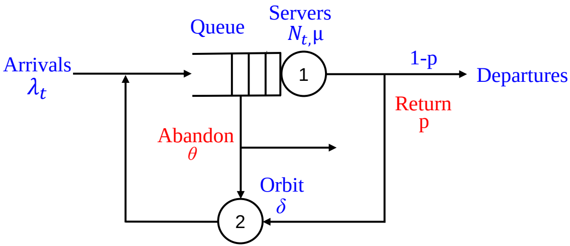
IL Telecom; June-September, 2004 (2205 30min intervals, over 13 weeks, week-days)




**e.g. Offered-load R $\overset{avg}{=}$ 5 calls per min $\times$ 3.2 min per call = 16 Erlangs**

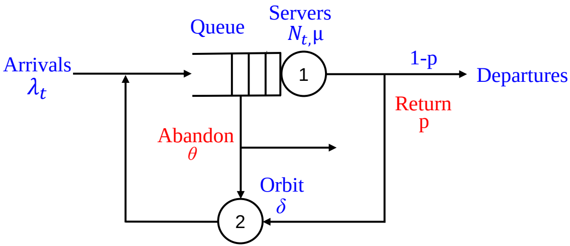
# Model Selection: As Simple as Possible but Not Simpler

**Service with Retrials and Abandonment**; **w/ Massey, Reiman, Stolyar**



# Model Selection: As Simple as Possible but Not Simpler

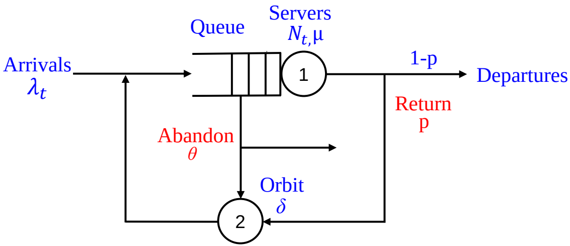
**Service with Retrials and Abandonment; w/ Massey, Reiman, Stolyar**




- Call centers: Visit durations naturally measured in minutes
  - Arrival rates are “constant" during visit
  - Returns occur hours after visit

## Model Selection: As Simple as Possible but Not Simpler

**Service with Retrials and Abandonment**; **w/ Massey, Reiman, Stolyar**

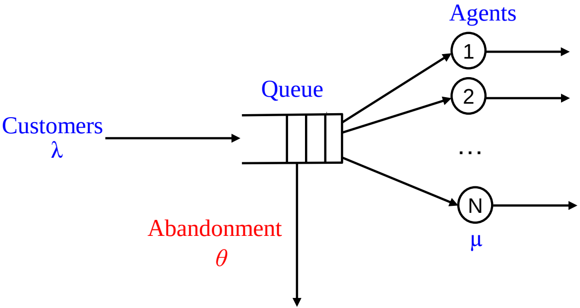



- Call centers: Visit durations naturally measured in minutes
  - Arrival rates are “constant" during visit
  - Returns occur hours after visit

$\Rightarrow$ “Select" Base Model (of 1/2 hour):

**Stationary, Abandonment**

# A Basic Staffing Model: Erlang-A




**w/ O. Garnett**

**"Birth & Death" Queue** = M/M/N + M (Palm 1940's):

- $\lambda$ – **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).

# Erlang-A: Is it Relevant?

**Experience:**

- Arrival process **not pure Poisson** (time-varying, $\sigma^2$ too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (behavior-dependent).

# Erlang-A: Is it Relevant?

**Experience:**

- Arrival process **not pure Poisson** (time-varying, $\sigma^2$ too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (behavior-dependent).

- Building Blocks need **not be independent** (eg. long wait associated with long service; **w/ M. Reich & Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills): **w/ R. Atar, G. Shaikhet; R. Atar, I. Gurvich,** . . .
- Customers return for service (after busy, abandonment; dependently: **P. Khudiakov, R. Ghebali, M. Gorfine, P. Feigin**)
- $\cdots$, and more.

# Erlang-A: Is it Relevant?

**Experience:**

- Arrival process **not pure Poisson** (time-varying, $\sigma^2$ too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (behavior-dependent).

- Building Blocks need **not be independent** (eg. long wait associated with long service; **w/ M. Reich & Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills): **w/ R. Atar, G. Shaikhet; R. Atar, I. Gurvich,** . . .
- Customers return for service (after busy, abandonment; dependently: **P. Khudiakov, R. Ghebali, M. Gorfine, P. Feigin**)
- $\cdots$, and more.

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

# Erlang-A: Is it Relevant?

**Experience:**

- Arrival process **not pure Poisson** (time-varying, $\sigma^2$ too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (behavior-dependent).

- Building Blocks need **not be independent** (eg. long wait associated with long service; **w/ M. Reich & Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills): **w/ R. Atar, G. Shaikhet; R. Atar, I. Gurvich,** . . .
- Customers return for service (after busy, abandonment; dependently: **P. Khudiakov, R. Ghebali, M. Gorfine, P. Feigin**)
- $\cdots$, and more.

Question: **Is Erlang-A Relevant? Robust enough? YES !**

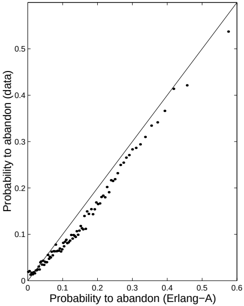
- **Practice**: Staffing engine of Work-Force Management software
- **Theory**: Theoretical engine of Operational Regimes

  **QD**, **ED**, **QED**

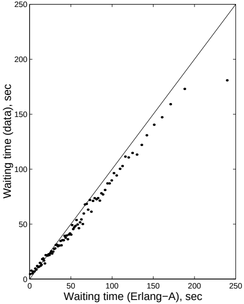
# Erlang-A: Fitting a Simple Model to a Complex Reality

## Hourly Performance vs. Erlang-A Predictions (1 year)

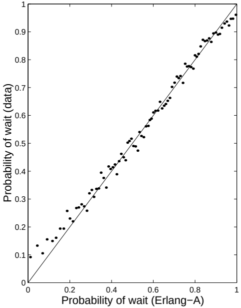
### % Abandon




### $\mathrm{E}[Wait]$




### $\%\{Wait > 0\}$

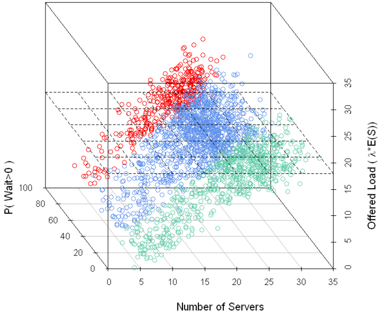



- Empirically-Based & Theoretically-Supported Estimation of (Im)Patience: $\hat{\theta} = P\{Ab\}/\mathrm{E}[W_q]$)
- Small Israeli Bank (more examples in progress)
- Hourly performance vs. Erlang-A predictions, 1 year: aggregated groups of 40 similar hours

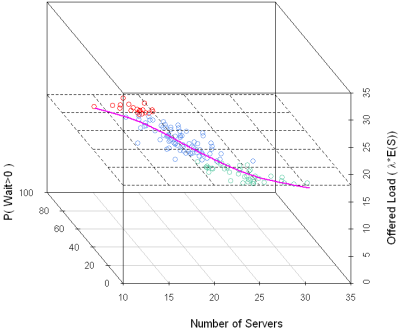
# Operational Regimes: Q(uality) vs. E(fficiency)

IL Telecom; June-September, 2004

**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**: $\geq 75\%$

# Asymptotic Erlang-X (Markovian Q's)

- 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(3): Avram, Janssen, van Leeuwaarden**
- **Refined Asymptotics (Erlang A/B/C, . . .), 2008-2013: Janssen, van Leeuwaarden, Zhang, Zwart**
- Production Q's, 2011: Reed & Zhang
- Universal Erlang-A, 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

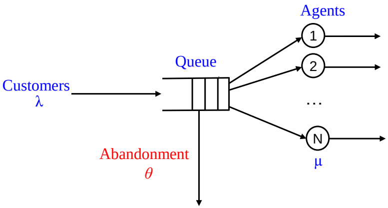
# Asymptotic Landscape: 9 Operational Regimes, and then some

## Erlang-A, w/ I. Gurvich & J. Huang

| Erlang-A $\mu$ & $\theta$ fixed | Conventional scaling | | | Many-Server scaling | | | NDS scaling | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sub | Critical | Over | QD | QED | ED | Sub | Critical | Over |
| Offered load per server | $\frac{1}{1+\delta}$ | $1-\frac{\beta}{\sqrt{n}}$ | $\frac{1}{1-\gamma}$ | $\frac{1}{1+\delta}$ | $1-\frac{\beta}{\sqrt{n}}$ | $\frac{1}{1-\gamma}$ | $\frac{1}{1+\delta}$ | $1-\frac{\beta}{n}$ | $\frac{1}{1-\gamma}$ |
| Arrival rate $\lambda$ | $\frac{\mu}{1+\delta}$ | $\mu-\frac{\beta}{\sqrt{n}}\mu$ | $\frac{\mu}{1-\gamma}$ | $\frac{n\mu}{1+\delta}$ | $n\mu-\beta\mu\sqrt{n}$ | $\frac{n\mu}{1-\gamma}$ | $\frac{n\mu}{1+\delta}$ | $n\mu-\beta\mu$ | $\frac{n\mu}{1-\gamma}$ |
| # servers | $1$ | | | $n$ | | | $n$ | | |
| Time-scale | $n$ | | | $1$ | | | $n$ | | |
| Impatience rate | $\theta/n$ | | | $\theta$ | | | $\theta/n$ | | |
| Staffing level | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu}(1+\frac{\beta}{\sqrt{n}})$ | $\frac{\lambda}{\mu}(1-\gamma)$ | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}$ | $\frac{\lambda}{\mu}(1-\gamma)$ | $\frac{\lambda}{\mu}(1+\delta)$ | $\frac{\lambda}{\mu}+\beta$ | $\frac{\lambda}{\mu}(1-\gamma)$ |
| Utilization | $\frac{1}{1+\delta}$ | $1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{\sqrt{n}}$ | $1$ | $\frac{1}{1+\delta}$ | $1-\sqrt{\frac{\theta}{\mu}}\frac{\hat{h}(\hat{\beta})}{\sqrt{n}}$ | $1$ | $\frac{1}{1+\delta}$ | $1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{n}$ | $1$ |
| $\mathbb{E}(Q)$ | $\frac{1}{\delta(1+\delta)}$ | $\sqrt{n}g(\hat{\beta})$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $\frac{1}{\delta}\varrho_n$ | $\sqrt{n}g(\hat{\beta})\alpha$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $o(1)$ | $ng(\hat{\beta})$ | $\frac{n^2\mu\gamma}{\theta(1-\gamma)}$ |
| $\mathbb{P}(Ab)$ | $\frac{1}{n}\frac{1}{\delta}\frac{\theta}{\mu}$ | $\frac{\theta}{\sqrt{n}\mu}g(\hat{\beta})$ | $\gamma$ | $\frac{1}{n}\frac{(1+\delta)}{\delta}\frac{\theta}{\mu}\varrho_n$ | $\frac{\theta}{\sqrt{n}\mu}g(\hat{\beta})\alpha$ | $\gamma$ | $o(\frac{1}{n^2})$ | $\frac{\theta}{n\mu}g(\hat{\beta})$ | $\gamma$ |
| $\mathbb{P}(W_q>0)$ | $\frac{1}{1+\delta}$ | $\approx 1$ | | $\varrho_n$ | $\alpha\in(0,1)$ | $\approx 1$ | $\approx 0$ | $\approx 1$ | |
| $\mathbb{P}(W_q>T)$ | $\frac{1}{1+\delta}e^{-\frac{\delta}{1+\delta}\mu T}$ | $1+O(\frac{1}{\sqrt{n}})$ | $1+O(\frac{1}{n})$ | $\approx 0$ | | $f(T)$ | $\approx 0$ | $\frac{\bar{\Phi}(\hat{\beta}+\sqrt{\theta\mu}T)}{\bar{\Phi}(\hat{\beta})}$ | $1+O(\frac{1}{n})$ |
| Congestion $\frac{\mathbb{E}W_q}{\mathbb{E}S}$ | $\frac{1}{\delta}$ | $\sqrt{n}g(\hat{\beta})$ | $n\mu\gamma/\theta$ | $\frac{1}{n}\frac{(1+\delta)}{\delta}\varrho_n$ | $\frac{\alpha}{\sqrt{n}}g(\hat{\beta})$ | $\frac{\mu\gamma}{\theta}$ | $o(\frac{1}{n})$ | $g(\hat{\beta})$ | $n\mu\gamma/\theta$ |

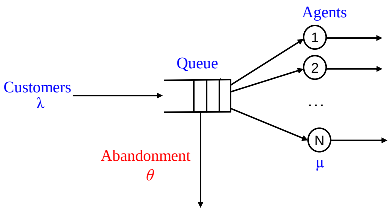
- Conventional: Ward & Glynn (03, $G/G/1+G$)
- Many-Server:
  - QED: Halfin-Whitt (81), w/ Garnett & Reiman (02)
  - ED: Whitt (04)
  - NDS: Atar (12)
- **“Missing”**: ED+QED; Hazard-rate scaling (M/M/N+G); Time-Varying, Non-Parametric; Moderate- and Large-Deviation; Networks (multi-regimes)

# Universal Approximations: Erlang-A (M/M/N + M)




**w/ I. Gurvich & J. Huang**

# Universal Approximations: Erlang-A (M/M/N + M)

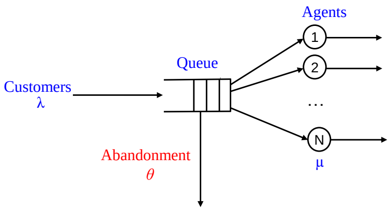



w/ **I. Gurvich & J. Huang**

- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^+, \quad q = 0, 1, \ldots$$

# Universal Approximations: Erlang-A (M/M/N + M)




w/ **I. Gurvich & J. Huang**

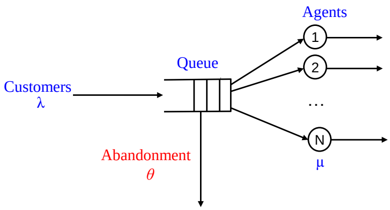
- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^+, \quad q = 0, 1, \ldots$$

- **FNet**: Dynamical (Deterministic) System – ODE

$$dx_t = F(x_t)dt, \quad t \geq 0$$

# Universal Approximations: Erlang-A (M/M/N + M)




w/ **I. Gurvich & J. Huang**

- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^+, \quad q = 0, 1, \ldots$$

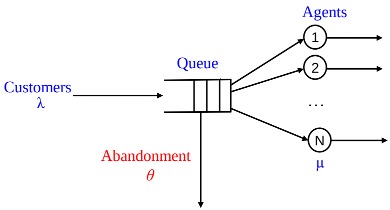
- **FNet**: Dynamical (Deterministic) System – ODE

$$dx_t = F(x_t)dt, \quad t \geq 0$$

- **DNet**: Universal (Stochastic) Approximation – SDE

$$dY_t = F(Y_t)dt + \sqrt{2\lambda}\, dB_t, \quad t \geq 0$$

# Universal Approximations: Erlang-A (M/M/N + M)




w/ **I. Gurvich & J. Huang**

- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^+, \quad q = 0, 1, \ldots$$

- **FNet**: Dynamical (Deterministic) System – ODE

$$dx_t = F(x_t)dt, \quad t \geq 0$$

- **DNet**: Universal (Stochastic) Approximation – SDE

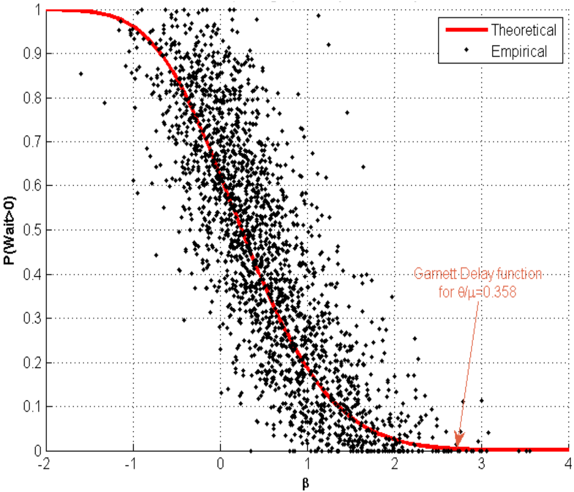
$$dY_t = F(Y_t)dt + \sqrt{2\lambda}\, dB_t, \quad t \geq 0$$

**eg.** $\mu = \theta$ : $\quad \dot{x} = \lambda - \mu \cdot x, \quad Y =$ OU process

**Parsimonious (Tractable, Robust), Accurate, Valuable**

# Erlang-A Value: DNet $P(W_q > 0)$ vs. Data

IL Telecom; June-September, 2004 (2205 30min intervals, weekdays)




- Approximations, w/ Patience ≈ 3× Service-Duration ($\mu/\theta \approx 3$)

## Accuracy: DNet vs. QNet

- $\Delta^\lambda$ is the "balancing" state, obtained by solving

$$\lambda = \mu(n \wedge \Delta^\lambda) + \theta(\Delta^\lambda - n)^+.$$

Solution: $\Delta^\lambda = \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - n\right)^+ \left(1 - \frac{\mu}{\theta}\right).$

Specifically: **QD** = $\frac{\lambda}{\mu}$; **ED** = $n + \frac{1}{\theta}(\lambda - n\mu)$; **QED** = $n + \mathcal{O}(\sqrt{\lambda}))$

- Centered processes (excursions):

$$\tilde{Q}^\lambda(\cdot) = Q(\cdot) - \Delta^\lambda, \quad \tilde{Y}^\lambda(\cdot) = Y(\cdot) - \Delta^\lambda.$$

# Accuracy: DNet vs. QNet

- $\Delta^\lambda$ is the "balancing" state, obtained by solving

$$\lambda = \mu(n \wedge \Delta^\lambda) + \theta(\Delta^\lambda - n)^+.$$

Solution: $\Delta^\lambda = \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - n\right)^+ \left(1 - \frac{\mu}{\theta}\right).$

Specifically: **QD** = $\frac{\lambda}{\mu}$; **ED** = $n + \frac{1}{\theta}(\lambda - n\mu)$; **QED** = $n + \mathcal{O}(\sqrt{\lambda})$)

- Centered processes (excursions):

$$\tilde{Q}^\lambda(\cdot) = Q(\cdot) - \Delta^\lambda, \quad \tilde{Y}^\lambda(\cdot) = Y(\cdot) - \Delta^\lambda.$$

**Theorem**: For $f$ bounded by an $m$-degree polynomial ($m \geq 0$):

$$\mathbb{E}f(\tilde{Q}^\lambda(\infty)) - \mathbb{E}f(\tilde{Y}^\lambda(\infty)) = \mathcal{O}(\sqrt{\lambda}^{m-1}).$$

- **Accurate**: more than heavy-traffic *limits*

# Simplicity: Why $2\lambda$?

- Semi-martingale representation of the B&D process: Fluid + Martingale
- Predictable quadratic variation:

$$\int_0^t [\lambda + \mu(Q_s \wedge n) + \theta(Q_s - n)^+] ds$$

- In steady-state, arrival rate $\equiv$ departure rate:

$$\lambda = \mathbb{E}[\mu(Q_s \wedge n) + \theta(Q_s - n)^+]$$

- Expectation of the predictable quadratic variation:

$$\mathbb{E} \int_0^t [\lambda + \mu(Q_s \wedge n) + \theta(Q_s - n)^+] ds = 2\lambda t$$

- **Simple** $\Rightarrow$ **Tractable, Robust**: $\text{dMartingale}_t \approx \sqrt{2\lambda} \cdot \text{dBrownian}_t$

# Reconciling Time-Varying and Steady-State Models

- **Rigid** (fixed) staffing level during a time-varying shift:
  Doomed to alternate between overloading and underloading
- **Flexible** staffing:
  Can design **time-varying staffing** that achieves, **at all times, Steady-State performance**
  via Square-Root Staffing (Modified Offered-Load)

# Reconciling Time-Varying and Steady-State Models

- **Rigid** (fixed) staffing level during a time-varying shift:
  Doomed to alternate between overloading and underloading
- **Flexible** staffing:
  Can design **time-varying staffing** that achieves, **at all times, Steady-State performance**
  via Square-Root Staffing (Modified Offered-Load)
- **History**:
  - Jennings, M., Reiman, Whitt (1996): Emergence of the phenomenon, with infinite-server heuristics
  - Feldman, M., Massey, Whitt (2008): Stabilize delay probability with QED staffing, with little theory
  - Liu and Whitt (2012): Stabilize abandonment probability, with ED theory
  - w/ Huang, Gurvich (ongoing): QED theory

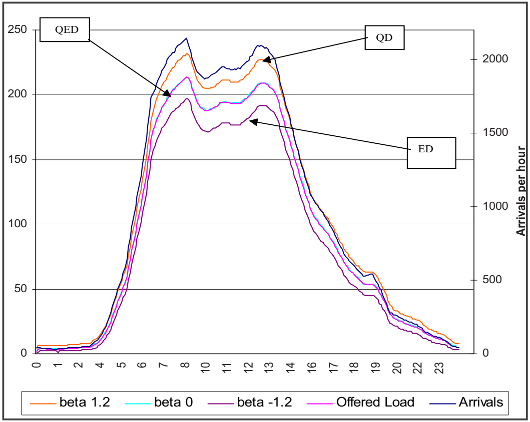
# Time-Varying Arrival Rates

**Square-Root Staffing:**

$$N(t) = R(t) + \beta\sqrt{R(t)}, \quad -\infty < \beta < \infty.$$

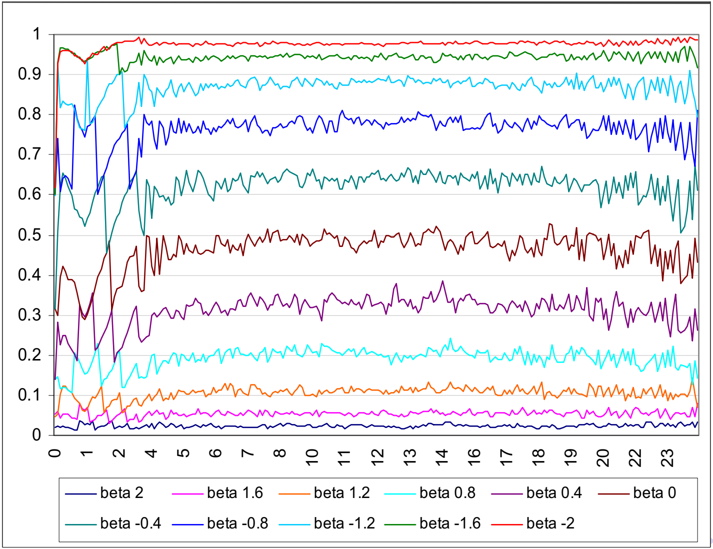
$R(t)$ is the **Offered-Load** at time $t$ $\quad ( R(t) \neq \lambda(t) \times E[S] )$

## Arrivals, Offered-Load and Staffing



# Time-Stable Performance of Time-Varying Systems

**Delay Probability** = as in the **Stationary Erlang-A / R**



# Calculating the Offered-Load $R(t)$, Theoretically

- Offered-Load Process: $\boldsymbol{L(\cdot)}$ = **Least** number of **servers** that guarantees **no delay**.
- **Offered-Load** Function $\boldsymbol{R(t)} = E[L(t)],\ t \geq 0$.

  Think $M_t/G/N_t^? + G$ vs. $\boldsymbol{M_t/G/\infty}$: **Ample-Servers**.

## Calculating the Offered-Load $R(t)$, Theoretically

- Offered-Load Process: $\boldsymbol{L(\cdot)}$ = **Least** number of **servers** that guarantees **no delay**.
- **Offered-Load** Function $\boldsymbol{R(t)} = E[L(t)]$, $t \geq 0$.

  Think $M_t/G/N_t^? + G$ vs. $\boldsymbol{M_t/G/\infty}$: **Ample-Servers**.

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

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

$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t - S_e)] \cdot E[S] \approx \dots .$$

- $\{\boldsymbol{A(t)},\ t \geq 0\}$ Arrival-Process, rate $\boldsymbol{\lambda(\cdot)}$;
- $\boldsymbol{S}$ ($\boldsymbol{S_e}$) generic Service-Time (Residual Service-Time).

# Calculating the Offered-Load $R(t)$, Theoretically

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

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

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

$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t - S_e)] \cdot E[S] \approx \ldots .$$

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

# Calculating the Offered-Load $R(t)$, Theoretically

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

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

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

$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t - S_e)] \cdot E[S] \approx \dots .$$

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

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

## Pause for a Commercial:

# Pause for a Commercial: The Technion SEELab

Technion
S.E.E
Center
Service Enterprise Engineering

# Technion SEE = Service Enterprise Engineering

## SEELab: Data-repositories for research and teaching

- Detailed **operational histories** (customers, servers), e.g.
  1. * Bank Anonymous: **1 year, 350K calls by 15 agents** - in 2000, which paved the way to:
  2. *U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**
  3. Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**
  4. Israeli Bank: **from January 2010, daily-deposit at a SEESafe**
  5. *Home (Rambam) Hospital: **4 years, 1000 beds**, ward-level flow
  6. Emergency Department (ED) patient flow:
     - 5 EDs in Israel: 1-2 years, **late David Sinreich**, ED arrivals & LOS
     - ED in Seoul: 2 months, **K. Song-Hee & W. Cha**, pilot
     - ED in Singapore: 2 years, **S. He & M. Sim**, pilot
  7. Service Engineering **internet site**: click-stream data (2 years)

# Technion SEE = Service Enterprise Engineering

## SEELab: Data-repositories for research and teaching

- Detailed **operational histories** (customers, servers), e.g.
  1. * Bank Anonymous : **1 year, 350K calls by 15 agents** - in 2000, which paved the way to:
  2. *U.S. Bank : **2.5 years, 220M calls, 40M by 1000 agents**
  3. Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**
  4. Israeli Bank: **from January 2010, daily-deposit at a SEESafe**
  5. *Home (Rambam) Hospital : **4 years, 1000 beds**, ward-level flow
  6. Emergency Department (ED) patient flow:
     - 5 EDs in Israel: 1-2 years, **late David Sinreich**, ED arrivals & LOS
     - ED in Seoul: 2 months, **K. Song-Hee & W. Cha**, pilot
     - ED in Singapore: 2 years, **S. He & M. Sim**, pilot
  7. Service Engineering **internet site**: click-stream data (2 years)

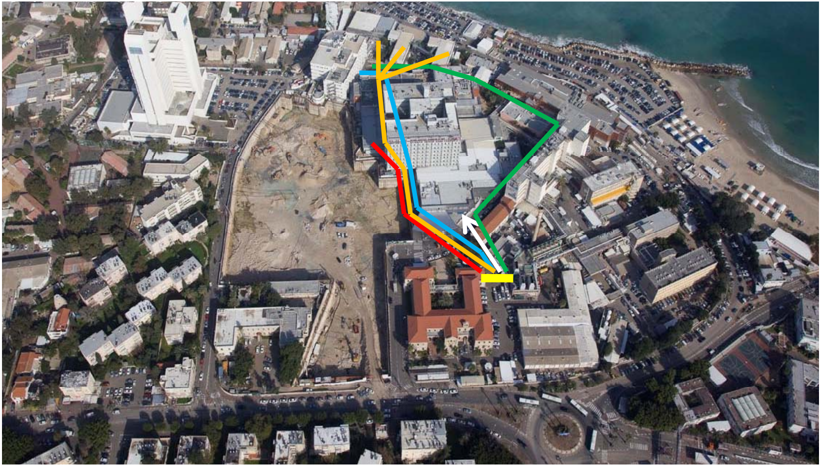
## Environment for graphical EDA in real-time:

- ***SEEStat** : primitives (arrivals, services, patience)
- **SEEGraph**: structure, animation (protocols $\Rightarrow$ simulation)

* **Open & Free for academic use**

# 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) !
- "Psychology" of Data Trust and Transfer (e.g. 2 years till transfer)

# Event-Logs in a Call Center (Bank Anonymous)

**A Data Sample (Excel worksheet)**

| vru+line | call_id | customer_id | priority | type | date | vru_entry | vru_exit | vru_time | q_start | q_exit | q_time | outcome | ser_start | ser_exit | ser_time | server |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AA0101 | 44749 | 27644400 | 2 | PS | 990901 | 11:45:33 | 11:45:39 | 6 | 11:45:39 | 11:46:58 | 79 | AGENT | 11:46:57 | 11:51:00 | 243 | DORIT |
| AA0101 | 44750 | 12887816 | 1 | PS | 990905 | 14:49:00 | 14:49:06 | 6 | 14:49:06 | 14:53:00 | 234 | AGENT | 14:52:59 | 14:54:29 | 90 | ROTH |
| AA0101 | 44967 | 58660291 | 2 | PS | 990905 | 14:58:42 | 14:58:48 | 6 | 14:58:48 | 15:02:31 | 223 | AGENT | 15:02:31 | 15:04:10 | 99 | ROTH |
| AA0101 | 44968 | 0 | 0 | NW | 990905 | 15:10:17 | 15:10:26 | 9 | 15:10:26 | 15:13:19 | 173 | HANG | 00:00:00 | 00:00:00 | 0 | NO_SERVER |
| AA0101 | 44969 | 63193346 | 2 | PS | 990905 | 15:22:07 | 15:22:13 | 6 | 15:22:13 | 15:23:21 | 68 | AGENT | 15:23:20 | 15:25:25 | 125 | STEREN |
| AA0101 | 44970 | 0 | 0 | NW | 990905 | 15:31:33 | 15:31:47 | 14 | 00:00:00 | 00:00:00 | 0 | AGENT | 15:31:45 | 15:34:16 | 151 | STEREN |
| AA0101 | 44971 | 41630443 | 2 | PS | 990905 | 15:37:29 | 15:37:34 | 5 | 15:37:34 | 15:38:20 | 46 | AGENT | 15:38:18 | 15:40:56 | 158 | TOVA |
| AA0101 | 44972 | 64185333 | 2 | PS | 990905 | 15:44:32 | 15:44:37 | 5 | 15:44:37 | 15:47:57 | 200 | AGENT | 15:47:56 | 15:49:02 | 66 | TOVA |
| AA0101 | 44973 | 3.06E+08 | 1 | PS | 990905 | 15:53:05 | 15:53:11 | 6 | 15:53:11 | 15:56:39 | 208 | AGENT | 15:56:38 | 15:56:47 | 9 | MORIAH |
| AA0101 | 44974 | 74780917 | 2 | NE | 990905 | 15:59:34 | 15:59:40 | 6 | 15:59:40 | 16:02:33 | 173 | AGENT | 16:02:33 | 16:26:04 | 1411 | ELI |
| AA0101 | 44975 | 55920755 | 2 | PS | 990905 | 16:07:46 | 16:07:51 | 5 | 16:07:51 | 16:08:01 | 10 | HANG | 00:00:00 | 00:00:00 | 0 | NO_SERVER |
| AA0101 | 44976 | 0 | 0 | NW | 990905 | 16:11:38 | 16:11:48 | 10 | 16:11:48 | 16:11:50 | 2 | HANG | 00:00:00 | 00:00:00 | 0 | NO_SERVER |
| AA0101 | 44977 | 33689787 | 2 | PS | 990905 | 16:14:27 | 16:14:33 | 6 | 16:14:33 | 16:14:54 | 21 | HANG | 00:00:00 | 00:00:00 | 0 | NO_SERVER |
| AA0101 | 44978 | 23817067 | 2 | PS | 990905 | 16:19:11 | 16:19:17 | 6 | 16:19:17 | 16:19:39 | 22 | AGENT | 16:19:38 | 16:21:57 | 139 | TOVA |
| AA0101 | 44764 | 0 | 0 | PS | 990901 | 15:03:26 | 15:03:36 | 10 | 00:00:00 | 00:00:00 | 0 | AGENT | 15:03:35 | 15:06:36 | 181 | ZOHARI |
| AA0101 | 44765 | 25219700 | 2 | PS | 990901 | 15:14:46 | 15:14:51 | 5 | 15:14:51 | 15:15:10 | 19 | AGENT | 15:15:09 | 15:17:00 | 111 | SHARON |
| AA0101 | 44766 | 0 | 0 | PS | 990901 | 15:25:48 | 15:26:00 | 12 | 00:00:00 | 00:00:00 | 0 | AGENT | 15:25:59 | 15:28:15 | 136 | ANAT |
| AA0101 | 44767 | 58859752 | 2 | PS | 990901 | 15:34:57 | 15:35:03 | 6 | 15:35:03 | 15:35:14 | 11 | AGENT | 15:35:13 | 15:35:15 | 2 | MORIAH |
| AA0101 | 44768 | 0 | 0 | PS | 990901 | 15:46:30 | 15:46:39 | 9 | 00:00:00 | 00:00:00 | 0 | AGENT | 15:46:38 | 15:51:51 | 313 | ANAT |
| AA0101 | 44769 | 78191137 | 2 | PS | 990901 | 15:56:03 | 15:56:09 | 6 | 15:56:09 | 15:56:28 | 19 | AGENT | 15:56:28 | 15:59:02 | 154 | MORIAH |
| AA0101 | 44770 | 0 | 0 | PS | 990901 | 16:14:31 | 16:14:46 | 15 | 00:00:00 | 00:00:00 | 0 | AGENT | 16:14:44 | 16:16:02 | 78 | BENSION |
| AA0101 | 44771 | 0 | 0 | PS | 990901 | 16:38:59 | 16:39:12 | 13 | 00:00:00 | 00:00:00 | 0 | AGENT | 16:39:11 | 16:43:35 | 264 | VICKY |
| AA0101 | 44772 | 0 | 0 | PS | 990901 | 16:51:40 | 16:51:50 | 10 | 00:00:00 | 00:00:00 | 0 | AGENT | 16:51:49 | 16:53:52 | 123 | ANAT |
| AA0101 | 44773 | 0 | 0 | PS | 990901 | 17:02:19 | 17:02:28 | 9 | 00:00:00 | 00:00:00 | 0 | AGENT | 17:02:28 | 17:07:42 | 314 | VICKY |
| AA0101 | 44774 | 32387482 | 1 | PS | 990901 | 17:18:18 | 17:18:24 | 6 | 17:18:24 | 17:19:01 | 37 | AGENT | 17:19:00 | 17:19:35 | 35 | VICKY |
| AA0101 | 44775 | 0 | 0 | PS | 990901 | 17:38:53 | 17:39:05 | 12 | 00:00:00 | 00:00:00 | 0 | AGENT | 17:39:04 | 17:40:43 | 99 | TOVA |

- Unsynchronized transition times, consistently

# Empirical Adventures at the SEELab

SEELab History suggests possible guidelines for ServNet Mining:

1. **Primitives**: arrivals, services, (im)patience
2. **Structure**: static process-maps
3. **Protocols**: Load Balancing, Dynamic Priority, Information

# Empirical Adventures at the SEELab

SEELab History suggests possible guidelines for ServNet Mining:

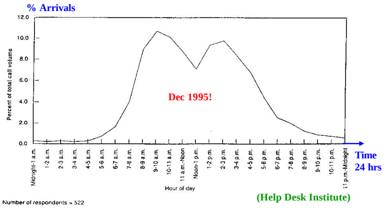
1. **Primitives**: arrivals, services, (im)patience
2. **Structure**: static process-maps
3. **Protocols**: Load Balancing, Dynamic Priority, Information

EDA $\Rightarrow$ open questions, new directions, uncharted territories

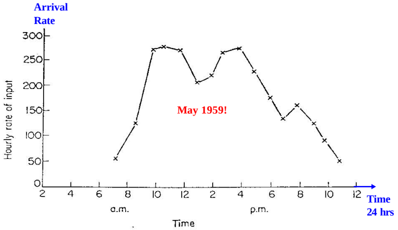
# Primitives: Arrival (Rates) to Service

## Why 2 Daily Peaks?

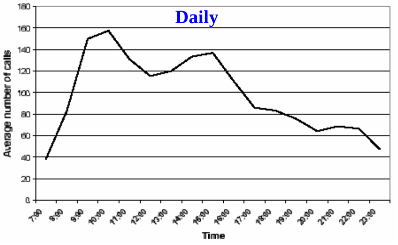
**CC**: Dec. **1995**, (USA, 700 Helpdesks)



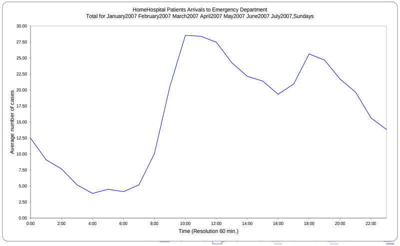

**CC**: May **1959** (England)




**CC**: Nov. 1999 (Israel)

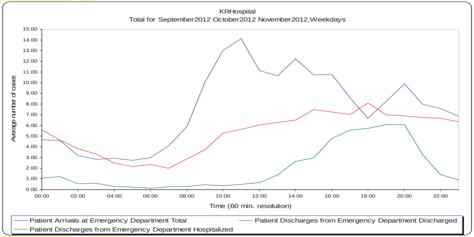



**ED**: Jan.–July 2007 (Israel)

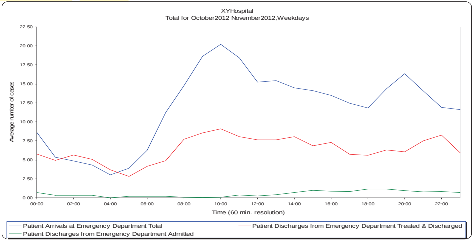


# Arrival (Discharge) Rates in Korea and Singapore

KRHospital, all ED patients



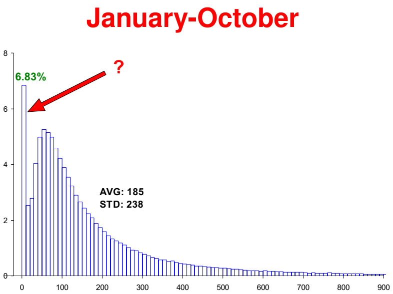

XYHospital, walking patients

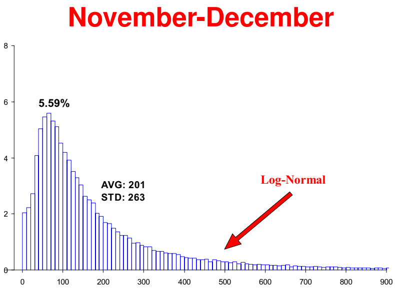


# Primitives: Services (Durations)

Histogram of Service-Duration in an Israeli Call Center, 1999

**Why short services? Why LogNormal?**



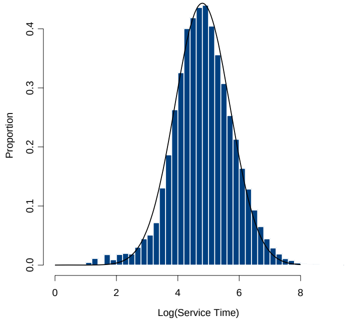




- January-October: **6.8% Short-Services** ($\leq$ 10 seconds) ?
- November-December: **LogNormal** durations (common) ?

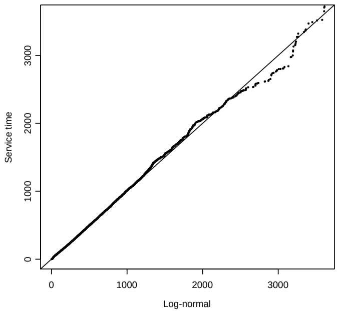
# Durations: Phone Calls (2 Surprises)

## Israeli Call Center, Nov–Dec, 1999

### Log(Service Times)



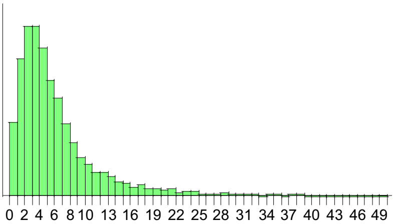

### LogNormal QQPlot




- **Practically Important**: (mean, std)(log) characterization
- **Theoretically Intriguing**: Why LogNormal ? Naturally multiplicative but, in fact, also **Infinitely-Divisible** (Generalized Gamma-Convolutions)

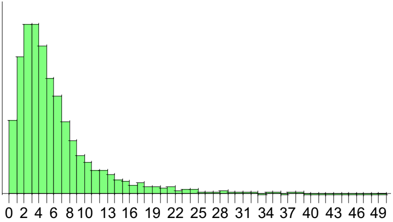
# Primitives, . . . : Length-of-Stay in a Hospital Ward

## Israeli Hospital, IW LOS, in Days:

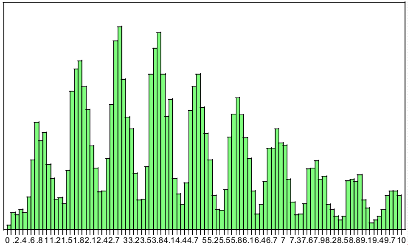


# Primitives, . . .: Length-of-Stay in a Hospital Ward

## Israeli Hospital, IW LOS, in Days:

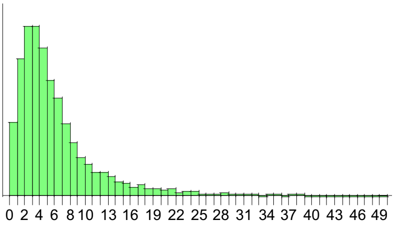



## In Hours

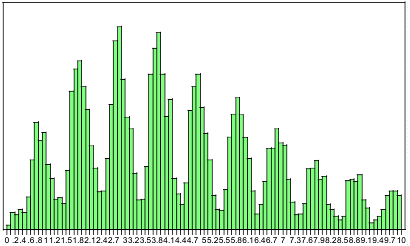


# Primitives, . . .: Length-of-Stay in a Hospital Ward

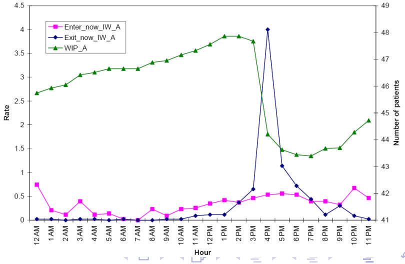
## Israeli Hospital, IW LOS, in Days:




## In Hours




**Explanation**: Patients released around **3pm** (2-3 in Singapore, 2-4 in UNC Hospital)



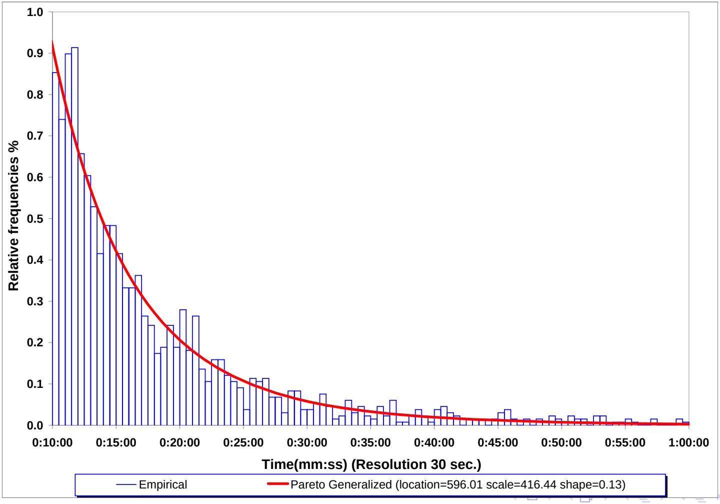

## Why Bother ?

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

# Primitives: (Im)Patience

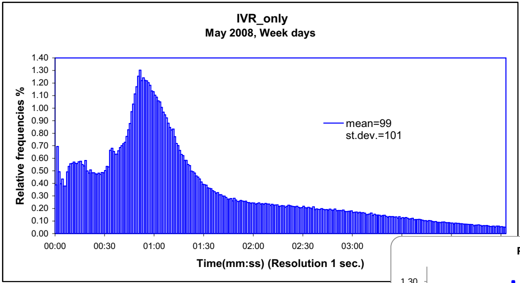
## Israeli Bank: Uncensored 13,000 Customers, 24/11/2008

## Patience $\geq 10min$: Why Pareto Tail?

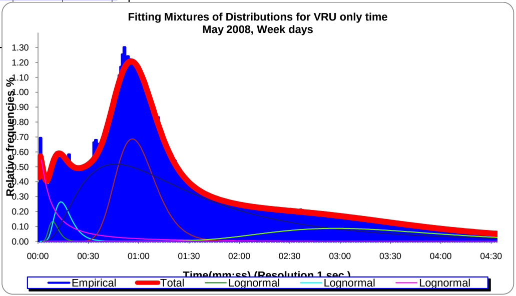


# Durations: Answering Machine (Mixture Fitting)

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



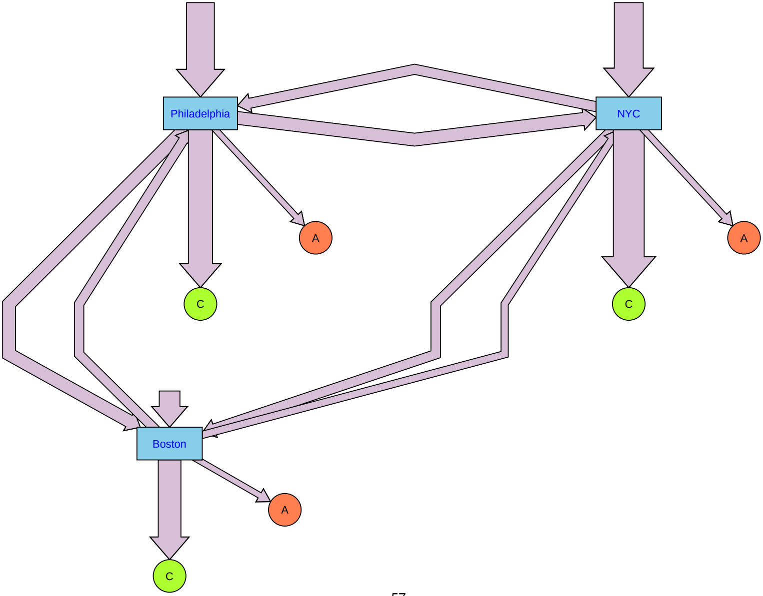

**Mixture: 7 LogNormals**




**w/ N. Carmeli, H. Kaspi**

Network (Tree) Tomography:
Reconstruct topology
from root and leaves; Optimize

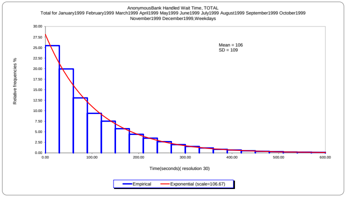
# Protocol Mining: Snapshots of Connectivity



# Protocols: Waiting Time in a Call Center

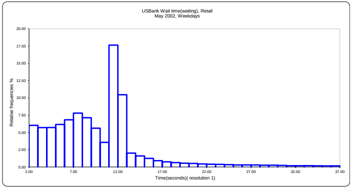
## Exponential in Heavy-Traffic (min.)

Small Israeli Bank



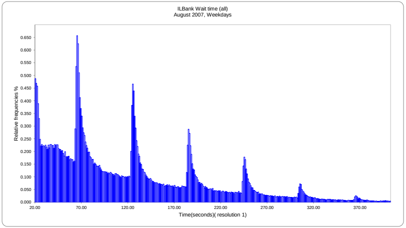

## Routing via Thresholds (sec.)

Large U.S. Bank



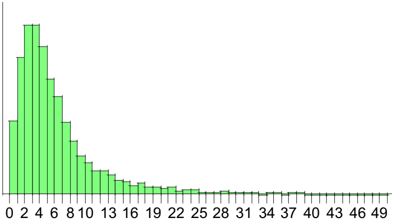

## Scheduling Priorities (sec.)    [compare Hospital LOS (hours)]

Medium Israeli Bank



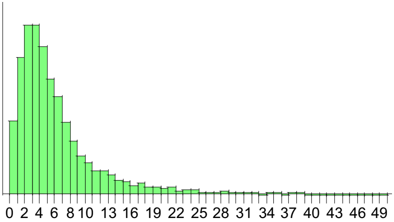
# Protocols: LOS in Hospitals - Beyond LogNormal

## Israeli Hospital, in Days: LN

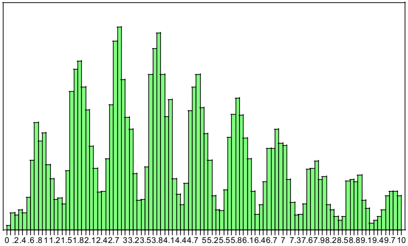


# Protocols: LOS in Hospitals - Beyond LogNormal

## Israeli Hospital, in Days: LN

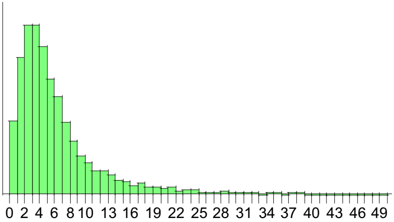



## In Hours: 2 Time Scales, Mixture

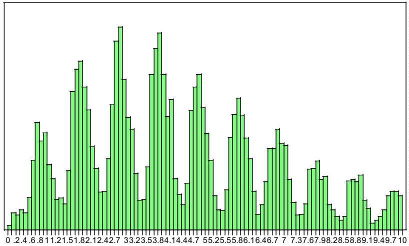


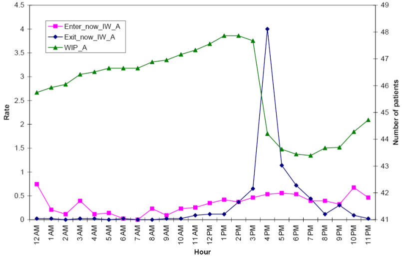
# Protocols: LOS in Hospitals - Beyond LogNormal

**Israeli Hospital, in Days: LN**




**In Hours: 2 Time Scales, Mixture**






**Explanation**: Patients released around **3pm** (2-3 in Singapore, 2-4 in UNC Hospital)

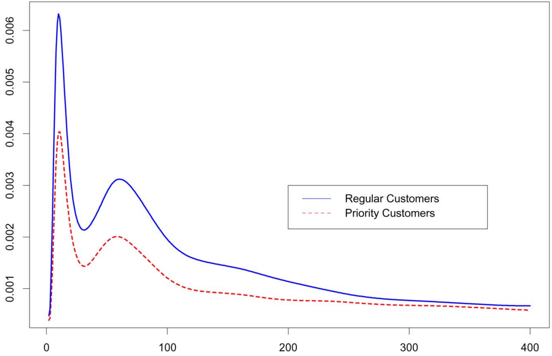
**Why Bother ?**

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

# Protocols: (Im)Patience while Waiting (Psychology)

Palm: (1943–53): **Irritation** $\propto$ **Hazard Rate**

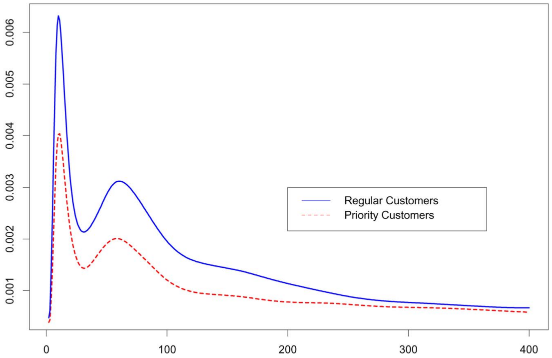
**Regular** over **VIP** Customers: VIP more patient here (Israeli Bank)



# Protocols: (Im)Patience while Waiting (Psychology)

Palm: (1943–53): **Irritation** $\propto$ **Hazard Rate**

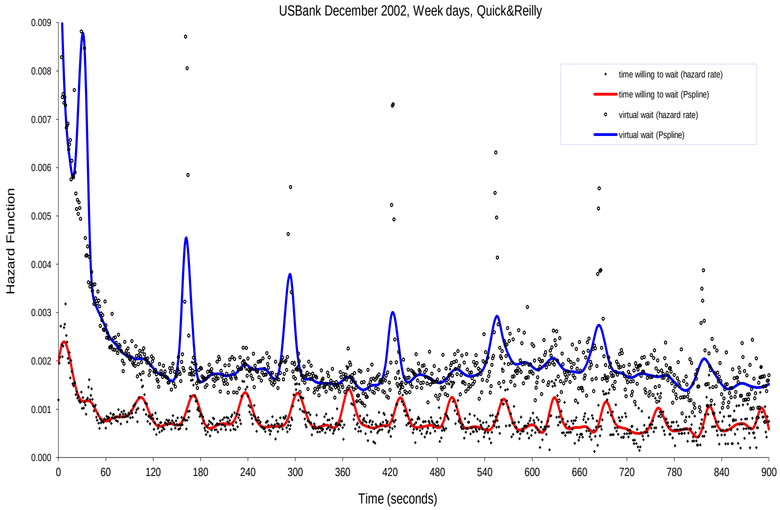
**Regular** over **VIP** Customers: VIP more patient here (Israeli Bank)




- Why **Peaks** of **abandonment**? **Announcement** epochs
  - Control abandonment w/ info: encourage, discourage
  - Technical Challenges, **w/ J. Huang, J. Zhang, H. Zhang**
- Statistical challenges: **Un-Censoring**, Smoothing

# Protocols + Psychology

## Patient Customers, Announcements, Priority Upgrades



# On Data-Based Research

w/ **V. Trofimov, E. Nadjharov, I. Gavako = Technion SEELab**

- **ServNets = P, C,. . . ; Q, Sim, F, D**
- **SimNets** of Service Systems = **Virtual Realities**, where "complex" models meet "simple" models, towards enhancement and credibility

# On Data-Based Research

w/ V. Trofimov, E. Nadjharov, I. Gavako = Technion SEELab

- **ServNets = P, C,. . . ; Q, Sim, F, D**
- **SimNets** of Service Systems = **Virtual Realities**, where “complex" models meet “simple" models, towards enhancement and credibility
- Beyond the prevalent: “single researcher (with a PhD student) obtaining small data for a single research project": unprofessional, no learning across generations, no sharing among researchers, irreproducible research
- Data-based Research: Tradition in Physics, Chemistry, Biology; Psychology (now also in Transportation (Science) and (Behavioral) Economics)
- **Why not in Service/Queueing** Science / Engineering / Management ?

## On Data-Based Research

w/ V. Trofimov, E. Nadjharov, I. Gavako = Technion SEELab

- **ServNets = P, C,. . . ; Q, Sim, F, D**
- **SimNets** of Service Systems = **Virtual Realities**, where "complex" models meet "simple" models, towards enhancement and credibility
- Beyond the prevalent: "single researcher (with a PhD student) obtaining small data for a single research project": unprofessional, no learning across generations, no sharing among researchers, irreproducible research
- Data-based Research: Tradition in Physics, Chemistry, Biology; Psychology (now also in Transportation (Science) and (Behavioral) Economics)
- **Why not in Service/Queueing** Science / Engineering / Management ?
- **Glad to see this happening in DSC/e**

## Data-Based Creation of ServNets: some Technicalities

- **ServNets** = QNets, SimNets, FNets, DNets
- **Graph Layout**: Adapted from but significantly extends Graphviz (AT&T, 90's); eg. *edge-width*, which must be restricted to *poly-lines*, since there are "no parallel Bezier (Cubic) curves ($B_n(p) = E_p F[B(n, p)], 0 \leq p \leq 1$)
- **Algorithm**: Dot Layout (but with cycles), based on Sugiyama, Tagawa, Toda ('81): "Visual Understanding of **Hierarchical** System Structures"

# Data-Based Creation of ServNets: some Technicalities

- **ServNets** = QNets, SimNets, FNets, DNets
- **Graph Layout**: Adapted from but significantly extends Graphviz (AT&T, 90's); eg. *edge-width*, which must be restricted to *poly-lines*, since there are "no parallel Bezier (Cubic) curves ($B_n(p) = E_p F[B(n,p)], 0 \leq p \leq 1$)
- **Algorithm**: Dot Layout (but with cycles), based on Sugiyama, Tagawa, Toda ('81): "Visual Understanding of **Hierarchical** System Structures"
- Draws data directly from **SEELab** data-bases:
  - Relational DBs (Large! eg. USBank Full Binary = 37GB, Summary Tables = 7GB)
  - Structure: Sequence of events/states, which (due to size) partitioned (yet integrated) into days (eg. call centers) or months (eg. hospitals)
  - Differs from industry DBs (in call centers, hospitals, websites)