# Data-Based Service Networks:

## A Research Framework for

## Asymptotic Inference, Analysis & Control of Service Systems


Avi Mandelbaum

Technion, Haifa, Israel

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


SAMSI Workshop, August 2012

- **Overheads available at SAMSI and my Technion websites**

# Research Partners

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

# Contents

- Service Networks: Call Centers, Hospitals, Websites, $\cdots$
- Redefine the paradigm of modeling/asymptotics via Data
- **ServNets: QNets, SimNets; FNets, DNets**
- **Ultimate Goal: Data-based creation and validation of ServNets, automatically in real-time**

# Contents

- Service Networks: Call Centers, Hospitals, Websites, $\cdots$
- Redefine the paradigm of modeling/asymptotics via Data
- **ServNets: QNets, SimNets; FNets, DNets**
- **Ultimate Goal: Data-based creation and validation of ServNets, automatically in real-time**
- Why be Optimistic? Pilot at the Technion SEELab
  - Lacking but Feasible: Dynamics, Durations, Protocols
  - Simple Models at the Service of Complex Realities
    - State-Space Collapse (Queues, Waiting Times)
    - Congestion Laws (LN, Little, ImPatience, Staffing)
    - Universal Approximations: Simplifying the Asymptotic Landscape
    - Stabilizing Time-Varying Performance (Offered-Load)
- Successes: Palm/Erlang-R (ED Feedback = FNet), Palm/Erlang-A (CC Abandonment = DNet)

# Contents

- Service Networks: Call Centers, Hospitals, Websites, $\cdots$
- Redefine the paradigm of modeling/asymptotics via Data
- **ServNets: QNets, SimNets; FNets, DNets**
- **Ultimate Goal: Data-based creation and validation of ServNets, automatically in real-time**
- Why be Optimistic? Pilot at the Technion SEELab
  - Lacking but Feasible: Dynamics, Durations, Protocols
  - Simple Models at the Service of Complex Realities
    - State-Space Collapse (Queues, Waiting Times)
    - Congestion Laws (LN, Little, ImPatience, Staffing)
    - Universal Approximations: Simplifying the Asymptotic Landscape
    - Stabilizing Time-Varying Performance (Offered-Load)
- Successes: Palm/Erlang-R (ED Feedback = FNet), Palm/Erlang-A (CC Abandonment = DNet)
- Elsewhere: Process Mining (Petri Nets, BPM), Networks (Social, Biological, Complex, ...), Simulation-based, ...
- Scenic Route: Open Problems, New Directions, Uncharted Territories

# On Asymptotic Research of Queueing Systems

Queueing asymptotics has grown to become a central research theme in Operations Research and Applied Probability, beyond just queueing theory. Its claim to fame has been the deep insights that it provides into the dynamics of Queueing Networks (**QNets**), and rightly so:

- Kingman's invariance principle in conventional heavy-traffic
- Whitt's sample-path (functional) framework
- Reiman's network analysis via oblique reflection
- Bramson-Williams' framework for state-space collapse
- Laws' resource pooling
- Harrison's paradigm for asymptotic control (Wein; van Mieghem's $Gc\mu$)
- Dai's fluid-based stability
- Halfin-Whitt's (QED regime) ($\sqrt{}$-staffing for many-server queues)
- $P = NP$ : Atar's equivalence of Preemptive and Non-Preemptive SBR; Stolyar, Gurvich
- Massey-Whitt's research of time-varying queues

# 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, that positive answers must have theory and applications nurture each other, which is good, and my approach to make this good happen is by marrying theory with data.***

# Prevalent Asymptotic Approximations

**System (Data)**

**Phenomenology**

**Accuracy**

**Models** ⟷ **Approximations**

**QNets** **SimNets**

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

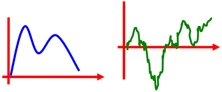
**System (Data)**

**Phenomenology**

**Accuracy**

**Models** ⟷ **Approximations**

**QNets** **SimNets** ⟷ **FNet** **DNet** **Models**



## Data–Based Prevalent Asymptotic Approximations Models

System (Data)

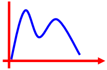
Phenomenology

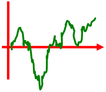
Value

Accuracy

Models ⟷ Approximations

QNets SimNets ⟷ FNet DNet Models





## Data–Based Prevalent Asymptotic Approximations Models

System (Data)

Phenomenology

Value

Accuracy

Models ⟷ Approximations

QNets SimNets ⟷ FNet DNet Models

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

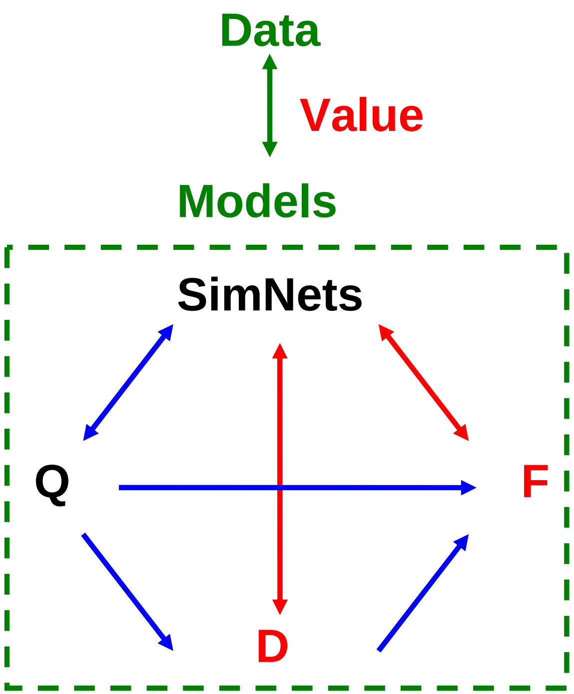
**Approx. / Models:** SLLN **(FNets)**, CLT **(DNets)** ; Laplace 1810

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

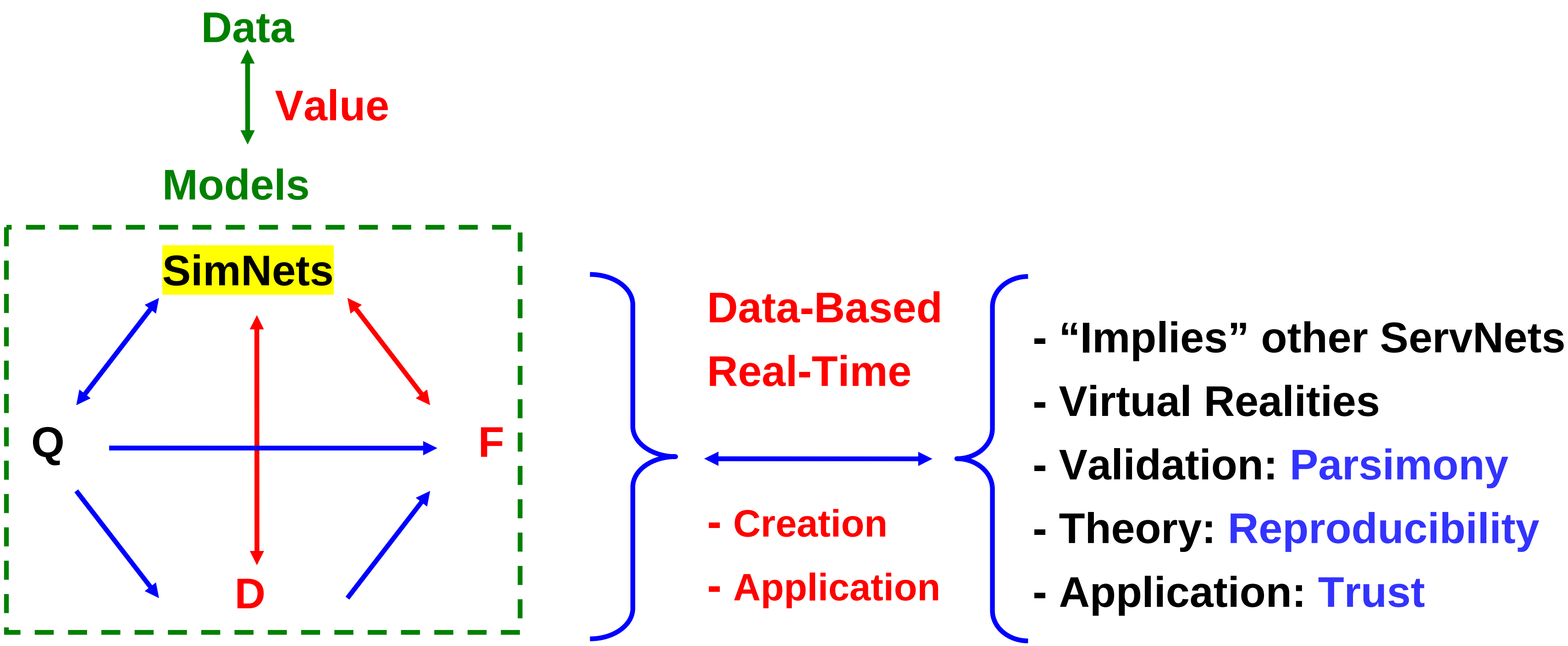
Normal, Brownian Motion ; Bachalier 1900

Poisson ; Poisson 1838

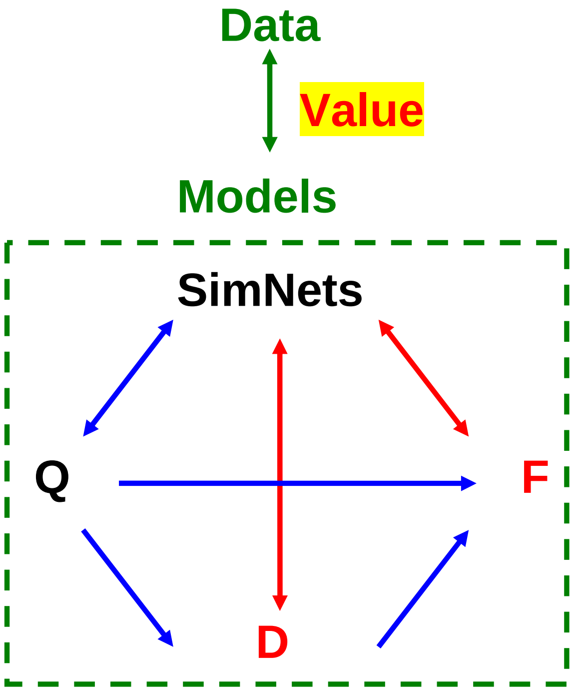
# Data–Based Framework: (Almost) All Models Born Equal



# Data–Based (Asymptotic) Framework: Simulation Mining



# Data–Based Asymptotic Framework: Added Value




Substitution Principle

Congestion Laws

Inference: Patience, Predictors

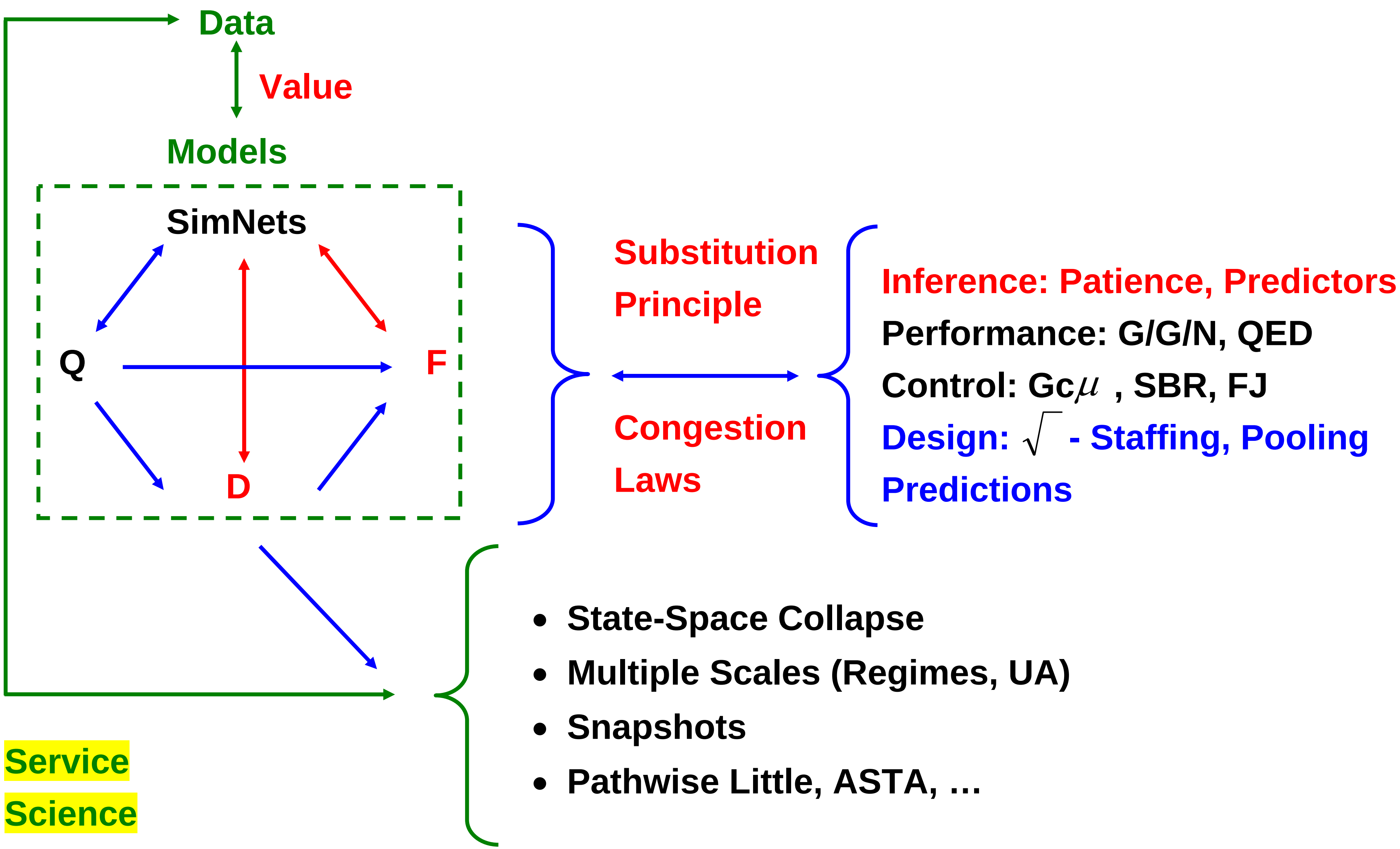
Performance: G/G/N, QED

Control: Gc$\mu$ , SBR, FJ

Design: $\sqrt{\ }$ - Staffing, Pooling

Predictions

# Ultimately: Automatic "Discovery, Conformance, Enhancement"



# Scope of the Service Industry

## Guangzhou Railway Station, Southern China

# Call Centers, Then Hospitals, Now Internet

## Call Centers - U.S. Stat.

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

# Call Centers, Then Hospitals, Now Internet

**Call Centers - U.S. Stat.**

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

**Healthcare** - similar, plus unique challenges:

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

**Internet -** . . .

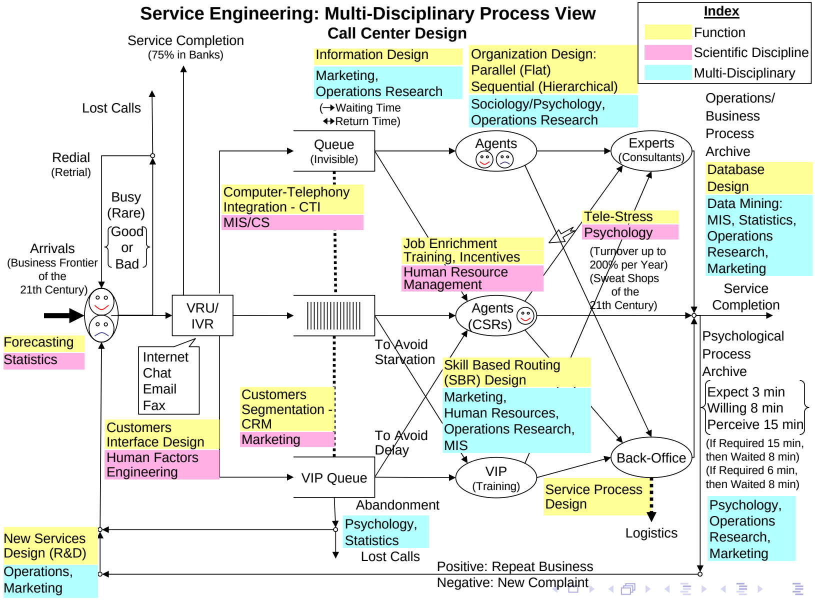
# Call-Center Environment: Service Network

# Operational Focus

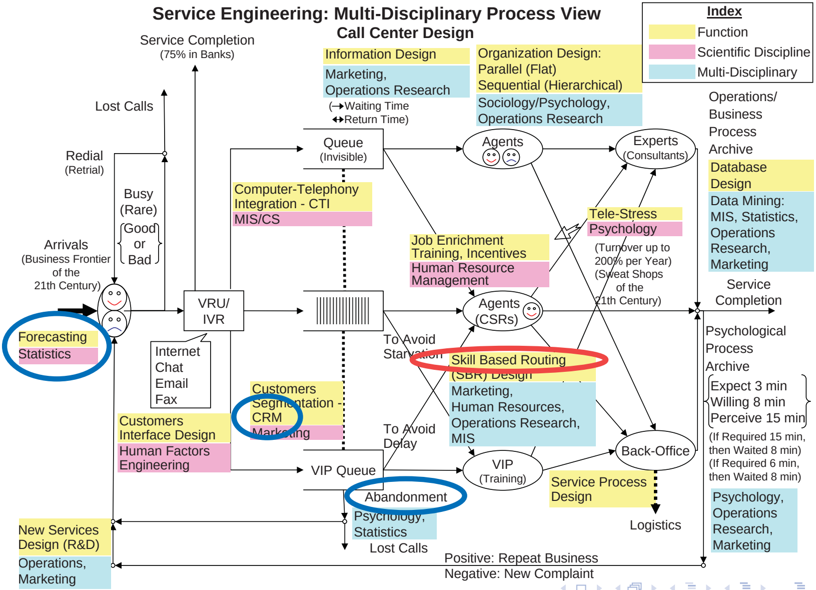
## Operational Measures:

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

# Call-Center Network: Gallery of Models



# Call-Center Network: Gallery of Models



# Skills-Based Routing in Call Centers

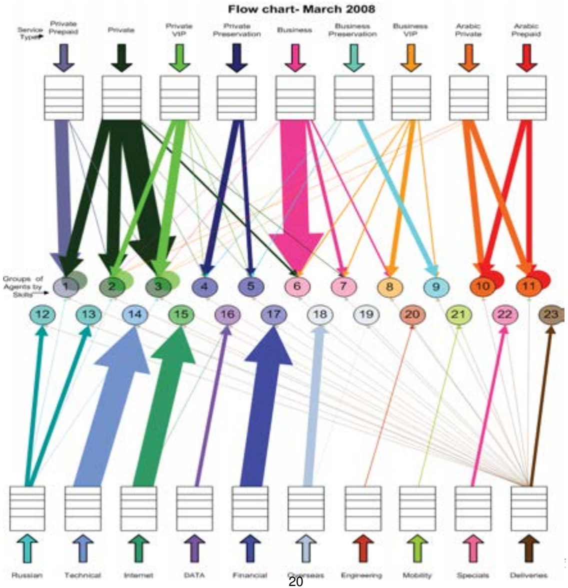
## EDA and OR, with I. Gurvich and P. Liberman

**Mktg.** $\Rightarrow$

**OR** $\Rightarrow$

**HRM** $\Rightarrow$

**MIS** $\Rightarrow$



# ER / ED Environment: Service Network

## Acute (Internal, Trauma)

## Walking

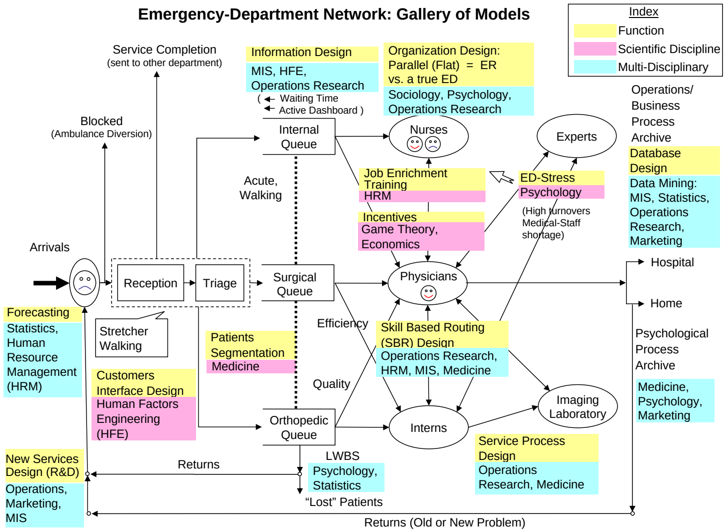
## Multi-Trauma

# ED-Environment in Israel

# Queueing in a “Good" Hospital

## Tong-ren Hospital at 6am, Beijing

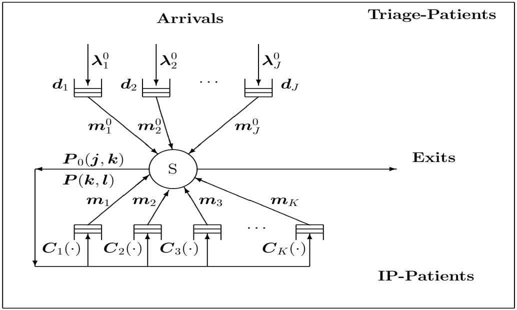
# Emergency-Department Network: Gallery of Models




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

# ED Patient Flow: The Physicians View

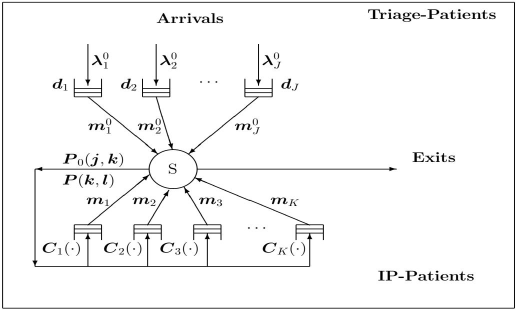
with **J. Huang, B. Carmeli**




- **Goal**: Adhere to **Triage-Constraints**, then **release In-Process** Patients
- **Model** = Multi-class Q with Feedback: Min. convex **congestion costs** of IP-Patients, s.t. **deadline constraints** on Triage-Patients.

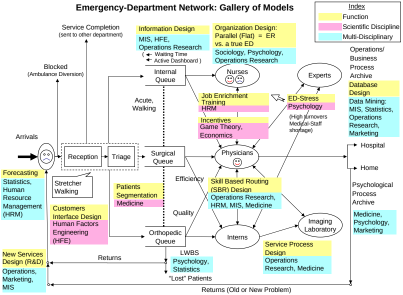
# ED Patient Flow: The Physicians View

with **J. Huang, B. Carmeli**




- **Goal**: Adhere to **Triage-Constraints**, then **release In-Process** Patients
- **Model** = Multi-class Q with Feedback: Min. convex **congestion costs** of IP-Patients, s.t. **deadline constraints** on Triage-Patients.
- **Solution**: In **conventional** heavy-traffic, **asymptotic least-cost** s.t. **asymptotic compliance** (as in Plambeck, Harrison, Kumar, who applied admission control):
  - Triage or IP? former, if some deadline is “too" close
  - Triage Priorities: Chose the closest deadline
  - IP-Priorities: $Gc\mu$, modified (simply) to account for feedback

# Emergency-Department Network: Flow Control




- *Queueing-Science, w/ **Armony, Marmor, Tseytlin, Yom-Tov**
- *Fair ED-to-IW Routing (Patients vs. Staff), w/ **Momcilovic, Tseytlin**
- *Triage vs. InProcess/Release (Plambeck et al, van Mieghem) in EDs, w/ **Carmeli, Huang; Shimkin**
- *Staffing Time-Varying Q's with Re-Entrant Customers (de Vericourt & Jennings), w/ **Yom-Tov**
- The Offered-Load in Fork-Join Nets (Adlakha & Kulkarni), w/ **Kaspi, Zaeid**
- Synchronization Control of Fork-Join Nets, w/ **Atar, Zviran**

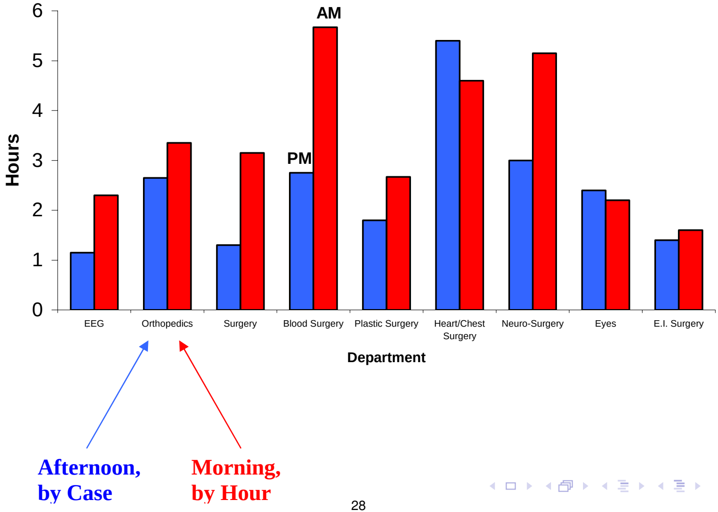
# Prerequisite I: Data

**Averages Prevalent** (and could be useful / interesting).

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

# Interesting Averages: The Human Factor, or Even "Doctors" Can Manage

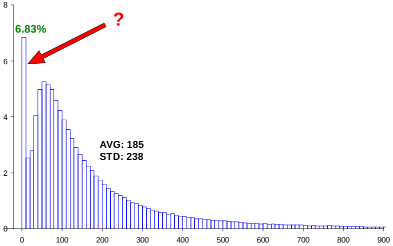
**Operations Time - Morning (by Hour) vs. Afternoon (by Case):**



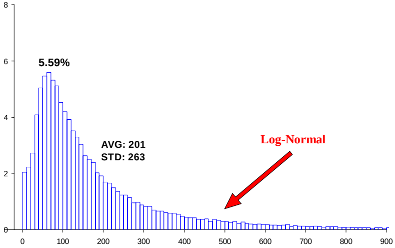
# Beyond Averages: The Human Factor

## Histogram of Service-Time in an Israeli Call Center, 1999

### January-October




### November-December

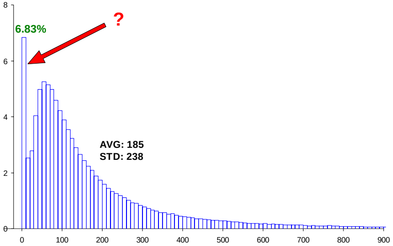



- 6.8% Short-Services:

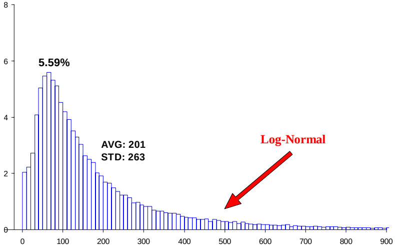
# Beyond Averages: The Human Factor

## Histogram of Service-Time in an Israeli Call Center, 1999

### January-October




### November-December




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

## Pause for a Commercial:

# Pause for a Commercial: The Technion SEE Center

# Technion SEE = Service Enterprise Engineering

## SEELab: Data-repositories for research and teaching

- For example:
  - Bank Anonymous: **1 year, 350K calls by 15 agents** - in 2000. **Brown, Gans, Sakov, Shen, Zeltyn, Zhao** (JASA), paved the way to:
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**
  - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**
  - Home (Rambam) Hospital: **4 years, 1000 beds**, ward-level flow
  - 5 EDs: **gathered by the late David Sinreich**, ED arrivals & LOS

# Technion SEE = Service Enterprise Engineering

## SEELab: Data-repositories for research and teaching

- For example:
  - Bank Anonymous: **1 year, 350K calls by 15 agents** - in 2000. **Brown, Gans, Sakov, Shen, Zeltyn, Zhao** (JASA), paved the way to:
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**
  - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**
  - Home (Rambam) Hospital: **4 years, 1000 beds**, ward-level flow
  - 5 EDs: **gathered by the late David Sinreich**, ED arrivals & LOS

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

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

# Technion SEE = Service Enterprise Engineering

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

- For example:
  - Bank Anonymous: **1 year, 350K calls by 15 agents** - in 2000.
    **Brown, Gans, Sakov, Shen, Zeltyn, Zhao** (JASA),
    paved the way to:
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents**
  - Israeli Bank: **from January 2010, daily-deposit at a SEESafe**
  - Home (Rambam) Hospital: **4 years, 1000 beds**, ward-level flow
  - 5 EDs: **gathered by the late David Sinreich**, ED arrivals & LOS

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

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

**SEEServer**: **Free for academic use**

- Register
- Access **U.S. Bank, Bank Anonymous, Home Hospital**

# 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)

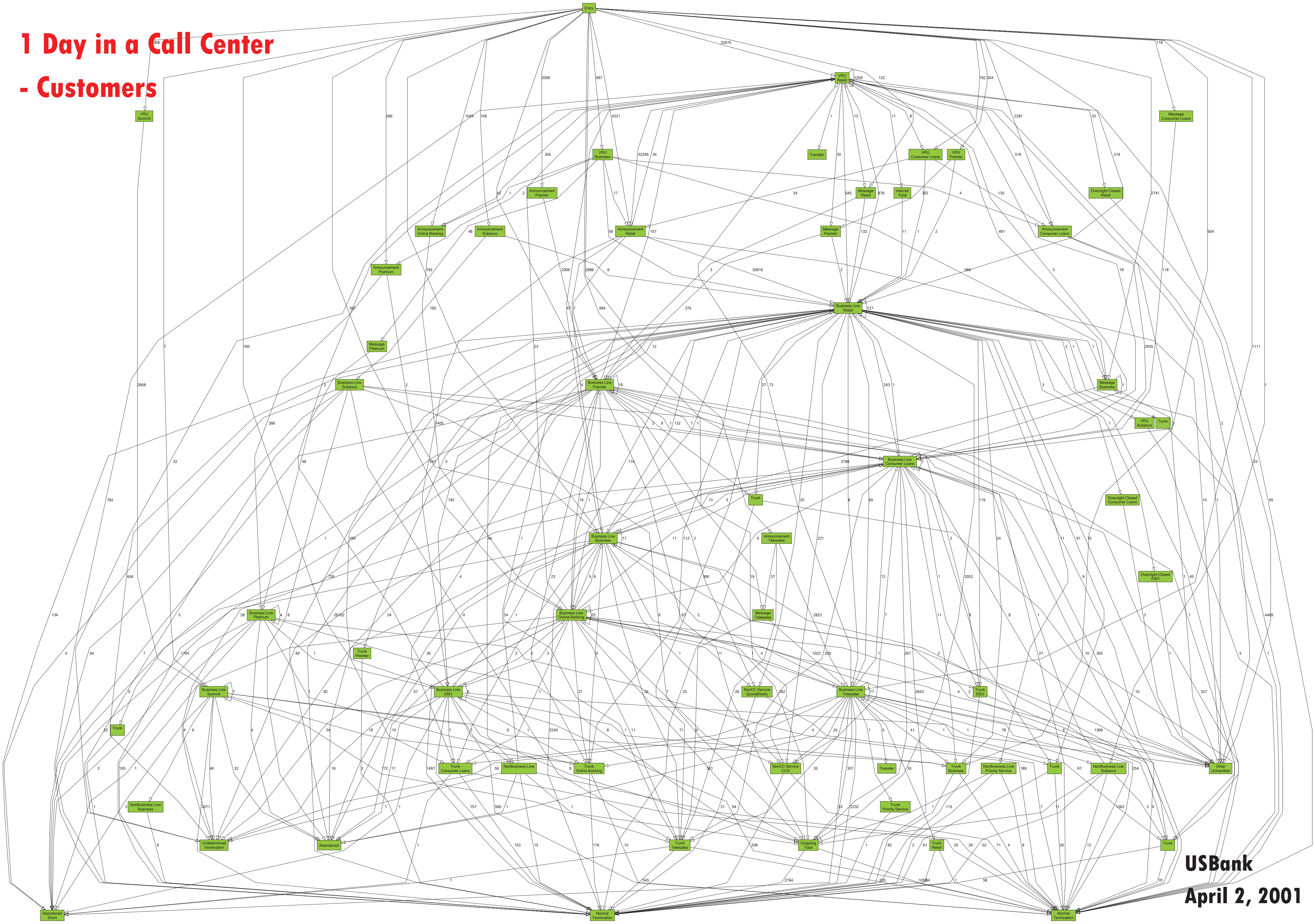
# Event-Logs in a Call Center (Bank Anonymous)

**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

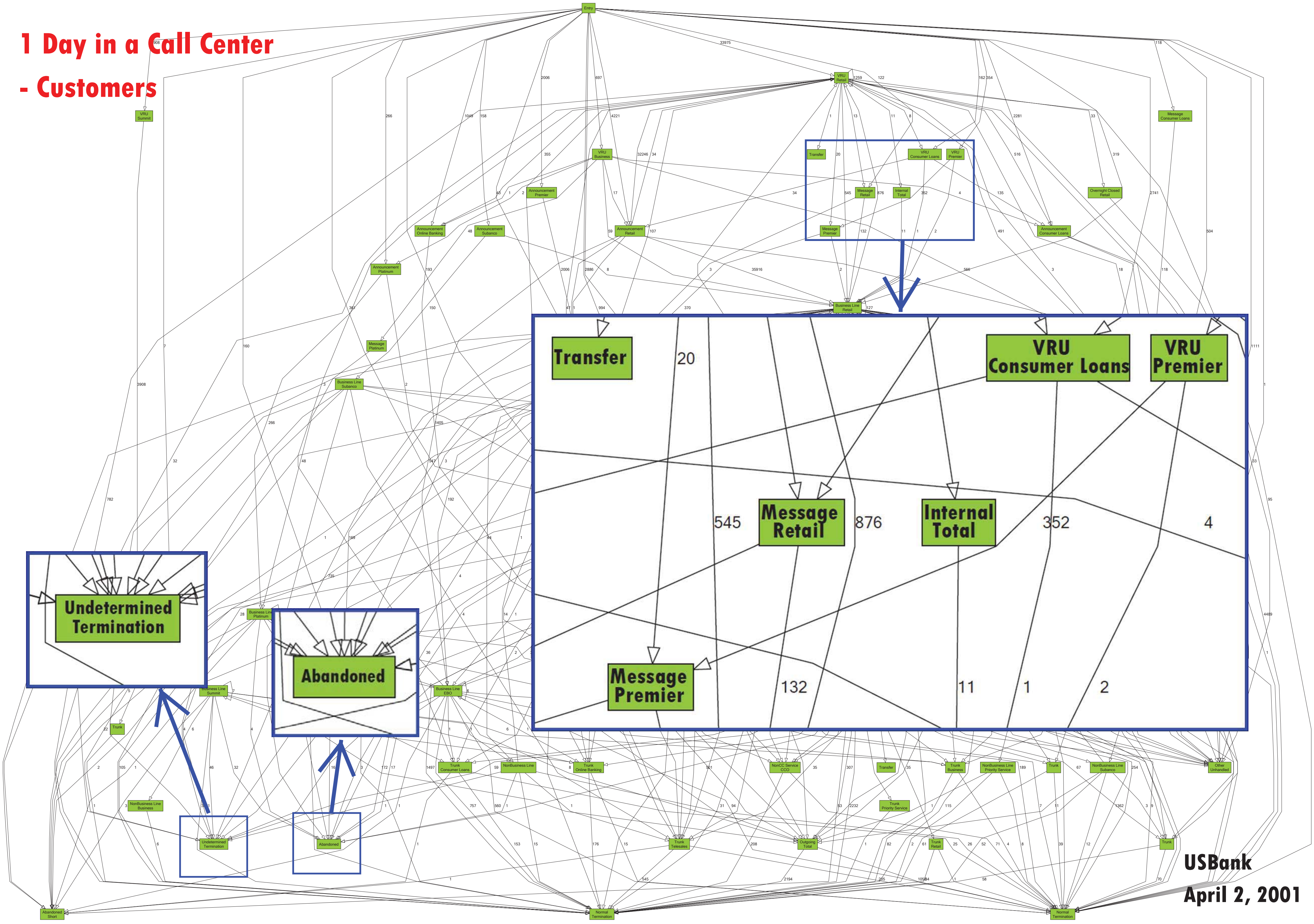
# 1 Day in a Call Center - Customers




USBank
April 2, 2001

# 1 Day in a Call Center

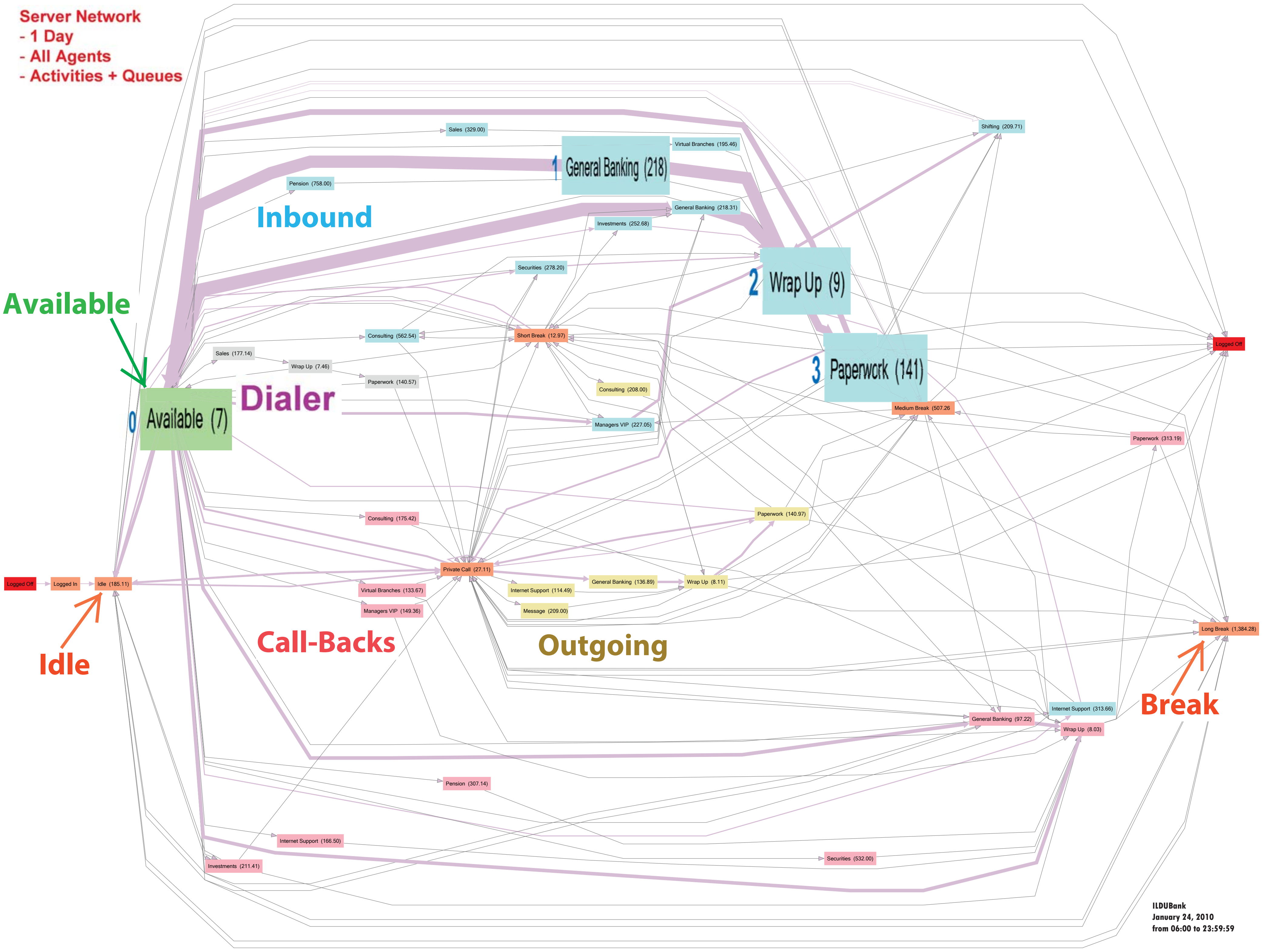
# - Customers




**USBank**

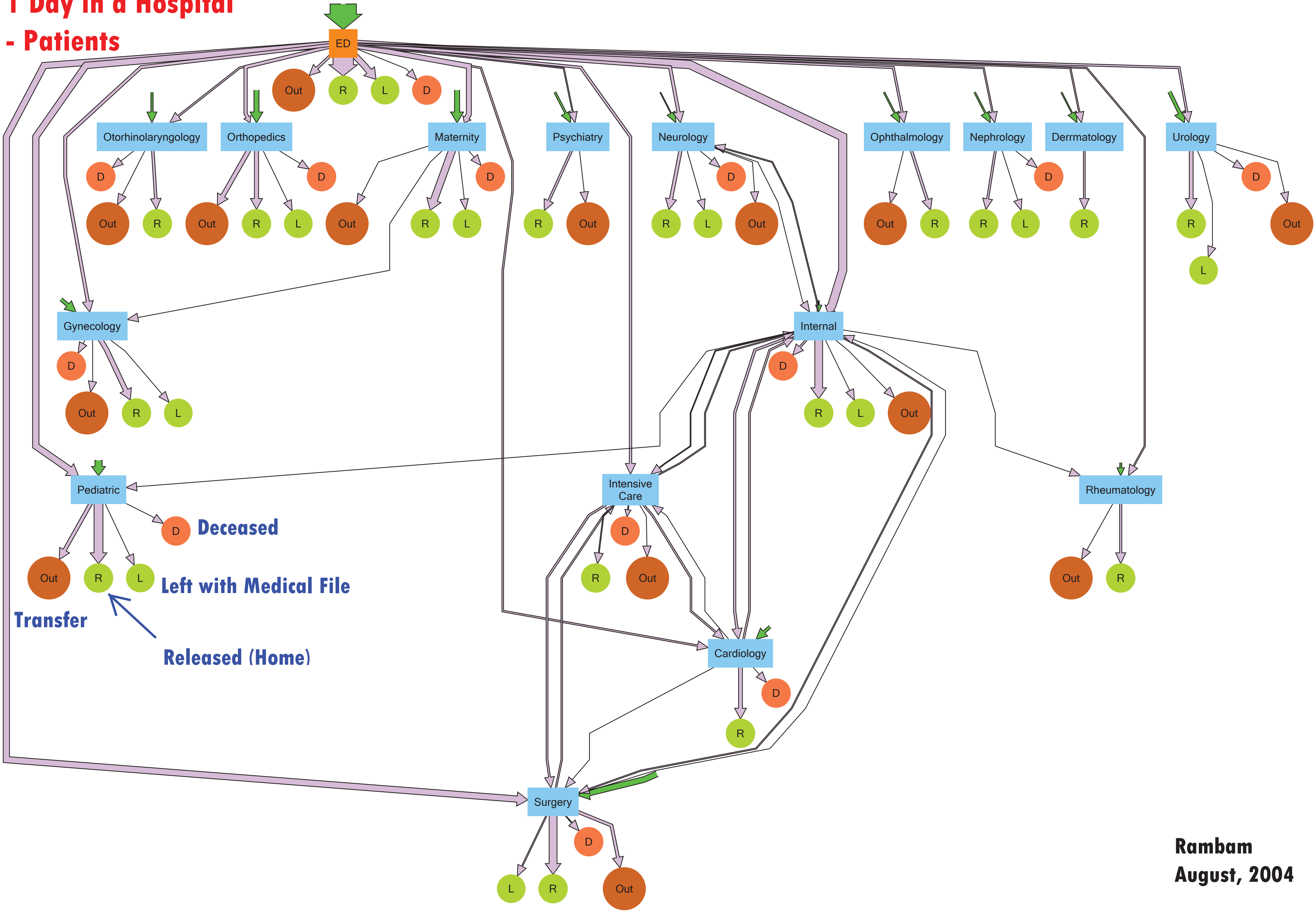
**April 2, 2001**

Server Network
- 1 Day
- All Agents
- Activities + Queues
Inbound
Available
Dialer
Idle
Call-Backs
Outgoing
Break
Logged Off
Logged In
Idle (185.11)
Available (7.36)
Sales (329.00)
Virtual Branches (195.46)
Pension (758.00)
General Banking (218.31)
Investments (252.68)
Wrap Up (8.71)
Securities (278.20)
Shifting (209.71)
Consulting (562.54)
Short Break (12.97)
Paperwork (140.62)
Sales (177.14)
Wrap Up (7.46)
Paperwork (140.57)
Consulting (208.00)
Managers VIP (227.05)
Medium Break (507.26
Logged Off
Paperwork (313.19)
Paperwork (140.97)
Consulting (175.42)
Private Call (27.11)
General Banking (136.89)
Wrap Up (8.11)
Virtual Branches (133.67)
Internet Support (114.49)
Managers VIP (149.36)
Message (209.00)
Long Break (1,384.28)
Internet Support (313.66)
General Banking (97.22)
Wrap Up (8.03)
Pension (307.14)
Internet Support (166.50)
Securities (532.00)
Investments (211.41)
ILDUBank
January 24, 2010
from 06:00 to 23:59:59



Server Network
- 1 Day
- All Agents
- Activities + Queues
Inbound
Available
Dialer
Idle
Call-Backs
Outgoing
Break
0 Available (7)
1 General Banking (218)
2 Wrap Up (9)
3 Paperwork (141)
Sales (329.00)
Virtual Branches (195.46)
Pension (758.00)
General Banking (218.31)
Investments (252.68)
Securities (278.20)
Shifting (209.71)
Consulting (562.54)
Short Break (12.97)
Sales (177.14)
Wrap Up (7.46)
Paperwork (140.57)
Consulting (208.00)
Managers VIP (227.05)
Medium Break (507.26
Paperwork (313.19)
Logged Off
Paperwork (140.97)
Consulting (175.42)
Private Call (27.11)
General Banking (136.89)
Wrap Up (8.11)
Internet Support (114.49)
Message (209.00)
Logged Off
Logged In
Idle (185.11)
Virtual Branches (133.67)
Managers VIP (149.36)
Long Break (1,384.28)
Internet Support (313.66)
General Banking (97.22)
Wrap Up (8.03)
Pension (307.14)
Internet Support (166.50)
Securities (532.00)
Investments (211.41)
ILDUBank
January 24, 2010
from 06:00 to 23:59:59

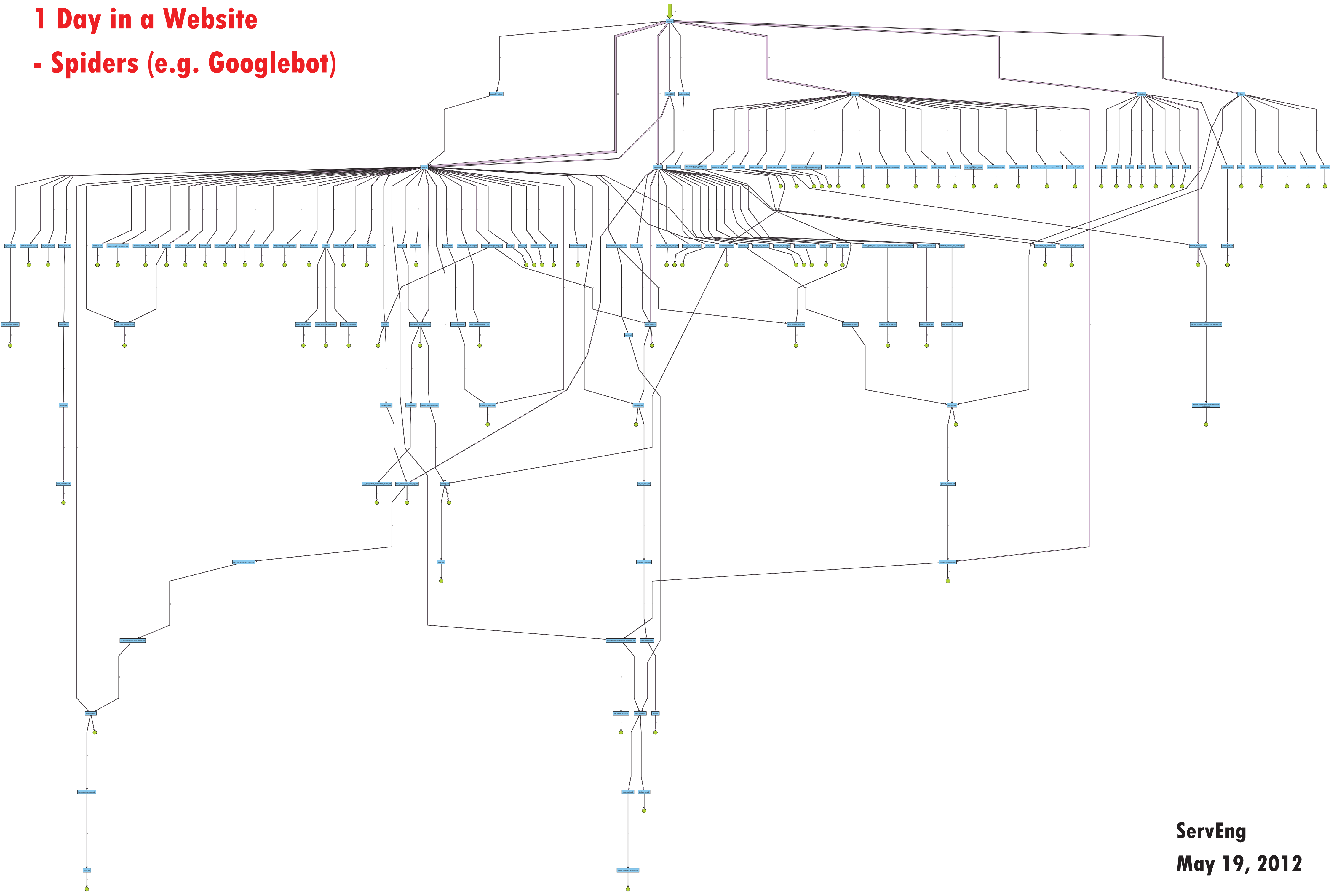
# 1 Day in a Hospital
# - Patients




Rambam
August, 2004

# 1 Day in a Website

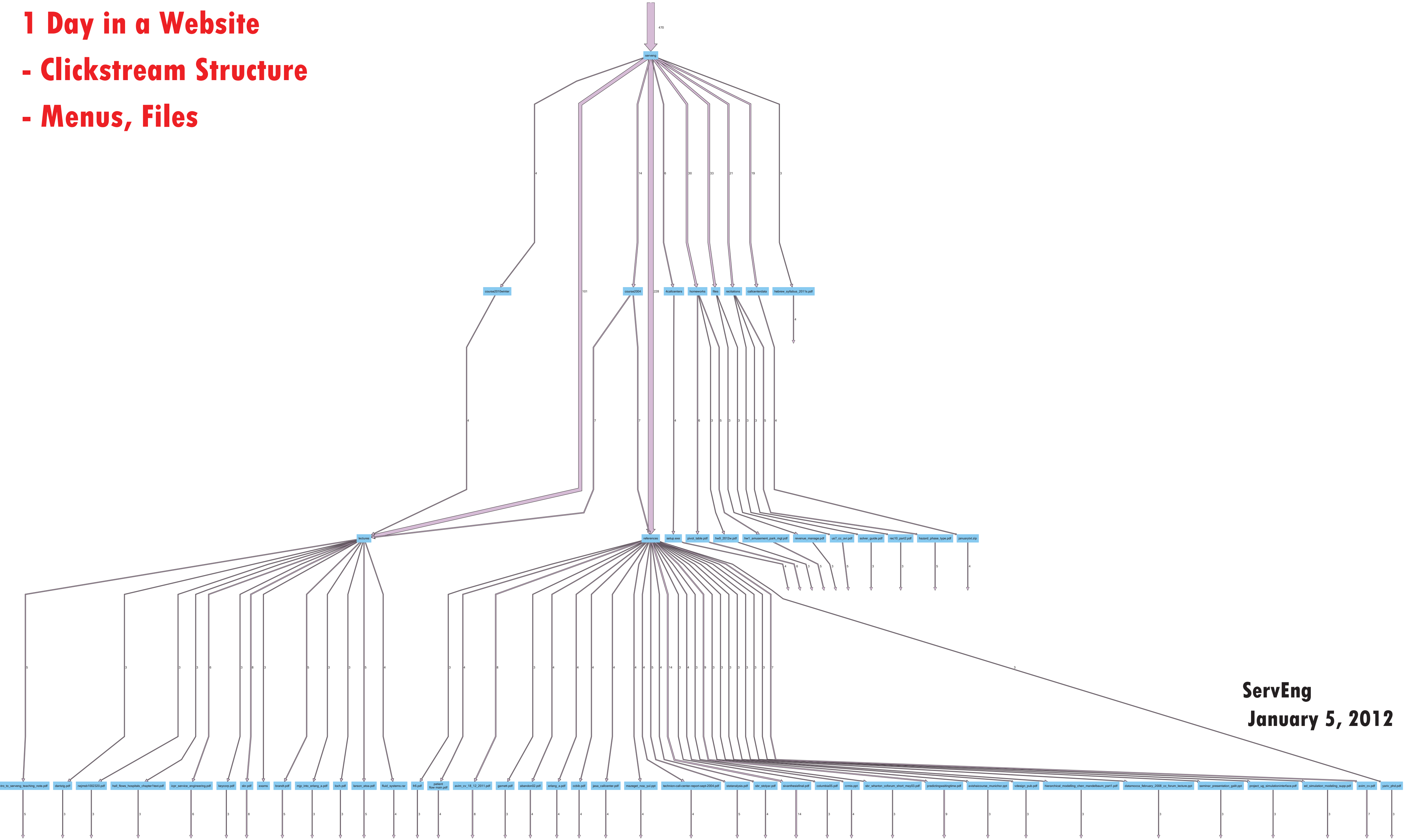
## - Spiders (e.g. Googlebot)



**ServEng**

**May 19, 2012**

# 1 Day in a Website
# - Clickstream Structure
# - Menus, Files

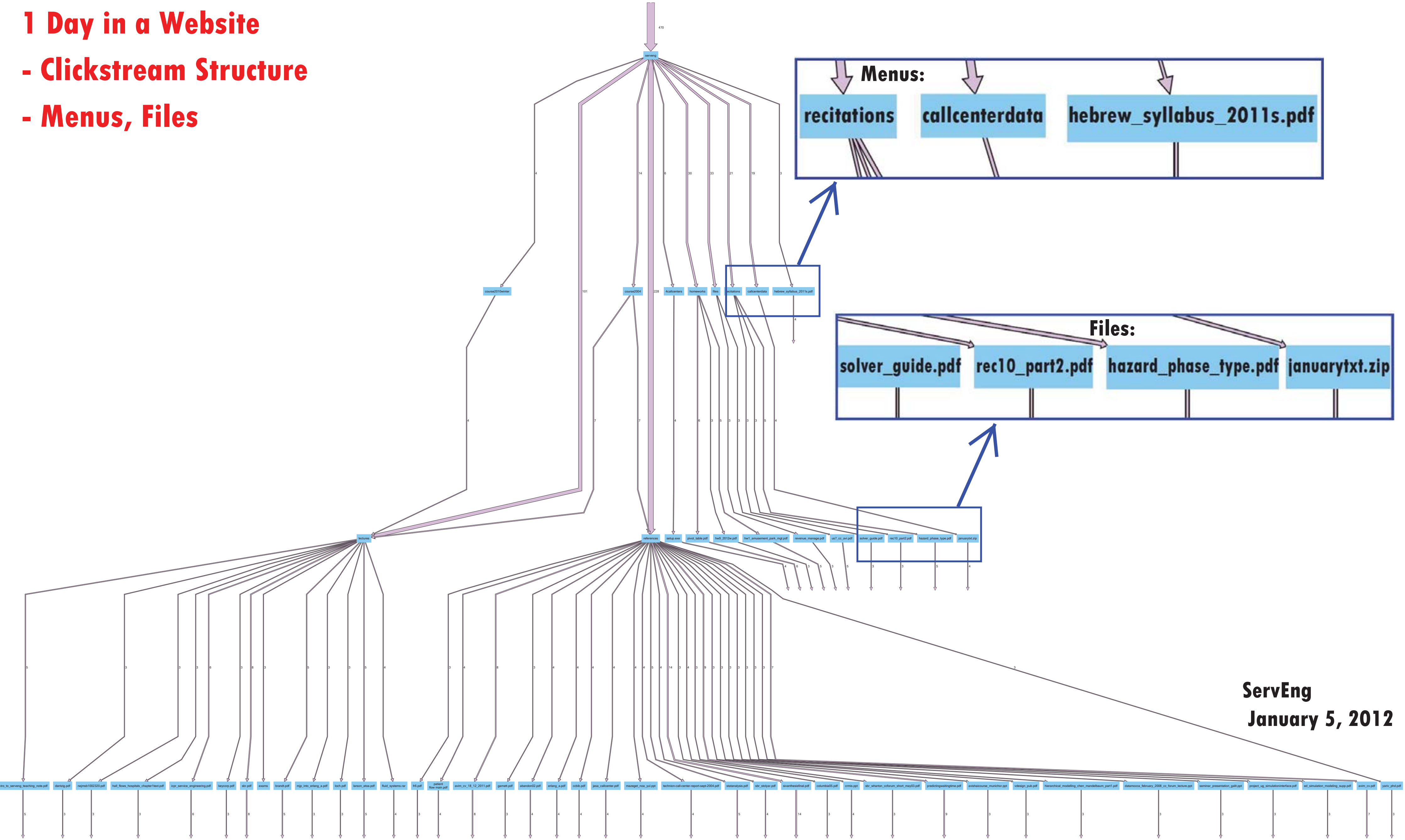


**ServEng**
**January 5, 2012**

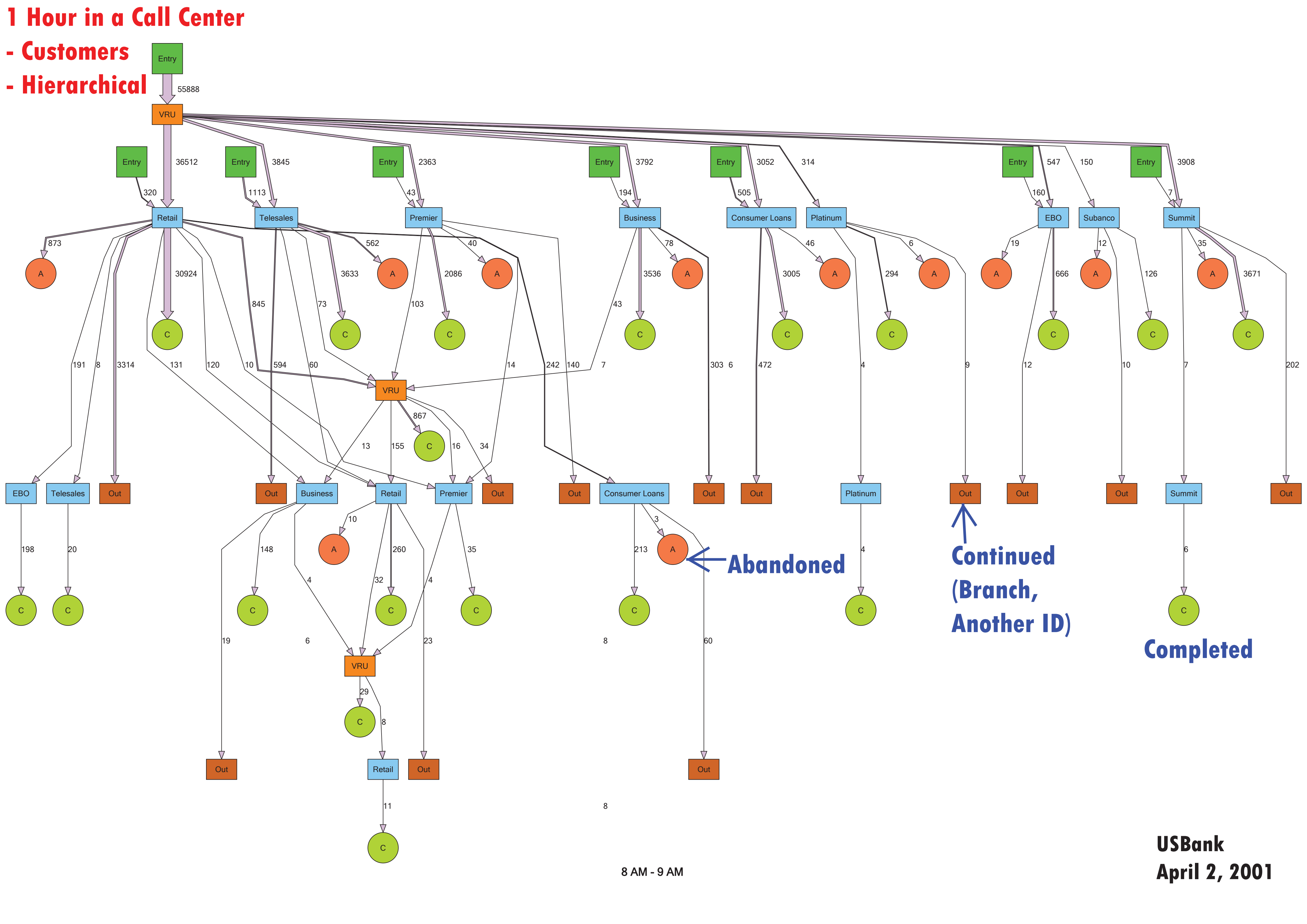
# 1 Day in a Website

# - Clickstream Structure

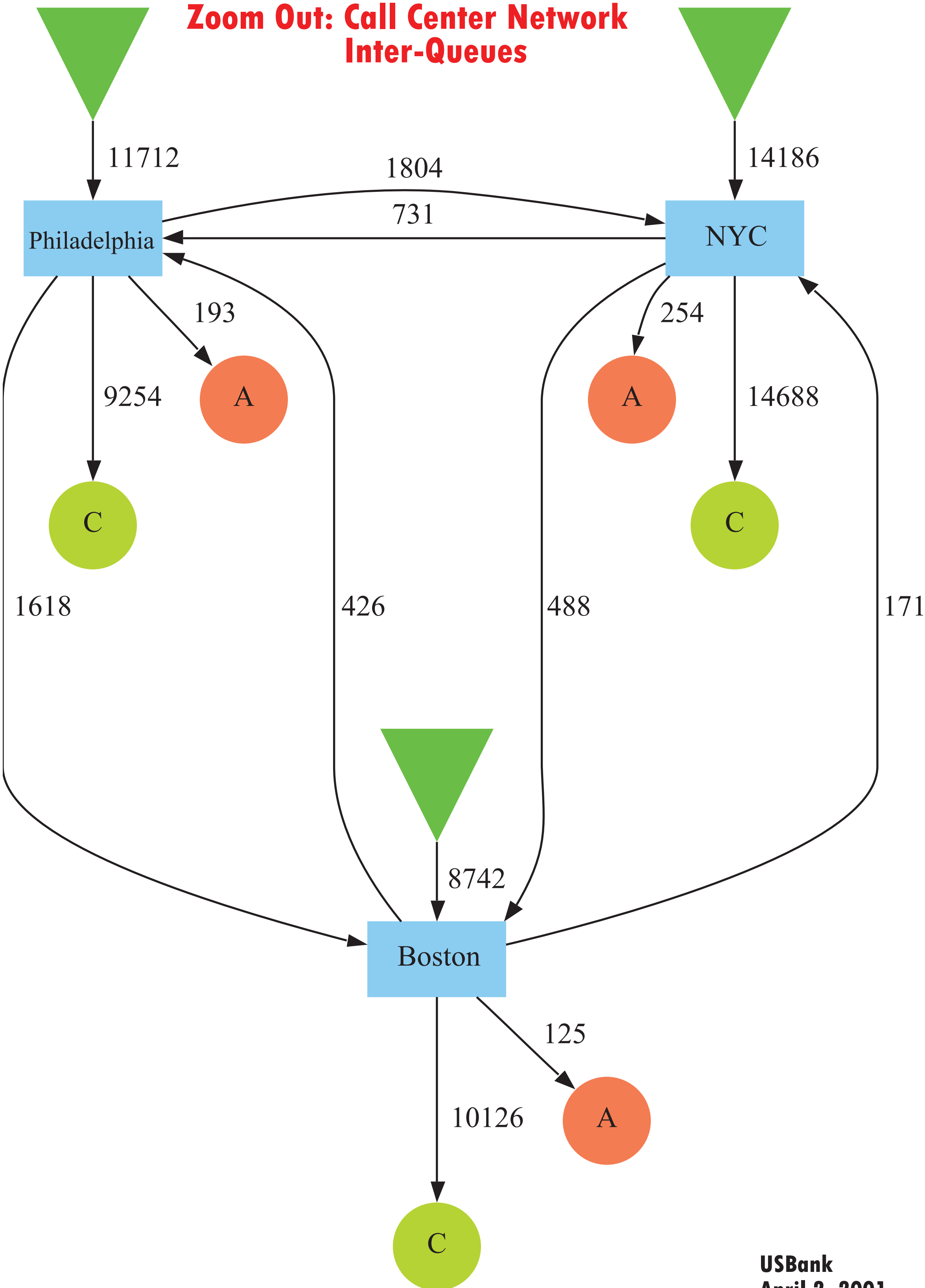
# - Menus, Files




**ServEng**
**January 5, 2012**

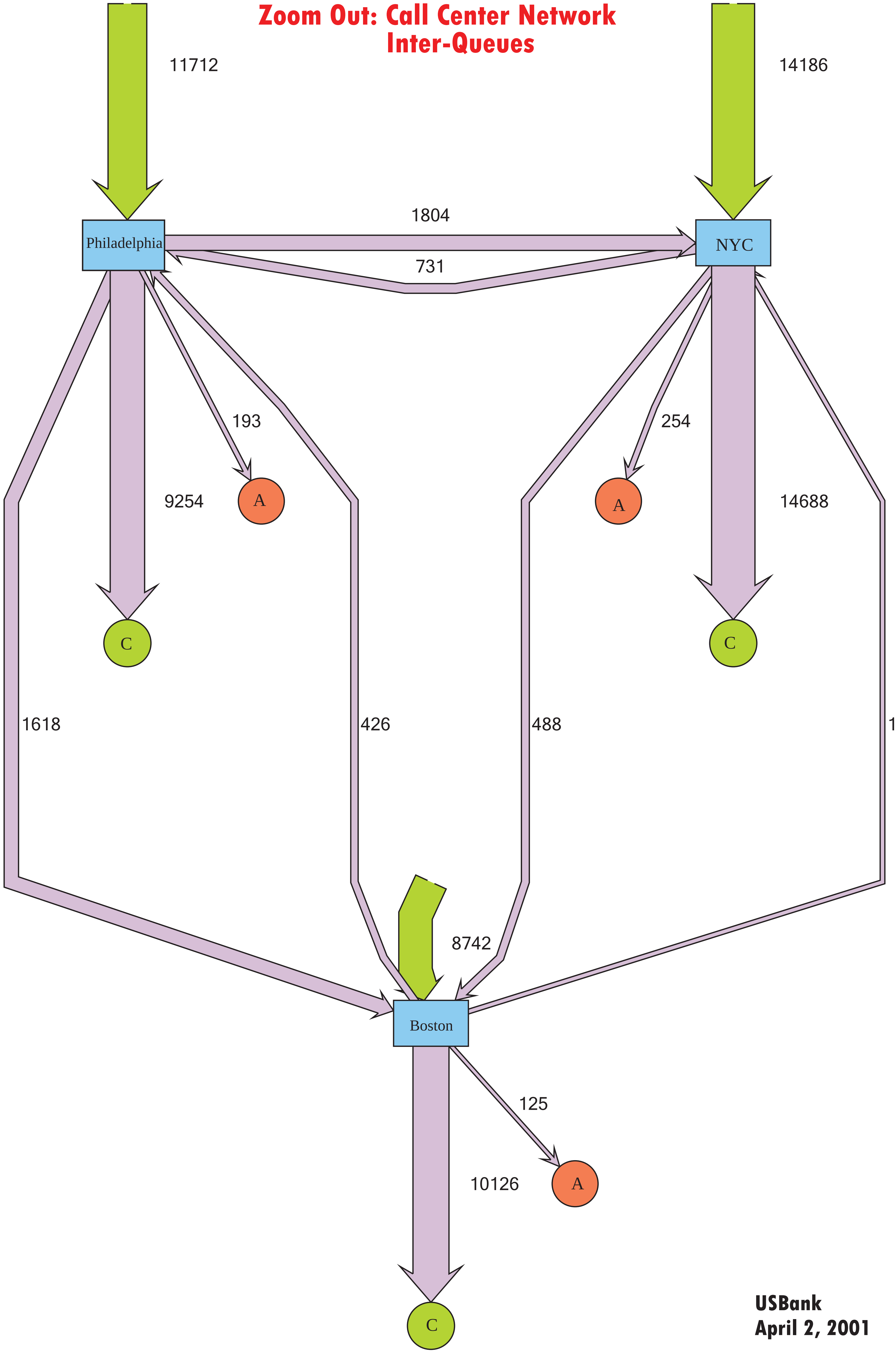

1 Hour in a Call Center
- Customers
- Hierarchical
Abandoned
Continued (Branch, Another ID)
Completed
8 AM - 9 AM
USBank
April 2, 2001

## Zoom Out: Call Center Network
## Inter-Queues



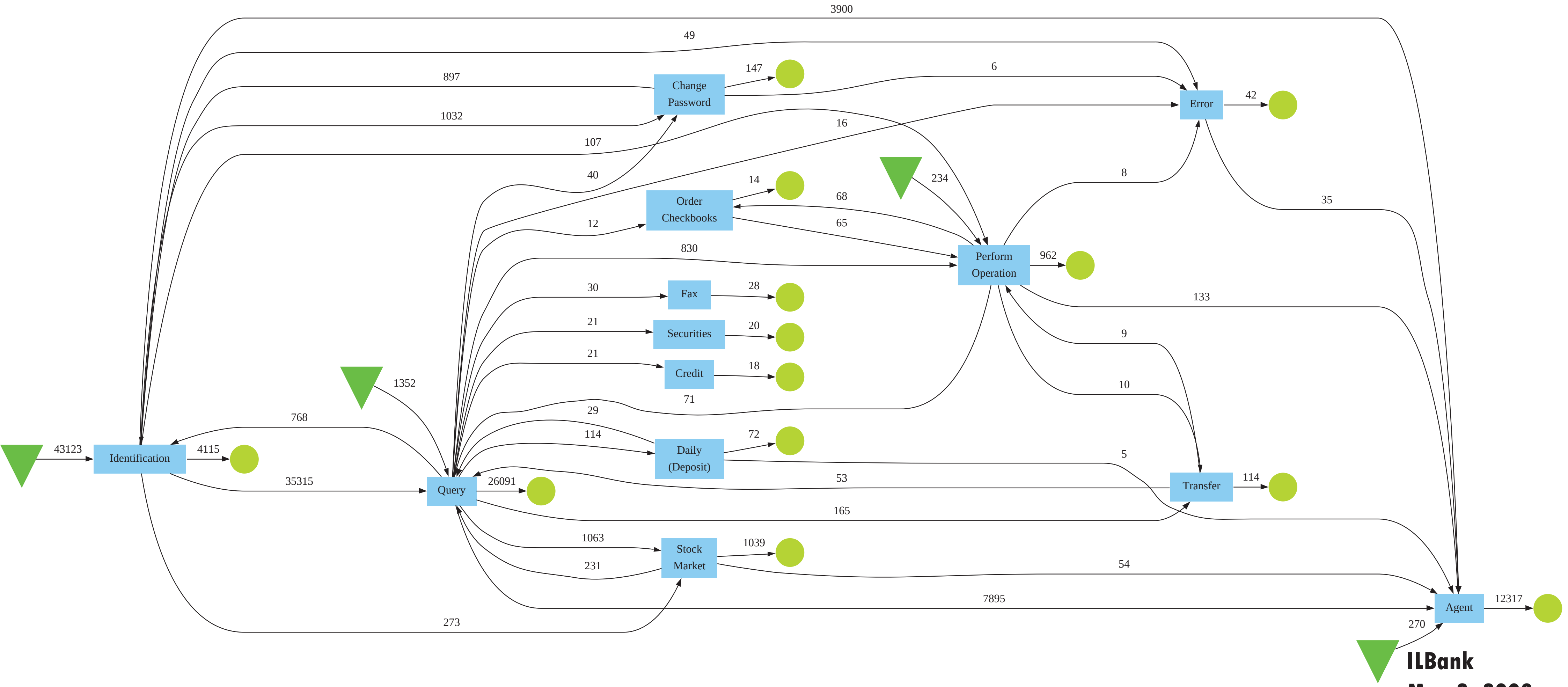

USBank
April 2, 2001

# Zoom Out: Call Center Network Inter-Queues

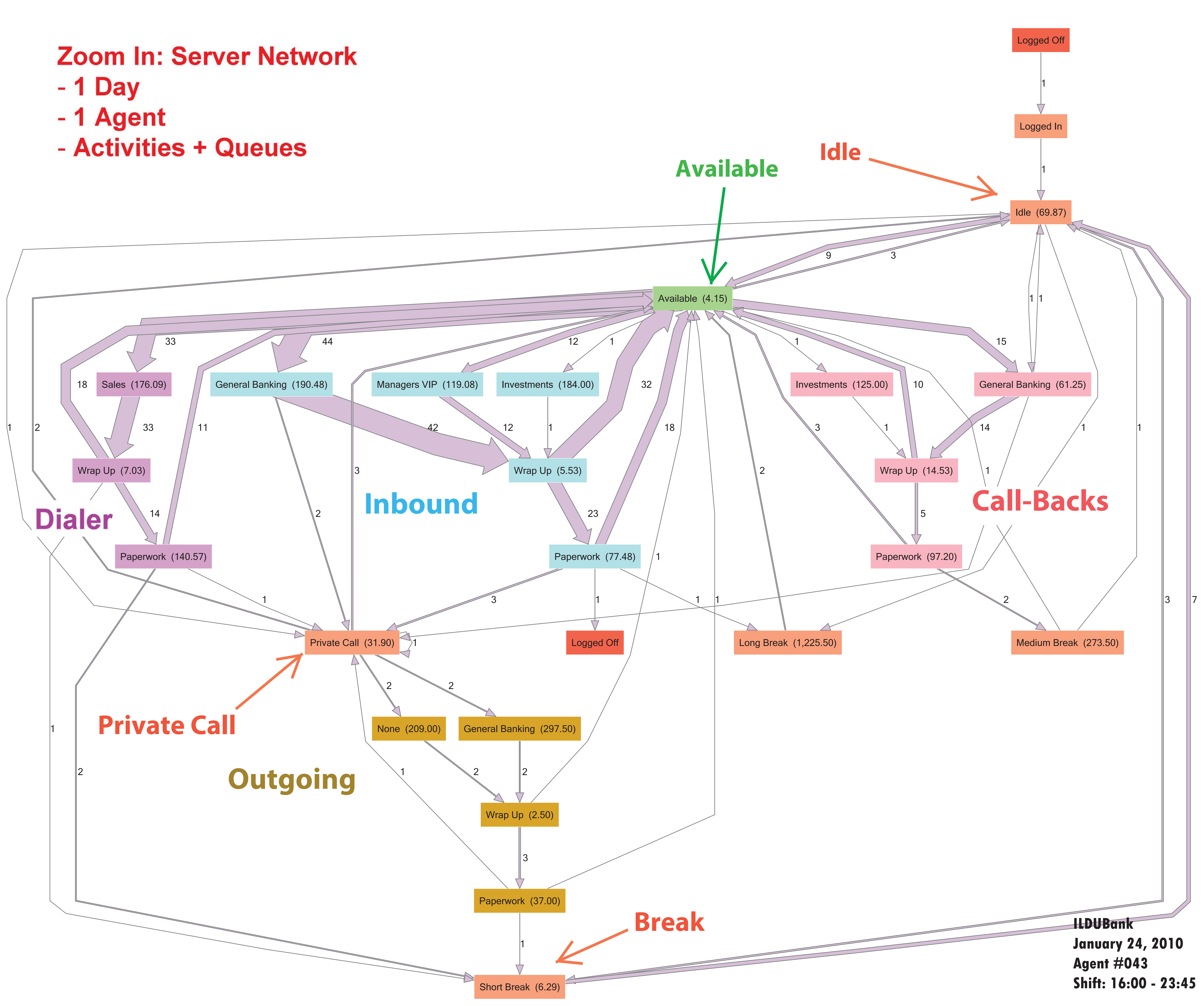



USBank
April 2, 2001

# Zoom In: Interactive Voice Response (IVR)



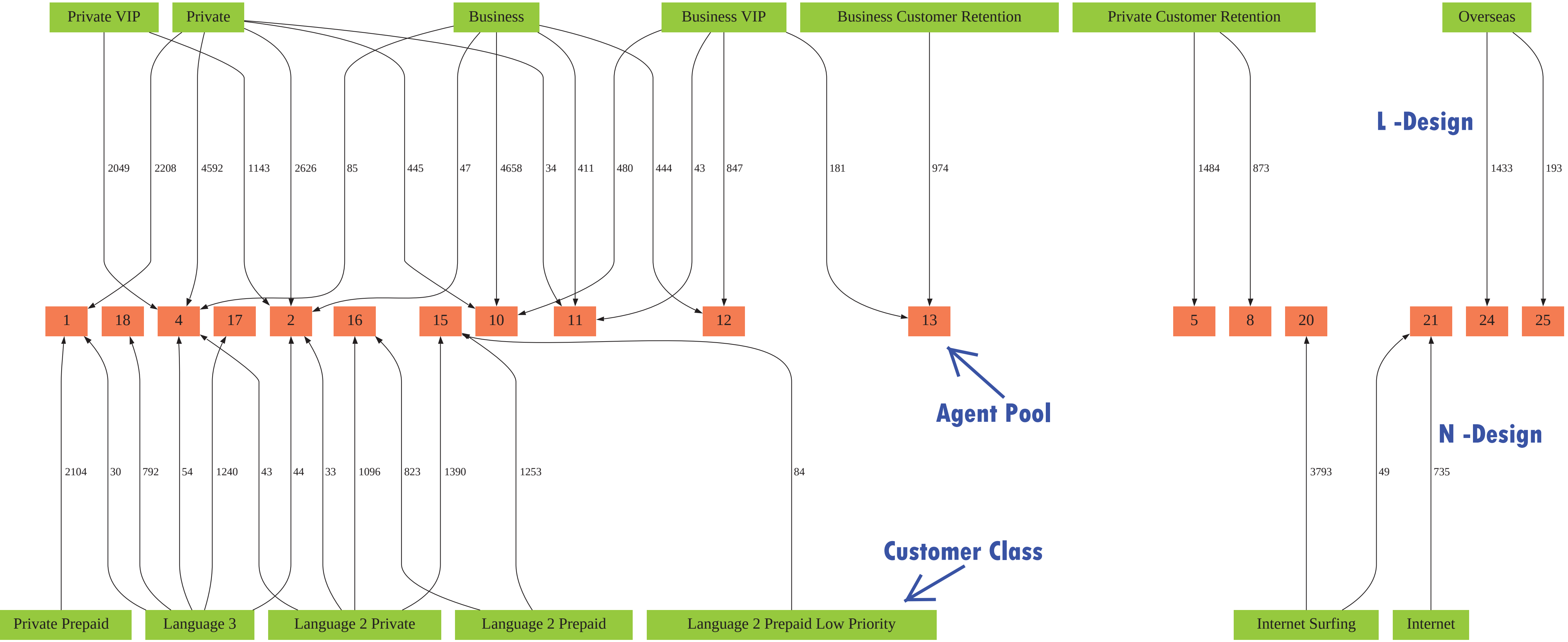

**ILBank**
**May 2, 2008**



Zoom In: Server Network
- 1 Day
- 1 Agent
- Activities + Queues
Logged Off
Logged In
Idle
Idle (69.87)
Available
Available (4.15)
Sales (176.09)
General Banking (190.48)
Managers VIP (119.08)
Investments (184.00)
Investments (125.00)
General Banking (61.25)
Wrap Up (7.03)
Wrap Up (5.53)
Wrap Up (14.53)
Dialer
Inbound
Call-Backs
Paperwork (140.57)
Paperwork (77.48)
Paperwork (97.20)
Private Call (31.90)
Logged Off
Long Break (1,225.50)
Medium Break (273.50)
None (209.00)
General Banking (297.50)
Private Call
Outgoing
Wrap Up (2.50)
Paperwork (37.00)
Break
Short Break (6.29)
ILDUBank
January 24, 2010
Agent #043
Shift: 16:00 - 23:45

# Zoom In: Skills - Based Routing (SBR)

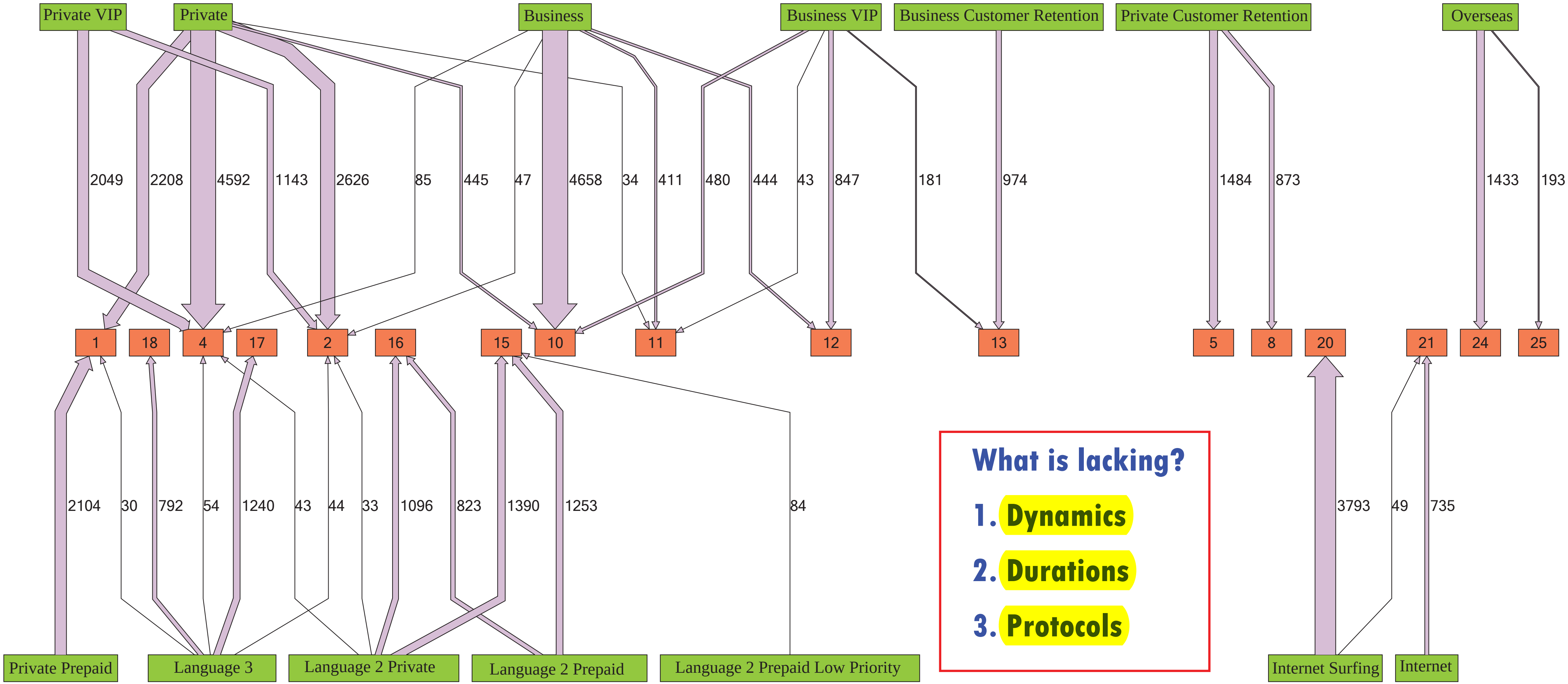
## Network Structure (Protocol)




**Connected Component**

**Israeli Telecom**
**February 10, 2008**

# Goal: Data-Based Real-Time Simulation (SimNet)




**What is lacking?**

1. Dynamics
2. Durations
3. Protocols

**Israeli Telecom**
**February 10, 2008**

# 4 Hours in a Call Center - Dynamics


USBank
April 2, 2001

8 AM - 9 AM


Entry
VRU
Entry
Entry
Entry
Entry
Entry
Entry
Entry
Subanco
Premier
Telesales
Business
Retail
Summit
Consumer Loans
EBO
Platinum
A
A
A
A
A
A
A
A
A
C
C
C
C
C
C
C
C
C
C
VRU
C
Out
Out
Out
Premier
Retail
Out
Out
Business
Consumer Loans
EBO
Telesales
Out
Summit
Out
Out
Out
Out
Platinum
A
A
C
C
C
C
C
C
C
C
VRU
C
Out
Retail
Out
Out
C

8 AM - 9 AM


Entry
VRU
Entry
Entry
Entry
Entry
Entry
Entry
Entry
Subanco
Premier
Telesales
Business
Retail
Summit
Consumer Loans
EBO
Platinum
A
A
A
A
A
A
A
A
A
C
C
C
C
C
C
C
C
C
VRU
C
Out
Out
Out
Premier
Retail
Out
Business
Consumer Loans
EBO
Telesales
Out
Summit
Out
Out
Out
Out
Platinum
A
A
C
C
C
C
C
C
C
C
VRU
C
Out
Retail
Out
Out
Business
C
C

9 AM - 10 AM


Entry
VRU
Subanco
Premier
Telesales
Business
Retail
Summit
Consumer Loans
EBO
Platinum
A
C
Out

10 AM - 11 AM

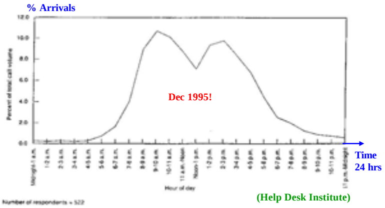

Entry
VRU
Subanco
Premier
Telesales
Business
Retail
Summit
Consumer Loans
EBO
Platinum
A
C
Out

11 AM - 12 PM

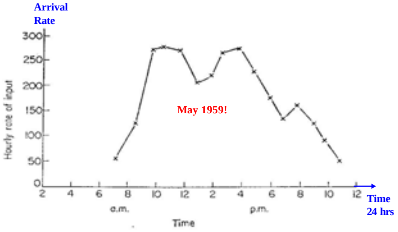
# Dynamics: Time-Varying Arrival-Rates

## 2 Daily Peaks

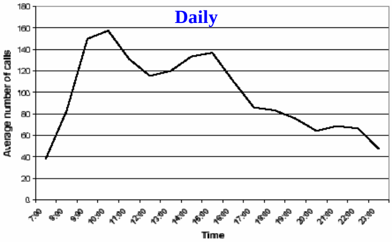
**CC**: Dec. **1995**, (USA, 700 Helpdesks)



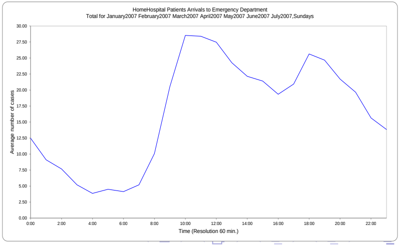

**CC**: May **1959** (England)




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

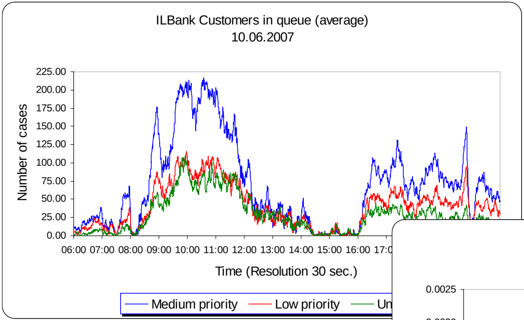



**ED**: Jan.–July 2007 (Israel)

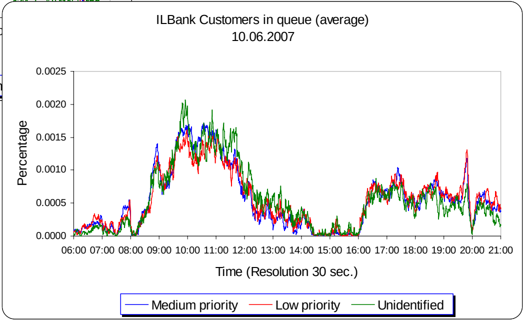


# Dynamics: Parsimonious Models (Congestion Laws)

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




## Queue "Shape"

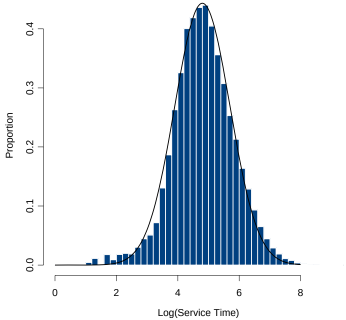



- Area normalized to 100%
- **State-Space Collapse**

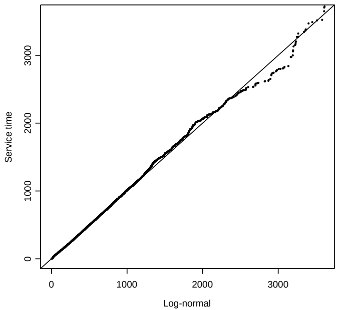
# Durations: Phone Calls (2 Surprises)

## Israeli Call Center, Nov–Dec, 1999

### Log(Service Times)



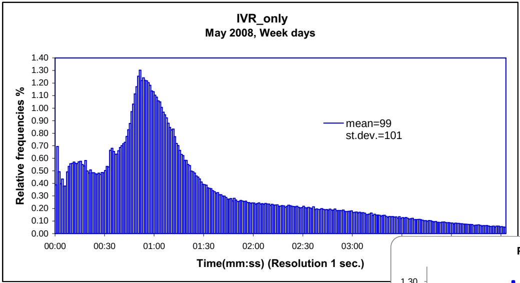

### LogNormal QQPlot



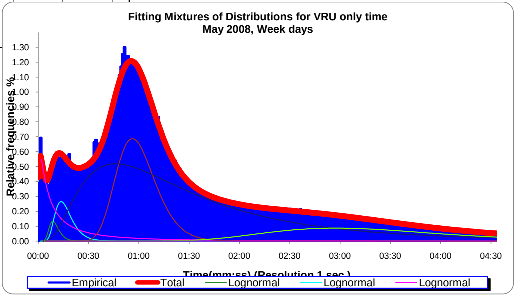

- **Practically Important**: (mean, std)(log) characterization
- **Theoretically Intriguing**: Why LogNormal ? Naturally multiplicative but, in fact, also **Infinitely-Divisible** (Generalized Gamma-Convolutions)

# Durations: Answering Machine

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




**Mixture: 7 LogNormals**

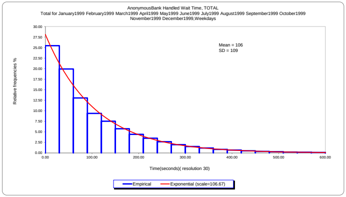



Tree-Network Tomography:
Reconstruct topology
from root and leaves

# Durations: Waiting Times in a Call Center
# ⇒ Protocols

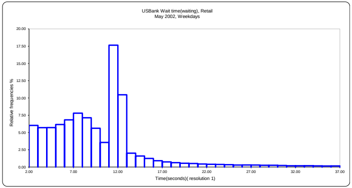
## Exponential in Heavy-Traffic (min.)

Small Israeli Bank



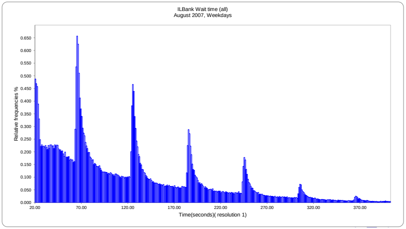

## Routing via Thresholds (sec.)

Large U.S. Bank



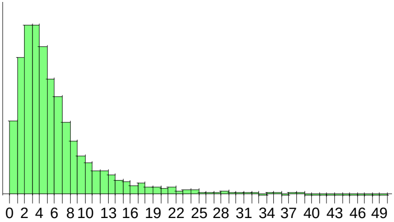

## Scheduling Priorities (sec.) **[compare Hospital LOS (hours)]**

Medium Israeli Bank



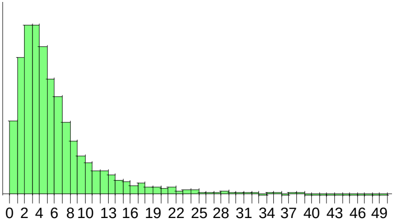
# LogNormal & Beyond: Length-of-Stay in a Hospital

## Israeli Hospital, in Days: LN

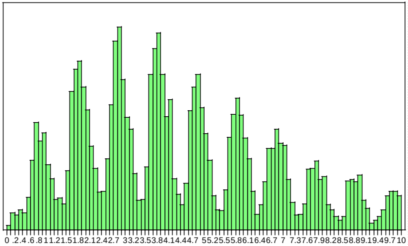


# LogNormal & Beyond: Length-of-Stay in a Hospital

## Israeli Hospital, in <u>Days</u>: LN

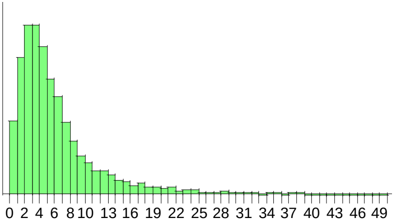



## Israeli Hospital, in <u>Hours</u>: Mixture

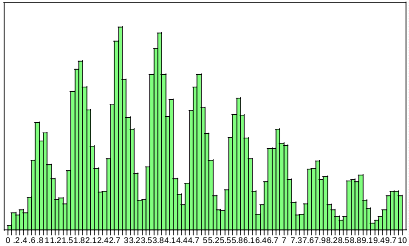


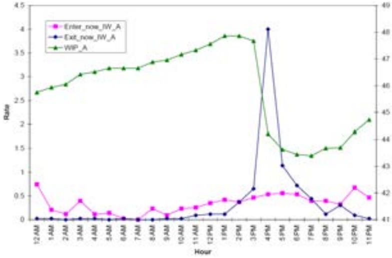
# LogNormal & Beyond: Length-of-Stay in a Hospital

## Israeli Hospital, in Days: LN




## Israeli Hospital, in Hours: Mixture






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

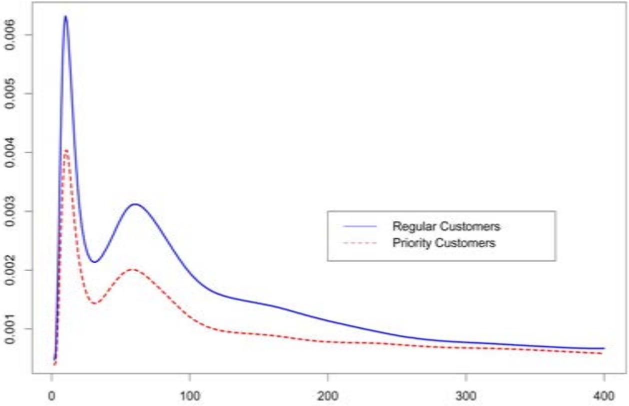
**Why Bother ?**

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

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

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

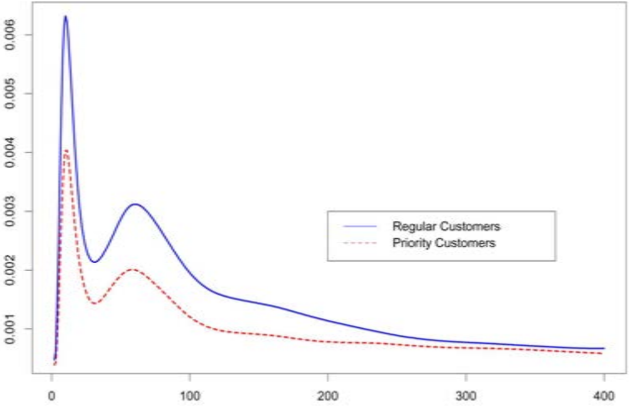
**Regular over VIP Customers – Israeli Bank**



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

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

**Regular over VIP Customers – Israeli Bank**

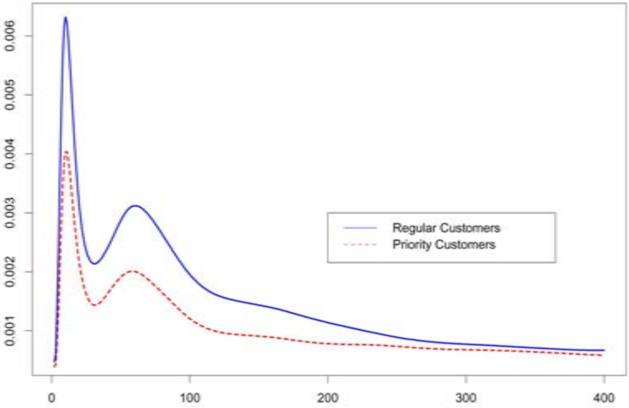



- Challenges: **Un-Censoring, Dependence, Smoothing** - requires **Call-by-Call Data**

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

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

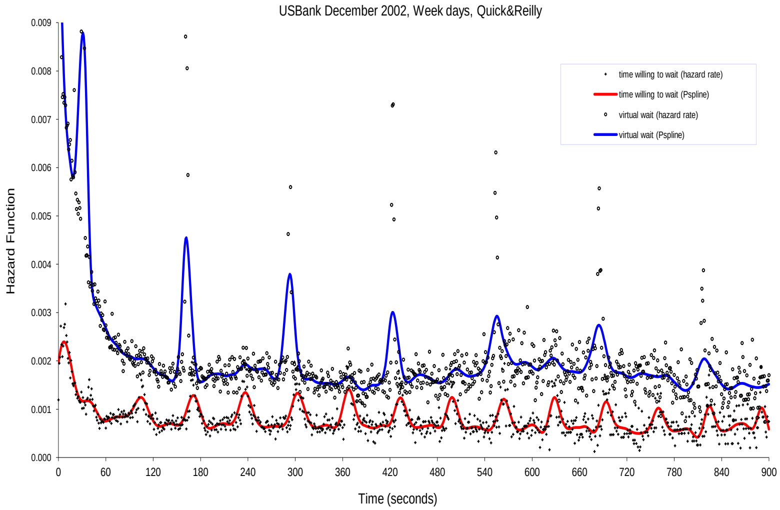
**Regular over VIP Customers – Israeli Bank**




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

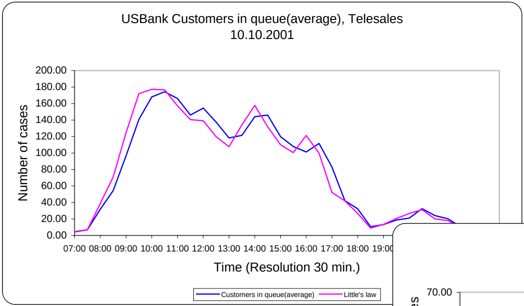
# Protocols + Psychology

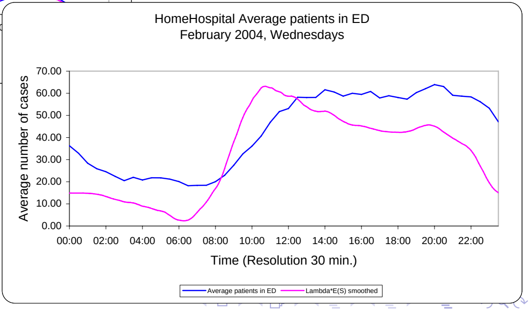
## Patient Customers, Announcements, Priority Upgrades



# Little's Law: Call Center & Emergency Department

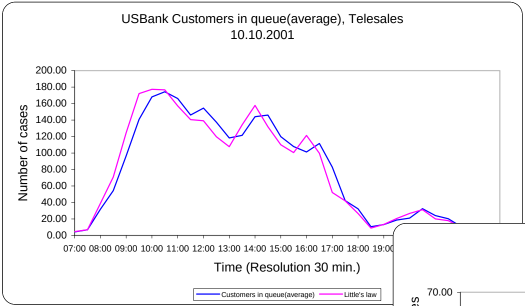
**Time-Gap**: **# in System** lags behind **Piecewise-Little** ($L = \lambda \times W$)

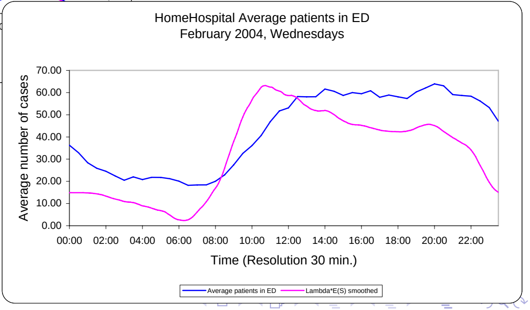


# Little's Law: Call Center & Emergency Department

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



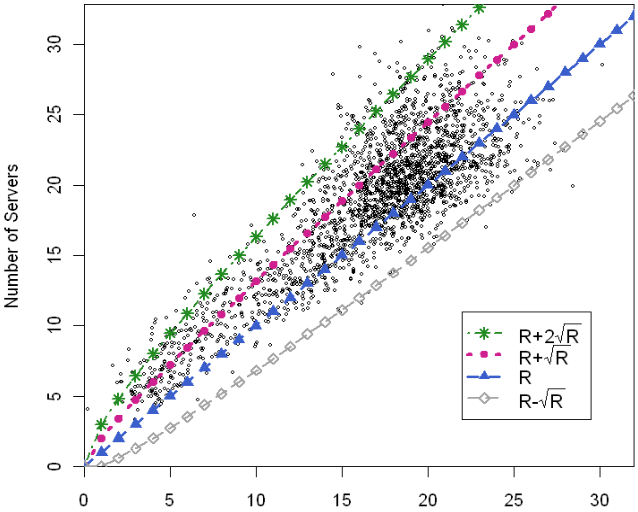




$\Rightarrow$ **Time-Varying Little's Law**

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

# Protocols: Staffing (N) vs. Offered-Load (R $= \lambda \times E(S)$)

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




**2205 half-hour intervals (13 summer weeks, week-days)**

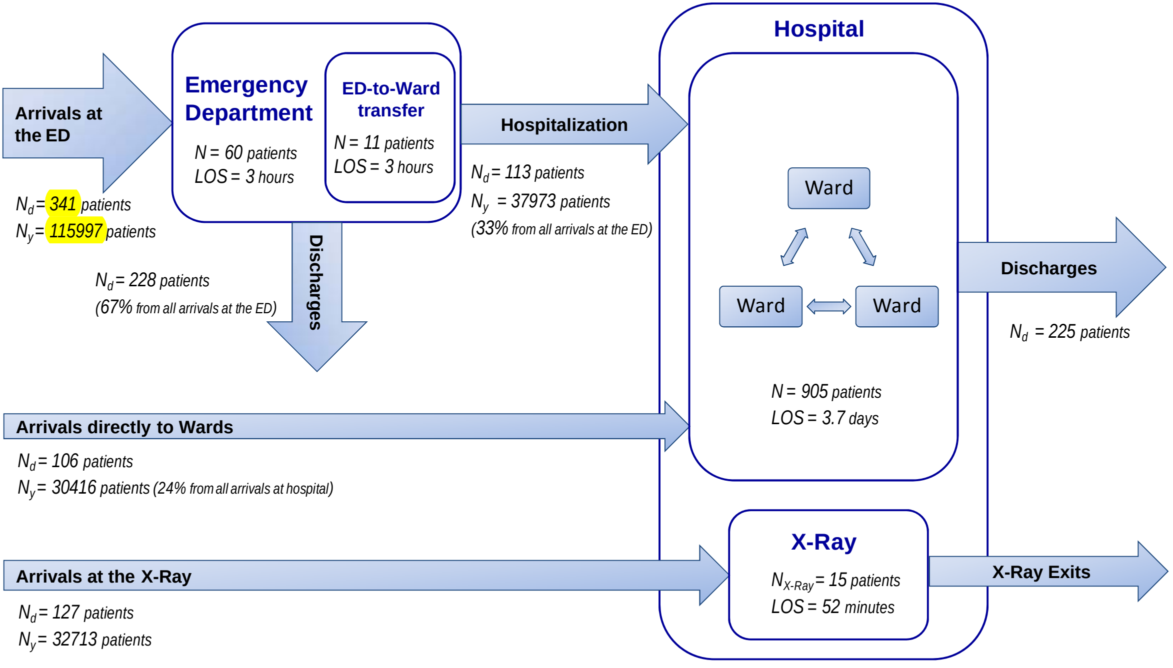
**Technion - Israel Institute of Technology**
The William Davidson Faculty of Industrial Engineering and Management
Center for Service Enterprise Engineering (SEE)
**http://ie.technion.ac.il/Labs/Serveng/**

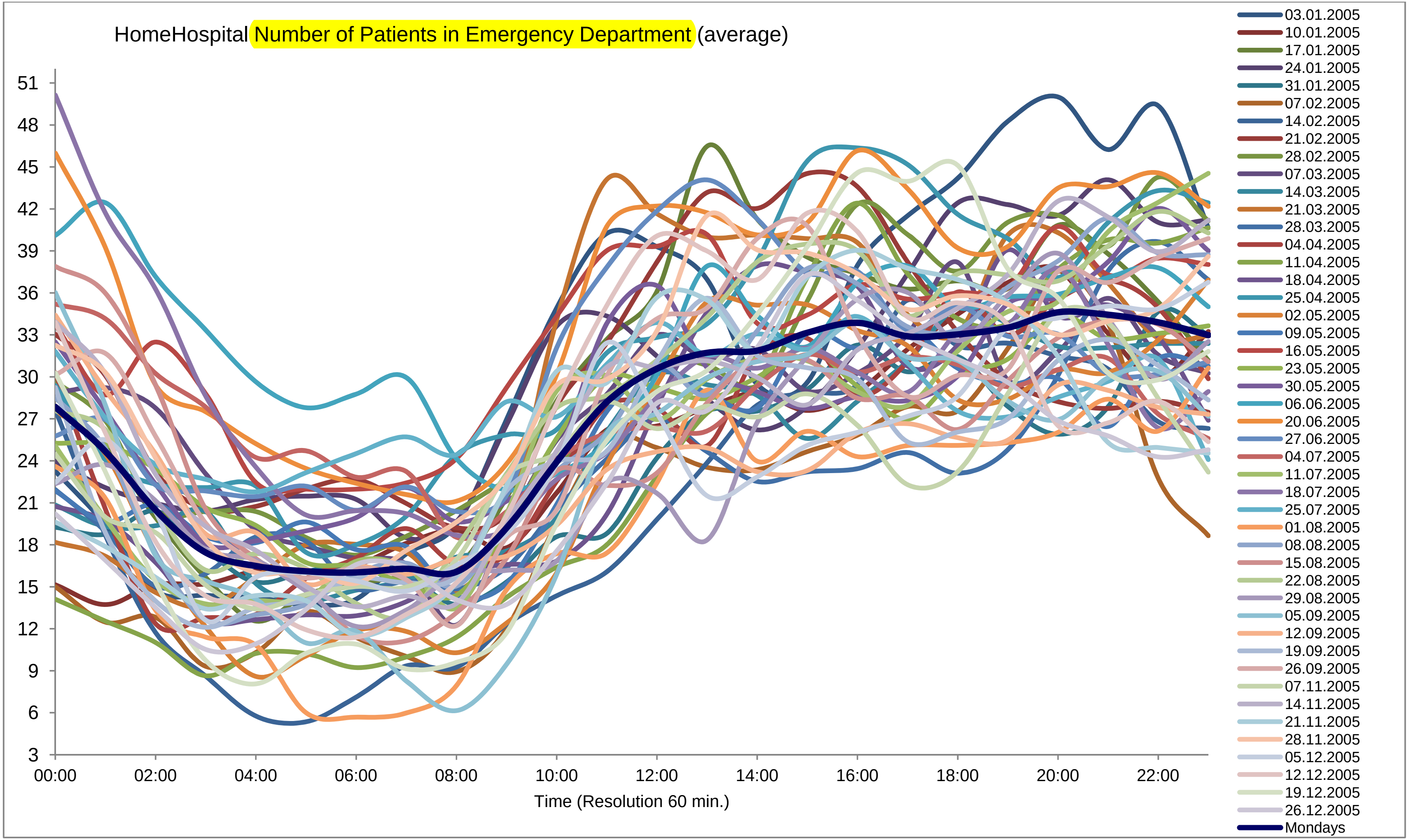
# SEEStat 3.0 Tutorial

August 23, 2012

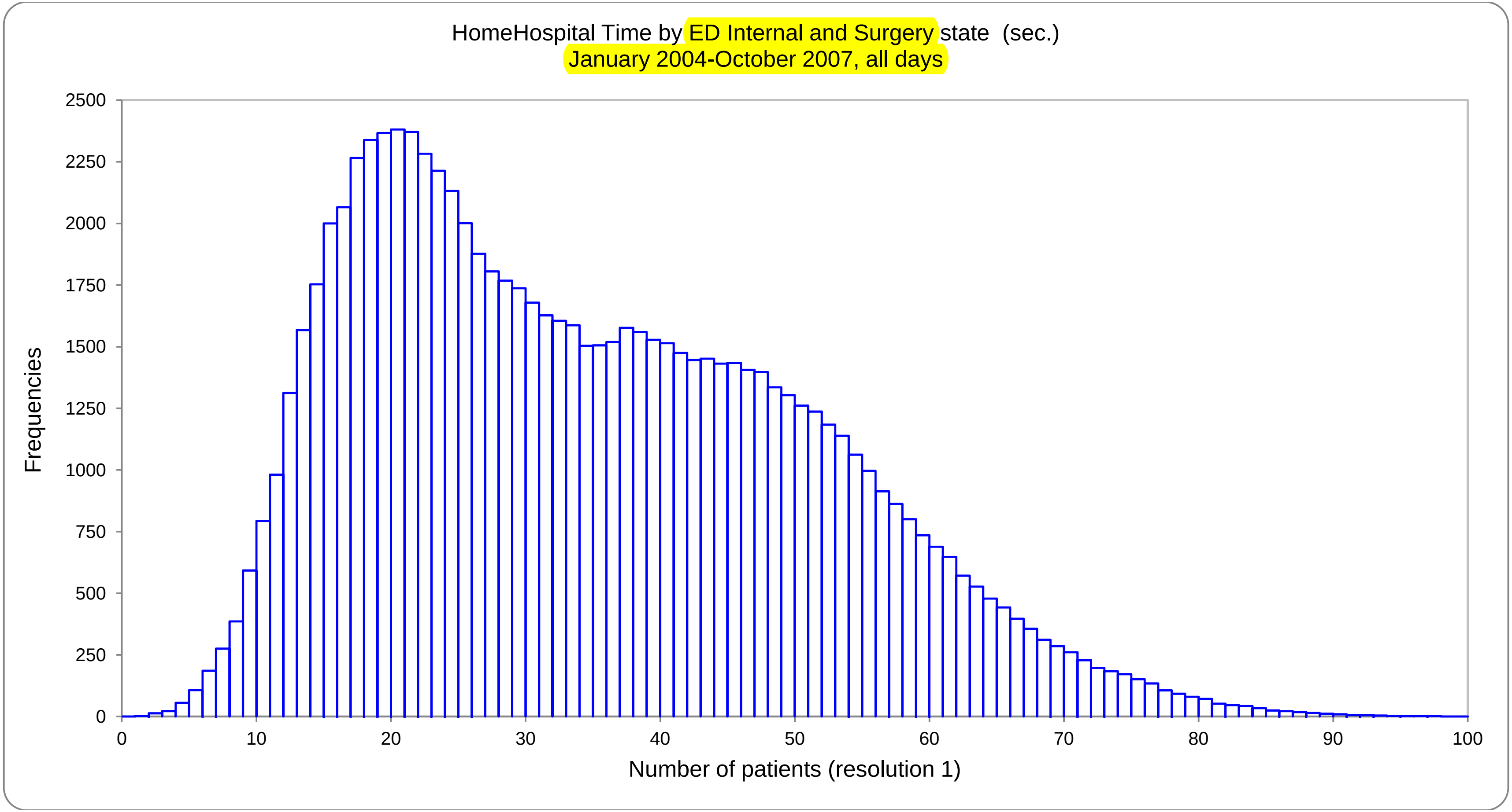
# HomeHospital Data

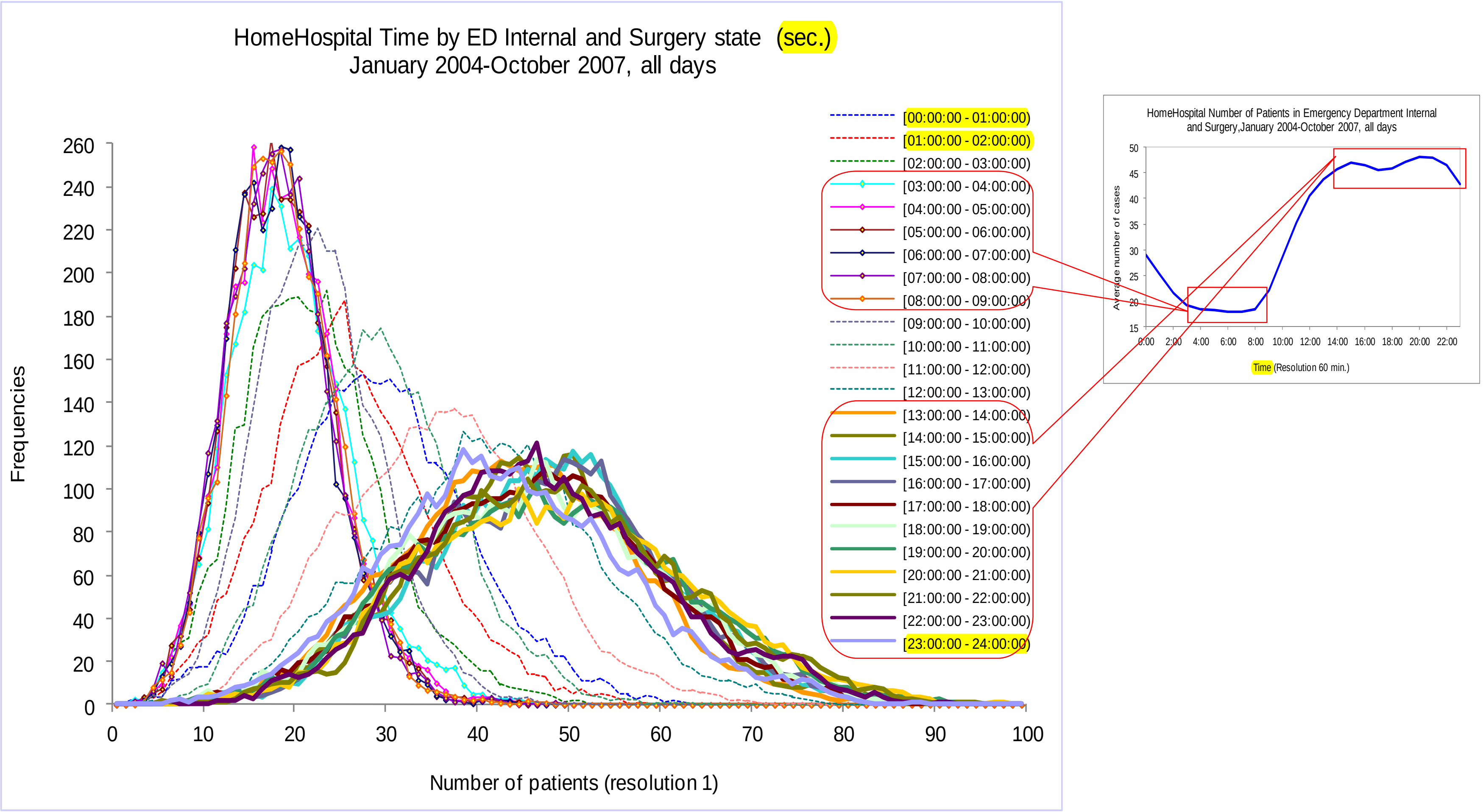
*Background: The data we rely on was collected at a large Israeli hospital. This hospital consists of about 1000 beds and 45 medical units. The data includes detailed information on patient flow throughout the hospital, over a period of several years (January 2004–October 2007). In particular, the data allows one to follow the paths of individual patients throughout their stay at the hospital, including admission, discharge, and transfers between hospital units. The data does not acknowledge resolutions within the ED or within wards.*

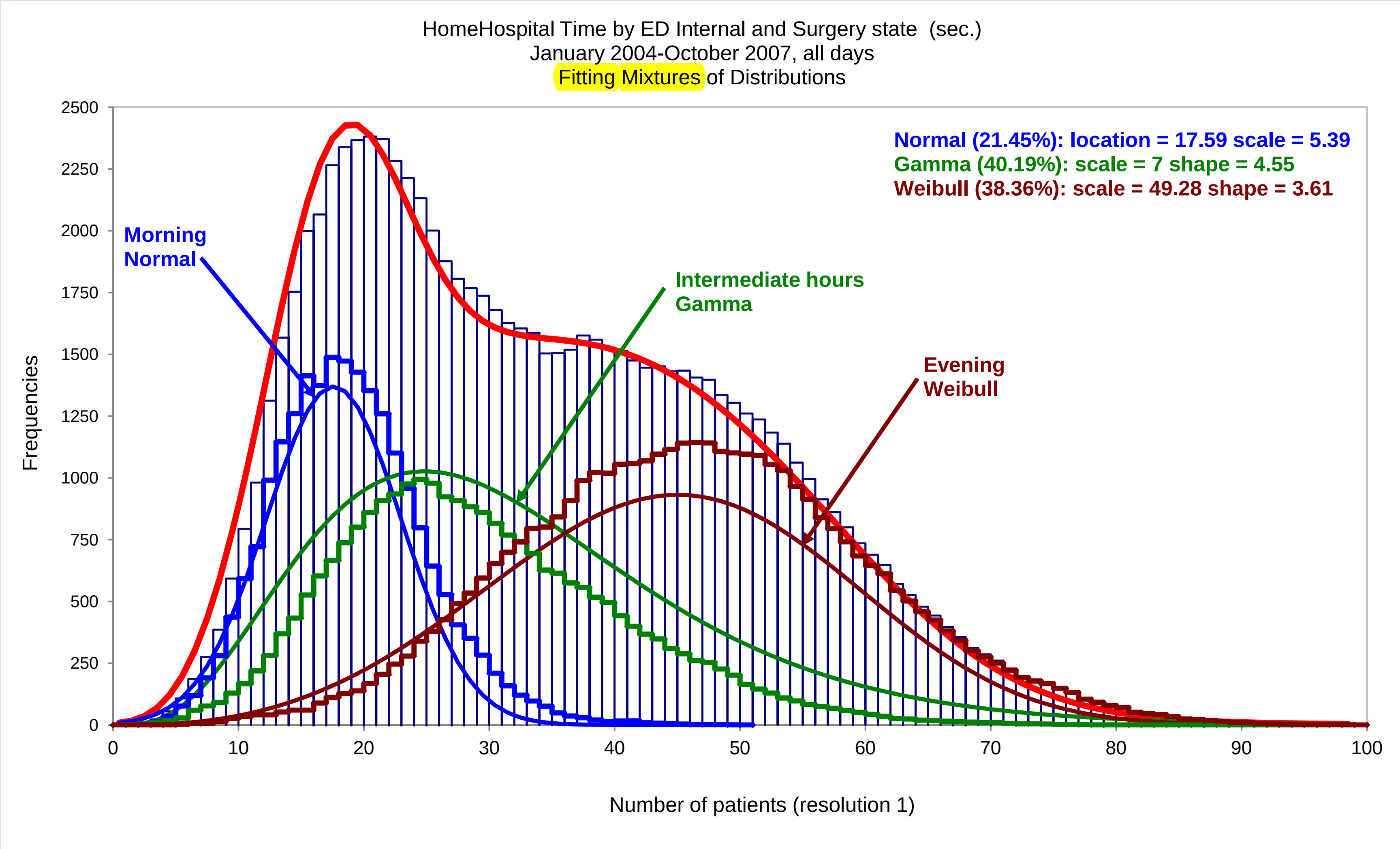



HomeHospital Number of Patients in Emergency Department (average)
51
48
45
42
39
36
33
30
27
24
21
18
15
12
9
6
3
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time (Resolution 60 min.)
03.01.2005
10.01.2005
17.01.2005
24.01.2005
31.01.2005
07.02.2005
14.02.2005
21.02.2005
28.02.2005
07.03.2005
14.03.2005
21.03.2005
28.03.2005
04.04.2005
11.04.2005
18.04.2005
25.04.2005
02.05.2005
09.05.2005
16.05.2005
23.05.2005
30.05.2005
06.06.2005
20.06.2005
27.06.2005
04.07.2005
11.07.2005
18.07.2005
25.07.2005
01.08.2005
08.08.2005
15.08.2005
22.08.2005
29.08.2005
05.09.2005
12.09.2005
19.09.2005
26.09.2005
07.11.2005
14.11.2005
21.11.2005
28.11.2005
05.12.2005
12.12.2005
19.12.2005
26.12.2005
Mondays

Number of patients in ED



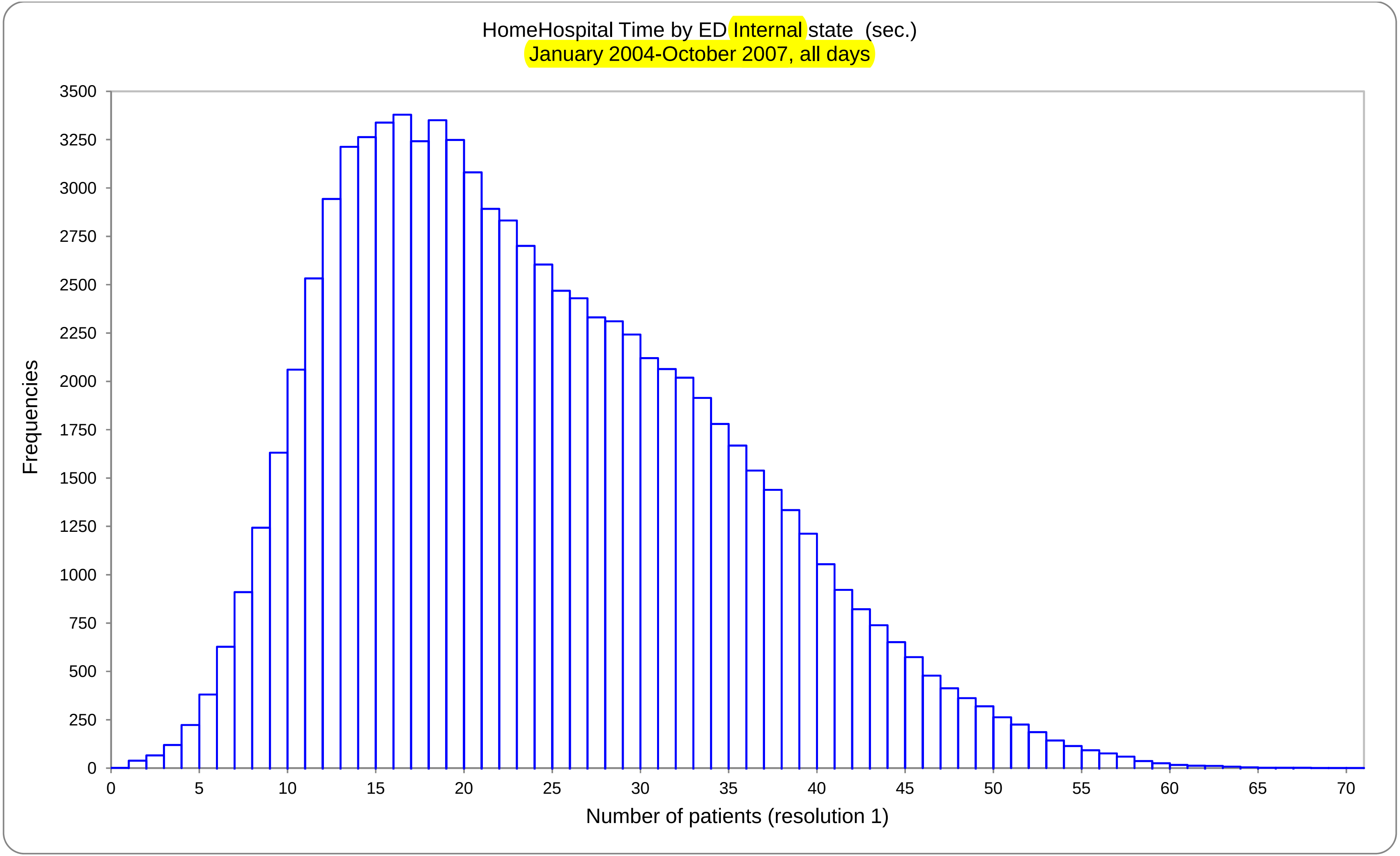

HomeHospital Time by ED Internal and Surgery state (sec.)
January 2004-October 2007, all days
Frequencies
Number of patients (resolution 1)
0
20
40
60
80
100
120
140
160
180
200
220
240
260
0
10
20
30
40
50
60
70
80
90
100
[00:00:00 - 01:00:00)
[01:00:00 - 02:00:00)
[02:00:00 - 03:00:00)
[03:00:00 - 04:00:00)
[04:00:00 - 05:00:00)
[05:00:00 - 06:00:00)
[06:00:00 - 07:00:00)
[07:00:00 - 08:00:00)
[08:00:00 - 09:00:00)
[09:00:00 - 10:00:00)
[10:00:00 - 11:00:00)
[11:00:00 - 12:00:00)
[12:00:00 - 13:00:00)
[13:00:00 - 14:00:00)
[14:00:00 - 15:00:00)
[15:00:00 - 16:00:00)
[16:00:00 - 17:00:00)
[17:00:00 - 18:00:00)
[18:00:00 - 19:00:00)
[19:00:00 - 20:00:00)
[20:00:00 - 21:00:00)
[21:00:00 - 22:00:00)
[22:00:00 - 23:00:00)
[23:00:00 - 24:00:00)
HomeHospital Number of Patients in Emergency Department Internal and Surgery,January 2004-October 2007, all days
Average number of cases
15
20
25
30
35
40
45
50
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time (Resolution 60 min.)

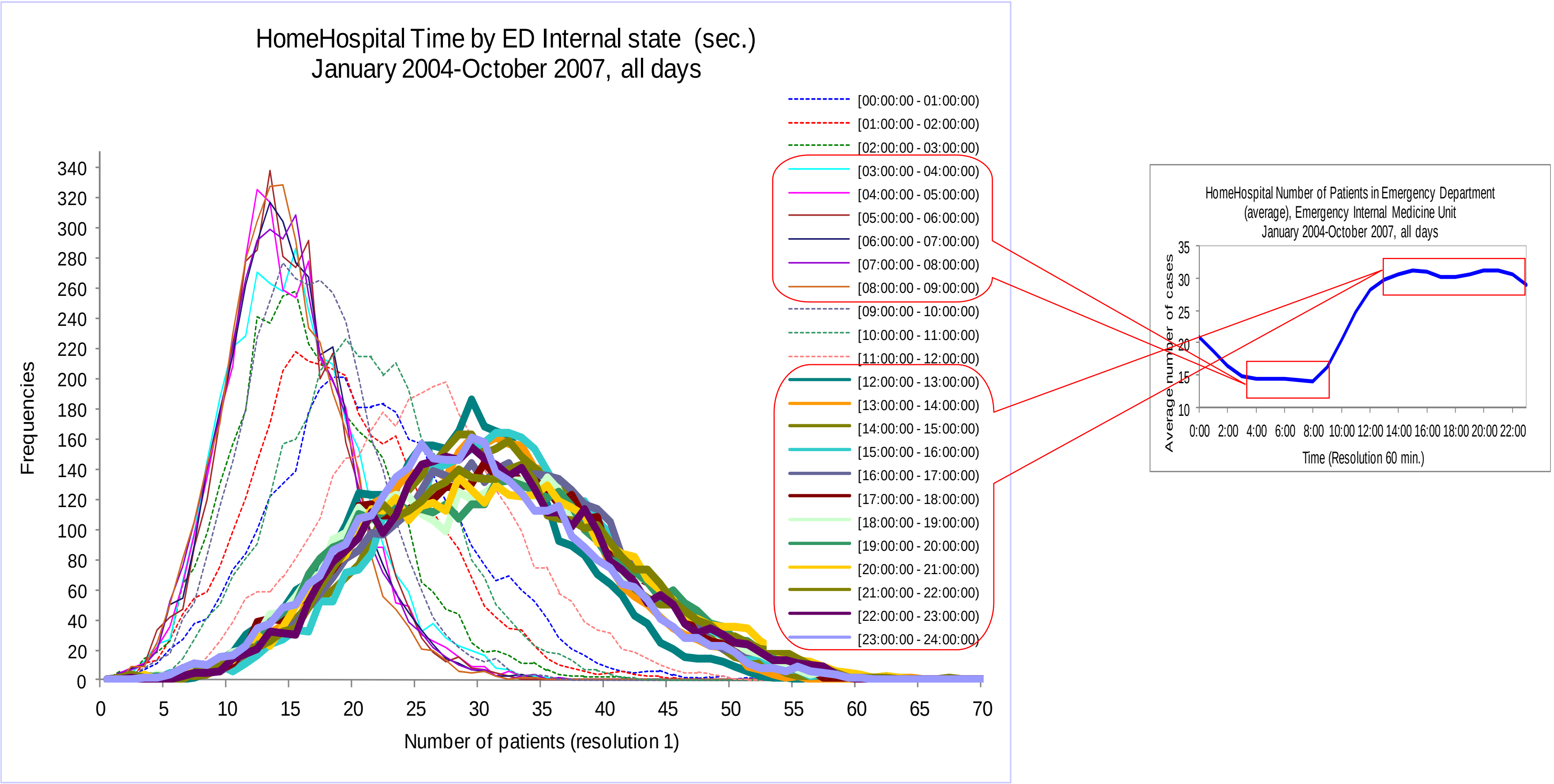

HomeHospital Time by ED Internal and Surgery state (sec.)
January 2004-October 2007, all days
Fitting Mixtures of Distributions
Normal (21.45%): location = 17.59 scale = 5.39
Gamma (40.19%): scale = 7 shape = 4.55
Weibull (38.36%): scale = 49.28 shape = 3.61
Morning
Normal
Intermediate hours
Gamma
Evening
Weibull
Frequencies
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
0
10
20
30
40
50
60
70
80
90
100
Number of patients (resolution 1)

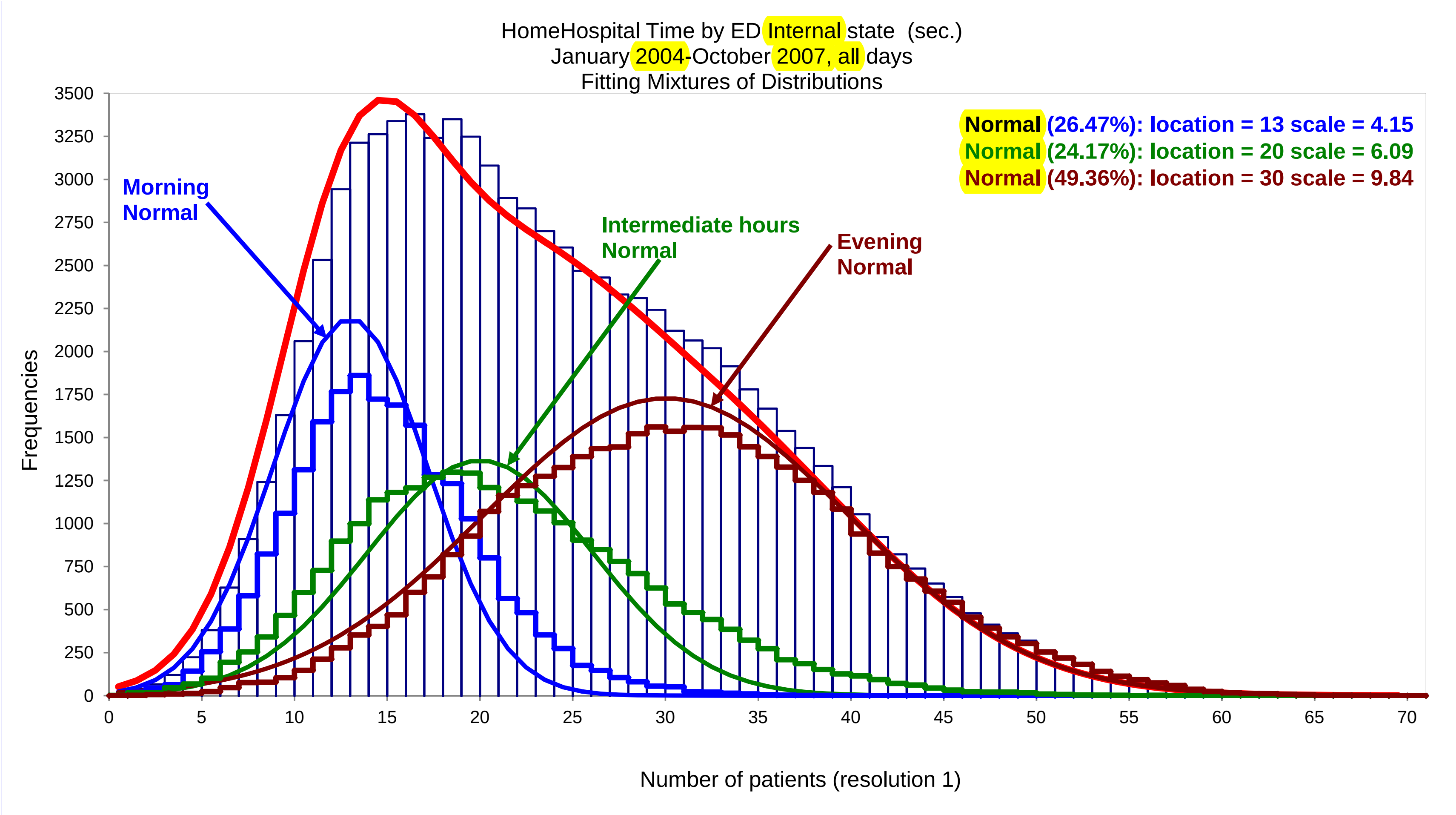
| *Time by ED Internal and Surgery state (sec.)* ***Statistics*** | |
|---|---|
| N | 120960000 |
| N(average per day) | 86400 |
| Mean | 34.1 |
| Standard Deviation | 16.34 |
| Variance | 266.87 |
| Median | 32 |
| Minimum | 0 |
| Maximum | 99 |
| Skewness | 0.524 |
| Kurtosis | -0.44452 |
| Standard Error Mean | 0.00149 |
| Interquartile Range | 25 |
| Mean Absolute Deviation | 13.7 |
| Median Absolute Deviation(MAD) | 12 |
| Coefficient of Variation (CV) (%) | 47.9 |
| L-moment 2 (half of Gini's Mean Difference) | 9.25 |
| L-Skewness | 0.121 |
| L-Kurtosis | 0.0561 |
| Coefficient of L-variation (L-CV)(%) (Gini's Coefficient) | 27.12 |

| ***Parameter Estimates*** | | | | | | |
|---|---|---|---|---|---|---|
| *Components* | *Mixing Proportions (%)* | *Location* | *Scale* | *Shape* | *Mean* | *Standard Deviation* |
| 1. Normal | 21.45 | 17.59 | 5.39 | | 17.59 | 5.39 |
| 2. Gamma | 40.19 | | 7.00 | 4.55 | 31.83 | 14.99 |
| 3. Weibull | 38.36 | | 49.28 | 3.61 | 44.42 | 13.679 |

| ***Goodness-of-Fit Tests*** | | | |
|---|---|---|---|
| *Tests* | *Statistic* | *DF* | *p Value* |
| Residuals Std | 0.011 | | |
| Kolmogorov-Smirnov | 0.028 | | <.0001 |
| Cramer-von Mises | 14953.04 | | <.0001 |
| Andersen-Darling | 96156.09 | | <.0001 |
| Chi-Square | 460741.3 | 94 | <.0001 |


HomeHospital Time by ED Internal state (sec.)
January 2004-October 2007, all days
Frequencies
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
Number of patients (resolution 1)

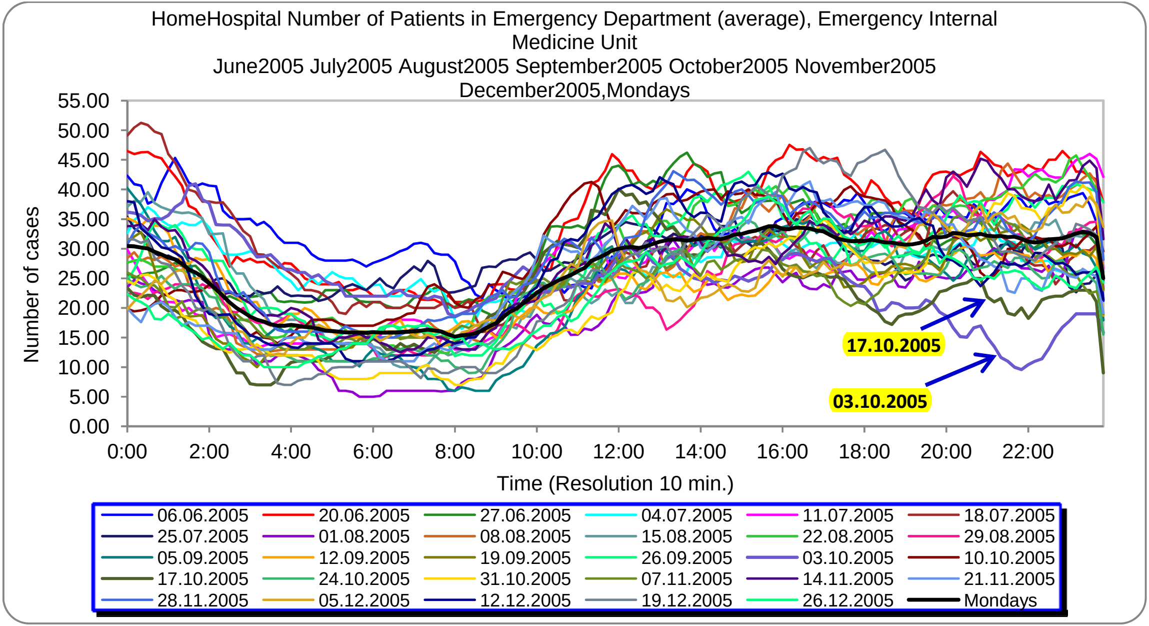

HomeHospital Time by ED Internal state (sec.)
January 2004-October 2007, all days
Frequencies
Number of patients (resolution 1)
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
[00:00:00 - 01:00:00)
[01:00:00 - 02:00:00)
[02:00:00 - 03:00:00)
[03:00:00 - 04:00:00)
[04:00:00 - 05:00:00)
[05:00:00 - 06:00:00)
[06:00:00 - 07:00:00)
[07:00:00 - 08:00:00)
[08:00:00 - 09:00:00)
[09:00:00 - 10:00:00)
[10:00:00 - 11:00:00)
[11:00:00 - 12:00:00)
[12:00:00 - 13:00:00)
[13:00:00 - 14:00:00)
[14:00:00 - 15:00:00)
[15:00:00 - 16:00:00)
[16:00:00 - 17:00:00)
[17:00:00 - 18:00:00)
[18:00:00 - 19:00:00)
[19:00:00 - 20:00:00)
[20:00:00 - 21:00:00)
[21:00:00 - 22:00:00)
[22:00:00 - 23:00:00)
[23:00:00 - 24:00:00)
HomeHospital Number of Patients in Emergency Department
(average), Emergency Internal Medicine Unit
January 2004-October 2007, all days
Average number of cases
Time (Resolution 60 min.)
10
15
20
25
30
35
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00


HomeHospital Time by ED Internal state (sec.)
January 2004-October 2007, all days
Fitting Mixtures of Distributions
Normal (26.47%): location = 13 scale = 4.15
Normal (24.17%): location = 20 scale = 6.09
Normal (49.36%): location = 30 scale = 9.84
Morning Normal
Intermediate hours Normal
Evening Normal
Frequencies
Number of patients (resolution 1)
0
250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70

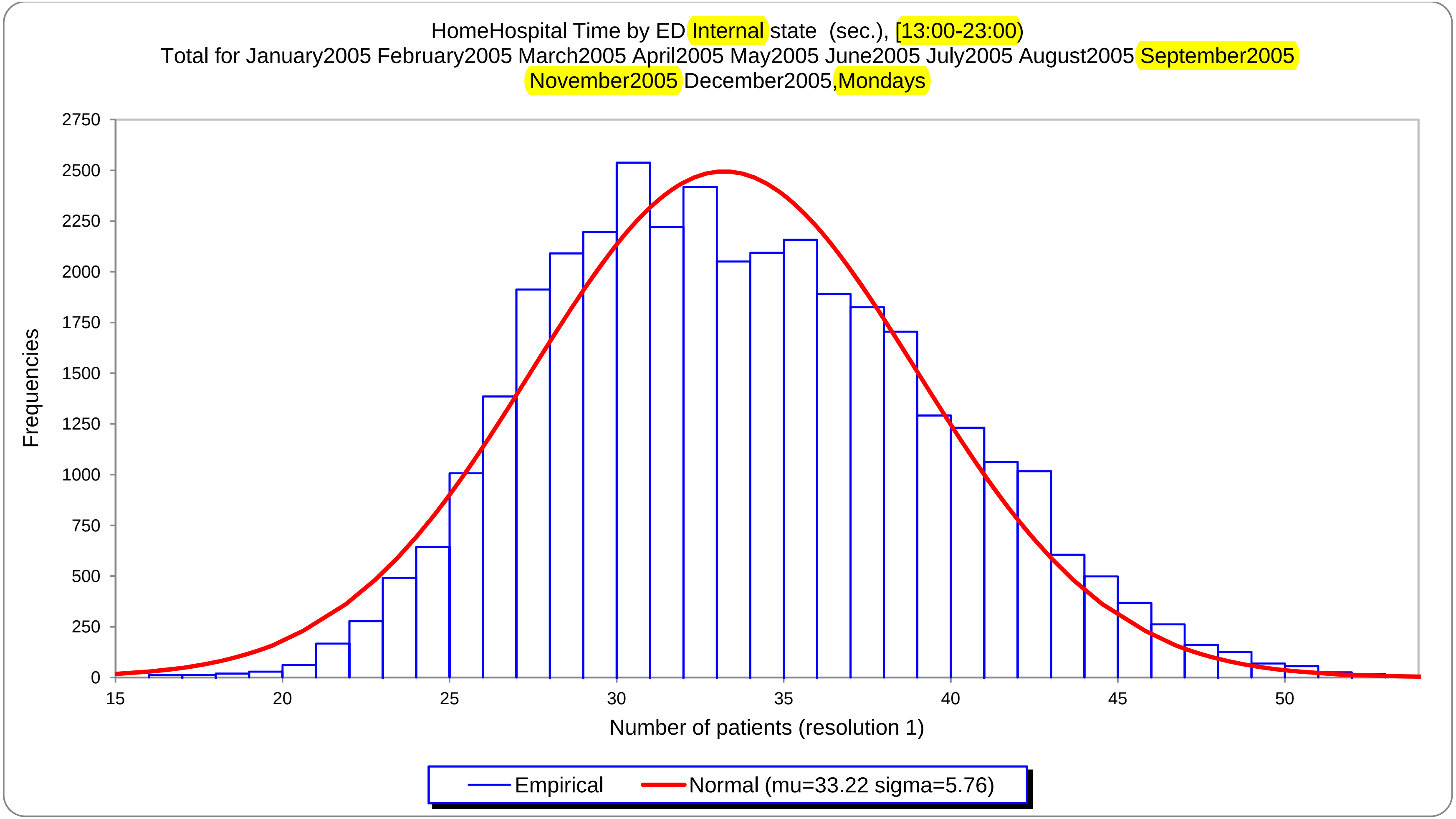
| *Time by ED Internal state (sec.)* **Statistics** | |
|---|---|
| N | 120960000 |
| N(average per day) | 86400 |
| Mean | 23.63 |
| Standard Deviation | 10.7 |
| Variance | 114.4 |
| Median | 22 |
| Minimum | 0 |
| Maximum | 70 |
| Skewness | 0.557 |
| Kurtosis | -0.23178 |
| Standard Error Mean | 0.00097 |
| Interquartile Range | 16 |
| Mean Absolute Deviation | 8.808 |
| Median Absolute Deviation(MAD) | 8 |
| Coefficient of Variation (CV) (%) | 45.27 |
| L-moment 2 (half of Gini's Mean Difference) | 6.031 |
| L-Skewness | 0.121 |
| L-Kurtosis | 0.0813 |
| Coefficient of L-variation (L-CV)(%) (Gini's Coefficient) | 25.53 |

| ***Parameter Estimates*** | | | | | | |
|---|---|---|---|---|---|---|
| *Components* | *Mixing Proportions (%)* | *Location* | *Scale* | *Shape* | *Mean* | *Standard Deviation* |
| 1. Normal | 26.47 | 13.01 | 4.15 | | 13.01 | 4.15 |
| 2. Normal | 24.17 | 20.00 | 6.09 | | 20.00 | 6.09 |
| 3. Normal | 49.36 | 30.04 | 9.84 | | 30.04 | 9.84 |

| ***Goodness-of-Fit Tests*** | | | |
|---|---|---|---|
| *Tests* | *Statistic* | *DF* | *p Value* |
| Residuals Std | 0.017 | | |
| Kolmogorov-Smirnov | 0.04 | | <.0001 |
| Cramer-von Mises | 34138.3 | | <.0001 |
| Andersen-Darling | 200657.9 | | <.0001 |
| Chi-Square | 584234.3 | 65 | <.0001 |



HomeHospital Number of Patients in Emergency Department (average), Emergency Internal Medicine Unit
June2005 July2005 August2005 September2005 October2005 November2005 December2005,Mondays
Number of cases
55.00
50.00
45.00
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time (Resolution 10 min.)
17.10.2005
03.10.2005
06.06.2005
20.06.2005
27.06.2005
04.07.2005
11.07.2005
18.07.2005
25.07.2005
01.08.2005
08.08.2005
15.08.2005
22.08.2005
29.08.2005
05.09.2005
12.09.2005
19.09.2005
26.09.2005
03.10.2005
10.10.2005
17.10.2005
24.10.2005
31.10.2005
07.11.2005
14.11.2005
21.11.2005
28.11.2005
05.12.2005
12.12.2005
19.12.2005
26.12.2005
Mondays


HomeHospital Time by ED Internal state (sec.), [13:00-23:00)
Total for January2005 February2005 March2005 April2005 May2005 June2005 July2005 August2005 September2005 November2005 December2005, Mondays
Frequencies
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
15
20
25
30
35
40
45
50
Number of patients (resolution 1)
Empirical
Normal (mu=33.22 sigma=5.76)

| *Time by ED Internal state (sec.), [13:00-23:00)* ***Statistics*** | |
|---|---|
| N | 1656000 |
| N(average per day) | 36000 |
| Mean | 33.22 |
| Standard Deviation | 5.756 |
| Variance | 33.13 |
| Median | 33 |
| Minimum | 16 |
| Maximum | 54 |
| Skewness | 0.282 |
| Kurtosis | -0.31791 |
| Standard Error Mean | 0.00447 |
| Interquartile Range | 8 |
| Mean Absolute Deviation | 4.712 |
| Median Absolute Deviation(MAD) | 4 |
| Coefficient of Variation(CV) (%) | 17.33 |
| L-moment 2 (half of Gini's Mean Difference) | 3.263 |
| L-Skewness | 0.0601 |
| L-Kurtosis | 0.0924 |
| Coefficient of L-variation(L-CV)(%) (Gini's Coefficient) | 9.82 |

| ***Parameters for Normal Distribution*** | |
|---|---|
| *Parameter* | *Estimate* |
| mu | 33.22 |
| sigma | 5.76 |
| mean | 33.22 |
| std | 5.756 |

| ***Goodness-of-Fit Tests for Normal Distribution*** | | | |
|---|---|---|---|
| *Test* | *Statistic* | *DF* | *p Value* |
| Residuals Std | 0.025 | | |
| Kolmogorov-Smirnov | 0.069 | | <.0001 |
| Cramer-von Mises | 1068.72 | | <.0001 |
| Anderson-Darling | 6193.27 | | <.0001 |
| Chi-Square | >1000 | 34 | <.0001 |

# Internal ED, non-holiday Mondays, 13:00-23:00

“Recall":

- Arrivals processes to EDs “are" inhomogeneous Poisson ($\approx$ piecewise Poisson)
- $M/M/\infty$ (ample-server queue) has Poisson steady-state
- Birth-Death processes are time-reversible hence steady-state invariant under state-conditioning

# Internal ED, non-holiday Mondays, 13:00-23:00

“Recall":

- Arrivals processes to EDs “are" inhomogeneous Poisson ($\approx$ piecewise Poisson)
- $M/M/\infty$ (ample-server queue) has Poisson steady-state
- Birth-Death processes are time-reversible hence steady-state invariant under state-conditioning

Our EDA “implies" (**w/ Armony, Marmor, Tseytlin, Yom-Tov**):

- **Israeli ED census** $\stackrel{d}{=} M/M/\infty$

# Internal ED, non-holiday Mondays, 13:00-23:00

“Recall":

- Arrivals processes to EDs “are" inhomogeneous Poisson ($\approx$ piecewise Poisson)
- $M/M/\infty$ (ample-server queue) has Poisson steady-state
- Birth-Death processes are time-reversible hence steady-state invariant under state-conditioning

Our EDA “implies" (**w/ Armony, Marmor, Tseytlin, Yom-Tov**):

- **Israeli ED census** $\stackrel{d}{=} M/M/\infty$
- ED $\stackrel{d}{=}$ Reversible Birth-Death process, hence conditioning on finite capacity, say $B$, gives rise to $M/M/B/B$ (Erlang-B)
- **U.S. ED census** $\stackrel{d}{=} M/M/B/B$, which can be (has been) used to analyze Ambulance Diversion (ED Blocking)

# Internal ED, non-holiday Mondays, 13:00-23:00

"Recall":

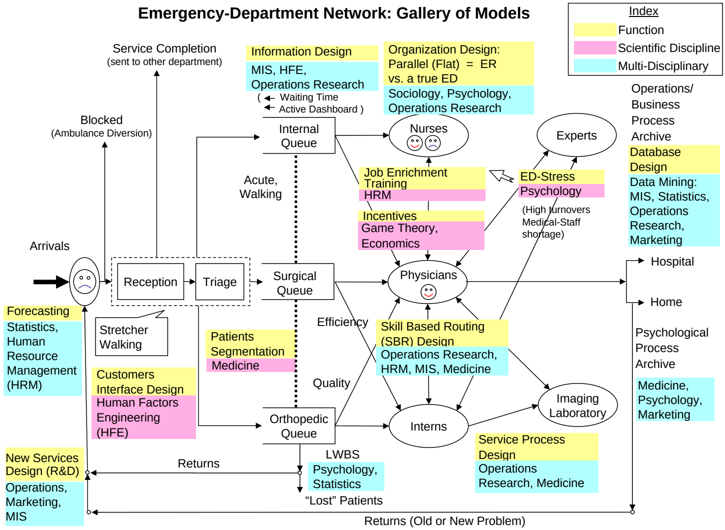
- Arrivals processes to EDs "are" inhomogeneous Poisson ($\approx$ piecewise Poisson)
- $M/M/\infty$ (ample-server queue) has Poisson steady-state
- Birth-Death processes are time-reversible hence steady-state invariant under state-conditioning

Our EDA "implies" (**w/ Armony, Marmor, Tseytlin, Yom-Tov**):

- **Israeli ED census** $\stackrel{d}{=} M/M/\infty$
- ED $\stackrel{d}{=}$ Reversible Birth-Death process, hence conditioning on finite capacity, say $B$, gives rise to $M/M/B/B$ (Erlang-B)
- **U.S. ED census** $\stackrel{d}{=} M/M/B/B$, which can be (has been) used to analyze Ambulance Diversion (ED Blocking)
- Puzzle (getting ahead): Is the ED an Erlang-A system with $\mu = \theta$? then the "effective number of servers" can be, perhaps, deduced via those LWBS = Left Without Being Seen (or LAMA)

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

# Emergency-Department Network



# Prerequisite II: Models (FNets)

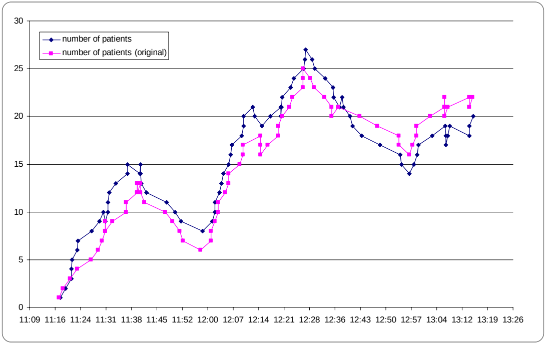
**“Laws of Large Numbers"** capture **Predictable** Variability

**Deterministic** Models: Scale Averages-out **Stochastic Individualism**

# Prerequisite II: Models (FNets)

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

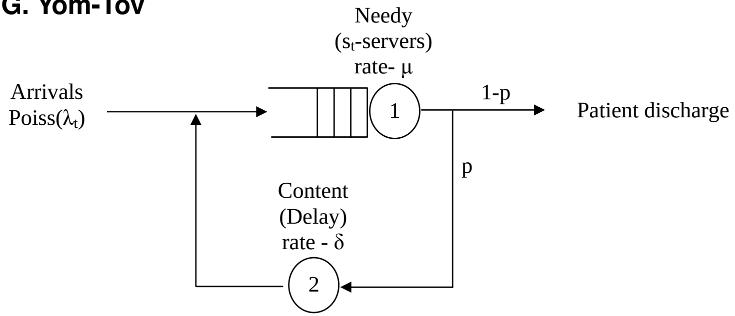
$\#$ **Severely-Wounded Patients, 11:00-13:00 (Censored LOS)**




- **Transient** Q's:
  - **Control** of Mass Casualty Events (w/ **I. Cohen, N. Zychlinski**)
  - **Staffing** Chemical MCE (w/ **G. Yom-Tov**)

# The Basic Service-Network Model: Erlang-R

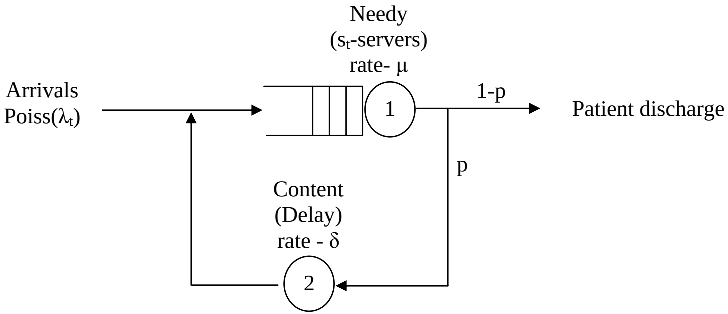
**w/ G. Yom-Tov**



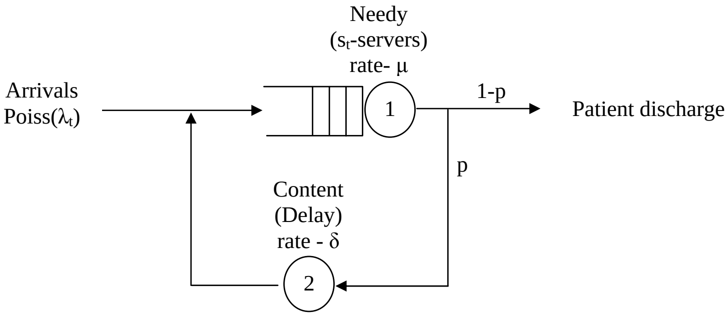

**Erlang-R** **(IE: Repairman Problem 50's; CS: Central-Server 60's)** =
**2-station "Jackson" Network** = (M/M/S, M/M/$\infty$) :

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

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



# Fluid Model $\leftrightarrow$ (Time-Varying) Erlang-R System



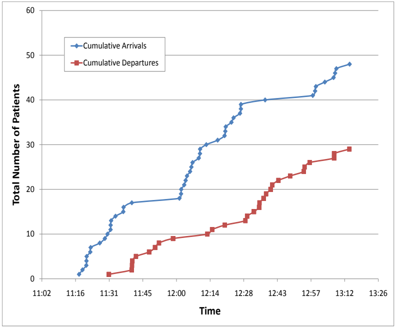

**FNet** of a 2-station "Jackson" Network:

$$\begin{aligned} \frac{d}{dt} q_t^1 &= \lambda_t - \mu \cdot (q_t^1 \wedge s_t) + \delta \cdot q_t^2, \\ \frac{d}{dt} q_t^2 &= p \cdot \mu \cdot (q_t^1 \wedge s_t) - \delta \cdot q_t^2. \end{aligned} \tag{1}$$

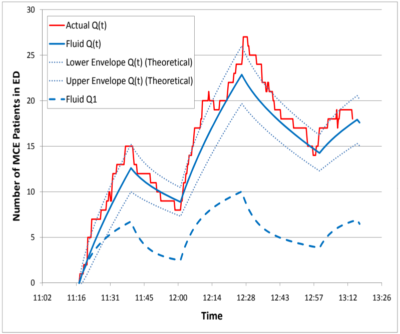
# Erlang-R: Fitting a Simple Model to a Complex Reality

## Chemical MCE Drill (Israel, May 2010)

**Arrivals & Departures (RFID)**




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

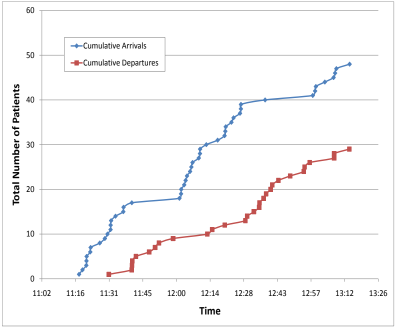



- **Recurrent/Repeated** services in MCE Events: eg. Injection every 15 minutes

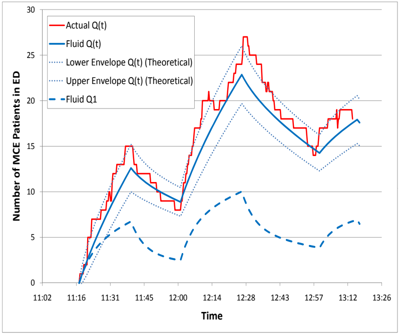
# Erlang-R: Fitting a Simple Model to a Complex Reality

## Chemical MCE Drill (Israel, May 2010)

**Arrivals & Departures (RFID)**




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




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

# An Asymptotic Framework: Erlang-R in the ED

**System** = Emergency Department (eg. Rambam Hospital)

- **SimNet** = Customized ED-Simulator (Marmor & Sinreich)
- **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)

## An Asymptotic Framework: Erlang-R in the ED

**System** = Emergency Department (eg. Rambam Hospital)

- **SimNet** = Customized ED-Simulator (Marmor & Sinreich)
- **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)

**Asymptotic Framework**

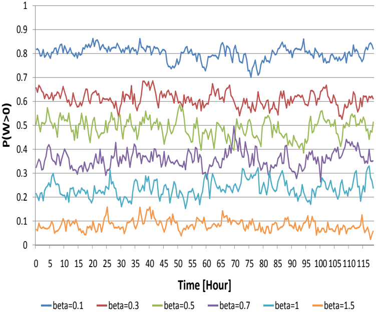
- Data and Measurements
- Fit a simple model (time-varying Erlang-R) to a complex reality (ED Physicians)
- Develop FNets (Offered-Load of Physicians) and (relevant) DNet (if needed)
- Use FNet / DNet for Design ($\sqrt{}$-Staffing), Analysis, ...
- Simulate reality (ED with $\sqrt{}$-staffing of Physicians)
- Validation: stable performance, confidence intervals, ...

# Case Study: Emergency Ward Staffing

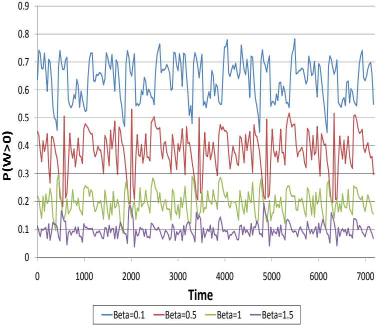
## Many-Server ($\uparrow \infty$) Approximations for Small Systems (1-7)

- Staffing resolution: 1 hour
- Lower bound: 1 doctor per type
- Flexible (time-varying square-root) staffing: **Yunan's Lecture**
- Rounding effects $\Rightarrow$ Not all performance levels achievable

**90 Servers**

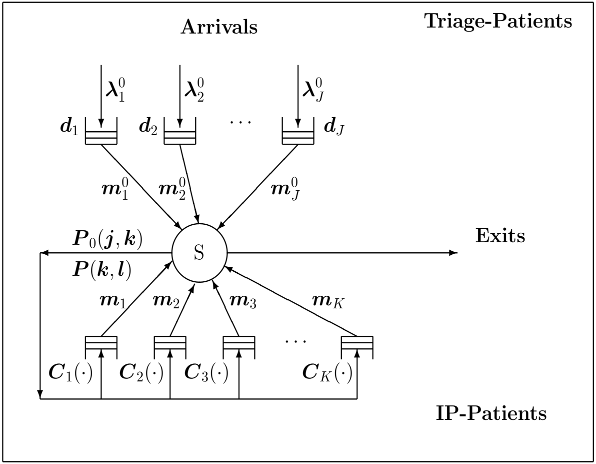



**1-7 Doctors**



# ED Patient Flow: The Physicians View

with **J. Huang, B. Carmeli; N. Shimkin**




**Goal**: Adhere to **Triage-Constraints**, then **release In-Process** Patients
**Models**: Time-Varying (FNet) - in process; Stationary (DNet) - v. soon

# Prerequisite II: Models (DNets, QED Q's)

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

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

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

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

# Prerequisite II: Models (DNets, QED Q's)

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

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

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

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

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

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

# Prerequisite II: Models (DNets, QED Q's)

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

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

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

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

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

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

Moreover, a significant fraction not delayed in queue: e.g. in well-run

- CCs: **50%** served "immediately" & **90%** utilization $\Rightarrow$ **QED**
- EDs + IWs: **?**

# Prerequisite II: Models (DNets, QED Q's)

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

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

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

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

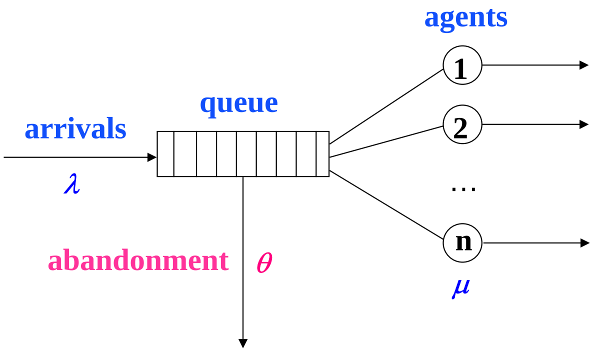
**Yet**, **heavily-loaded** queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

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

Moreover, a significant fraction not delayed in queue: e.g. in well-run

- CCs: **50%** served "immediately" & **90%** utilization $\Rightarrow$ **QED**
- EDs + IWs: **?** Multiple scales! IW-"Beds" (10's) are QED while IW-Doctors (1's) are in conventional heavy-traffic (hours wait for minutes service), hence the bottlenecks

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




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

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

# Erlang-A: Practical Relevance?

**Experience:**

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

# Erlang-A: Practical Relevance?

**Experience:**

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

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

# Erlang-A: Practical Relevance?

**Experience:**

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

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

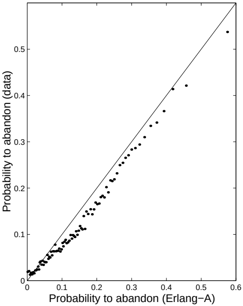
Question: **Is Erlang-A Relevant?**

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

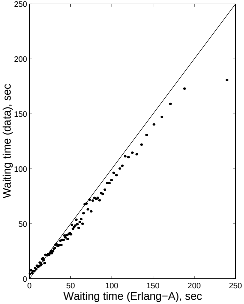
# Erlang-A: Fitting a Simple Model to a Complex Reality

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

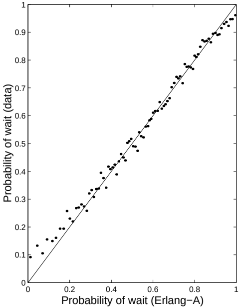
### % Abandon




### 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

## QED Theory (Erlang '13; Halfin & Whitt '81; w/Garnett & Reiman '02)

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

- **QED** $\quad$ $\%\{\text{Cust Wait} > 0\} \approx \alpha,$ $\quad$ $0 < \alpha < 1;$

  or $\quad$ $\%\{\text{Serv Idle} > 0\} \approx 1 - \alpha$

- **Customers** $\quad$ $\{\text{Abandon}\} \approx \frac{\gamma}{\sqrt{N}},$ $\quad$ $0 < \gamma;$

- **Agents** $\quad$ $\text{OCC} \approx 1 - \frac{\beta+\gamma}{\sqrt{N}}$ $\quad$ $-\infty < \beta < \infty;$

- **Managers** $\quad$ $N \approx R + \beta\sqrt{R}$, $\ R = \lambda \times E(S)$ $\quad$ not small;

Here $\ \boldsymbol{R}$ = **Offered Load**
eg. $\quad R = 25$ call/min. $\times$ 4 min./call $= 100$

# Erlang-A: QED Approximations (Examples)

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

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

- **Delay Probability:**

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

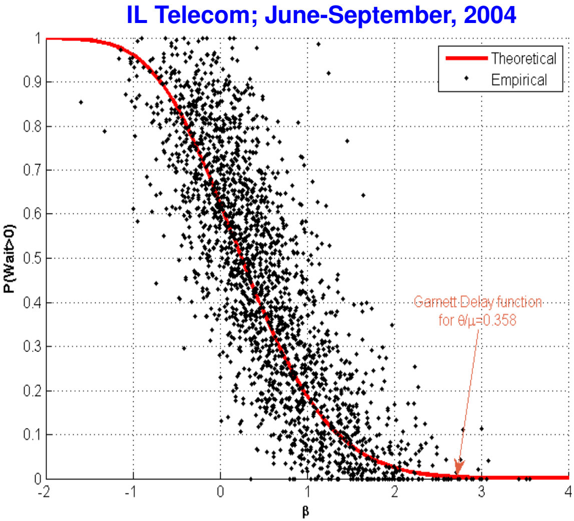
- **Probability to Abandon:**

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

- $\boldsymbol{P\{Ab\} \propto \mathrm{E}[W_q]}$ , both order $\frac{1}{\sqrt{N}}$ :

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W_q]} = \theta \quad (\approx g(0) > 0).$$

# QED Theory vs. Data: $P(W_q > 0)$




- 2205 half-hour intervals (13 summer weeks, week-days)
- Erlang-A approximations for the appropriate $\mu/\theta \approx 3$

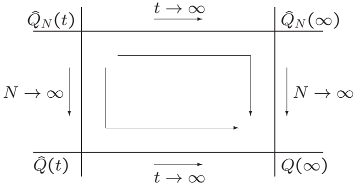
# Process Limits (Queueing, Waiting)

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

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

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

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




Approximating (Virtual) Waiting Time

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

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

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $N \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$, $P(I > 0) \uparrow 1$.

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

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $N \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$, $P(I > 0) \uparrow 1$.
   - Fix $N$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$, $P(I > 0) \downarrow 0$.

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

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $N \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$, $P(I > 0) \uparrow 1$.
   - Fix $N$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$, $P(I > 0) \downarrow 0$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $N$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $N \approx R + \beta\sqrt{R}$ $\left(\lambda \approx N\mu - \beta\sqrt{N\mu}\right)$

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

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

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

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $N \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$, $P(I > 0) \uparrow 1$.
   - Fix $N$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$, $P(I > 0) \downarrow 0$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $N$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $N \approx R + \beta\sqrt{R}$ $\left(\lambda \approx N\mu - \beta\sqrt{N\mu}\right)$

2. **Erlang-A** (M/M/N+M), with parameters $\lambda, \mu, \theta; N$, in which $\mu = \theta$: (Im)Patience and Service-times are equally distributed.
   - Steady-state: $L(M/M/N+M) \stackrel{d}{=} L(M/M/\infty) \stackrel{d}{=} Poisson(R)$, with $R = \lambda/\mu$ (Offered-Load)

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

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $N \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$, $P(I > 0) \uparrow 1$.
   - Fix $N$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$, $P(I > 0) \downarrow 0$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $N$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $N \approx R + \beta\sqrt{R}$ $\left(\lambda \approx N\mu - \beta\sqrt{N\mu}\right)$

2. **Erlang-A** (M/M/N+M), with parameters $\lambda, \mu, \theta$; $N$, in which $\mu = \theta$: (Im)Patience and Service-times are equally distributed.
   - Steady-state: $L(M/M/N+M) \stackrel{d}{=} L(M/M/\infty) \stackrel{d}{=} Poisson(R)$, with $R = \lambda/\mu$ (Offered-Load)
   - Poisson$(R) \stackrel{d}{\approx} R + Z\sqrt{R}$, with $Z \stackrel{d}{=} N(0, 1)$.

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

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $N \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$, $P(I > 0) \uparrow 1$.
   - Fix $N$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$, $P(I > 0) \downarrow 0$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $N$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $N \approx R + \beta\sqrt{R}$ $\left(\lambda \approx N\mu - \beta\sqrt{N\mu}\right)$

2. **Erlang-A** (M/M/N+M), with parameters $\lambda, \mu, \theta$; $N$, in which $\mu = \theta$: (Im)Patience and Service-times are equally distributed.
   - Steady-state: $L(M/M/N+M) \stackrel{d}{=} L(M/M/\infty) \stackrel{d}{=} Poisson(R)$, with $R = \lambda/\mu$ (Offered-Load)
   - Poisson$(R) \stackrel{d}{\approx} R + Z\sqrt{R}$, with $Z \stackrel{d}{=} N(0,1)$.
   - $P\{W_q(M/M/N+M) > 0\} \stackrel{PASTA}{=} P\{L(M/M/N+M) \geq N\} \stackrel{\mu=\theta}{=}$

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

     $P\{Z \geq (N-R)/\sqrt{R}\} \stackrel{\sqrt{\cdot}\text{ staffing}}{\approx} P\{Z \geq \beta\} = 1 - \Phi(\beta).$

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

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix $\lambda$ and let $N \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$, $P(I > 0) \uparrow 1$.
   - Fix $N$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$, $P(I > 0) \downarrow 0$.
   - $\Rightarrow$ **Must** have <u>both</u> $\lambda$ and $N$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $N \approx R + \beta\sqrt{R}$ $\left(\lambda \approx N\mu - \beta\sqrt{N\mu}\right)$

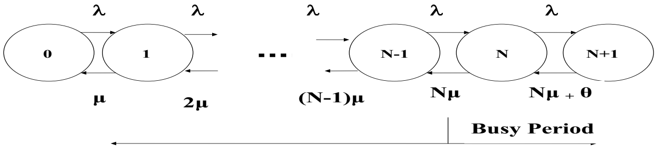
2. **Erlang-A** (M/M/N+M), with parameters $\lambda, \mu, \theta$; $N$, in which $\mu = \theta$: (Im)Patience and Service-times are equally distributed.
   - Steady-state: $L(M/M/N+M) \stackrel{d}{=} L(M/M/\infty) \stackrel{d}{=} Poisson(R)$, with $R = \lambda/\mu$ (Offered-Load)
   - Poisson$(R) \stackrel{d}{\approx} R + Z\sqrt{R}$, with $Z \stackrel{d}{=} N(0,1)$.
   - $P\{W_q(M/M/N+M) > 0\} \stackrel{PASTA}{=} P\{L(M/M/N+M) \geq N\} \stackrel{\mu=\theta}{=}$

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

     $P\{Z \geq (N-R)/\sqrt{R}\} \stackrel{\sqrt{\cdot}\ staffing}{\approx} P\{Z \geq \beta\} = 1 - \Phi(\beta)$.

3. QED **Excursions**

# QED Intuition via Excursions: Busy-Idle Cycles




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

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

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

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

# QED Intuition via Excursions: Asymptotics

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

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

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

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

$$P(\text{Server Wait} > 0) = P(\text{Customer Wait} = 0)$$

# QED Intuition via Excursions: Asymptotics

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

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

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

Hence, $P(\text{Customer Wait} > 0) \sim \left[1 + \dfrac{h(\delta)/\delta}{h(-\beta)/\beta)}\right]^{-1}$, and

$$P(\text{Server Wait} > 0) = P(\text{Customer Wait} = 0)$$

Special case: $\boldsymbol{\mu = \theta}$ (Impatient):

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

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

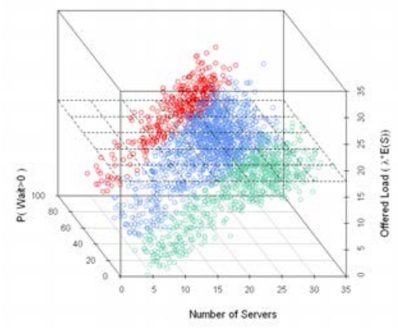
# QED Erlang-X (Markovian Q's: Performance Analysis)

- Pre-History, 1914: **Erlang** (Erlang-B = M/M/n/n, Erlang-C = M/M/n)
- Pre-History, 1974: Jagerman (Erlang-B)
- History Milestone, 1981: **Halfin-Whitt** (Erlang-C, GI/M/n)
- Erlang-A (M/M/N+M), 2002: w/ **Garnett** & Reiman
- Erlang-A with General (Im)Patience (M/M/N+G), 2005: w/ Zeltyn
- Erlang-C (ED+QED), 2009: w/ Zeltyn
- Erlang-B with Retrial, 2010: Avram, Janssen, van Leeuwaarden
- Refined Asymptotics (Erlang A/B/C), 2008-2011: Janssen, van Leeuwaarden, Zhang, Zwart
- 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

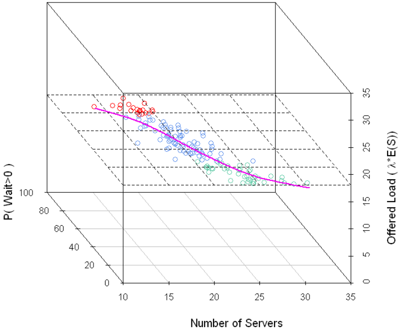
# Operational Regimes: Q(uality) vs. E(fficiency)

## CC in a Large Israeli Bank

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




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




**3 Operational Regimes**:

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

# Operational Regimes: Conceptual Framework

**$R$: Offered Load**

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

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

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

# Operational Regimes: Conceptual Framework

***R*: Offered Load**

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

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

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

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

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

# Operational Regimes: Conceptual Framework

**$R$: Offered Load**

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

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

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

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

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

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

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

# Operational Regimes: Conceptual Framework

**$R$: Offered Load**

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

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

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

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

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

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

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

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

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

# 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 | $\frac{1}{1+\delta}$ | $1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat\beta)}{\sqrt{n}}$ | $1$ | $\frac{1}{1+\delta}$ | $1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat\beta)}{\sqrt{n}}$ | $1$ | $\frac{1}{1+\delta}$ | $1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat\beta)}{n}$ | $1$ |
| $\mathbb{E}(Q)$ | $\frac{1}{\delta(1+\delta)}$ | $\sqrt{n}g(\hat\beta)$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $\frac{1}{\delta}\varrho_n$ | $\sqrt{n}g(\hat\beta)\alpha$ | $\frac{n\mu\gamma}{\theta(1-\gamma)}$ | $o(1)$ | $ng(\hat\beta)$ | $\frac{n^2\mu\gamma}{\theta(1-\gamma)}$ |
| $\mathbb{P}(Ab)$ | $\frac{1}{n}\frac{1}{\delta}\frac{\theta}{\mu}$ | $\frac{\theta}{\sqrt{n\mu}}g(\hat\beta)$ | $\gamma$ | $\frac{1}{n}\frac{(1+\delta)}{\delta}\frac{\theta}{\mu}\varrho_n$ | $\frac{\theta}{\sqrt{n\mu}}g(\hat\beta)\alpha$ | $\gamma$ | $o(\frac{1}{n^2})$ | $\frac{\theta}{n\mu}g(\hat\beta)$ | $\gamma$ |
| $\mathbb{P}(W_q>0)$ | $\frac{1}{1+\delta}$ | $\approx 1$ | | $\varrho_n$ | $\alpha\in(0,1)$ | $\approx 1$ | $\approx 0$ | $\approx 1$ | |
| $\mathbb{P}(W_q>T)$ | $\frac{1}{1+\delta}e^{-\frac{\delta}{1+\delta}\mu T}$ | $1+O(\frac{1}{\sqrt{n}})$ | $1+O(\frac{1}{n})$ | $\approx 0$ | | $f(T)$ | $\approx 0$ | $\frac{\bar\Phi(\hat\beta+\sqrt{\theta\mu}T)}{\bar\Phi(\hat\beta)}$ | $1+O(\frac{1}{n})$ |
| 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), Garnett-M-Reiman (02)
  - ED: Whitt (04)
  - NDS: Atar (12)

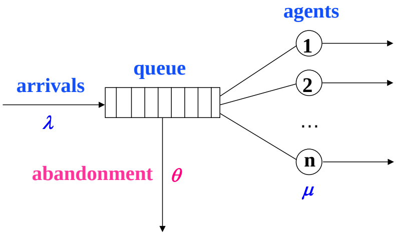
# Asymptotic Landscape: 9 Operational Regimes, and then some

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

| Erlang-A<br>$\mu$ & $\theta$ 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 $\frac{\mathbb{E}W_q}{\mathbb{E}S}$ | $\frac{1}{\delta}$ | $\sqrt{n}g(\hat\beta)$ | $n\mu\gamma/\theta$ | $\frac{1}{n}\frac{(1+\delta)}{\delta}\varrho_n$ | $\frac{\alpha}{\sqrt{n}}g(\hat\beta)$ | $\frac{\mu\gamma}{\theta}$ | $o(\frac{1}{n})$ | $g(\hat\beta)$ | $n\mu\gamma/\theta$ |

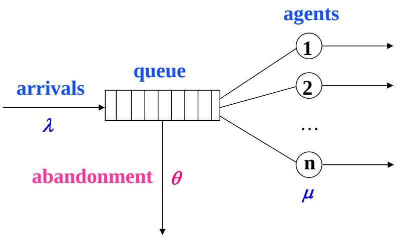
- Conventional: Ward & Glynn (03, $G/G/1+G$)
- Many-Server:
  - QED: Halfin-Whitt (81), Garnett-M-Reiman (02)
  - ED: Whitt (04)
  - NDS: Atar (12)
- **"Missing"**: ED+QED; Hazard-rate scaling (M/M/N+G); Time-Varying, Non-Parametric; Moderate- and Large-Deviation; Networks; Control

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




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

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

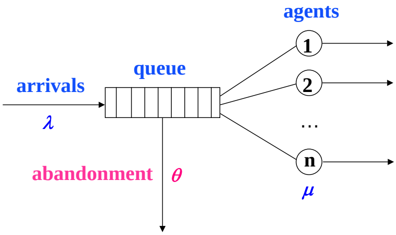



w/ **I. Gurvich & J. Huang**

- **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**

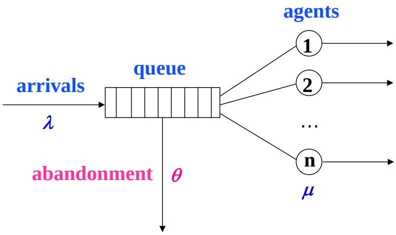
- **QNet**: Birth & Death Queue, with B - D rates

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

- **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

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

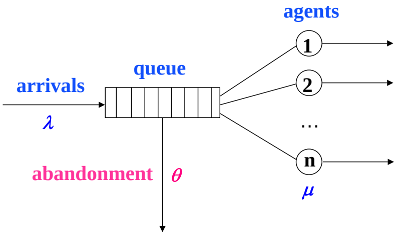
- **FNet**: Dynamical (Deterministic) System – ODE

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

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

**Accuracy** increases as $\lambda \uparrow \infty$ (no additional assumptions)

# Value of Universal Approximation

- Tractable - closed-form stable expressions
- Accurate - more than heavy traffic limits
- Robust - all many-server regimes, and beyond, with hardly any assumptions
- Value
  - Performance Analysis
  - Optimization (Staffing)
  - Inference (**w/ G. Pang**)
  - Simulation (**w/ J. Blanchet**)
- Limitation: Steady-State (but working on it)

Why does it work so well?

Coupling **"Busy" + "Idle" Excursions** of B&D and the corresponding Diffusion (durations order $\frac{1}{\sqrt{\lambda}}$)

# Universal Diffusion: Tractability

- Density function of $Y(\infty) - n$:

$$\pi(x) = \begin{cases} \frac{\sqrt{\mu}}{\sqrt{\lambda}} \frac{\phi(\sqrt{\mu}(x/\sqrt{\lambda}+\beta/\mu))}{\Phi(\beta/\sqrt{\mu})} p(\beta,\mu,\theta), & \text{if } x \le 0, \\ \frac{\sqrt{\theta}}{\sqrt{\lambda}} \frac{\phi(\sqrt{\theta}(x/\sqrt{\lambda}+\beta/\theta))}{1-\Phi(\beta/\sqrt{\theta})} (1 - p(\beta,\mu,\theta)), & \text{if } x > 0, \end{cases}$$

Here $\beta := (n\mu - \lambda)/\sqrt{\lambda}$
and

$$p(\beta,\mu,\theta) = \left[ 1 + \sqrt{\frac{\mu}{\theta}} \frac{\phi(\beta/\sqrt{\mu})}{\Phi(\beta/\sqrt{\mu})} \frac{1 - \Phi(\beta/\sqrt{\theta})}{\phi(\beta/\sqrt{\theta})} \right]^{-1}.$$

# Universal Approximation: Accuracy

- $\Delta^\lambda$ 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 = $n + \mathcal{O}(\sqrt{\lambda})$)

# Universal Approximation: Accuracy

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

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

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

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 ($m \geq 0$):

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

- **Accuracy**: higher than heavy-traffic *limits*

# Universal Approximation: 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:

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

- $\text{dMartingale}_t \approx \sqrt{2\lambda} \cdot \text{dBrownian}_t$

# Reconciling Steady-State and Time-Varying Models

- **Challenge**: Accommodate time-varying demand (routine)
- **Prerequisite**: **Flexible Capacity**
  - As in Call Centers and to a degree in Healthcare,
  - In contrast to **rigid** (fixed) staffing level during a shift: doomed to alternate between overloading and underloading

# Reconciling Steady-State and Time-Varying Models

- **Challenge**: Accommodate time-varying demand (routine)
- **Prerequisite**: **Flexible Capacity**
  - As in Call Centers and to a degree in Healthcare,
  - In contrast to **rigid** (fixed) staffing level during a shift: doomed to alternate between overloading and underloading
- **Idea/Goal**: In the face of time-varying demand, design time-varying staffing which accommodates demand such that performance is stable over time
- **Solution**: In fact, a time-varying system with **Steady-State** performance, at all times, via (Modified) **Offered-Load** (Square-Root) Staffing.

# Reconciling Steady-State and Time-Varying Models

- **Challenge**: Accommodate time-varying demand (routine)
- **Prerequisite**: **Flexible Capacity**
  - As in Call Centers and to a degree in Healthcare,
  - In contrast to **rigid** (fixed) staffing level during a shift: doomed to alternate between overloading and underloading
- **Idea/Goal**: In the face of time-varying demand, design time-varying staffing which accommodates demand such that performance is stable over time
- **Solution**: In fact, a time-varying system with **Steady-State** performance, at all times, via (Modified) **Offered-Load** (Square-Root) Staffing.
- **History**:
  - **Jennings, M., Reiman, Whitt** (1996): Emergence of the phenomenon, via infinite-server heuristics
  - **Feldman, M., Massey, Whitt** (2008): Stabilize delay probability via QED staffing (justified theoretically only for Erlang-A with $\mu = \theta$)
  - **Liu and Whitt** (ongoing): Stabilize abandonment probability by ED staffing, via a corresponding network, theoretically and empirically
  - **Huang, Gurvich, M.** (ongoing): QED theory

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

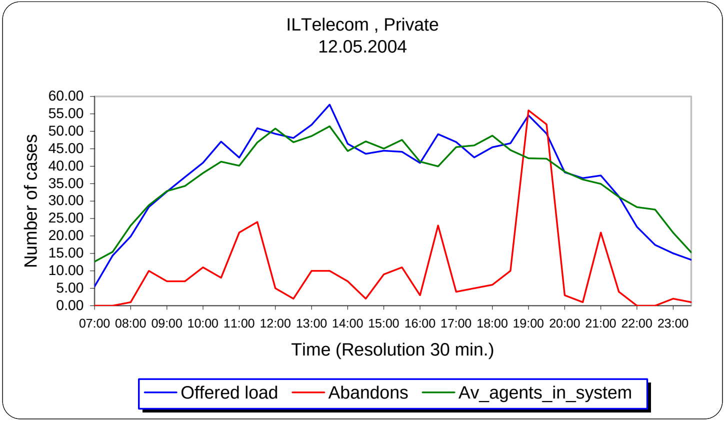
**Empirically** (in SEEStat):

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

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

**Empirically** (in SEEStat):

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



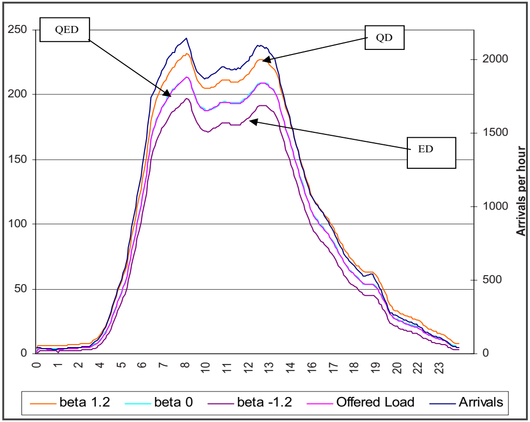
# Time-Varying Arrival Rates

## Square-Root Staffing:

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

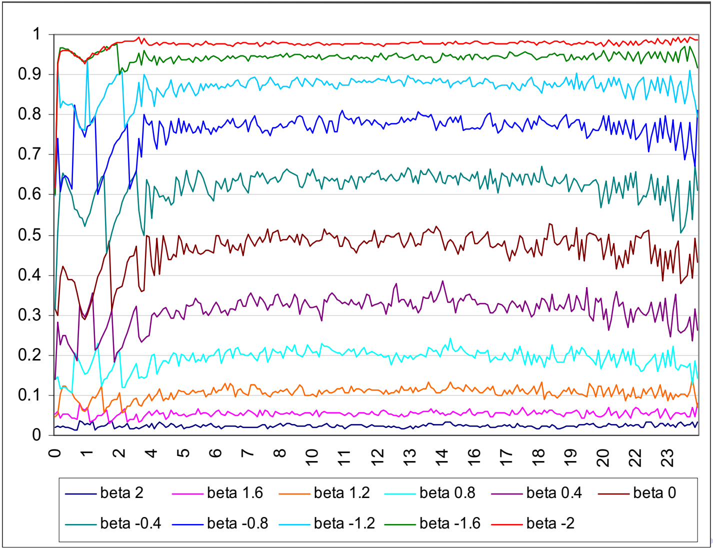
$R(t)$ is the **Offered-Load** at time $t$ $\quad$ ( $R(t) \neq \lambda(t) \times$ E[S] )

### Arrivals, Offered-Load and Staffing



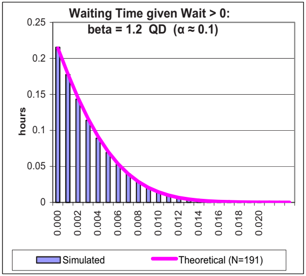
# Time-Stable Performance of Time-Varying Systems

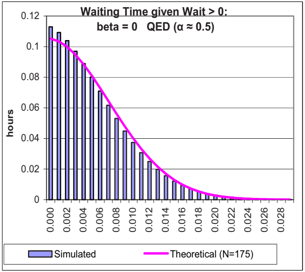
**Delay Probability** = as in the **Stationary Erlang-A / R**

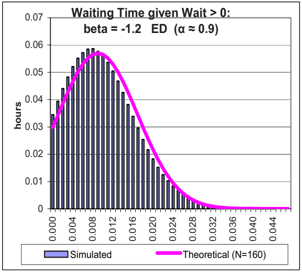


# Time-Stable Performance of Time-Varying Systems

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




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

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

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

- $\{\boldsymbol{A(t)},\ t \geq 0\}$ 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 $R(t) \equiv \boldsymbol{\lambda \times E[S]}$.

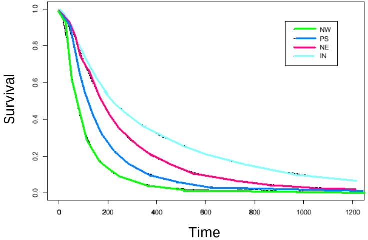
**QED-c**: $\boldsymbol{N_t = R_t + \beta R_t^c}$, $1/2 \leq c < 1$; ($c = 1$ separate analysis).

# Estimating / Predicting the Offered-Load

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

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

## Survival Functions by Type (Small Israeli Bank)



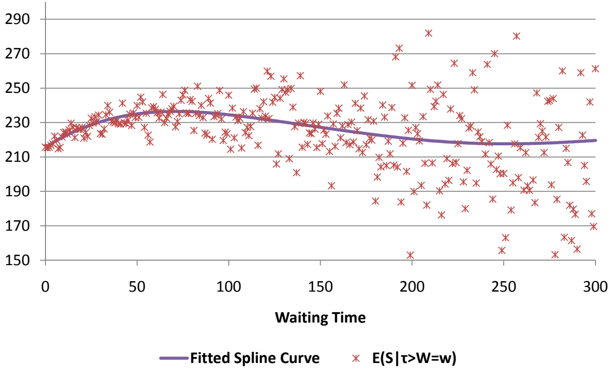

**Service times** stochastic order: $S_{New} \overset{st}{<} S_{Reg} \overset{st}{<} S_{VIP}$

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

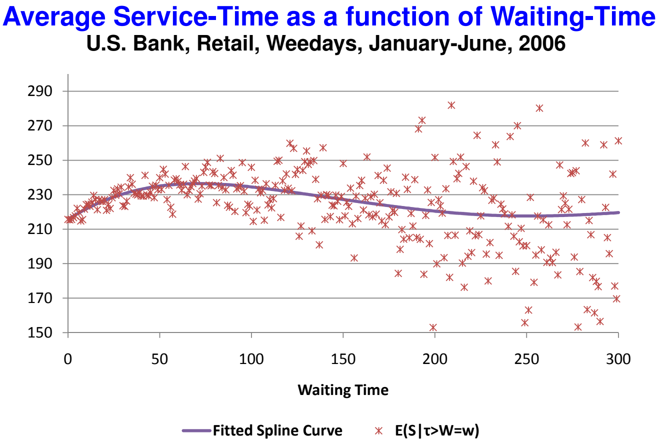
# Dependent Primitives: Service- vs. Waiting-Time

## Average Service-Time as a function of Waiting-Time

**U.S. Bank, Retail, Weedays, January-June, 2006**



# Dependent Primitives: Service- vs. Waiting-Time




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

# (Imputing) Service-Times of Abandoning Customers

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

1. **Estimate** $g(w) = E[S \,|\, \text{Patience} > \text{Wait} = w],\ w \geq 0$:
   Mean service time of those **served after waiting exactly** $w$ units of time (via non-linear regression: $S_i = g(W_i) + \varepsilon_i$);

2. **Calculate**

$$E[S \,|\, \text{Patience} = w] = g(w) - \frac{g'(w)}{h_\tau(w)}\ ;$$

   $h_\tau(w)$ = hazard-rate of (im)patience (via un-censoring);

3. **Offered-load** calculations: Impute $E[S \,|\, \text{Patience} = w]$ (or the conditional distribution).

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

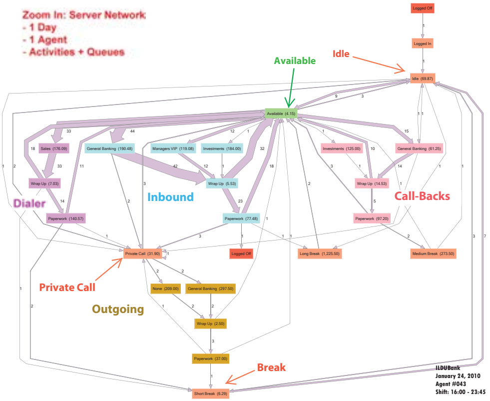
# Extending the Notion of the “Offered-Load"

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

# Server Networks

w/ **A. Senderovic: Planning**

then joined **A. Gal, M. Weid, Y. Goldberg: Process Mining**



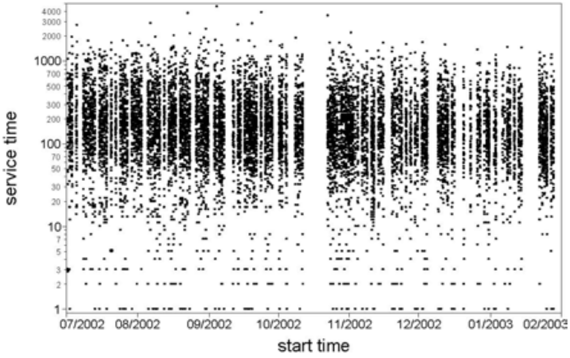

- As **Challenging** - in theory and practice - as Customer Networks
- **Uncharted territory**: e.g. Gnedenko/Palm's Machine Repair model, **Armony and Ward** (Asymptotic Little & ASTA)

# Individual Agents: Service-Duration, Variability

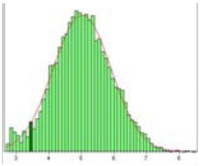
w/ **Gans, Liu, Shen & Ye**

**Agent 14115**

**Service-Time Evolution: 6 month**




**Log(Service-Time)**

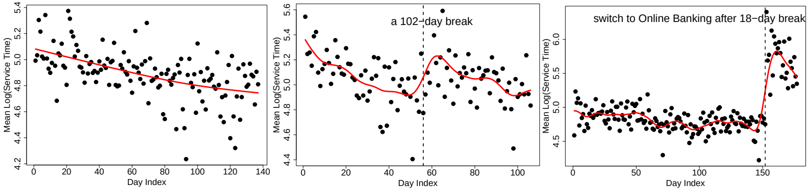


- **Learning**: Noticeable decreasing-trend in service-duration
- **LogNormal** Service-Duration, individually and collectively

# Individual Agents: Learning, Forgetting, Switching

## Daily-Average Log(Service-Time), over 6 months

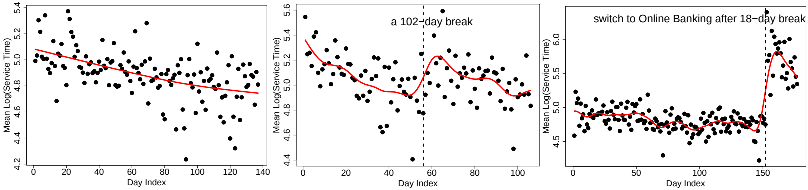
## Agents 14115, 14128, 14136



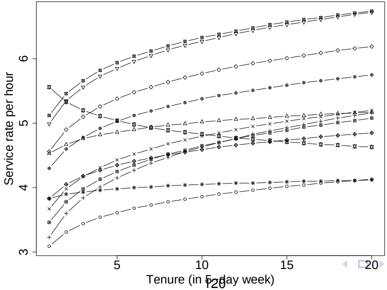
# Individual Agents: Learning, Forgetting, Switching

## Daily-Average Log(Service-Time), over 6 months

## Agents 14115, 14128, 14136

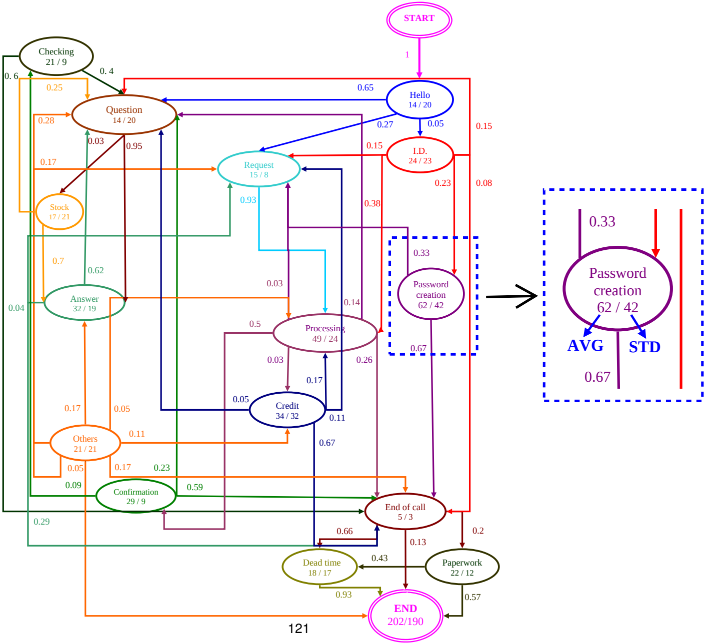



## Weakly Learning-Curves for 12 Homogeneous(?) Agents



# Tele-Service Process (Beyond present SEE Data)

**Retail Service (Israeli Bank)**




**Statistics**
**OR**
**IE**
**Psychol.**
**MIS**

# Why Bother?

- Server Networks: Unchartered Research Territory
- In large call centers:
  **+One Second** to Service-Time implies **+Millions** in costs, annually

  $\Rightarrow$ **Time and "Motion" Studies** (**Classical IE** with New-age IT)

# Why Bother?

- Server Networks: Unchartered Research Territory
- In large call centers:
  **+One Second** to Service-Time implies **+Millions** in costs, annually

  $\Rightarrow$ **Time and "Motion" Studies** (**Classical IE** with New-age IT)

- **Service-Process Model**: Customer-Agent Interaction
  - **Work Design** (w/ **Khudiakov**)
    eg. **Cross-Selling**: higher profit vs. longer (costlier) services;
    Analysis yields (congestion-dependent) cross-selling protocols
  - **“Worker" Design** (w/ **Gans, Liu, Shen & Ye**)
    eg. **Learning, Forgetting,** . . . : Staffing & individual-performance prediction, in a heterogenous environment

# Why Bother?

- Server Networks: Unchartered Research Territory
- In large call centers:
  **+One Second** to Service-Time implies **+Millions** in costs, annually

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

- **Service-Process Model**: Customer-Agent Interaction
  - **Work Design** (w/ **Khudiakov**)
    eg. **Cross-Selling**: higher profit vs. longer (costlier) services;
    Analysis yields (congestion-dependent) cross-selling protocols
  - **“Worker" Design** (w/ **Gans, Liu, Shen & Ye**)
    eg. **Learning, Forgetting,** . . . : Staffing & individual-performance prediction, in a heterogenous environment

- **IVR-Process Model**: Customer-Machine Interaction
  **75% bank-services**, poor design, yet scarce research;
  Same approach, automatic (easier) data (**w/ N. Yuviler**)

# ServNets: Data-Based Online Automatic Creation

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

- **ServNets = QNets, SimNets, FNets, DNets**
- **SimNets** of Service Systems = **Virtual Realities**
- SimNets **also** of QNEts, FNets, DNets
  eg. ED MD (Physics): Where are the Differential Equations?

# ServNets: Data-Based Online Automatic Creation

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

- **ServNets = QNets, SimNets, FNets, DNets**
- **SimNets** of Service Systems = **Virtual Realities**
- SimNets **also** of QNEts, FNets, DNets
  eg. ED MD (Physics): Where are the Differential Equations?
- Ultimately: **Research Labs**, offering universal access to data and ServNets, will become necessary (hence must be funded!)

# ServNets: Data-Based Online Automatic Creation

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

- **ServNets = QNets, SimNets, FNets, DNets**
- **SimNets** of Service Systems = **Virtual Realities**
- SimNets **also** of QNEts, FNets, DNets

  eg. ED MD (Physics): Where are the Differential Equations?
- Ultimately: **Research Labs**, offering universal access to data and ServNets, will become necessary (hence must be funded!)
- Data-based Research: Tradition in Physics, Chemistry, Biology; Psychology (now also in Transportation (Science) and (Behavioral) Economics)
- **Why not in Service Science / Engineering / Management ?**

# ServNets: Data-Based Online Automatic Creation

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

- **ServNets = QNets, SimNets, FNets, DNets**
- **SimNets** of Service Systems = **Virtual Realities**
- SimNets **also** of QNEts, FNets, DNets
  eg. ED MD (Physics): Where are the Differential Equations?
- Ultimately: **Research Labs**, offering universal access to data and ServNets, will become necessary (hence must be funded!)
- Data-based Research: Tradition in Physics, Chemistry, Biology; Psychology (now also in Transportation (Science) and (Behavioral) Economics)
- **Why not in Service Science / Engineering / Management ?**
- **Moreover, address the Reproducibility Crisis in Scientific Research (Computation, Massive-data, . . .)**

## Data-Based Creation 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 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)