Empirical Adventures in Call Centers Emergency Departments, ...

Avishai Mandelbaum
Technion IE&M

with Graduate Students, Research Partners
Technion SEE Center, IBM+Rambam+Technion OCR

WITOR, September 2009

Contents of Talk

Contents of Talk

DataMOCCA

Data **MO**dels for **C**all **C**enter **A**nalysis

Project Collaborators:

Technion:	Paul Feigin, Avi Mandelbaum Technion SEElab: Valery Trofimov, Ella Nadjharov, Igor Gavako, Katya Kutsy, Polyna Khudyakov, Shimrit Maman, Pablo Liberman Students (PhD, MSc, BSc), RAs
Wharton:	Larry Brown, Noah Gans, Haipeng Shen (N. Carolina), Students, Wharton Financial Institutions Center
Companies:	U.S. Bank, Israeli Telecom, 2 Israeli Banks, Israeli Hospitals, ...

The SEE Center - Project DataMOCCA

Goal: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing, displaying and interacting with transaction-based data.

The SEE Center - Project DataMOCCA

Goal: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing, displaying and interacting with transaction-based data.

Enable the Study of:

- Customers (Callers, Patients)	Waiting, Abandonment, Returns
- Servers (Agents, Nurses)	Service Duration, Activity Profile
- Managers (System)	Loads, Queue Lengths, Trends

Call-Center: Hidden Complex Service Network

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

A "Good" Hospital in Beijing

DataMOCCA History: The Data Challenge

- Queueing Research lead to Service Operations (Early 90s)
- Services started with Call Centers which, in turn, created data-needs
- Queueing Theory had to expand to Queueing Science: Fascinating
- WFM was Erlang-C based, but customers abandon! (Im)Patience?
- (Im)Patience censored hence Call-by-Call data required: 4-5 years saga
- Finally Data: a small call center in a small IL bank (15 agents, 4 service types, 350K calls per year)
- Technion Stat. Lab, guided by Queueing Science: Descriptive Analysis
- Building blocks (Arrivals, Services, (Im)Patiece): even more Fascinating

DataMOCCA History: Research & Teaching

- Large well-run call centers beyond conventional Queueing Theory:
  - Both **Q**uality and **E**fficiency **D**riven (vs. tradeoff)
  - Multi-Disciplinary view: OR/OM, HRM (Psychology), Marketing, MIS
- Research: Asymptotic analysis of the Palm/Erlang-A model, in the Halfin-Whitt regime = QED Regime (Many-Server limits); Fork-Join Networks; Queueing Laws; Data-based Simulations.
- Teaching: Service Management + Industrial Engineering = Service Engineering / Science
- INSEAD + Wharton Mini-course (Zeynep Aksin, Morris Cohen), then Wharton Seminar (Statistics, Larry Brown) + Call Center Forum (Noah Gans) = cooperation with a large(r) banking call center (1000 agents), ...

DataMOCCA: System Components

1. Clean Databases: Operational histories of **individual** customers and servers (mostly with IDs).
   - In Call Centers: from IVR to Exit;
   - In Hospitals: from ED to Exit (or just ED).

2. **SEEStat: Online** GUI (friendly, flexible, powerful)
   - Queueing-Science perspective;
   - Operational data (vs. financial, contents or clinical);
   - Flexible customization (e.g. seconds to months);

3. Tools:
   - Online statistics (survival analysis, mixtures, smoothing);
   - Dynamic Graphs (flow-charts, work-flows)
   - Simulators (CC, ED; data-driven).

Current Databases

1. U.S. Bank (**PUBLIC**): 220M calls, 40M agent-calls, 1000 agents, 2.5 years, 7-40GB.
2. Israeli Banks:
   - Small (**PUBLIC**): 350K calls, 15 agents, 1 year. Started it all in 1999 (JASA), now "romancing" again (Medium, with 300 agents);
   - Large (ongoing): 500 agents, 1.5 years, 3-8GB.
3. Israeli Telecom (ongoing): 800 agents, 3.5 years; 5-55GB.
4. Israeli Hospitals:
   - Six ED's (to be made **PUBLIC**);
   - Large (ongoing): 1000 beds, 45 medical units, 75,000 patients hospitalized yearly, 4 years, 7GB.
5. Website (pilot).

DataMOCCA: Future

- Operational (ACD) data with Business (CRM) data, Contents/Medical
- Contact Centers: IVR, Chats, Emails; Websites
- Daily update (as opposed to montly DVDs)
- Web-access (Research; Applications, e.g. CC/ED Simulation; Teaching)
- Nurture Research, for example
  - Skills-Based Routing: Control, staffing, design, online; HRM
  - The Human Factor: Service-anatomy, agents learning, incentives
- Hospitals (OCR: with IBM, Haifa hospital): Operational, Human-Factors, Medical & Financial data; RFIDs for flow-tracking

DataMOCCA Interface: SEEStat

- Daily / monthly / yearly reports & flow-charts for a complete operational view.

- Graphs and tables, in customized resolutions (month, days, hours, minutes, seconds) for a variety of (pre-designed) operational measures (arrival rates, abandonment counts, service- and wait-time distribution, utilization profiles,).

- Graphs and tables for new user-defined measures.

- Direct access to the raw (cleaned) data: export, import.

- Online Statistics: Survival Analysis, Mixtures, Smoothing, Graphics.

Data-Based Research: Must (?) & Fun

- Contrast with "EmpOM": Industry / Company / Survey Data (Social Sciences)

- Converge to: Measure, Model, Validate, Experiment, Refine (Physics, Biology, ...) - The Scientific Paradigm

- Prerequisites: OR/OM, (Marketing) for Design; Computer Science, Information Systems, Statistics for Implementation

- Outcomes: Relevance, Credibility, Interest; Pilot (eg. Healthcare, Web). Moreover,

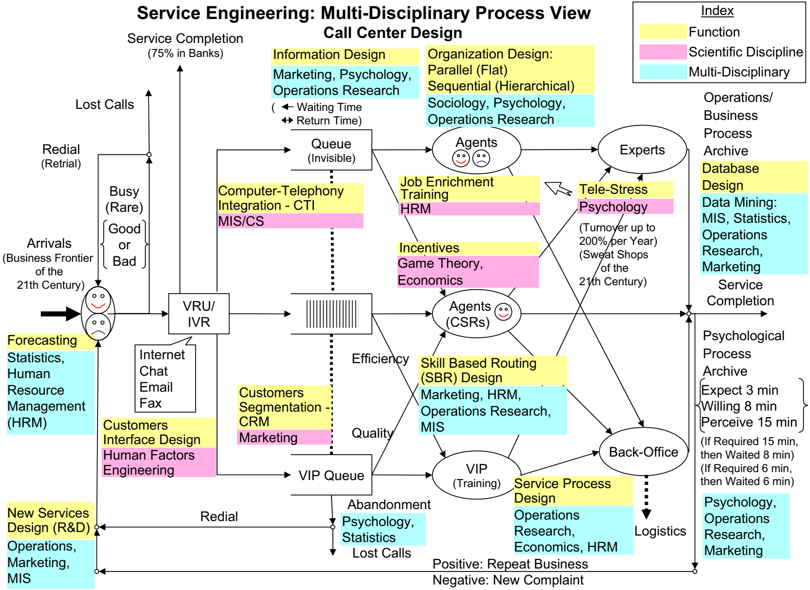
  Teaching: Class, Homework (Experimental Data Analysis); Cases.

  Research: Test (Queueing) Theory / Laws, Stimulate New Models / Theory.

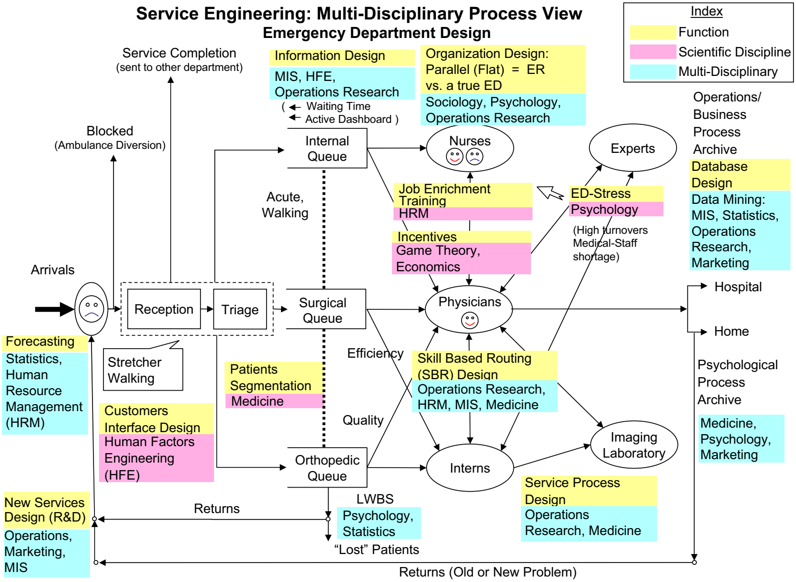
  Practice: OM Tools (Scenario Analysis), Mktg (Trends, Benchmarking).

Service Engineering: Multi-Disciplinary Process View

Call Center Design



Service Engineering: Multi-Disciplinary Process View
Emergency Department Design

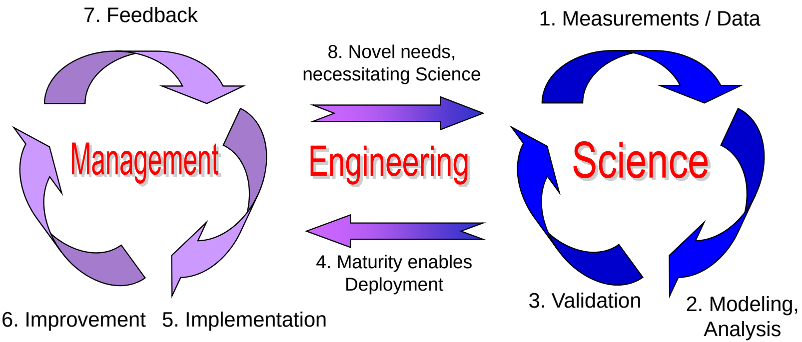


Expanding the Scientific Paradigm (OCR)

- Physics, Biology, . . . : Measure, Model, Experiment, Validate, Refine.
- Human-complexity triggered the above in Transportation, Economics.

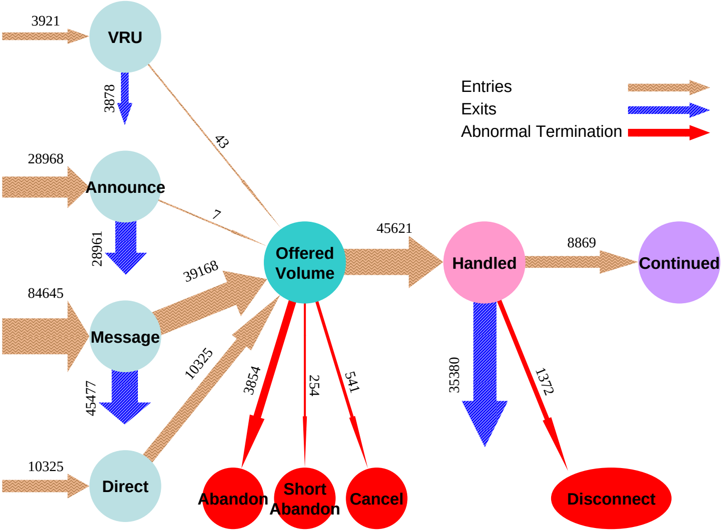
Expanding the Scientific Paradigm (OCR)

- Physics, Biology, . . . : Measure, Model, Experiment, Validate, Refine.
- Human-complexity triggered the above in Transportation, Economics.
- Expand to:



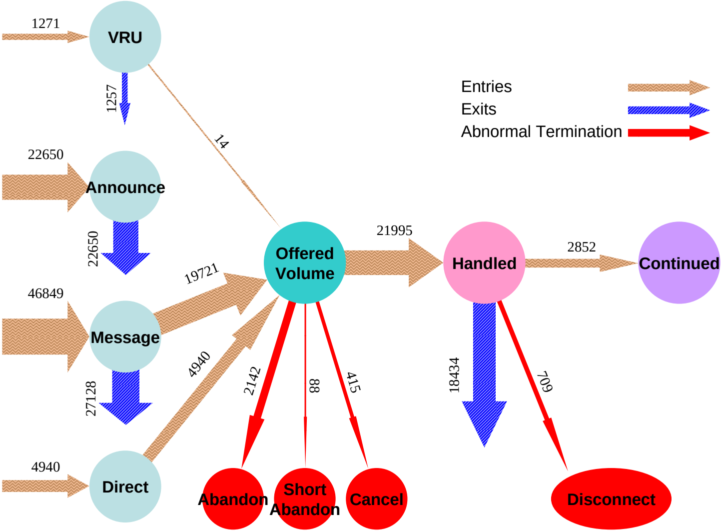
Flow Chart: Daily Report (SEEStat)

Call Center: April 13, 2004 Regular Day



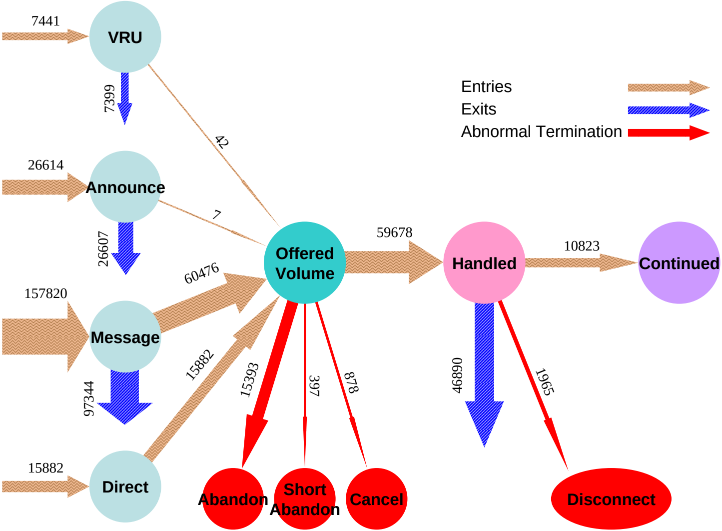
Flow Chart: Daily Report

Call Center: April 27, 2004 - Holiday

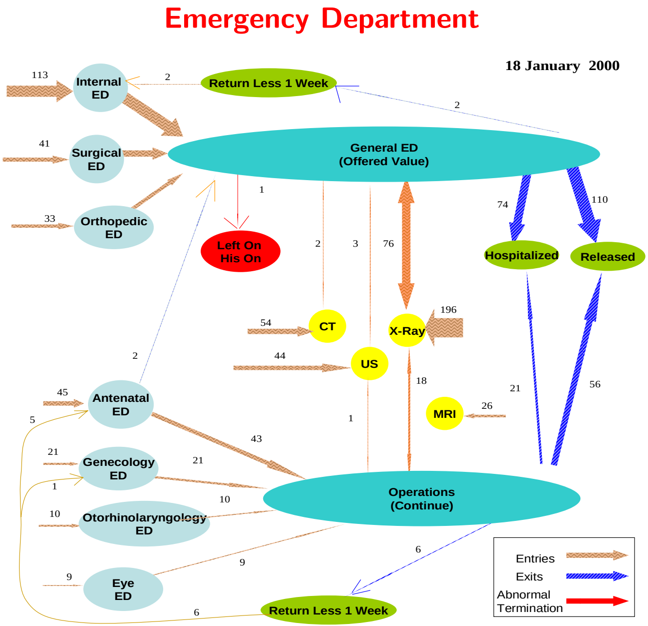


Flow Chart: Daily Report

Call Center: April 20, 2004 - Heavily Loaded Day



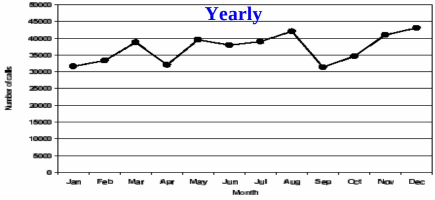
Flow Chart: Daily Report



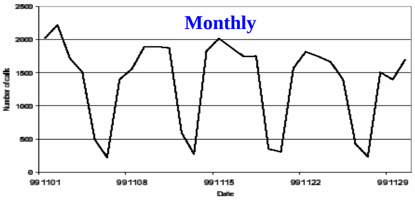
Contents of Talk

Arrivals to a Call Center (Israel, 1999): Time Scales

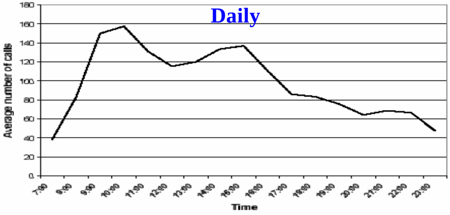
Strategic



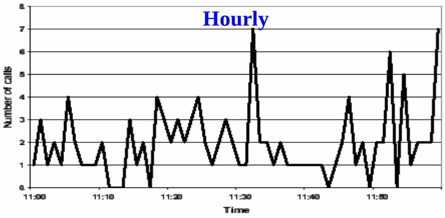
Tactical



Operational (Predictable Var.)



Regulatory (Stochastic)



Arrivals to a Call Center (U.S., 1976): Queueing Science

(E. S. Buffa, M. J. Cosgrove, and B. J. Luce, "An Integrated Work Shift Scheduling System")

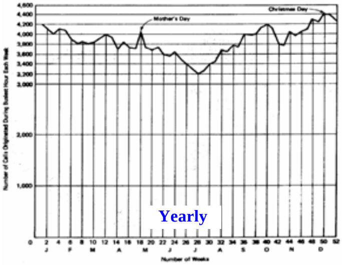


Figure 1 Typical distribution of calls during the busiest hour for each week during a year.

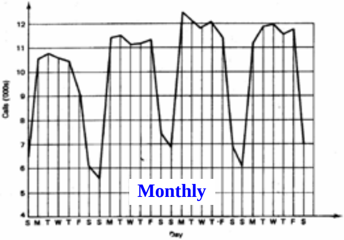


Figure 2 Daily call load for Long Beach, January 1972.

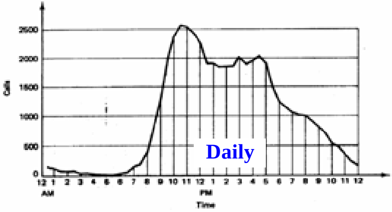


Figure 3 Typical half-hourly call distribution (Bundy D A).

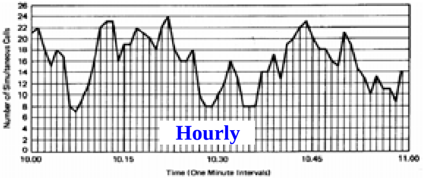
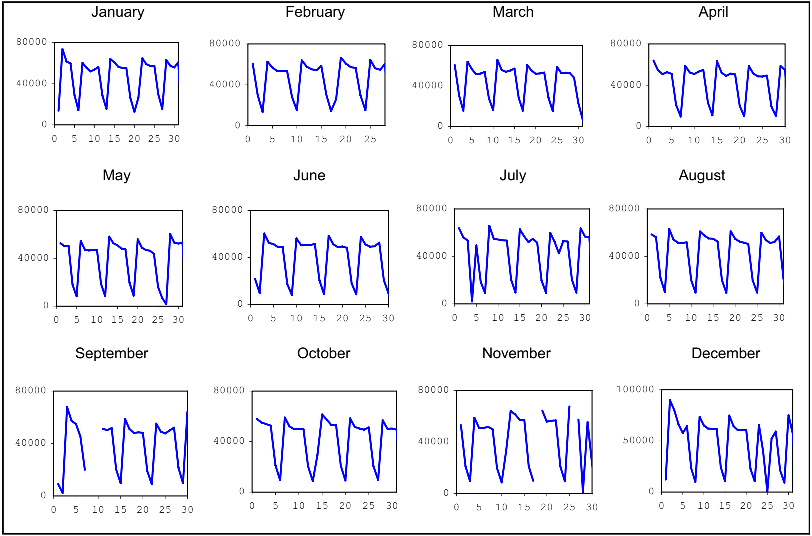


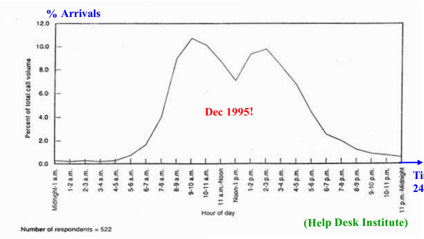
Figure 4 Typical intrahour distribution of calls, 10:00–11:00 A.M.

U.S. Bank: Daily Arrival-Rates, over a Month, 2002

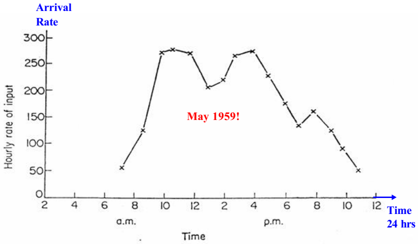


Intraday Arrival-Rates (per hour) to Call Centers

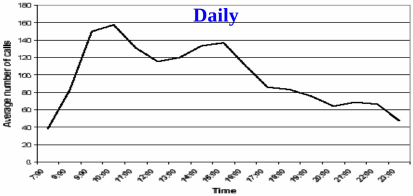
December 1995 (700 U.S. Helpdesks)



May 1959 (England)

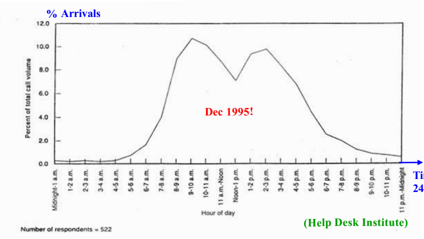


November 1999 (Israel)

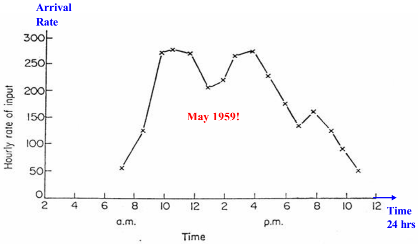


Intraday Arrival-Rates (per hour) to Call Centers

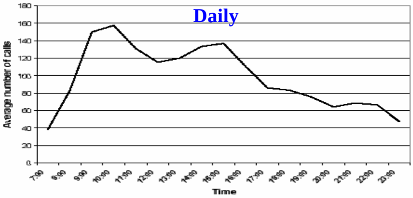
December 1995 (700 U.S. Helpdesks)



May 1959 (England)



November 1999 (Israel)



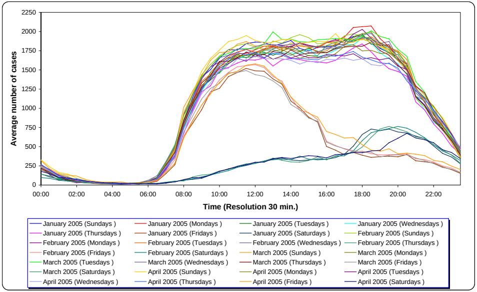
Observation:

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

Intraday Arrival Rates: Does a Day have a Shape ?

Arrival Patterns, Israeli Telecom

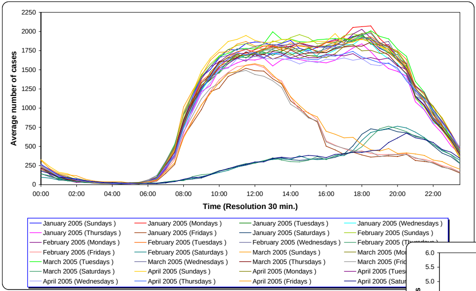
Arrivals, Avg. Weekdays/1-4/2005



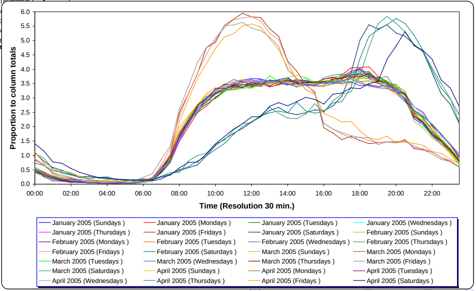
Intraday Arrival Rates: Does a Day have a Shape ?

Arrival Patterns, Israeli Telecom

Arrivals, Avg. Weekdays/1-4/2005



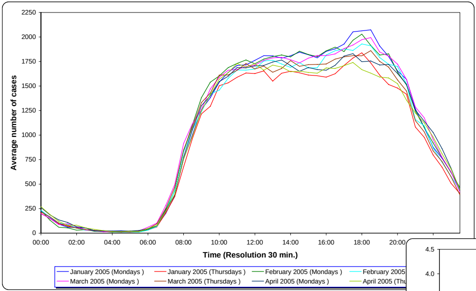
Percent to Total, 30 min.



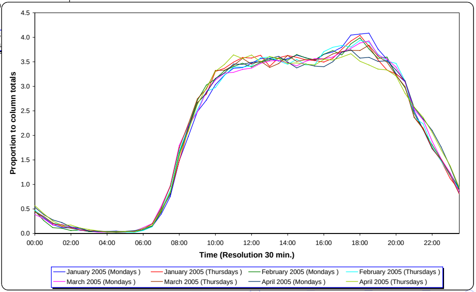
Shape Stability of Intraday Arrival Rates

Mondays (Busiest) and Thursdays (Lightest), 2005

Arrivals



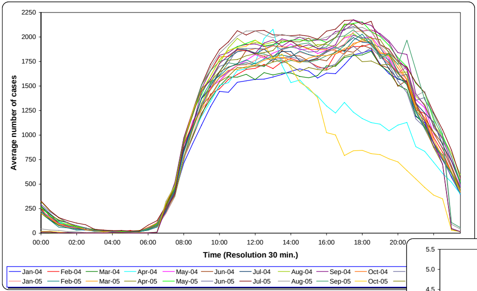
Percent to Total



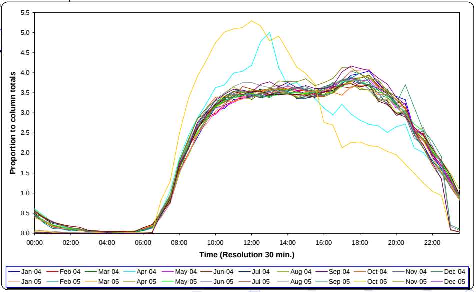
Shape Stability of Intraday Arrival Rates

Mondays, 2004-5 (Averages)

Arrivals



Percent to Total



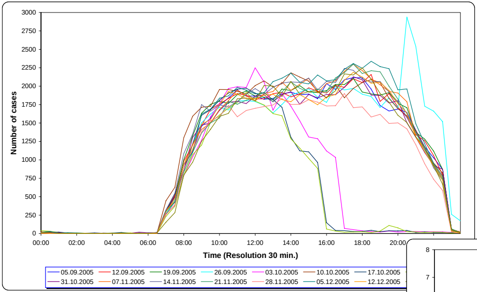
Unusual patterns:

October and April with many of the **Jewish holidays**.

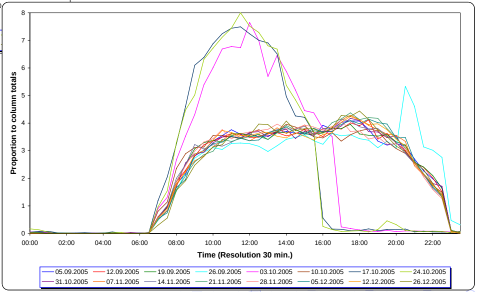
Shape Stability of Intraday Arrival Rates

Mondays, 2005 (Individual Days, Oct-Dec)

Arrivals



Percent to Total



Unusual patterns:

Jewish holidays in October.

Exogenous Arrivals to Service: How to Model?

- Axiomatically, “completely random arrivals” are **Poisson**.
- Arrivals over the day are not time-homogeneous.
- Hence, arrivals over the day are non-homogeneous Poisson.
- Arrivals over small intervals (15, 30, 60 min) are close to time-homogeneous Poisson.

Practically:
Test (L. Brown), then model, as a **Poisson process with piecewise-constant arrival rates**.

A (Common) Model for Call Arrivals

Whitt (99'), Brown et. al. (05'), Gans et. al. (09'), and others:

Doubly-stochastic (Cox, Mixed) Poisson with instantaneous rate

$$\boldsymbol{\Lambda(t) = \lambda(t) \cdot X} \ ,$$

where $\int_0^T \lambda(t)dt = 1$.

- $\boldsymbol{\lambda(t)}$ = "Shape" of weekday [**Predictable variability**]
- $\mathbf{X}$ = Total # arrivals [**Unpredictable variability**]

w/ Maman & Zeltyn (09'):
Above assumes **"too-much" stochastic variability!**

Unpredictable Variability: The Multi-Class Case

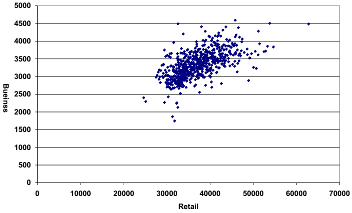
- Research w/ I. Gurvich & P. Liberman, ongoing.

Unpredictable variability: $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_I)$

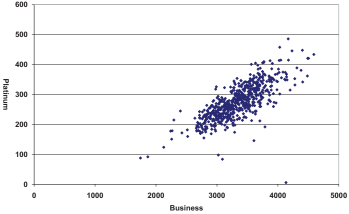
Pairs: $(\mathbf{X}_{Retail}, \mathbf{X}_{Business})$ and $(\mathbf{X}_{Business}, \mathbf{X}_{Platinum})$

US Bank: Correlations, 600 weekdays

Business vs. Retail

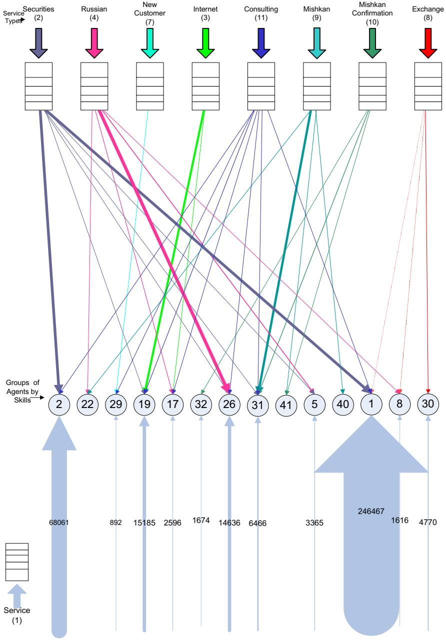


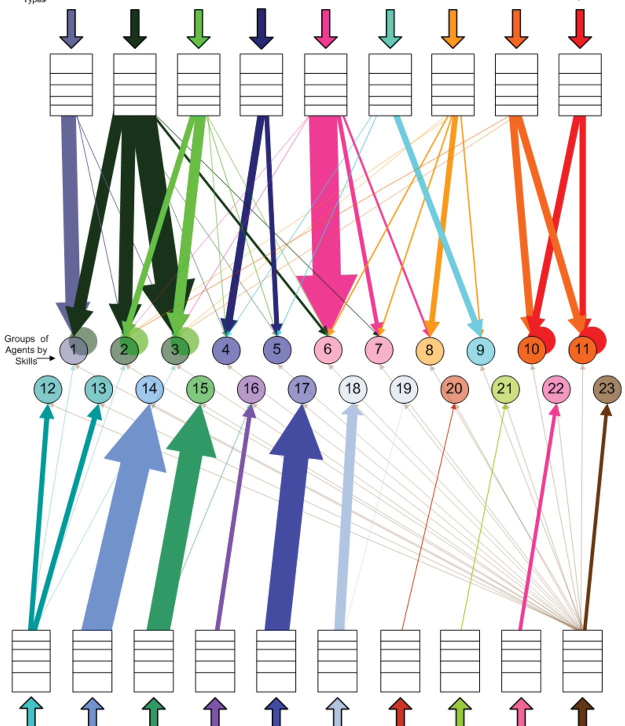
Business vs. Platinum

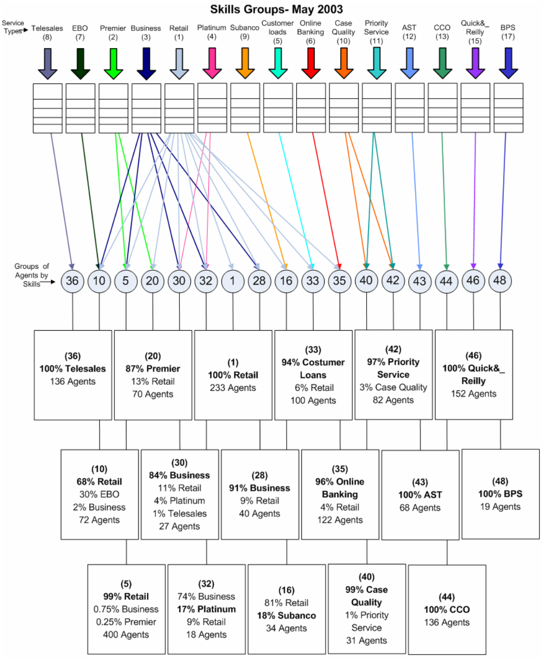


- Positive correlation (vs. independent in existing research)
- Research: Empirical, then Impact on design and control ?

Flow chart- May 2007




Groups of Agents by Skills
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


Skills Groups- May 2003
Service Types
Telesales (8)
EBO (7)
Premier (2)
Business (3)
Retail (1)
Platinum (4)
Subanco (9)
Customer loads (5)
Online Banking (6)
Case Quality (10)
Priority Service (11)
AST (12)
CCO (13)
Quick&_ Reilly (15)
BPS (17)
Groups of Agents by Skills
36 10 5 20 30 32 1 28 16 33 35 40 42 43 44 46 48
(36) 100% Telesales 136 Agents
(20) 87% Premier 13% Retail 70 Agents
(1) 100% Retail 233 Agents
(33) 94% Costumer Loans 6% Retail 100 Agents
(42) 97% Priority Service 3% Case Quality 82 Agents
(46) 100% Quick&_ Reilly 152 Agents
(10) 68% Retail 30% EBO 2% Business 72 Agents
(30) 84% Business 11% Retail 4% Platinum 1% Telesales 27 Agents
(28) 91% Business 9% Retail 40 Agents
(35) 96% Online Banking 4% Retail 122 Agents
(43) 100% AST 68 Agents
(48) 100% BPS 19 Agents
(5) 99% Retail 0.75% Business 0.25% Premier 400 Agents
(32) 74% Business 17% Platinum 9% Retail 18 Agents
(16) 81% Retail 18% Subanco 34 Agents
(40) 99% Case Quality 1% Priority Service 31 Agents
(44) 100% CCO 136 Agents

System Design: Simplification via State-Space Collapse

- Theory: R. Atar; PhD's: G. Shaikhet, T. Tezcan, I. Gurvich.

Service Rate: Class or Pool Dependent?

Agents Group\Service Class	Private Prepaid	Private	Private VIP
Private Prepaid	163.1	236.1	
Private - Private VIP (1)		243.5	195.1
Private - Private VIP (2)		244	201.4

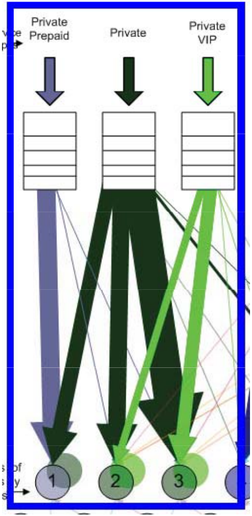
$\Rightarrow$ **Class-dependent** service rate

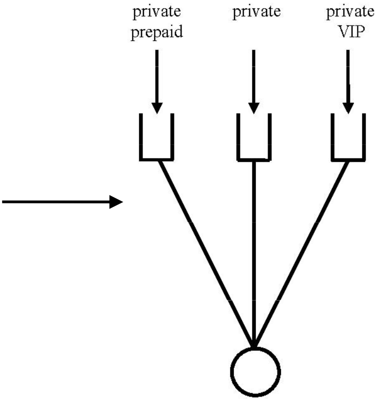
Agents Group\Service Class	Business	Business VIP	Business Preservation
Business (1)	276.9	261.5	
Business (2)	336.7	334.5	
Business VIP	315.9	280.5	
Business Preservation		386.2	634.1

$\Rightarrow$ **Pool-dependent** service rate

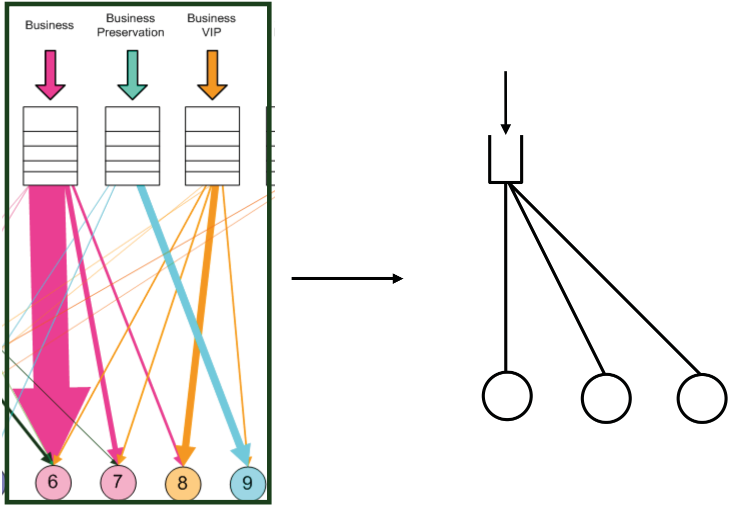
Many-Server Approximations: State-Space Collapse

Class-Dependent $\approx$ V-Model



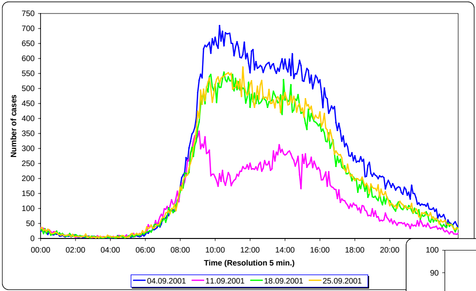


Pool-Dependent $\approx$ Λ-Model

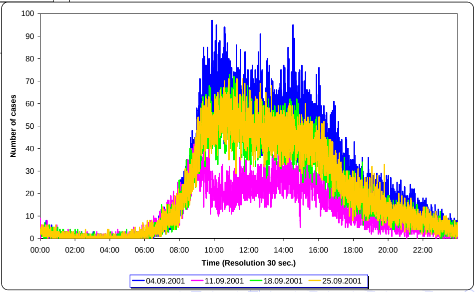


US Bank: Arrival-Rates on Tuesdays in a September

5 Minutes Resolution



30 Seconds Resolution

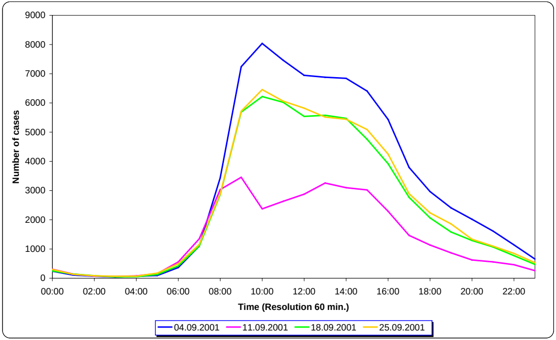


Stochastic Variability ←

Arrivals to Service: Predictable vs. Random

US Bank: Arrival-Rates on Tuesdays in a September

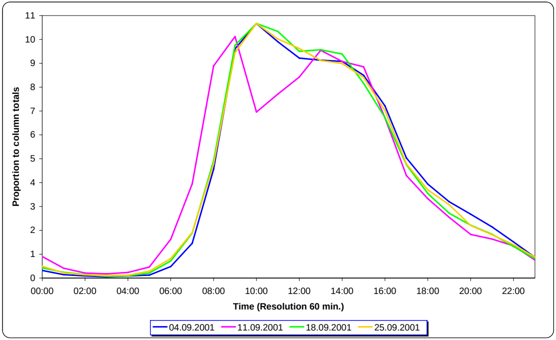
One Hour Resolution



- **Tuesday**, September 4th: **Heavy**, following Labor Day
- **Tuesdays**, September 18 & 25: **Normal**
- **Tuesday, September 11th, 2001**

US Bank: Arrival-Rates on Tuesdays in a September

Percent to Total

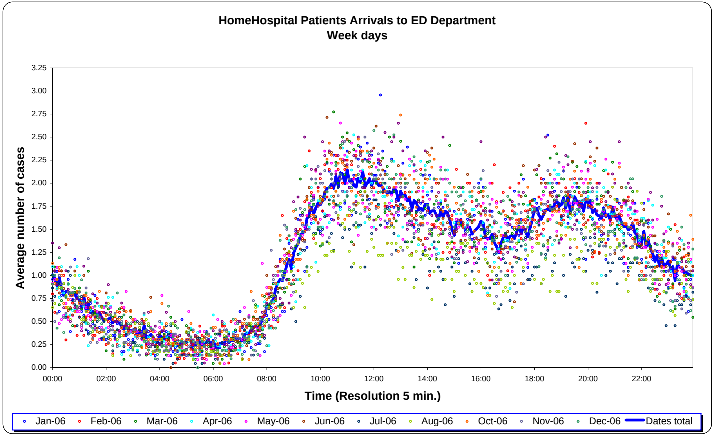


September 11th:

- Beginning, until 7:30-8:00: perfect fit of shape, left-shifted.
- After 13:00 - perfect fit.

Arrivals to an Emergency Department (ED)

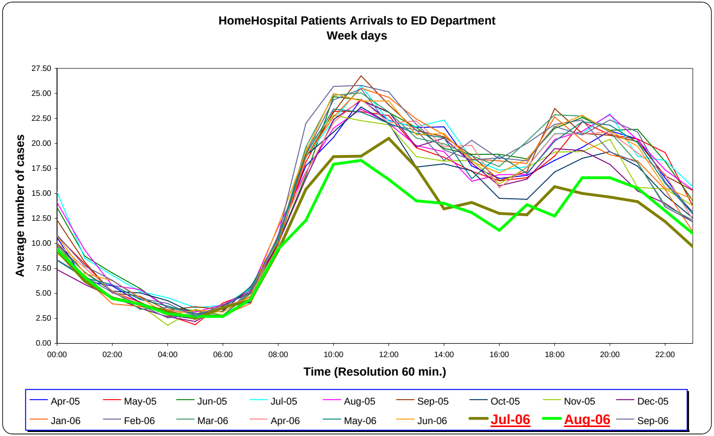
Large Israeli ED, 2006



- **Second** peak at 19:00 (vs. 15:00 in call centers).
- How much stochastic variability ?

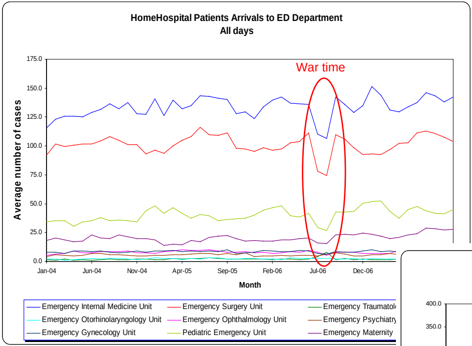
Arrivals to ED: Environment Dependence

Large Israeli ED, 2005-6

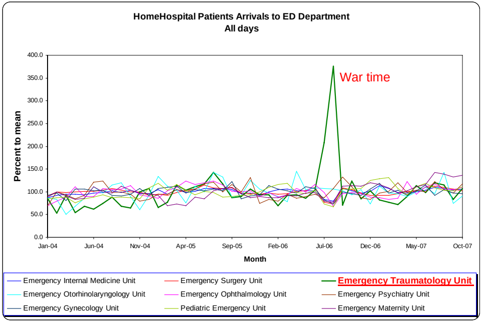


Arrivals to ED: Environment Dependence

Number of Arrivals



Percent to Mean

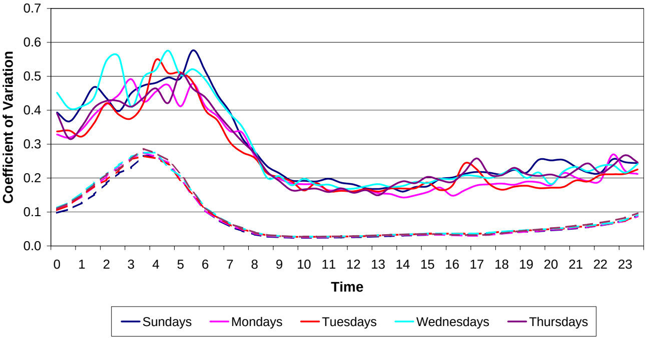


Contents of Talk

Israeli-Bank Call-Center

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

Sampled CV - solid line, **Poisson** CV - dashed line

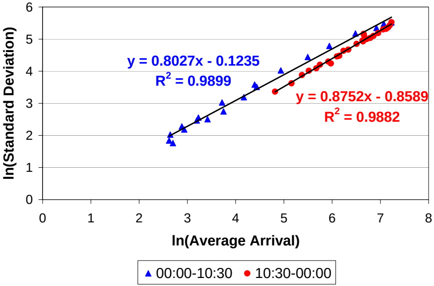


- 263 regular **days**, 4/2007 - 3/2008.
- Poisson CV $= 1/\sqrt{\text{mean arrival-rate}}$.
- Sampled CV's $\gg$ Poisson CV's $\Rightarrow$ **Over-Dispersion**.

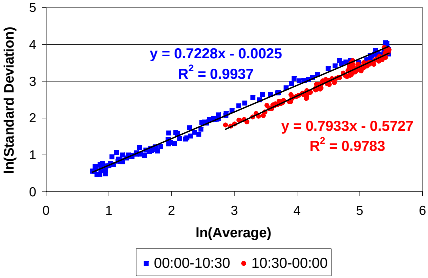
Over-Dispersion: Fitting a Regression Model

ln(STD) vs. ln(AVG)

Tue-Wed, 30 min resolution



Tue-Wed, 5 min resolution



Significant linear relations (Aldor & Feigin):

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

Over-Dispersion: Random Arrival-Rate Model

The **linear relation** between ln(STD) and ln(AVG) motivates the following model:

Arrivals distributed **Poisson with a Random Rate**

$$\mathbf{\Lambda = \lambda + \lambda^c \cdot X, \quad 0 \le c \le 1 ;}$$

- X is a random-variable with $E[X] = 0$, capturing the magnitude of **stochastic deviation** from mean arrival-rate.
- c determines **scale-order** of the over-dispersion:
  $c = 1$, proportional to λ;
  $c = 0$, Poisson-level, same as $0 \le c \le 1/2$.

In **call centers**, over-dispersion (per 30 min.) is of order λ^c, $\mathbf{c \approx 0.8 - 0.85}$.

Over-Dispersion: Distribution of X ?

- Fitting a **Gamma Poisson** mixture model to the data:
  Assume a (conjugate) prior Gamma distribution for the arrival rate $\Lambda \stackrel{d}{=} Gamma(a, b)$.
  Then, $Y \stackrel{d}{=} Poiss(\Lambda)$ is Negative Binomial.

- Very good fit of the Gamma Poisson mixture model, to data of the Israeli Call Center, for the majority of time intervals .

- Relation between our c-based model and Gamma-Poisson mixture is established.

- Distribution of X derived, under the Gamma prior assumption:
  X **is asymptotically normal**, as $\lambda \to \infty$.

Over-Dispersion: The QED-c Regime

QED-c Staffing: Under offered-load $\mathbf{R} = \boldsymbol{\lambda} \cdot \mathbf{E[S]}$,

$$\mathbf{n} \;=\; \mathbf{R} \;+\; \boldsymbol{\beta} \cdot \mathbf{R^c}\,, \qquad \mathbf{0.5 < c < 1}$$

Performance measures:

a. Delay probability: $P\{W_q > 0\} \;\sim\; 1 - F(\beta)$

b. Abandonment probability: $P\{Ab\} \;\sim\; \dfrac{E[X-\beta]_+}{n^{1-c}}$

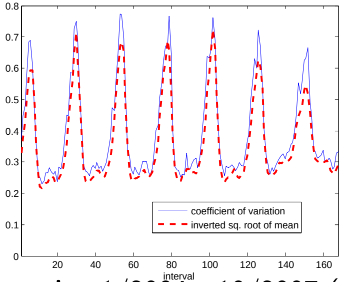
c. Average offered wait: $E[V] \;\sim\; \dfrac{E[X-\beta]_+}{n^{1-c} \cdot g_0}$

d. Average actual wait: $E_{\Lambda,n}[W] \;\sim\; E_{\Lambda,n}[V]$

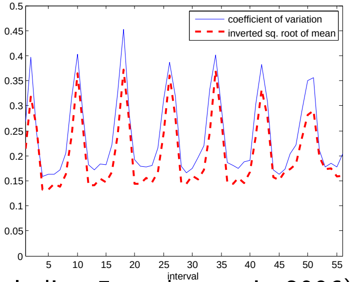
Israeli-Hospital Emergency-Department

Arrival Counts - Coefficient of Variation, per 1-hr. & 3-hr.

One-hour resolution



Three-hour resolution



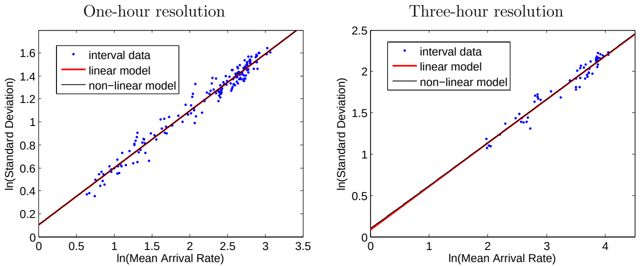
- 194 **weeks**, 1/2004 - 10/2007 (excluding 5 weeks war in 2006).
- Moderate over-dispersion: $\mathbf{c = 0.5}$ reasonable for hourly resolution.
- **ED beds in conventional QED** (**Less** var. than call centers ! ?).

Over-Dispersion: Fitting a Regression Model

Arrival Process: $\mathbf{Y \sim \text{Poisson}(\lambda + \lambda^c X), \quad c \le 1}$

Linear Regression ($c > 0.5$): $\ln(\sigma(Y)) = c \cdot \ln(\lambda) + \ln(\sigma(X))$

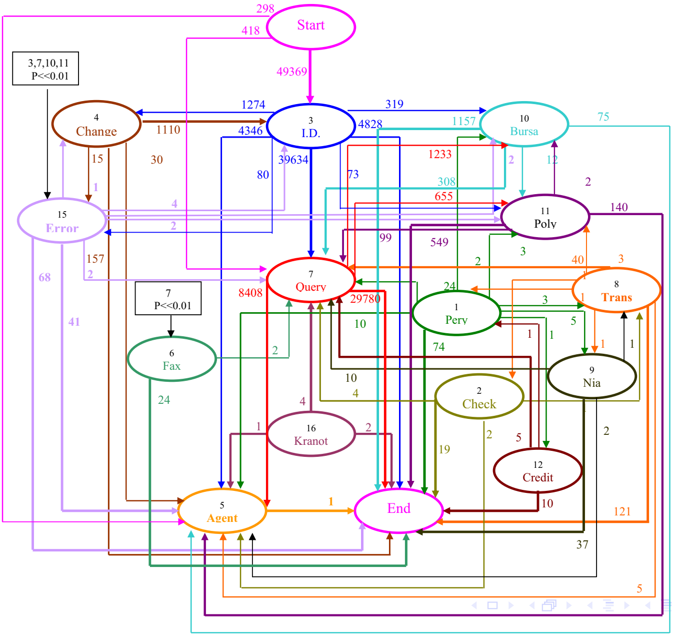
Non-Linear Regression: $\ln(\sigma(Y)) = 0.5 \cdot \ln(\lambda^{2c}\sigma^2(X) + \lambda)$



- Over-dispersion is of order $\boldsymbol{\lambda^c}$, $\mathbf{c \approx 0.5}$.
- Resolution: c depends on interval-length (1 vs. 3-hours).
- **Less** variability in ED's than in call centers ! ?

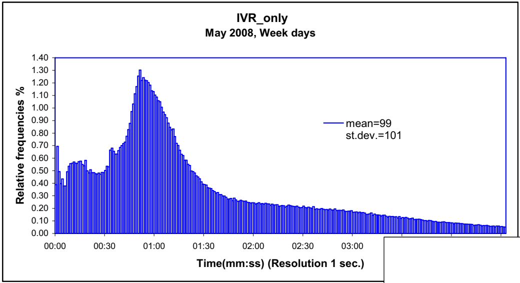
Contents of Talk

Call Transitions in the IVR - Phase Type

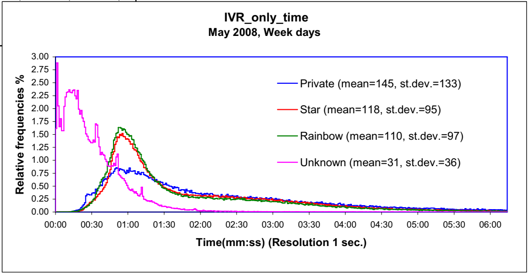


Israeli Bank: Served only by IVR, May 2008

All Customers



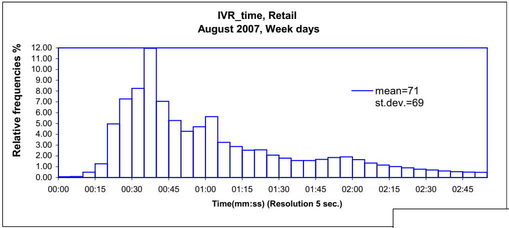
By Service Type



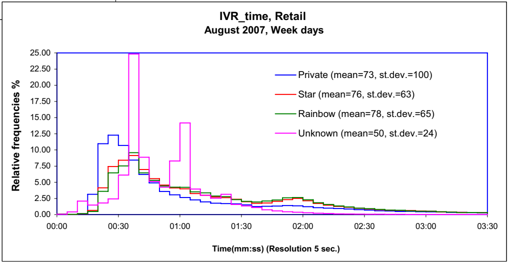
IVR Times: Histograms

Israeli Bank: Served by an Agent, May 2008

All Customers

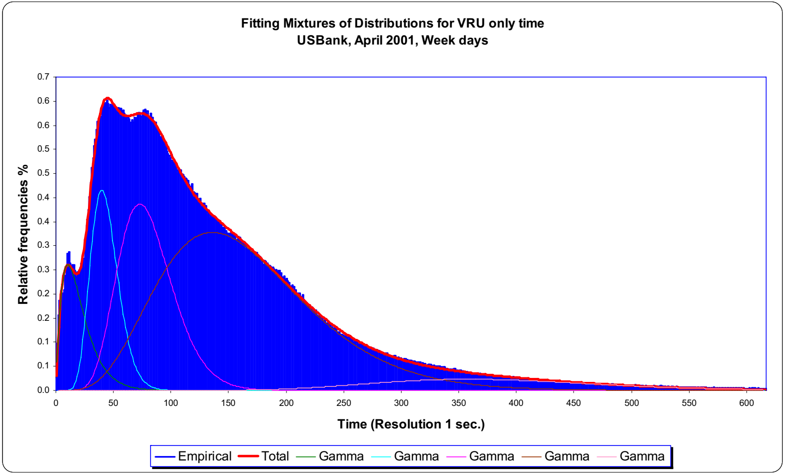


By Service Type



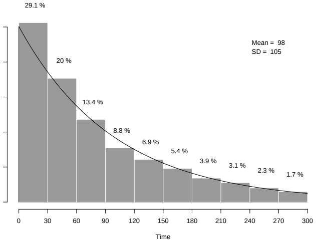
Service Times: Fitting Distribution

Fitting Mixture of 5 Gamma Components

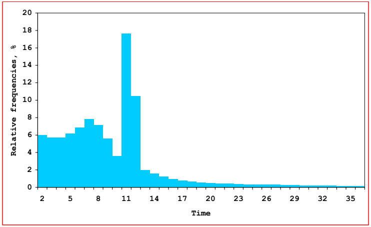


Contents of Talk

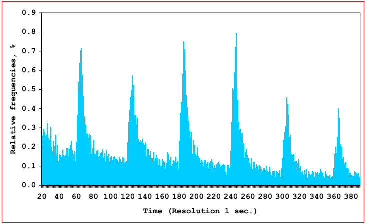
Small Israeli Bank



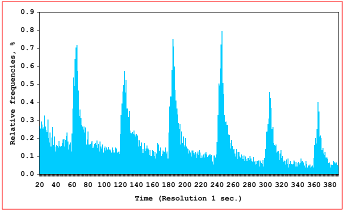
Large U.S. Bank



Medium Israeli Bank



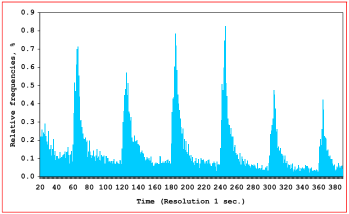
"Waiting-Times" Puzzle at a Large Israeli Bank



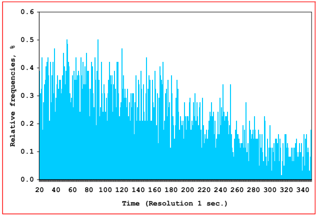
Peaks Every 60 Seconds. Why?

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

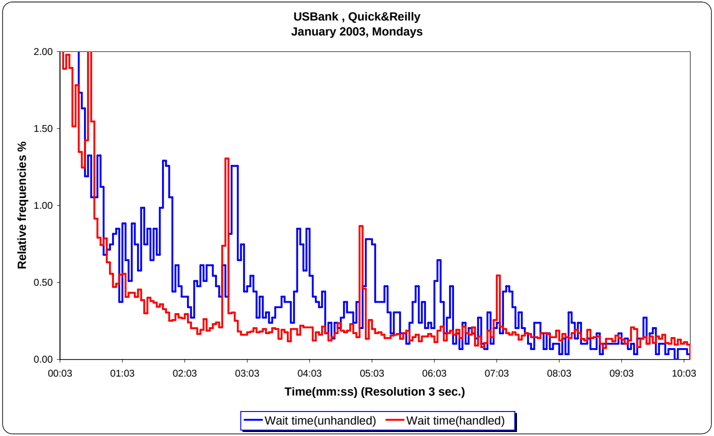
Served Customers



Abandoning Customers



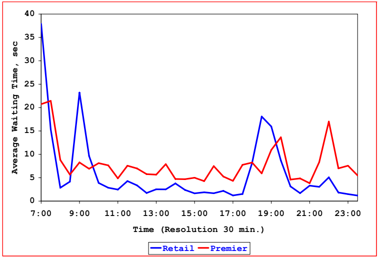
Still a Puzzle at a US Bank



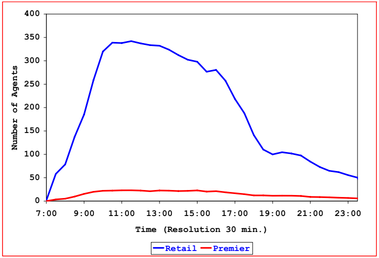
- Different **cycles of peaks** in the waiting times of both served (protocol?) and abandoning (psychology?) customers.
- **No theory** for periodic updates of either priorities or information.

US Bank: Regular vs. VIP Customers, December 2002

Average Wait



Staffing Level

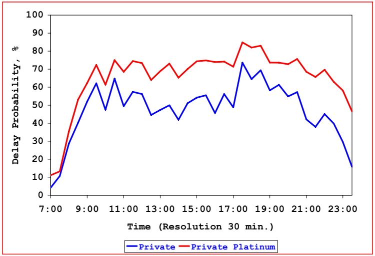


Premier (VIP of Retail) customers do not get a better service level.

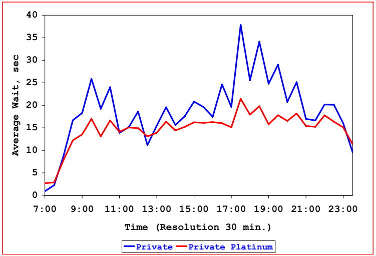
Number of agents assigned to Premier is small and they do not get enough help from regular agents.

Israeli Telecom: Regular vs. VIP Customers, October 2004

Delay Probability



Average Wait

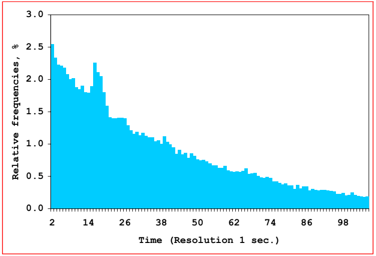


More **Platinum** customers have to wait, **but** their average wait is shorter.

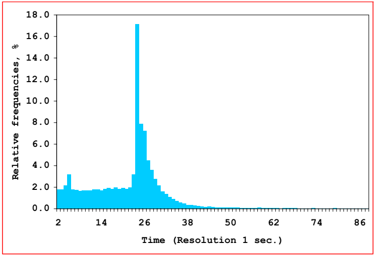
How to explain?

Histograms of Waiting Times, October 2004

Private



Private Platinum



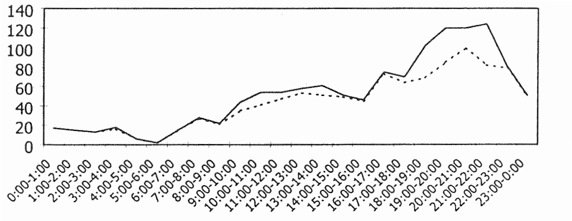
After 25 seconds of wait, Platinum are routed to Regular agents getting **high priority**. Hence, almost no long waiting times for Platinum.

Contents of Talk

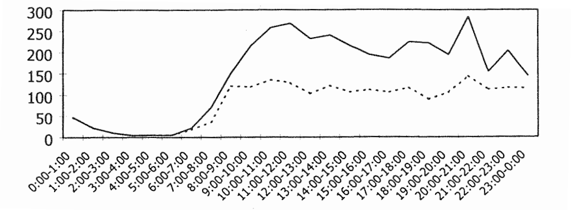
Example: "A Catastrophic Situation"

Marketing Campaign at a Call Center

Average wait 72 sec, 81% calls answered (Saturday)



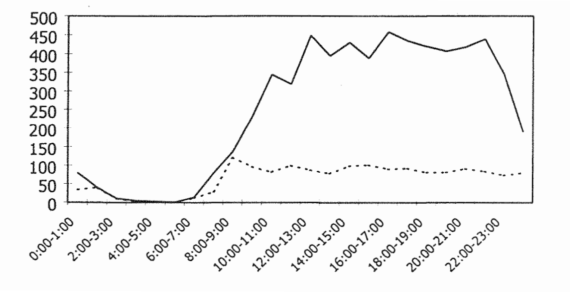
Average wait 217 sec, 53% calls answered



Example: "A Catastrophic Situation"

Avg. wait **376** sec, Max wait **1214** sec, **24% calls answered** (Sunday)

Note: Systems's capacity about 100 customers per hour.



The "Phases of Waiting" for Service

Common Experience:

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

An attempt at "Modeling the Experience":

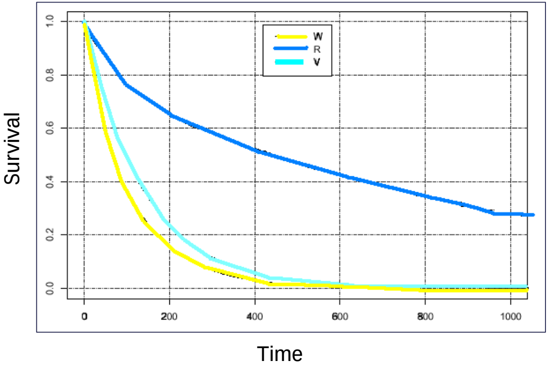
1.	Time that a customer	**expects** to wait	
2.		**willing** to wait	((Im)Patience: $\boldsymbol{\tau}$)
3.		**required** to wait	(Offered Wait:**V**)
4.		**actually** waits	($\mathbf{W_q = min(\boldsymbol{\tau}, V)}$)
5.		**perceives** waiting.	

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

Thus **left with** $(\boldsymbol{\tau}, \mathbf{V})$.

Stochastic Order: Patience vs. Offered Wait

Small Israeli Bank: Survival Functions

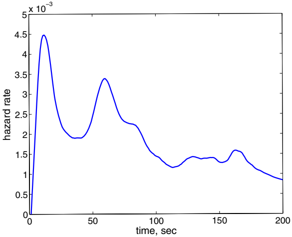


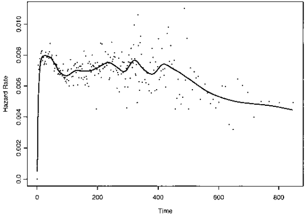
$$\mathbf{W} \overset{st}{<} \mathbf{V} \overset{st}{<} \tau$$

Call Center Data: Hazard Rates (Un-Censored)

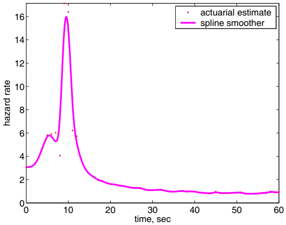
(Im)Patience Time | **Required/Offered Wait**

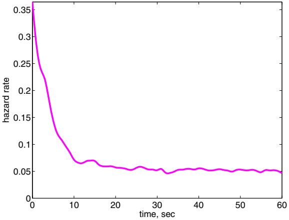
Israel





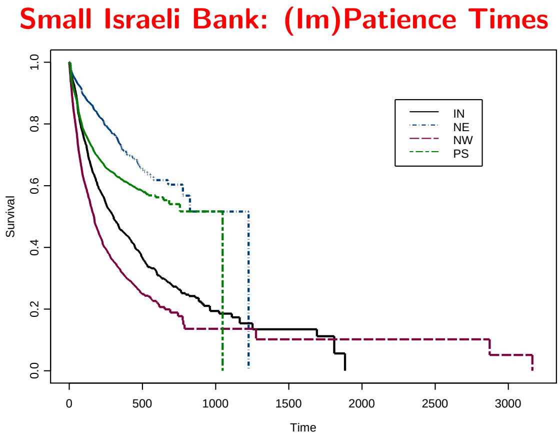
U.S.





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

(Im)Patience: Examples of Survival Function



Given Time < **750** seconds,

$$\tau_{NW} \overset{st}{<} \tau_{IN} \overset{st}{<} \tau_{PS} \overset{st}{<} \tau_{NE}$$

A Patience Index

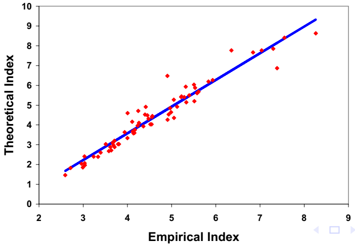
How to quantify (Im)Patience?

$$\textbf{Theoretical} \text{ Patience Index} \triangleq \frac{\textbf{Willing} \text{ to wait}}{\textbf{Expected} \text{ to wait}} = \frac{E[\tau]}{E[V]},$$

the last = if Experienced: then calculable but complex, error-prone.
Simple (but not too simple) model suggests the easily-measurable:

$$\textbf{Empirical} \text{ Patience Index} \triangleq \frac{\% \text{ Served}}{\% \text{ Abandoning}}$$

Patience index - Theoretical vs. Empirical



A Patience Index

$$\textbf{Mean Patience} = \begin{array}{c}\text{Mean Wait}\\ \text{of } \textbf{Abandoning}\\ \text{customers}\end{array} + \begin{array}{c}\text{Mean Wait}\\ \text{of } \textbf{Served}\\ \text{customers}\end{array} \times \begin{array}{c}\textbf{Patience}\\ \textbf{Index}\end{array}$$

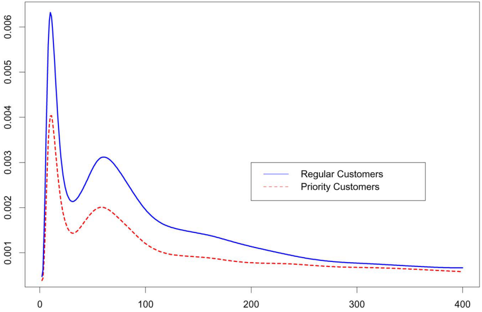
Example: Israeli Bank Data

Statistics	Average wait
360K served (80%)	2 min
90K abandoned (20%)	1 min

$$\textbf{Mean Patience} = 1 + 2 \times \frac{80\%}{20\%} = \textbf{9}!$$

If the average patience is **9 minutes**, why customers abandon in **1 minute**?

Hazard Rates of (Im)Patience in an Israeli Small Bank: Regular over VIP Customers



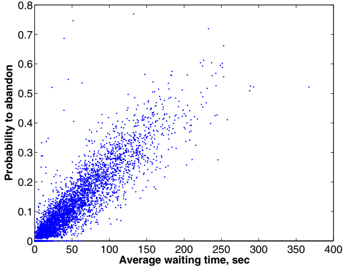
- **VIP** customers are **more patient** (needy).
- Why peaks in abandonment? **Announcements!**
- **Call-by-call data** required to obtain this graph (+Uncensoring).
- Triggered **Research**: M/M/n+GI (w/ Zeltyn, '05), G/GI/n+GI (w/ Momcilovic '09); Info while waiting (Munichor & Rafaeli '08)

Estimating Patience: $P\{Ab\} \propto E[Wq]$ Relation

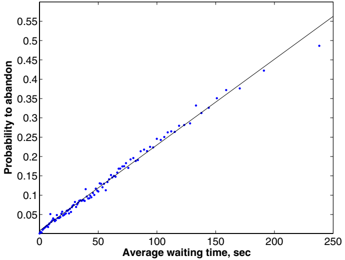
In queues with $\mathbf{exp(\theta)}$ patience: $\mathbf{P\{Ab\} = \theta \cdot E[Wq]}$.

Israeli Bank: Yearly Data

Hourly Data



Aggregated

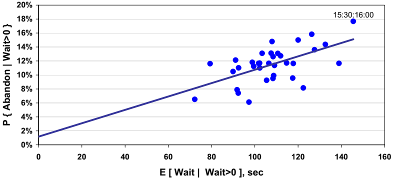


Graphs are based on 4158 hour intervals.

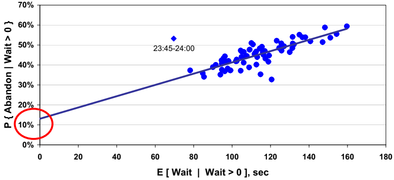
Estimate of mean patience: $250/0.55 \approx$ **450 seconds**.

Small Israeli Bank, 1999

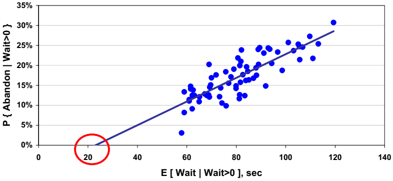
Standard Exponential (NE)



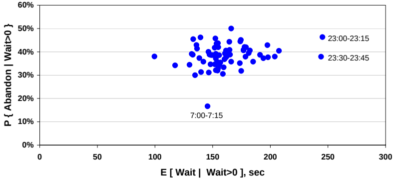
Exponential with Balking (NW)



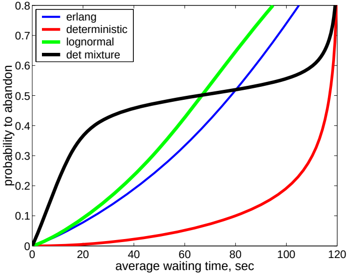
Two Phases (PS)

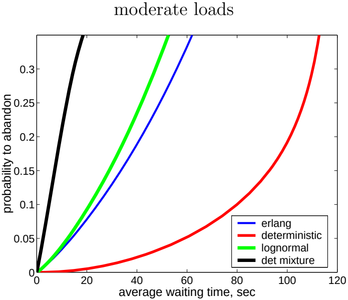


Learning (IN)



Theoretical Examples of Non-Linear Relations



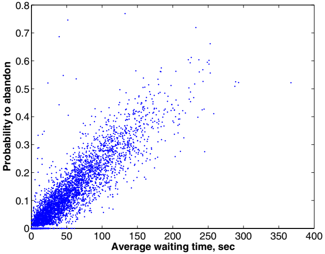


Patience distributions:

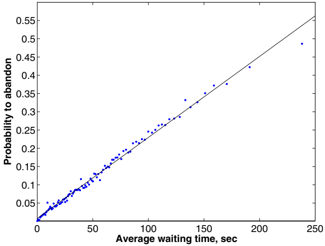
- **D**: Deterministic, 2 minutes exactly;
- **E**: Erlang with two exp(mean=1) phases;
- **LN**: Lognormal, both average and standard deviation equal to 2;
- **D-Mix**: 50-50% mixture of two constants: 0.2 and 3.8.

The Impact of Customers Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/N+G Queue

Hourly Data



Aggregated



Theory:

Erlang-A: $P\{Ab\} = \theta \cdot E[W_q]$ **M/M/N+G:** $P\{Ab\} \approx g_0 \cdot E[W_q]$.

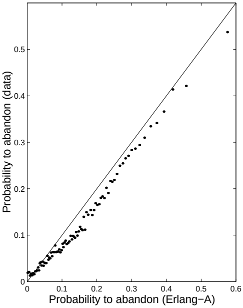
Recipe:

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

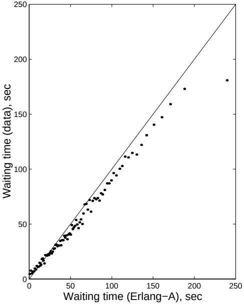
Erlang-A: Fitting a Simple Model to a Complex Reality

- Small Israeli bank (10 agents)
- Patience estimated via $P\{Ab\}/E[Wq]$
- Graphs: **hourly performance vs. Erlang-A predictions**, over 1 year, aggregating groups with 40 similar hours.

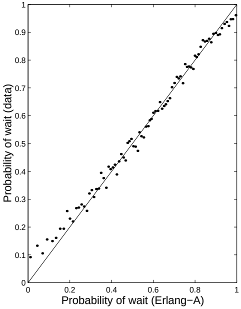
$\mathbf{P\{Ab\}}$



$\mathbf{E[W_q]}$



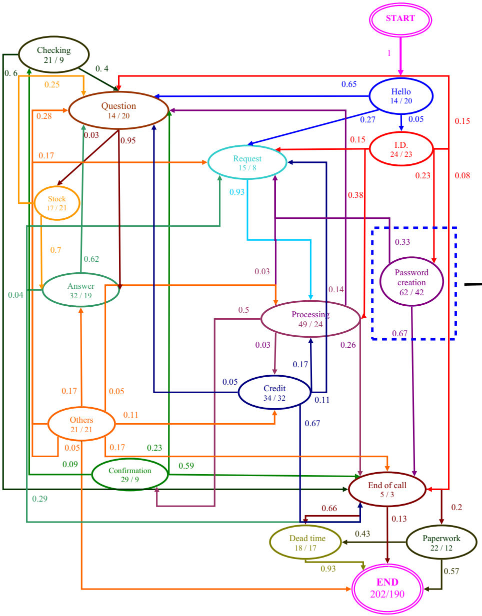
$\mathbf{P\{W_q > 0\}}$

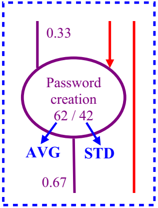


Contents of Talk

Call Transitions in the Service

Israeli Bank,
Retail Service

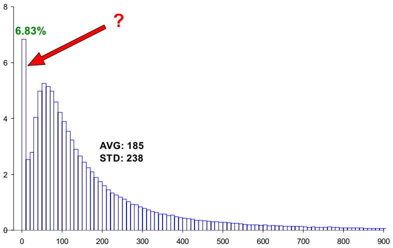




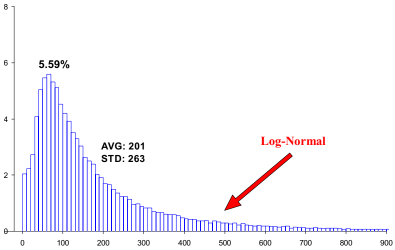
Service Times: Distribution and Psychology

Histogram of Service Times in a Small Israeli Bank

January-October



November-December

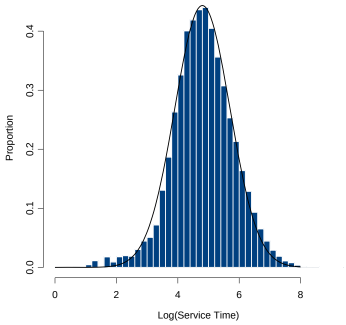


- **Lognormal** service times prevalent in call centers
- **6.8% Short-Services:** Agents' "Abandon" (improve bonus, rest)
- **Distributions**, not only Averages, must be measured.

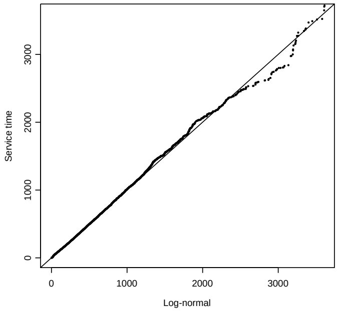
Validating LogNormality of Service Times

Israeli Call Center, Nov-Dec, 1999

Log(Service Times)

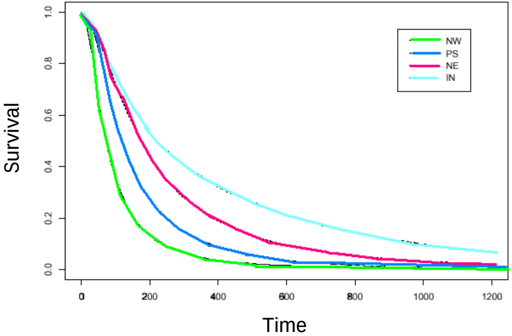


LogNormal QQPlot



Service Times: Stochastic Order

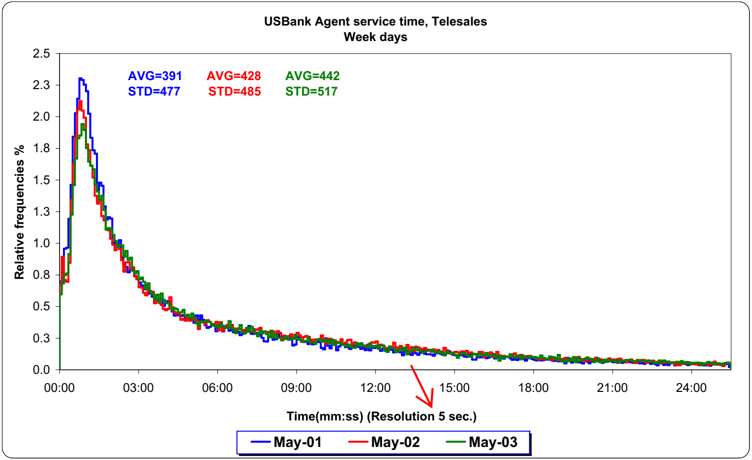
Small Israeli Bank: Survival Functions by Type



Service times stochastic order: $\mathbf{S_{NW}} \overset{st}{<} \mathbf{S_{PS}} \overset{st}{<} \mathbf{S_{NE}} \overset{st}{<} \mathbf{S_{IN}}$

Patience times stochastic order: $\tau_{NW} \overset{st}{<} \tau_{IN} \overset{st}{<} \tau_{PS} \overset{st}{<} \tau_{NE}$

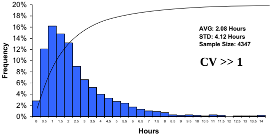
US Bank: Service Time Histograms for Telesales, 2001-3



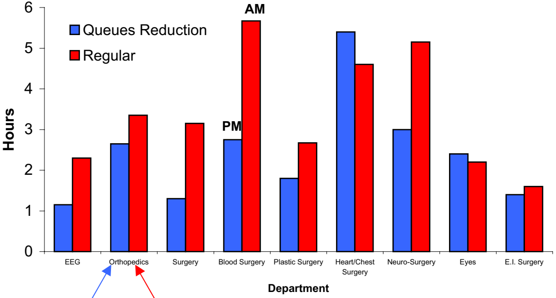
Service Times: Management

Operations Time In a Hospital

Histogram



Morning (by Hour) vs. Afternoon (by Case)



Afternoon, by Case

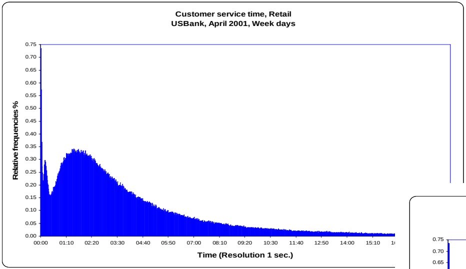
Morning, by Hour

Ethical? **Even Doctors Can Manage!**

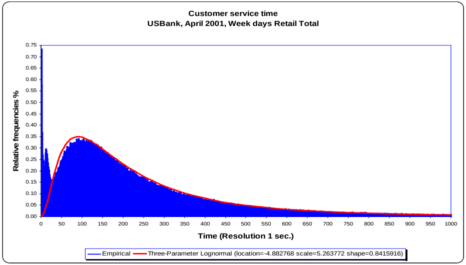
Service Times: Fitting Distribution

US Bank: Service Time of Retail, April 2001

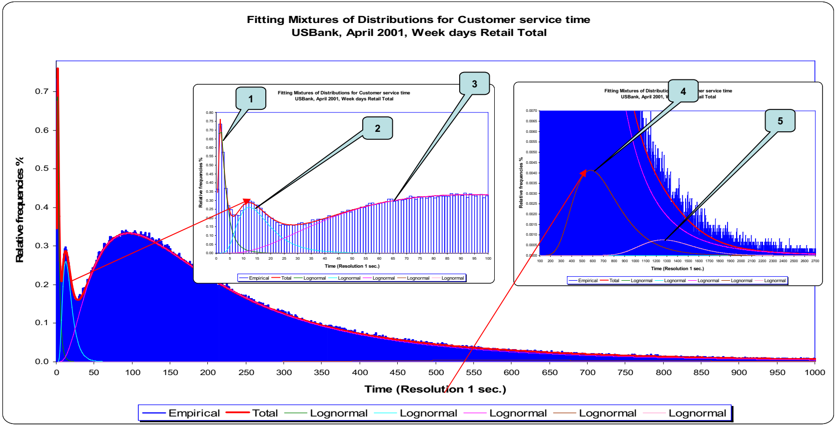
Histogram



Fitting Distribution

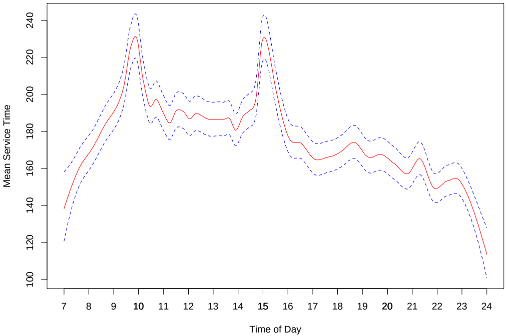


Fitting Mixture of 5 Lognormal Components



Service Times: Time and/or State-Dependence

Israeli Bank: Mean Service Time vs. Time over the Day



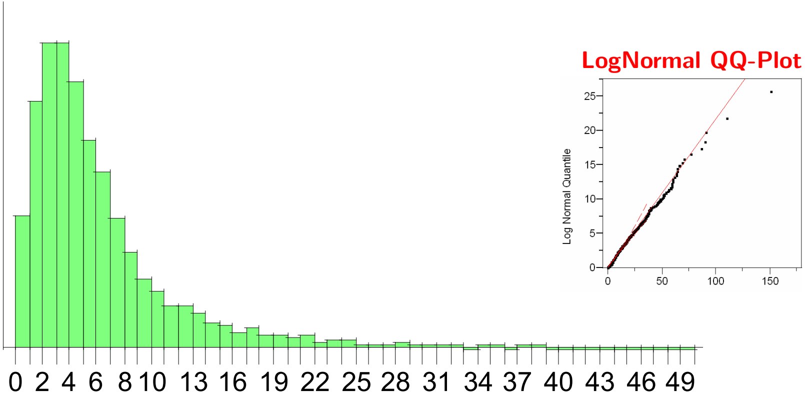
Prevalent: **Longest services** at **peak-loads** (10:00, 15:00). **Why?**

Explanations:

- Common: Service protocol different (longer) at congestion.
- Operational: The **needy** abandon less during peak loads; hence the **VIP** remain on line, with their **longer** service times.

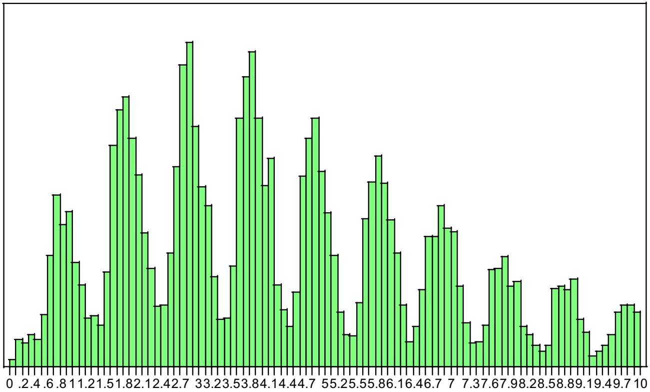
Israeli Large Hospital: LOS in IW

Days Resolution



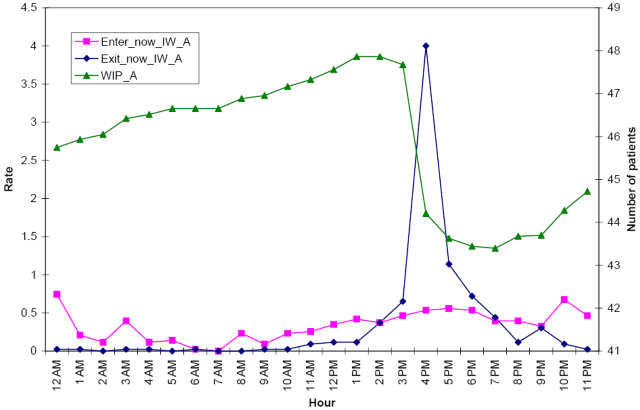
Length of Stay: Resolution Dependence

Hours Resolution: LN = Normal Mixture



Length of Stay: Resolution Dependence

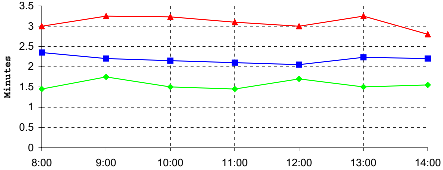
Internal Ward A: **Arrivals** / **Departures** / **# Patients** , by hour



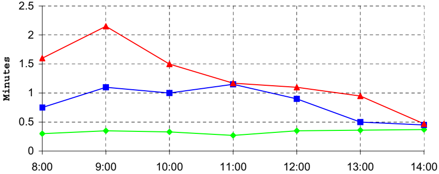
Ongoing: Empirical Analysis of an ED, IW and Everything In Between, w/ Y. Marmor, Y. Tseytlin, G. Yom-Tov, M. Armony.

Three Israeli Call Centers, Doing the Same

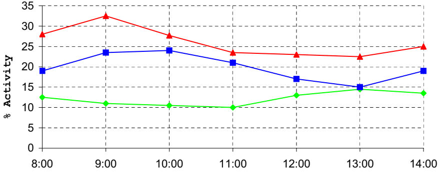
Average Service Time



Average Waiting Time



% Abandonment

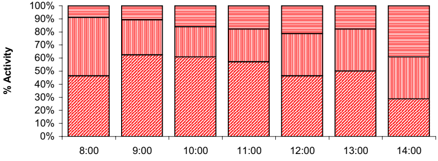


- Tel Aviv
- Raanana
- Jerusalem

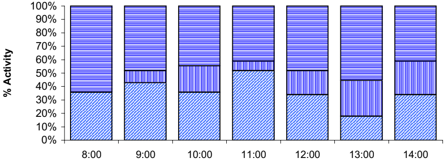
Utilization Profile

Three Israeli Call Centers, Doing the Same

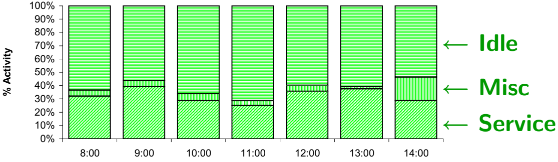
Tel Aviv



Raanana



Jerusalem



Operational challenge:
managing idleness

Calculating (Mean) Service Time

First approach: Sum up components of the "service time", then add related activities of servers.

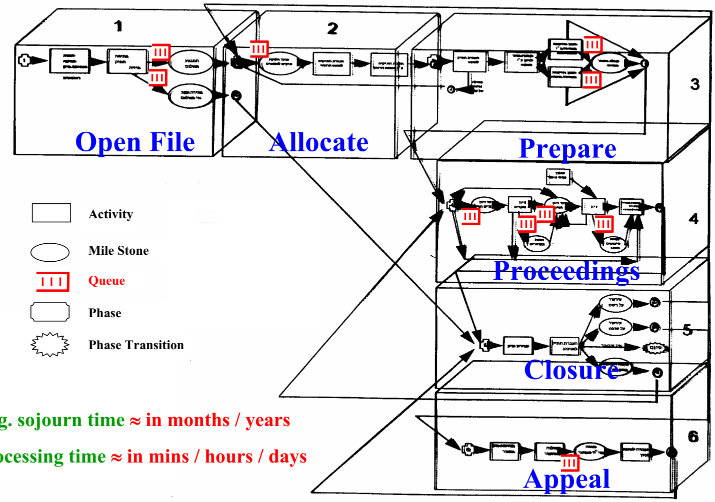
Second approach (Avoids Ambiguities):
Fix a time interval (eg. a shift).

$$\textbf{Mean Service Time} = \frac{\textbf{Available Time - Idle Time}}{\textbf{Number of Calls}},$$

where
Available Time = # Agents $\times$ Interval Duration,
and
Idle Time is summed over all agents.

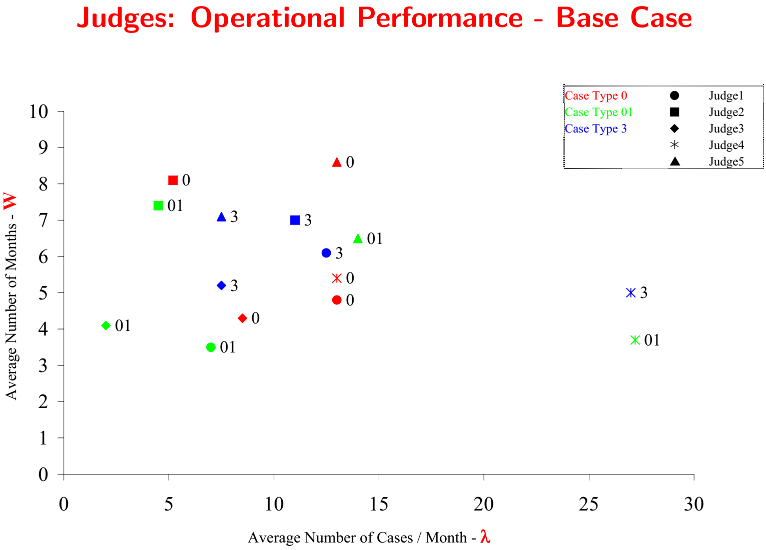
Conceptual Model: The "Production of Justice"

The Labor-Court Process in Haifa, Israel

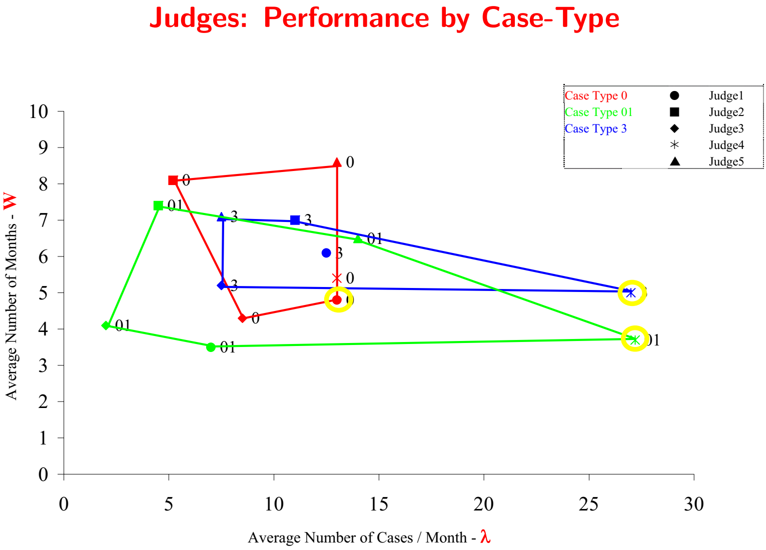


Avg. sojourn time ≈ in months / years

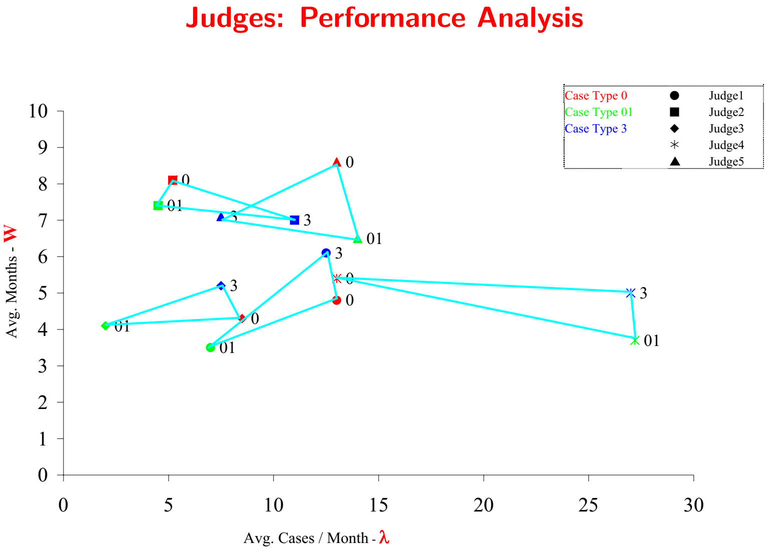
Processing time ≈ in mins / hours / days


Judges: Operational Performance - Base Case
Case Type 0
Case Type 01
Case Type 3
Judge1
Judge2
Judge3
Judge4
Judge5
Average Number of Months - W
Average Number of Cases / Month - λ
0
1
2
3
4
5
6
7
8
9
10
0
5
10
15
20
25
30
0
01
3

Analytical Model: Little's Law in Court (II)

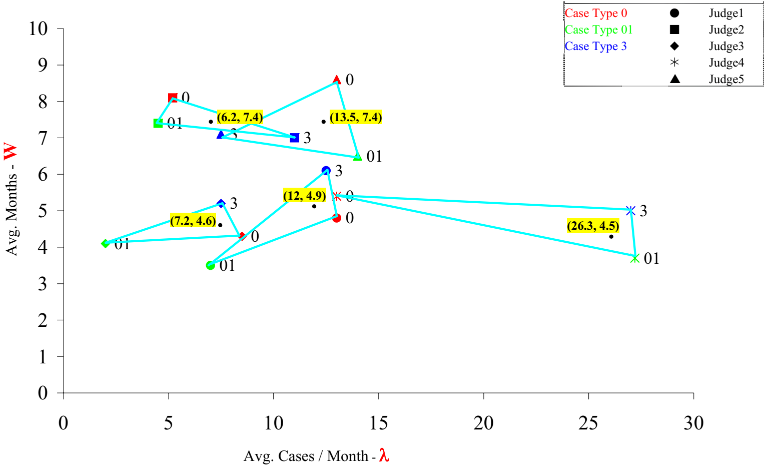


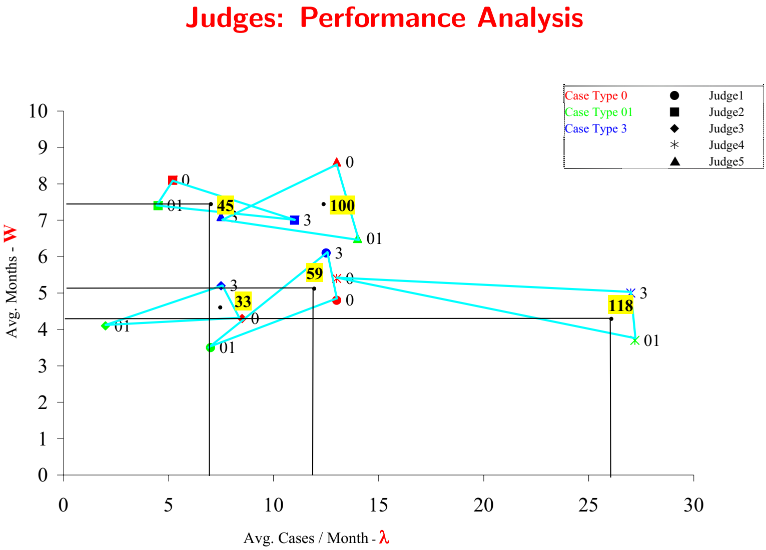
Analytical Model: Little's Law in Court (III)



Analytical Model: Little's Law in Court (IV)

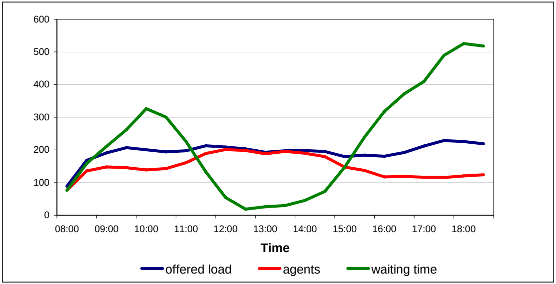
Judges: Performance Analysis

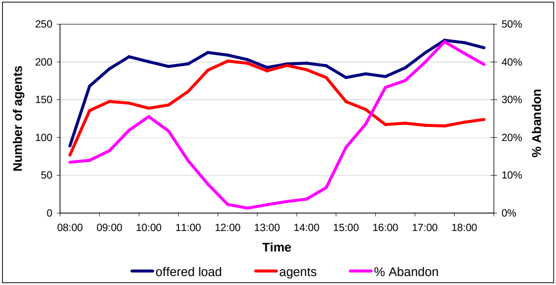



Judges: Performance Analysis
Case Type 0
Case Type 01
Case Type 3
Judge1
Judge2
Judge3
Judge4
Judge5
Avg. Months - W
Avg. Cases / Month - λ
45
100
59
33
118
0
01
3
0 5 10 15 20 25 30
0 1 2 3 4 5 6 7 8 9 10

Offered-Load vs. # Agents

Israeli Cable Company,
Retail Service,
January 2009

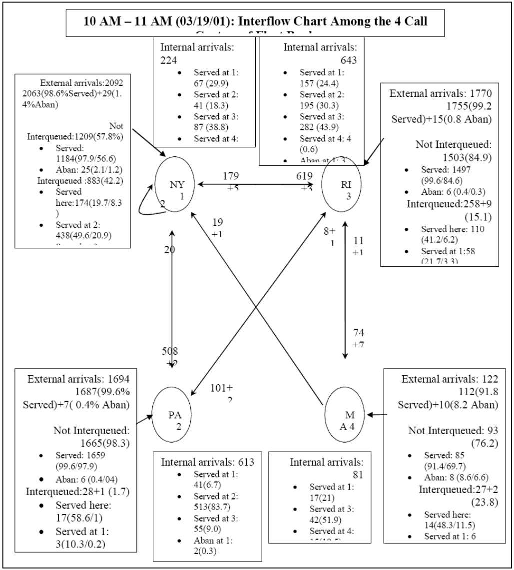




Contents of Talk

System Design: Inter-queue Model

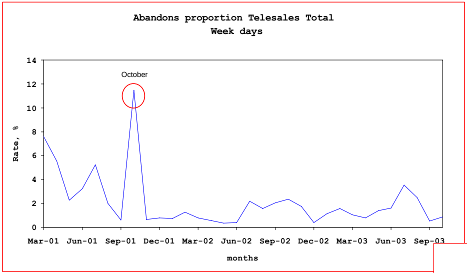
US Bank



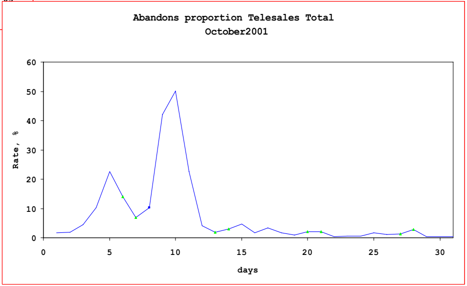
Contents of Talk

1. Peak of Telesales Abandonment, US Bank

Monthly Abandonment Rate



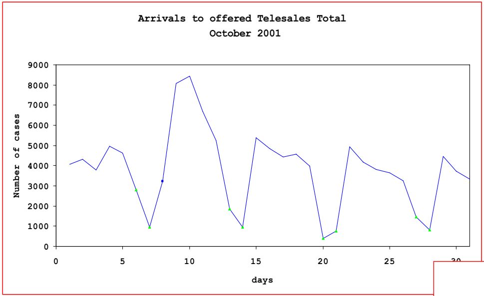
Daily Abandonment Rate, October 2001



Unusual large abandonment rates on October 9-10th ($\approx$ 40-50%)

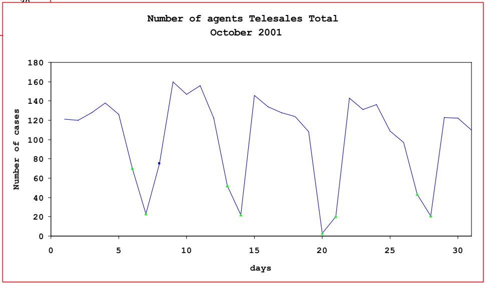
1. Peak of Telesales Abandonment, US Bank

Daily Arrivals, October 2001



October 9th: Heavy, following the Columbus day

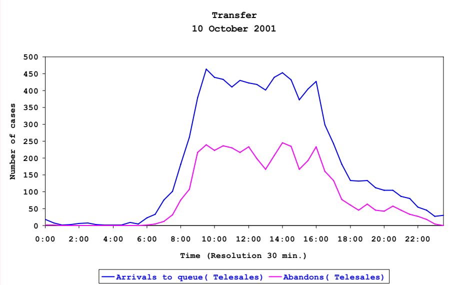
Number of Agents, October 2001



Slightly larger number of agents on October 9-11th

1. Peak of Telesales Abandonment, US Bank

Arrivals and abandonments, October 10th

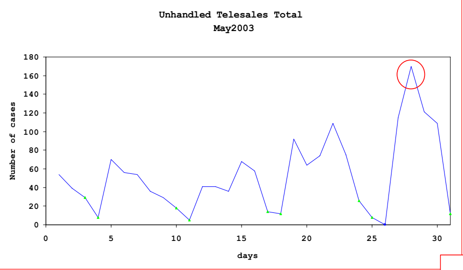


Nearly 50% abandonments along the working day

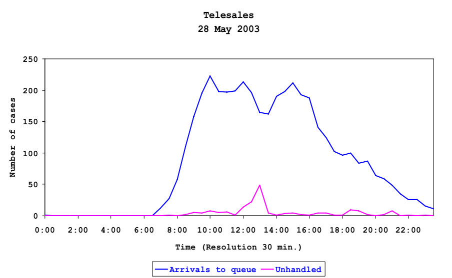
The somewhat increased number of agents on October 9-11th is insufficient for sustaining the usual service level

2. Peak of Telesales Unhandled, US Bank

Unhandled, May 2003

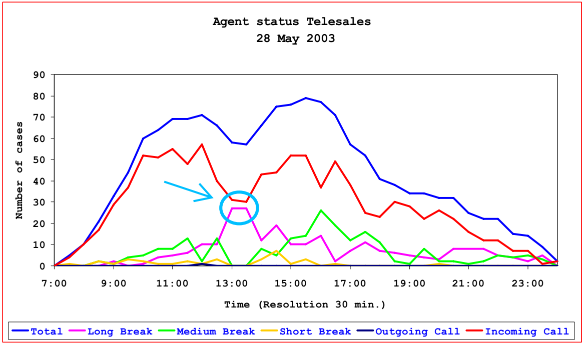


Arrivals and unhandled, 28 May 2003



13:00: Peak of unhandled calls with a significant decrease of the arrivals

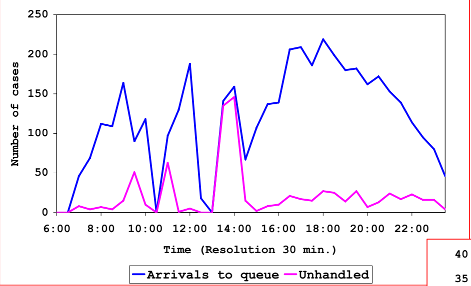
Agents Status, 28 May 2003



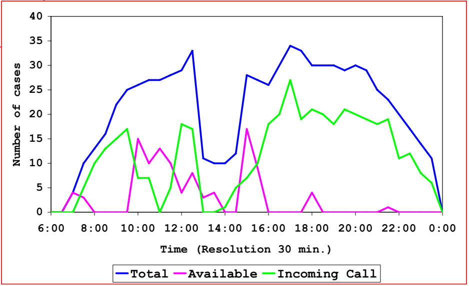
12:00-13:00: Significant decrease of agents serving incoming calls and sharp increase in the number of agents who were on a long-break

3. Peak of Technical Unhandled, Israeli Telecom

Unhandled Calls, May 24th, 2005



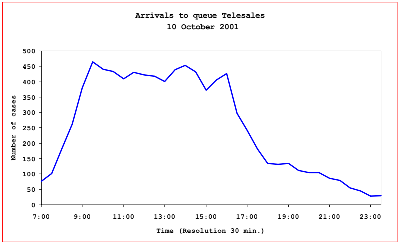
Agents Status, May 24th, 2005



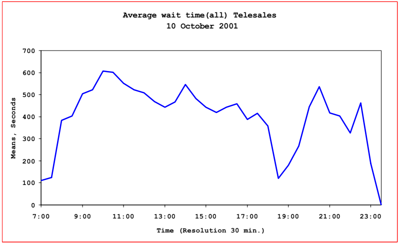
Contents of Talk

US Bank: Telesales Calls, October 10, 2001

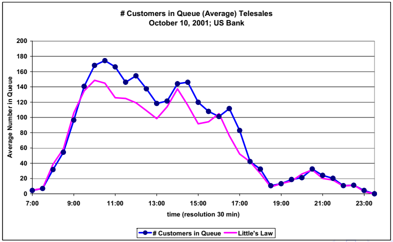
λ, Throughput Rate



W, Average Waiting Time

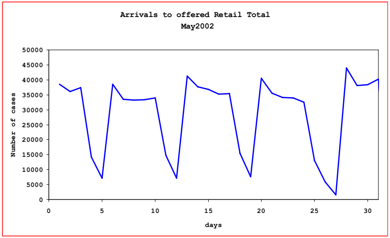


L, Average Queue Length

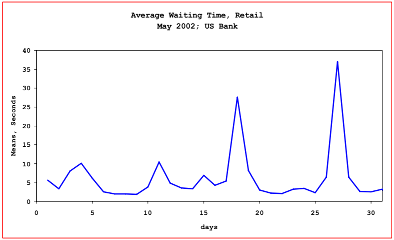


US Bank: Retail calls, May 2002

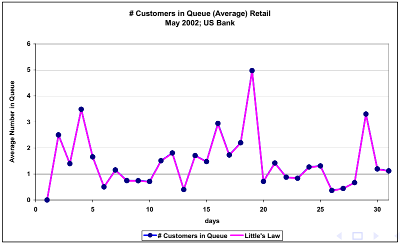
λ, Throughput Rate



W, Average Waiting Time

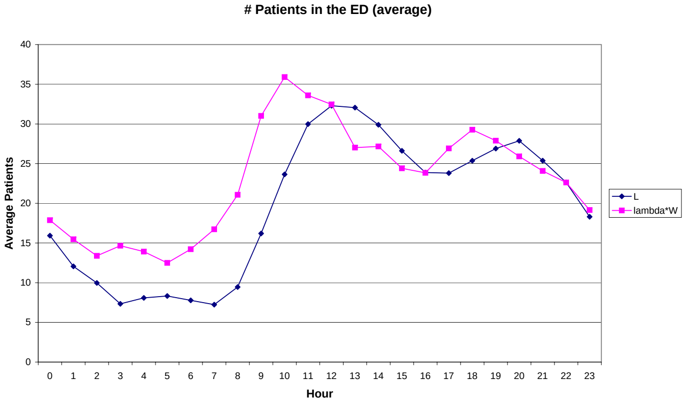


L, Average Queue Length



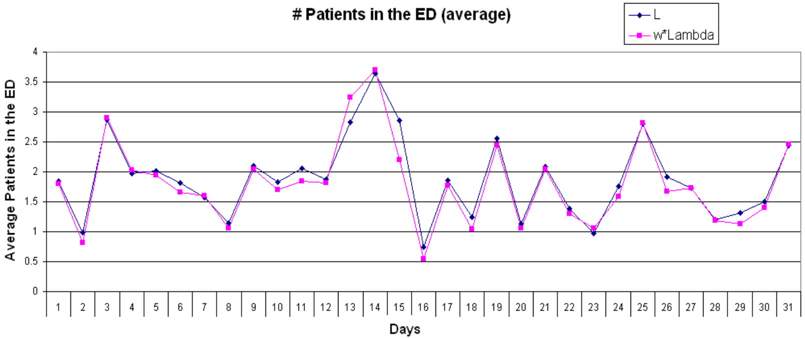
Little's Law, $L = \lambda \cdot W$

Israeli ED, Hour Resolution

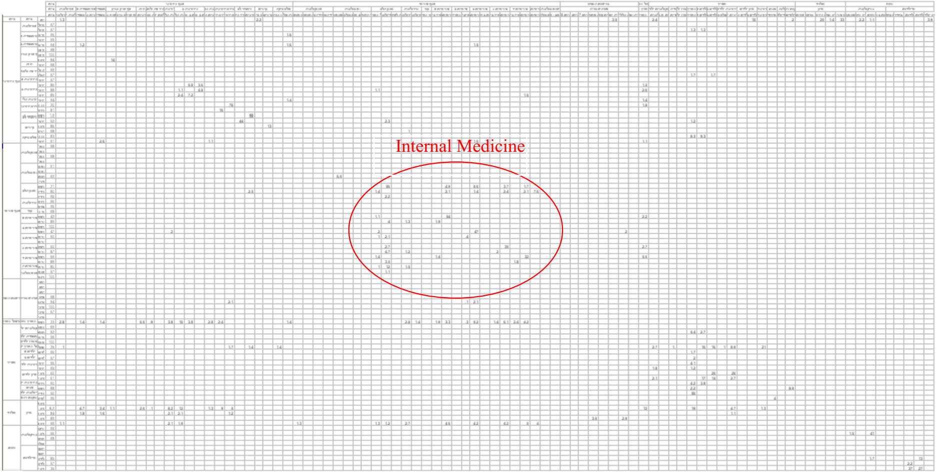


Little's Law, $L = \lambda \cdot W$

Israeli ED, October 1999, Day Resolution



90 × 90 Matrix, Sub-Ward Resolution



8×8 Matrix, Division Resolution

Including Arrivals and Releases

	Home	Surgery	Internal	Psychology	Intensive Care	Pediatrics	Emergency Dep.	Gynecology
Home		8.4	3.2	0.1		18.3	**60.3**	9.7
Surgery	90	7.9	1.3		0.7	0.1		
Internal	84.4	1.9	13	0.1	0.5			0.1
Psychology	94.3	1.9	3.8					
Intensive Care	17.2	**40.9**	**38.4**			0.9		2.6
Pediatrics	78.8	0.6				20.6		
Emergency Dep.	69.9	8.9	**19.2**	0.2	0.3	1		0.5
Gynecology	55.3	0.3	0.2		0.1			44.1

Transitions Inside the Hospital

	Surgery	Internal	Psychology	Intensive Care	Pediatrics	Emergency Dep.	Gynecology
Surgery	**78.3**	12.7	0.2	7	1.4		0.4
Internal	12	**83.3**	0.6	3.4	0.2		0.5
Psychology	33.3	66.7					
Intensive Care	**49.5**	**46.4**			1		3.1
Pediatrics	2.6	0.2		0.1	**96.9**		0.2
Emergency Dep.	**29.7**	**63.7**	0.6	0.9	3.4		1.7
Gynecology	0.7	0.4		0.1			**98.8**

- About 50% of transitions between ED and internal wards.
- Most transitions are inside the specific hospitalized unit.

IW Operational Measures, or Efficiency vs. Fairness

Israeli Large Hospital (1/5/06 to 30/10/08, excluding 1-3/07)

	Ward A	Ward B	Ward C	Ward D
ALOS (days)	6.37	4.47	5.36	5.56
Avg Occupancy Rate	97%	95%	86%	92%
Avg # Patients per Month	206	187	210	210
Standard capacity	45	30	44	42
Avg # Patients /Bed/Month	4.57	6.25	4.77	4.77
Return Rate	15.4%	15.6%	16.2%	14.8%

"ED-to-IW" Routing

IW Operational Measures, or Efficiency vs. Fairness

Israeli Large Hospital (1/5/06 to 30/10/08, excluding 1-3/07)

	Ward A	Ward B	Ward C	Ward D
ALOS (days)	6.37	4.47	5.36	5.56
Avg Occupancy Rate	97%	95%	86%	92%
Avg # Patients per Month	206	187	210	210
Standard capacity	45	30	44	42
Avg # Patients /Bed/Month	4.57	6.25	4.77	4.77
Return Rate	15.4%	15.6%	16.2%	14.8%

- The "fastest" + smallest Ward B subject to highest workload: **occupancy, flux**: **unfair**.
- Calls for ED-to-IW routing, which is both efficient and fair (w/ Tseytlin (MSc), Tseytlin & Momcilovic, Tseytlin & Zviran): **exact analysis, QED approximation (natural - hours wait for days service), partial bed-information**.

What is "Fair" Allocation?

Each nurse/doctor should have the same workload.

- Take care of an equal number of patients.
- Number of nurses/doctors is proportional to standard number of beds.

$\Rightarrow$ Balance occupancy rates among the wards.

- But then, by Littles law, wards with shorter ALOS will have a higher turnover rate.
- And the load on the wards staff is not uniform during a patient's stay - extra work involved in reception and dischage.

$\Rightarrow$ Balance number of patients per bed per time unit (**flux**) among the wards.

% Block (Galit)

Definitions:

- The part of ward i in the system's **dynamic capacity**: $a_i = \dfrac{N_i \mu_i}{\lambda}$.
- The part of ward i in the system's **static capacity**: $q_i = \dfrac{N_i}{N}$.

Routing Policies

- Randomized Most-Idle (**RMI**): A customer is routed to ward i with probability $\dfrac{\mathrm{I}_1}{\mathrm{I}_1 + \mathrm{I}_2}$.
  - If $\mu_1 > \mu_2$:
    - $\rho_1 < \rho_2$ (occupancy)
    - $\gamma_1 > \gamma_2$ (flux)
  - Asymptotically, $\dfrac{\mathrm{I}_1}{\mathrm{I}_2} \approx \dfrac{a_1}{a_2}$.

- Most Idle (**MI**), the naive non-random equivalent to RMI: A customer is routed to the most vacant ward.
  - Larger ward has higher occupancy.
  - Asymptotically, $\dfrac{\mathrm{I}_1}{\mathrm{I}_2} \approx 1$.

Routing Policies

- Weighted Most-Idle (**WMI**): A customer is routed to the ward with the number of idle servers multiplied by the ward's weight is maximal.
  - Weight vector: (w_1, w_2), $w_i \in (0,1)$, $w_1 + w_2 = 1$.
  - Interesting cases:
    - $w_1 = w_2 = 1/2$: MI routing policy.
    - $w_1 = a_2$, $w_2 = a_1$: Non-random Equivalent to RMI - NERMI routing policy.
    - $w_1 = q_2$, $w_2 = q_1$: Occupancy-Balancing policy - routing an arriving customer to the least utilized ward.
  - Asymptotically, $\dfrac{\mathrm{I}_1}{\mathrm{I}_2} \approx \dfrac{w_1}{w_2}$.

Comparison: WMI vs. RMI

		Idleness-Ratio	Flux-ratio	P(block)
$w_1 q_1 = w_2 q_2$		WMI	RMI	WMI
$w_1 q_1 > w_2 q_2$	$\frac{\mu_1}{\mu_2} < \frac{w_1 q_1}{w_2 q_2}$	RMI	RMI	WMI
	$\frac{\mu_1}{\mu_2} = \frac{w_1 q_1}{w_2 q_2}$	equal		
	$\frac{\mu_1}{\mu_2} > \frac{w_1 q_1}{w_2 q_2}$	WMI		
$w_1 q_1 < w_2 q_2$	$w_1 a_1 < w_2 a_2$	RMI	WMI	RMI
	$w_1 a_1 = w_2 a_2$	equal	equal	equal
	$w_1 a_1 > w_2 a_2$	WMI	RMI	WMI

For different sets of parameters and different target functions, a different policy is superior.

⋆ Idle-ratio: ratio between proportion of idle servers in the wards, $\frac{\mathrm{I}_1/N_1}{\mathrm{I}_2/N_2}$.
⋆ Flux-ratio: ratio between flux through the wards.

Rules-of-Thumb

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

ED: $\mathbf{n \approx R - \gamma R} \quad (0.1 \leq \gamma \leq 0.25)$.

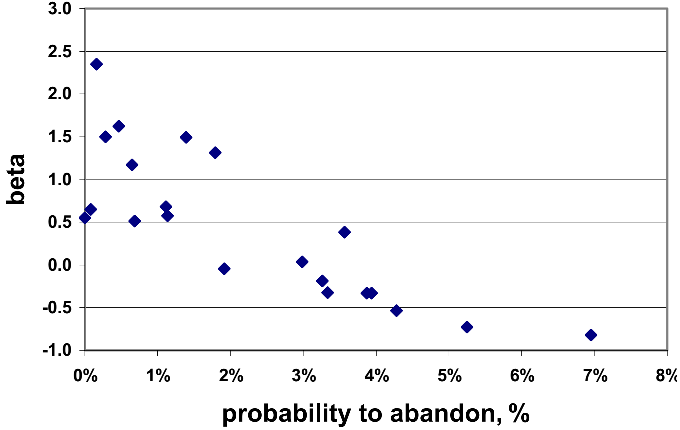
QD: $\mathbf{n \approx R + \delta R} \quad (0.1 \leq \delta \leq 0.25)$.

QED: $\mathbf{n \approx R + \beta\sqrt{R}} \quad (-1 \leq \beta \leq 1)$.

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

QED: Practical Support

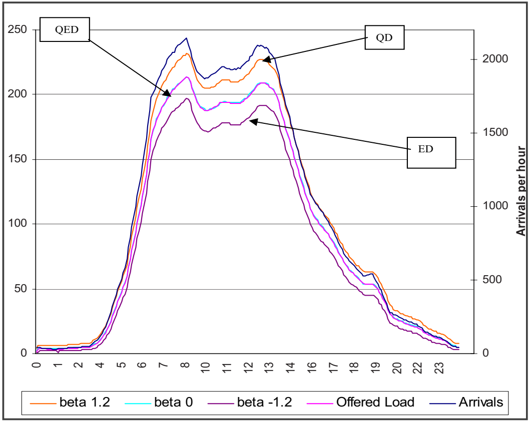
QOS parameter $\beta = (n - R)/\sqrt{R}$ vs. %Abandonment



Square-Root Staffing:

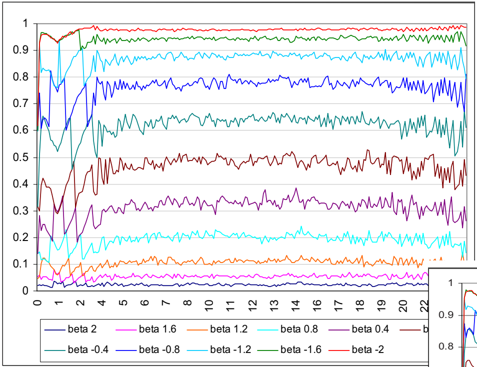
$$\mathbf{n(t)} \;=\; \mathbf{R(t)} \;+\; \beta\sqrt{\mathbf{R(t)}}, \qquad -\infty < \beta < \infty$$

Arrivals, Offered Load and Staffing

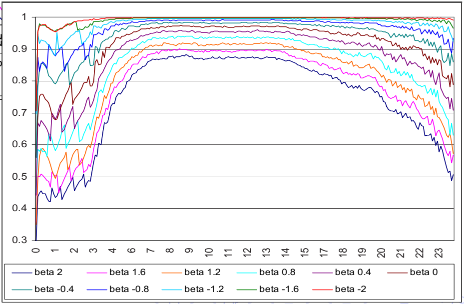


Time-Stable Performance of Time-Varying Systems

Delay Probability

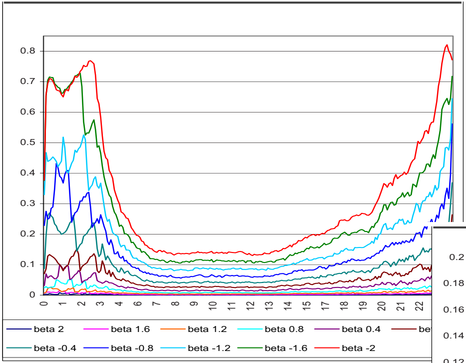


Servers' Utilization

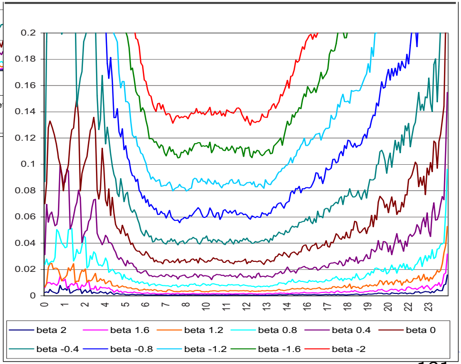


Time-Stable Performance of Time-Varying Systems

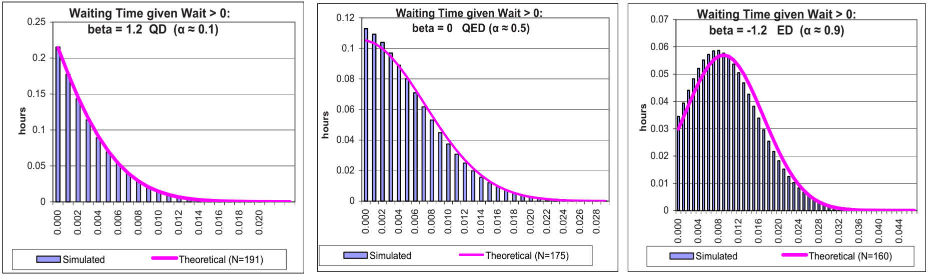
Abandonment Probability



(Zoom)



Waiting Time, Given Waiting: Empirical vs. Theoretical Distribution



Contents of Talk

Workload and Offered-Load

- **Workload**: Stochastic process, representing the amount of work present at time t, under the assumptions of **infinitely many resources** (service commences immediately upon arrival).
- **Offered-Load**: Function of time $t \geq 0$, representing the average of the workload at time t.

The Offered-Load, $R(t)$, determines staffing level via c-staffing ($c = 0.5$ is conventional square-root staffing):

$$N(t) = R(t) + \beta \cdot [R(t)]^c$$

Notations and Assumptions

Notations:

- S - Service time of a customer.
- τ - (Im)patience of a customer, i.e. the time willing to wait before abandoning.
- V - Virtual-waiting-time (or offered-waiting-time), i.e. time required to wait.
- W - Waiting time of a customer, i.e. the minimum between τ and V.

Notations and Assumptions

Notations:

- S - Service time of a customer.
- τ - (Im)patience of a customer, i.e. the time willing to wait before abandoning.
- V - Virtual-waiting-time (or offered-waiting-time), i.e. time required to wait.
- W - Waiting time of a customer, i.e. the minimum between τ and V.

Assumptions:

- W is observable for all customers.
- S observed only for customers who are served ($\tau > V$, in which case also $\tau > W$.)

Offered-Load Representations (or Time-Varying Little)

For the $M_t/GI/N_t + GI$ queue, the **offered-load** $R = \{R(t),\ t \geq 0\}$, has the following representations:

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

where
$A = \{A(t),\ t \geq 0\}$ is the Arrival process;
S is a generic service time;
S_e is a generic excess (residual) service.

In stationary models, where $\lambda(t) \equiv \lambda$, the offered-load $R(t)$ is the familiar $\lambda \cdot E[S]$ (or λ/μ), measured in Erlangs.

First Method: via Averaging Workload

- Estimate (say daily) sample-paths of the workload process. Then, average these over i.i.d. days.
- To estimate workload, calculate the number of customers in service (equivalently, the number of busy servers) at any time t, in a corresponding (virtual) $M_t/GI/\infty$ queue.

Difficulties:

- Must eliminate customers' waiting times. Then left to calculate the number of served customers in the virtual system.
- Must impute **service times of abandoning customers**.

Estimation of the Offered-Load

Second method: via time-varying Little

- Approximate the integral in the representation:
  $R(t) = \int_{-\infty}^{t} \lambda(u) \cdot P(S > t - u)du$,
  over all t.
- Must first to estimate the survival function of the service time, $P(S > t)$, $t \geq 0$, and the arrival rate $\lambda(t)$.

Difficulties:

- Approximating the integral.
- Estimating the survival function of service-time S for all customers (including abandoning - will be discussed momentarily).

Imputing Service Times of Abandoning Customers

In calculating the offered-load, one must account for service-times of abandoning customers.

A prevalent assumptions is that service times and (im)patience times are independent. Experience suggests that this assumption is often violated.
For example, it is not unreasonable that customers who anticipate longer service times, will be willing to wait more for service before abandoning.

Ongoing research (w/ M. Reich, Y. Ritov) develops a procedure for calculating the function $E(S|\tau = w)$:

1. Introduce $g(w) = E(S|\tau > W = w)$, which is the mean service time of those who waited exactly w units of time **and were served**. Then calculate g via the non-linear regression:

$$S_i = g(W_i) + \varepsilon_i \ ,$$

   where i indexing served customers.

2. Calculate $E(S|\tau = w)$ via the (established) relation

$$E(S|\tau = w) = g(w) - \frac{g'(w)}{h_\tau(w)} \, ,$$

   where $h_\tau(w)$ is the hazard-rate function of (im)patience, to be estimated via un-censoring.

Finally, extend the above to calculate the distribution of S, given w, which is then used to impute service-times for calculating the offered-load.