"Theompirical" OR (IE/OM/MS), or Theoretical & Empirical Journeys through Service Systems (Hospitals, Call Centers, Banks,…)

Avishai Mandelbaum

IE&M & SEELab, Technion

link

Research Partners (30+ years)

Students:
Aldor*, Baron Yonit*, Carmeli-Yuviler Nitzan, Carmeli Boaz*, Chen Hong*, Cohen Izik*, Feldman Zohar*, Garnett*, Ghebali*, Gurvich*, Khudiakov*, Koren*, Maman*, Marmor*, Reich*, Rosenshmidt*, Shaikhet*, Senderovich*, Tseytlin*, Yom-Tov*, Zaied*, Zeltyn*, Zychlinski, Zohar Eti*, Zviran*, …

Theory:
Armony, Atar, Azriel, Chen Hong, Cohen Izik*, Gurvich*, Feigin, Gal, Huang Junfei*, Jelenkovic, Kaspi, Massey, Momcilovic, Reiman, Shimkin, Stolyar, Trichakis, Trofimov, Wasserkrug, Whitt, Yom-Tov*, Zeltyn*, Zhang Jiheng, Zhang Hanqin, …

Exploratory Data Analysis, Data Sources, Statistics, Projects:
Brown, Gans, Shen Haipeng*, Sakov, Zhao Linda; Zeltyn*; Ritov, Goldberg*; Gurvich*, Huang Junfei*, Liberman*; Liu Nan, Ye Han; Armony, Marmor*, Tseytlin*, Yom-Tov*; Gorfine, Ghebali*; Tezcan; Kim Song-Hee, Won Chul Cha; Feigin, Azriel*; Rafaeli; Momcilovic, Trichakis; Bunnell, Kadish, Leib; …

Industry:
Mizrahi Bank, Fleet Bank, Rambam Hospital, IBM Research, Hapoalim Bank, Pelephone Cellular, Samsung Hospital, Dana Farber Cancer Institute, LivePerson, CheetahLabs,…

Technion SEE Laboratory (SEELab):
Feigin; Trofimov, Nadjharov, Gavako; Kutsy; Senderovic*, Carmeli*; Liberman*, Koren*, Plonsky*;
Research Assistants, Visitors, Postdocs, …

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

2 Prerequisites: Data, Models

- Data: The language of
  - ✓ Multi-disciplinary research (e.g. OR + Psychology + NLP)
  - ✓ Academia-Industry partnerships (e.g. university & bank)

- Models: Simple models at the service of complex realities (not too simple)
  - ✓ Insights often rooted in deep mathematics
  - ✓ Stochastic networks: Empirical; QNets, FNets, DNets; SimNets

Service Engineering (& Science)

Background: Math, CS, Stat, OR;
Science, Management, Engineering

Playground: Service Systems (Call Centers, Hospitals, Banks,…)

Theoretical Framework: Asymptotics (of Queueing/Stochastic Networks)

Empirical Support: Technion SEELab

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) ($\surd$-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: Accuracy vs. Value

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

- I am of the opinion that the answer to such questions **could and should have been "yes"** more often than it has been.

- How to change this? And why be optimistic?

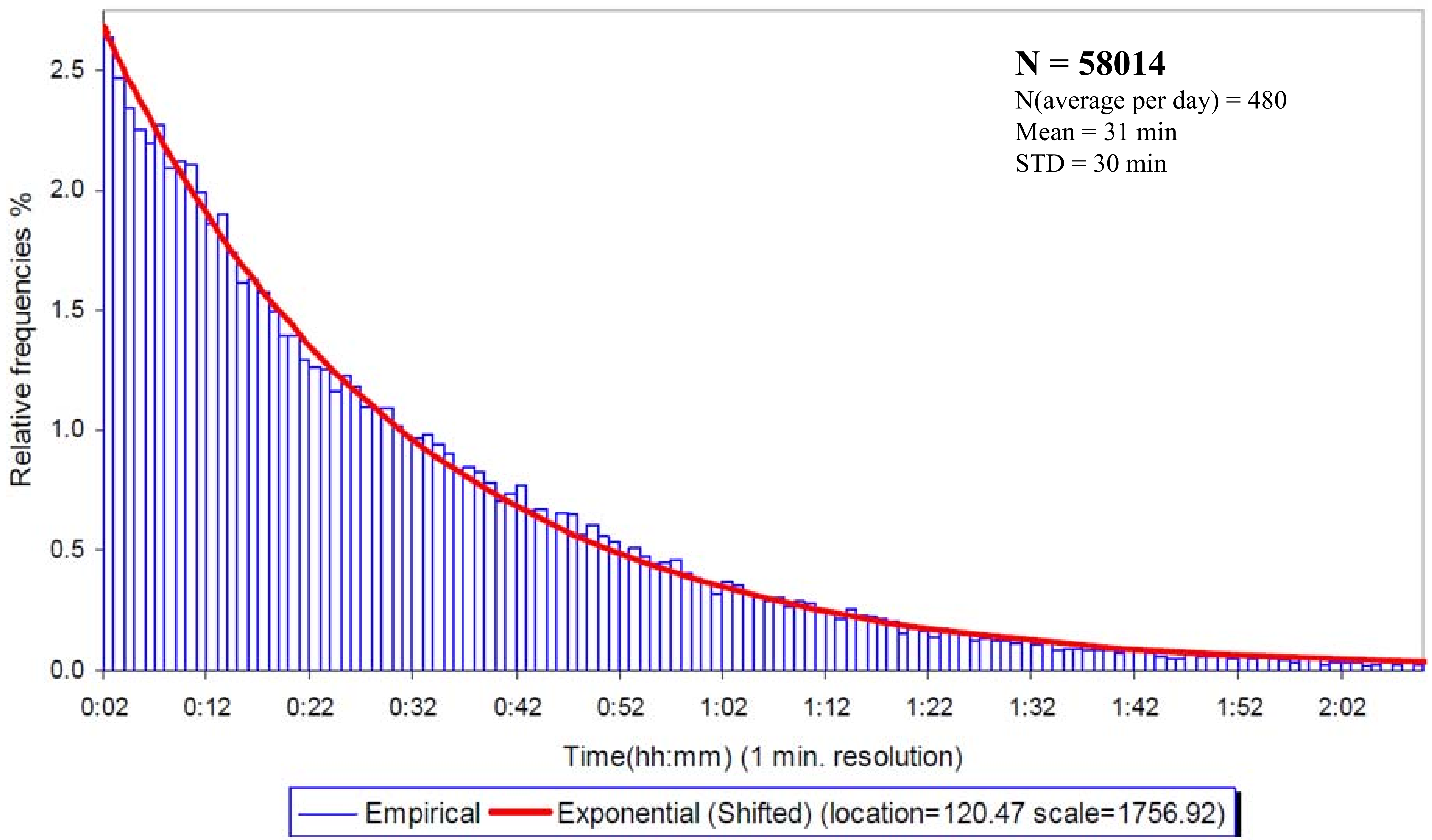
  My approach has been to **marry theory with data (theompirical research)**, supported by (what I only recently came to realize is) **Process Mining (= of Stochastic Networks**: their building-blocks, structure, <u>protocols</u>; flows and laws).

Pause for a Commercial (Behavioral OR) Mining Service Protocols

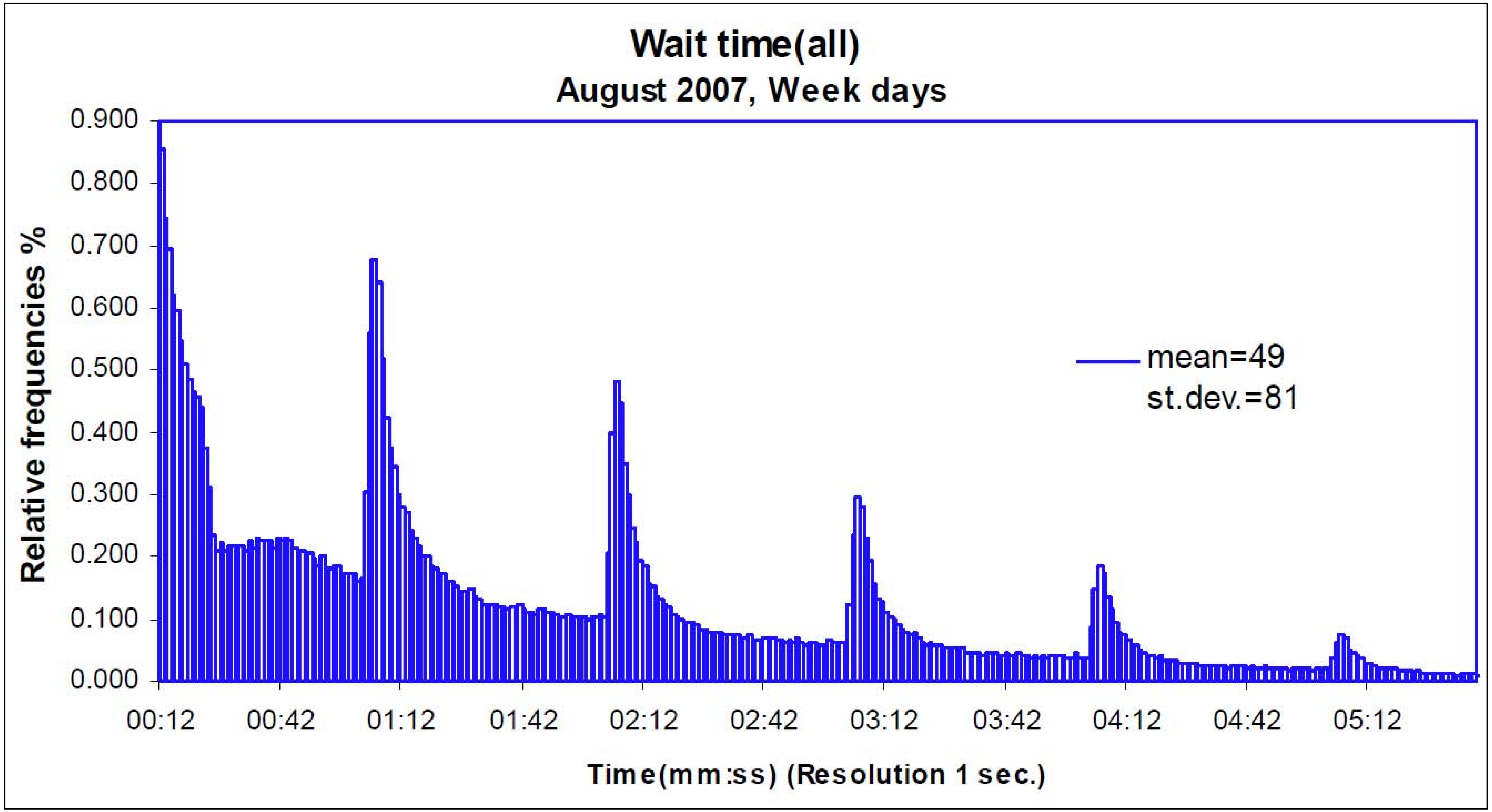
- FCFS
- Dynamic priority
- Random order (mixed with FCFS: exposes data limitations - must be there to figure it out)

Waiting-Time for Exam: Single-Server Queue (in HT)

November2013–May2014



Distribution of Waiting Time on the Phone



Back to Queueing Asymptotics: Accuracy vs. Value

Prevalent Asymptotic Approximations

System (Data)

Phenomenology

Accuracy

Models ⟷ Approximations

QNets SimNets

Data–Based Prevalent Asymptotic Approximations Models

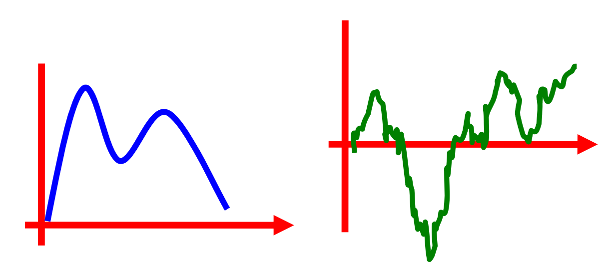
System (Data)

Phenomenology

Accuracy

Models ⟷ Approximations

QNets SimNets ⟷ FNet DNet Models



Data–Based Prev~~a~~lent Asymptotic Approx~~i~~mations Models

System (Data)

Phenomenology

Value

Acc~~u~~racy

Models

Appro~~x~~imations

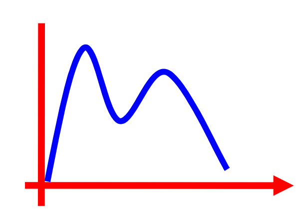
QNets

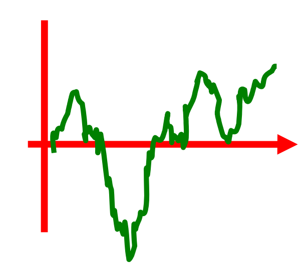
SimNets

FNet

DNet

Models





Data–Based Prev~~a~~lent Asymptotic Approx~~i~~mations Models

System (Data)

Phenomenology

Value

Acc~~u~~racy

Models ⟷ Approx~~i~~mations

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

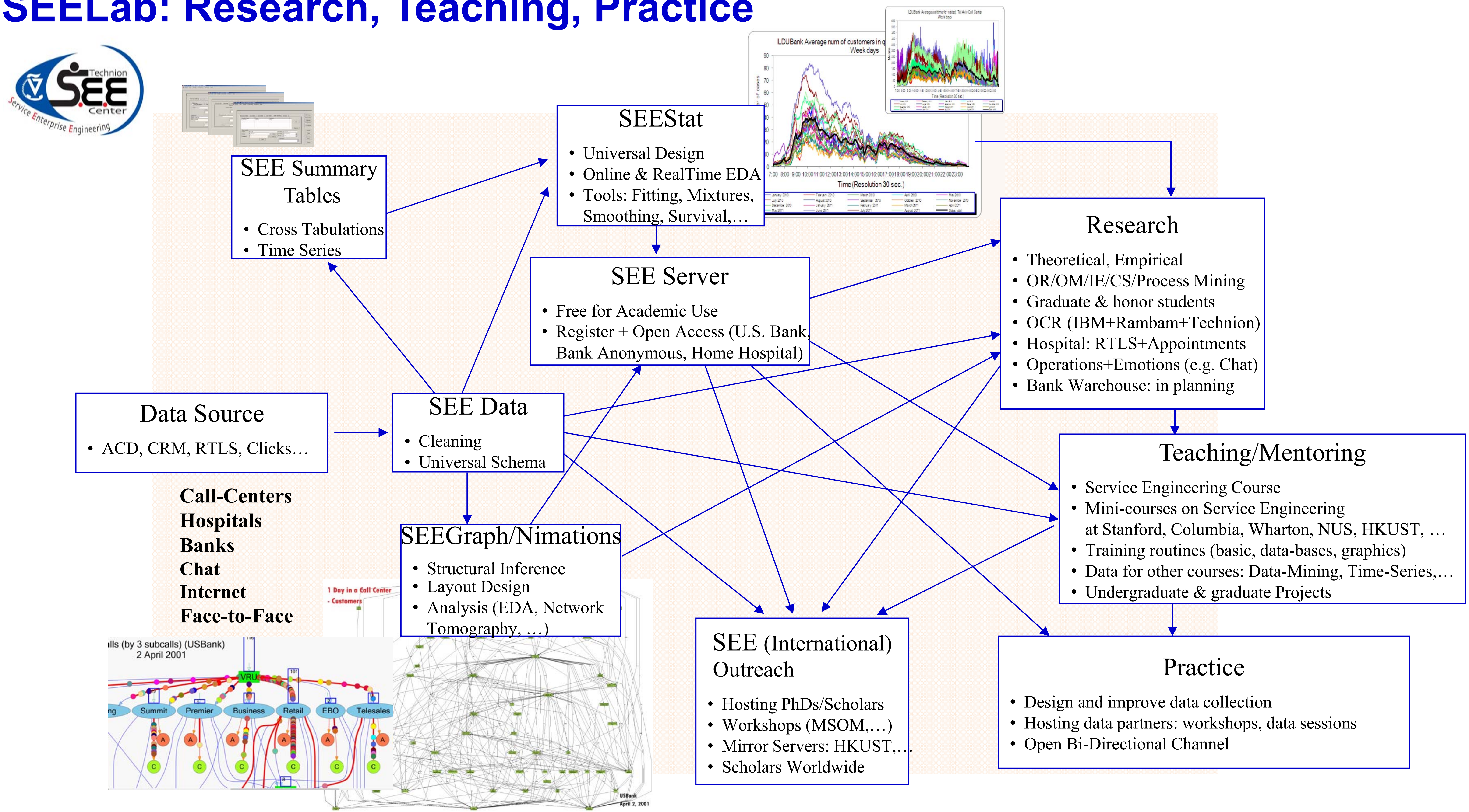
Normal, Brownian Motion ; Bachalier 1900

Poisson ; Poisson 1838

Technion

SEELab (2007- …)
Service Enterprise Engineering

SEELab: Research, Teaching, Practice



SEELab: Environment for graphical EDA

Operational histories (customers, servers) at the **individual-transaction level**, e.g.

1. ***Bank Anonymous Call-Center**: **1 year, 350K calls by 15 agents** - in 2000, **which paved the way to:**
2. ***U.S. Bank Call-Center** : **2.5 years, 220M calls, 40M by 1000 agents**
3-4. Israeli Cellular Company: **2.5 years, 110M calls, 25M calls by 750 agents;** ILBank (IVR, SBR): 2 years
5. Back to Bank Anonymous: Call-Center and more - **from January 2010, daily-deposit at a SEESafe**

6. Service Engineering **internet site**: click-stream data (2 years)

7. ***Home (Rambam) Hospital** : **4 years, 1000 beds**, inter-ward patient flow

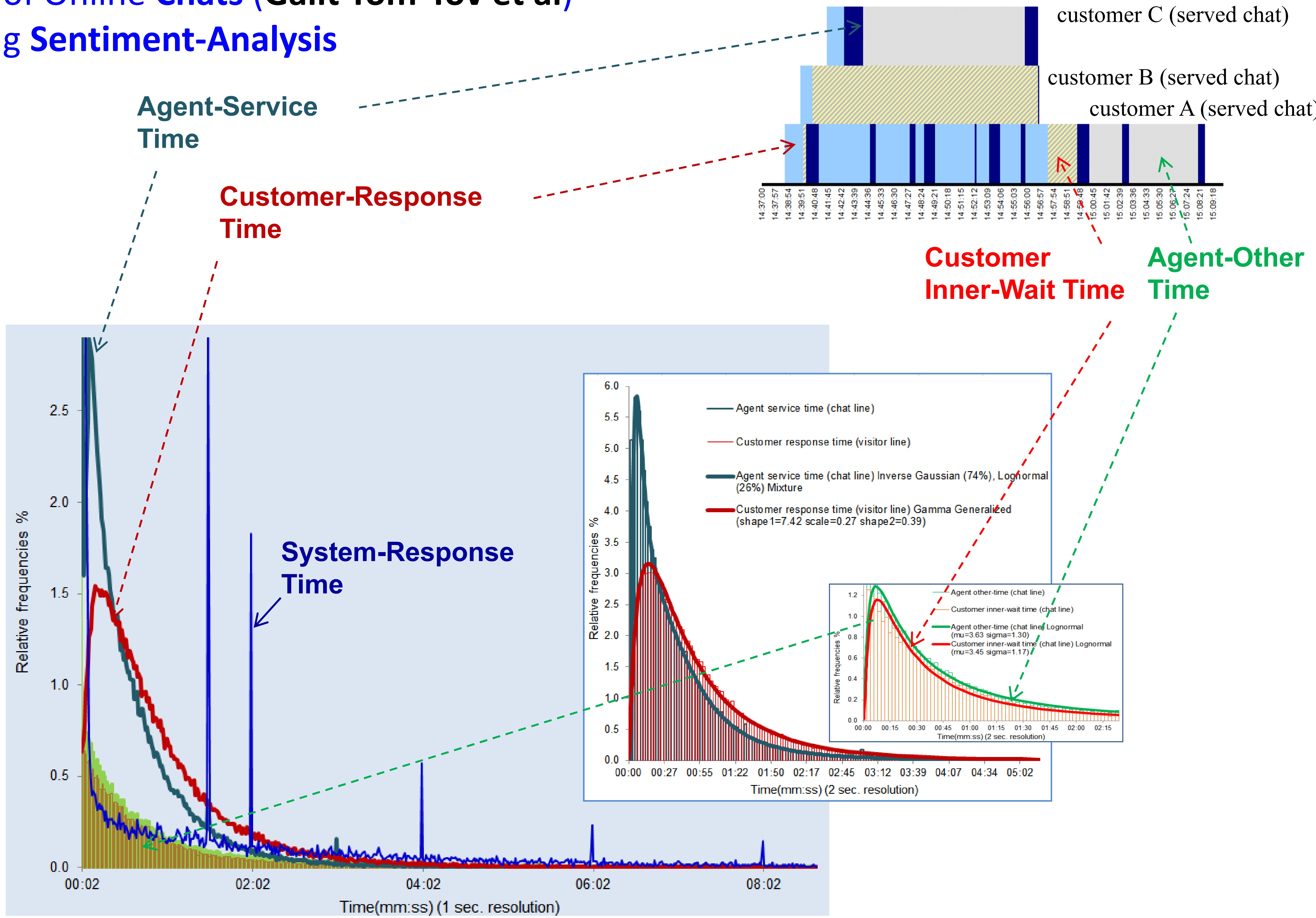
8. **Emergency Departments** (ED) patient flow:
   - 5 EDs in Israel: 1-2 years, **late David Sinreich**, ED arrivals & LOS
   - ED in Seoul: 2 months, **K. Song-Hee & W. Cha**, pilot
   - ED in Singapore: 2 years, pilot

9. **U.S. Ambulatory Hospital RTLS (Real-Time Location System): Since November 2013**
   - **250K events/day (1GB/week): 1000 patients, 350 staff (1500 tagged entities), every 3 sec's**
   - Infrastructure: **900 readers (sensors) over ceilings of 7 (now 8) clinical floors**
   - Both actual and planned (**appointment book** of resources: staff, patients, rooms)

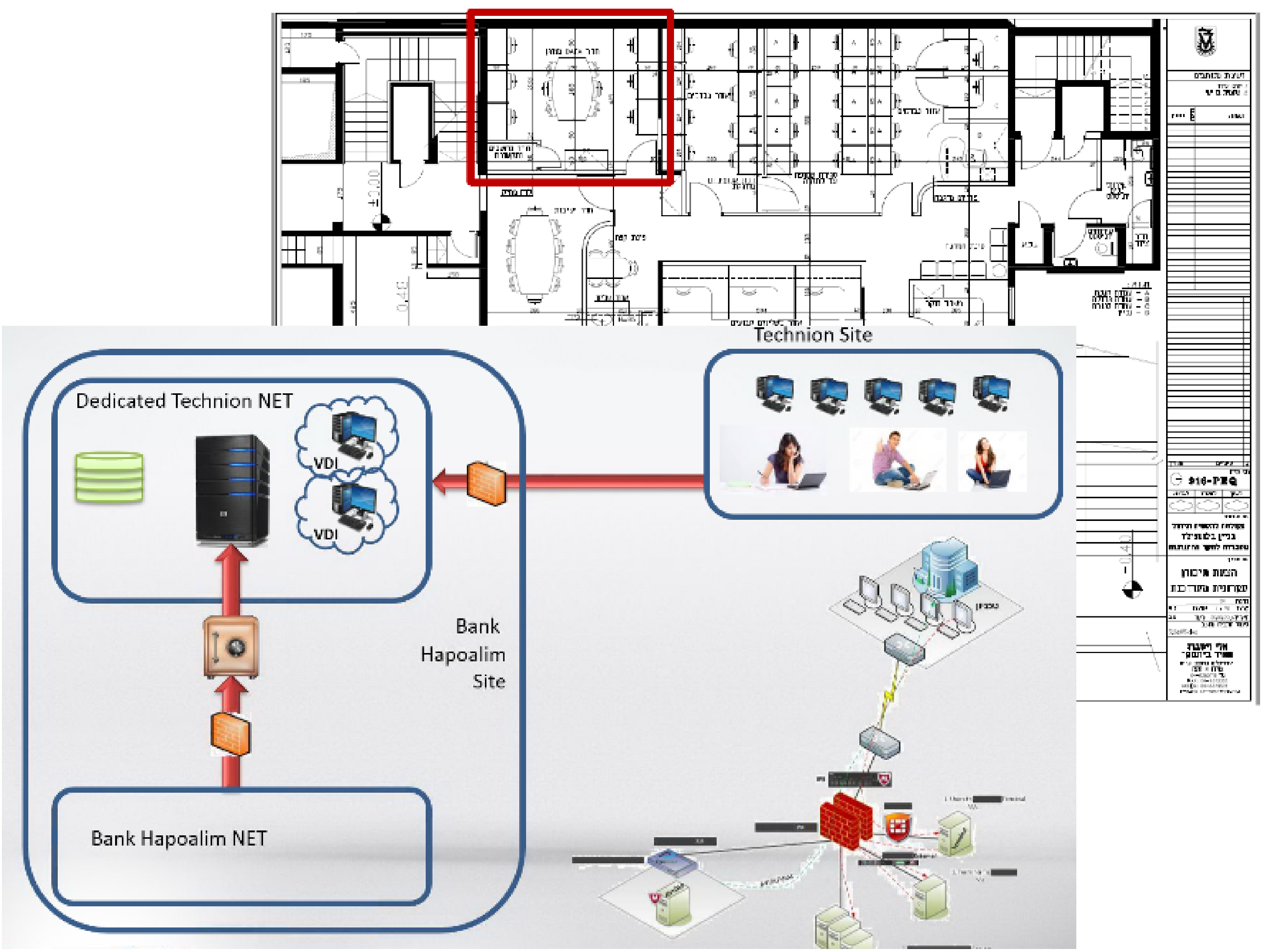
10-13:
Chat Services (Europe); ILBank Warehouse; Smart-City Simulator (Haifa, Boston,…); Justice System

***Open & Free for (reproducible) research and teaching**

Anatomy of Online **Chats** (**Galit Yom-Tov et al**)
Supporting **Sentiment-Analysis**



Data-Room (Secured): Accessing 80% of Bank Warehouse



- T-PADS מעבדה: איך, מי, איפה, ומתי
- מיקוד ראשוני Data Science & Engineering
- אפשרות הרחבה ל Blockchain, IoT וכו'
  - חוזה 5X5
  - בעלות משותפת IP -
  - פרסומים משותפים בהסכמה
  - דרך פרויקטים, לאימות המחקר POC
- סטטוס: 1ח, ועדת היגוי - 6ח, יום אירוע מרוכז- 1ש

Smart-City "Simulator" (40K sq ft "playground", 10K offices)

Emergency services

Traffic lights

Lightening

Smart parking

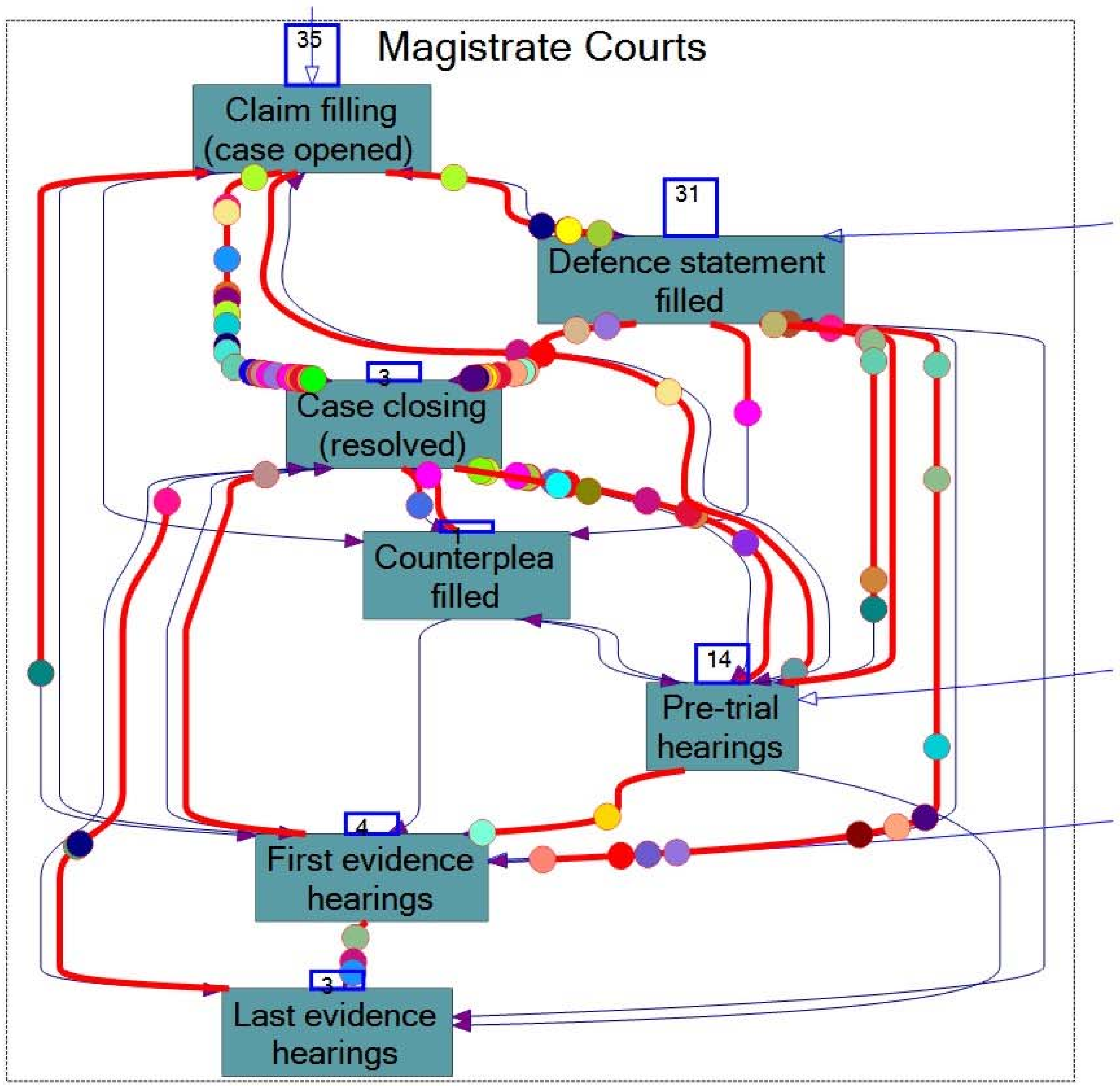
Water, sewage

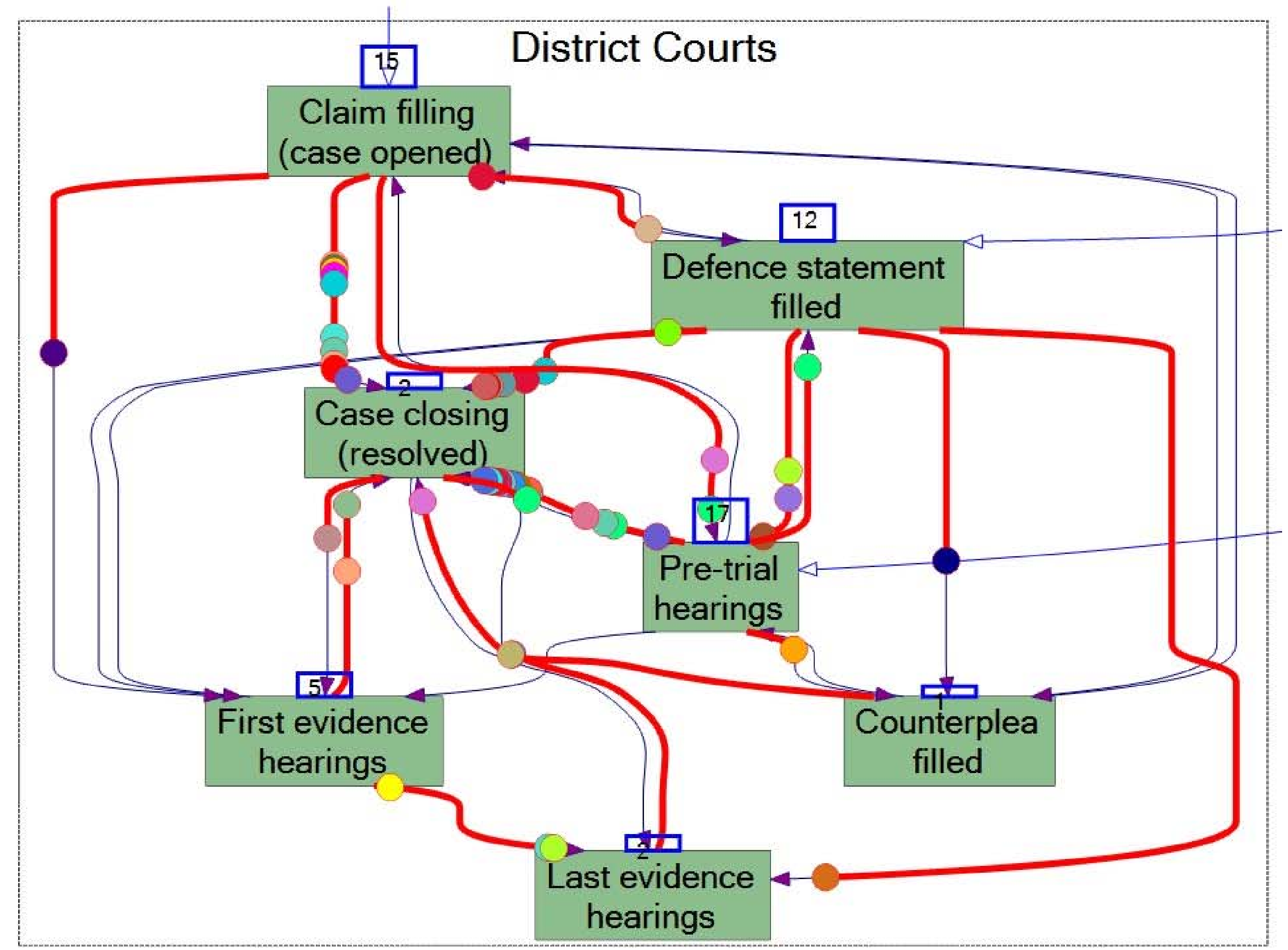
Energey systems

600 מ"ר משרדים + **2.5 דונם** של "מגרש משחקים" + 2 דונם הרחבה עתידית

CHEETAHLABS

“Production of Justice”: Case-Flow in the Justice System





Pause for a Commercial: from the Skeptics

The Skeptics

ERC: on using data to motivate theory (as in say Physics, Biology,…)

How can breakthrough mathematics come out of so much data?

NSF: on funding data-collection and maintenance in OR research grants

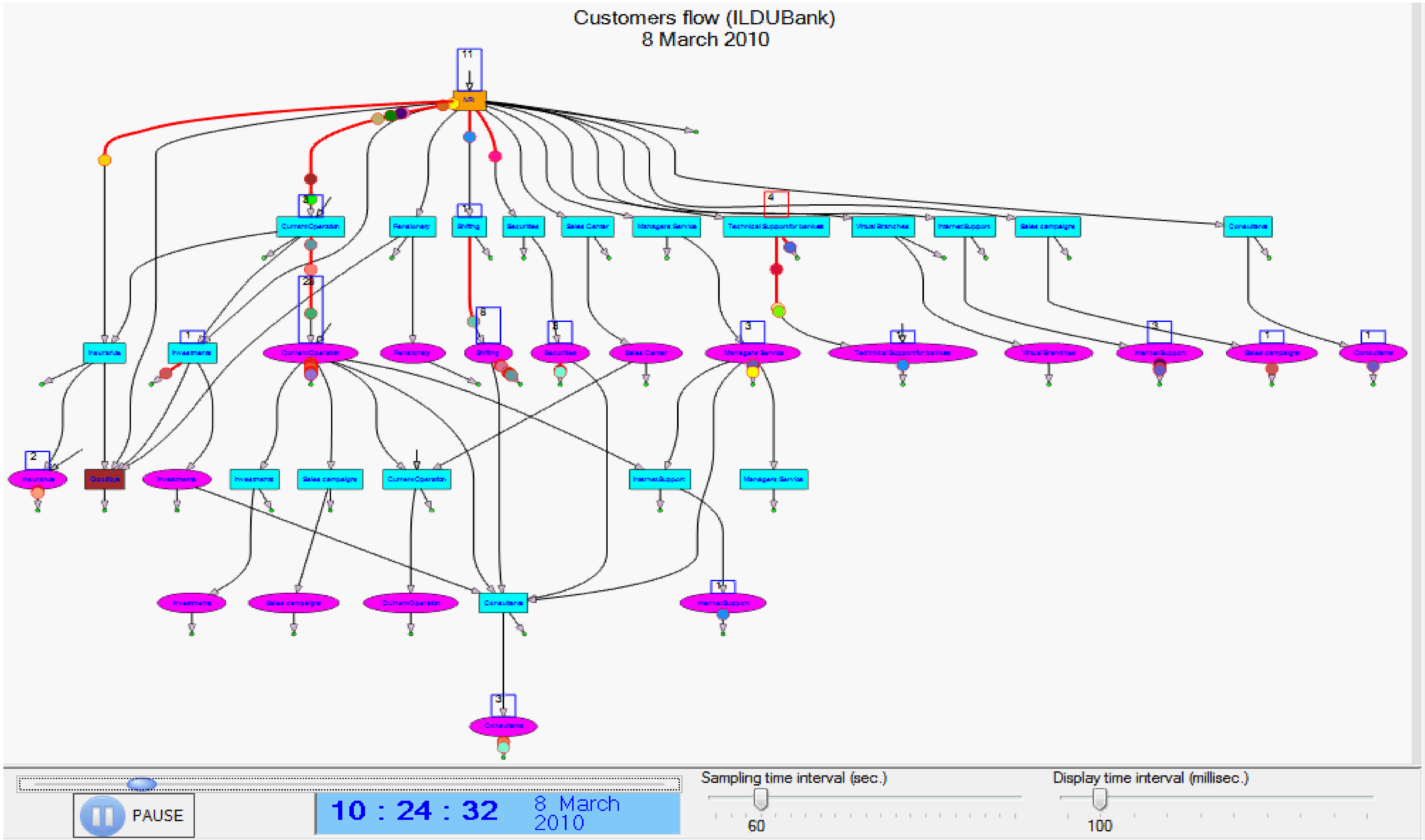
Finding an interested industry-partner w/ data $\Rightarrow$ problem solved

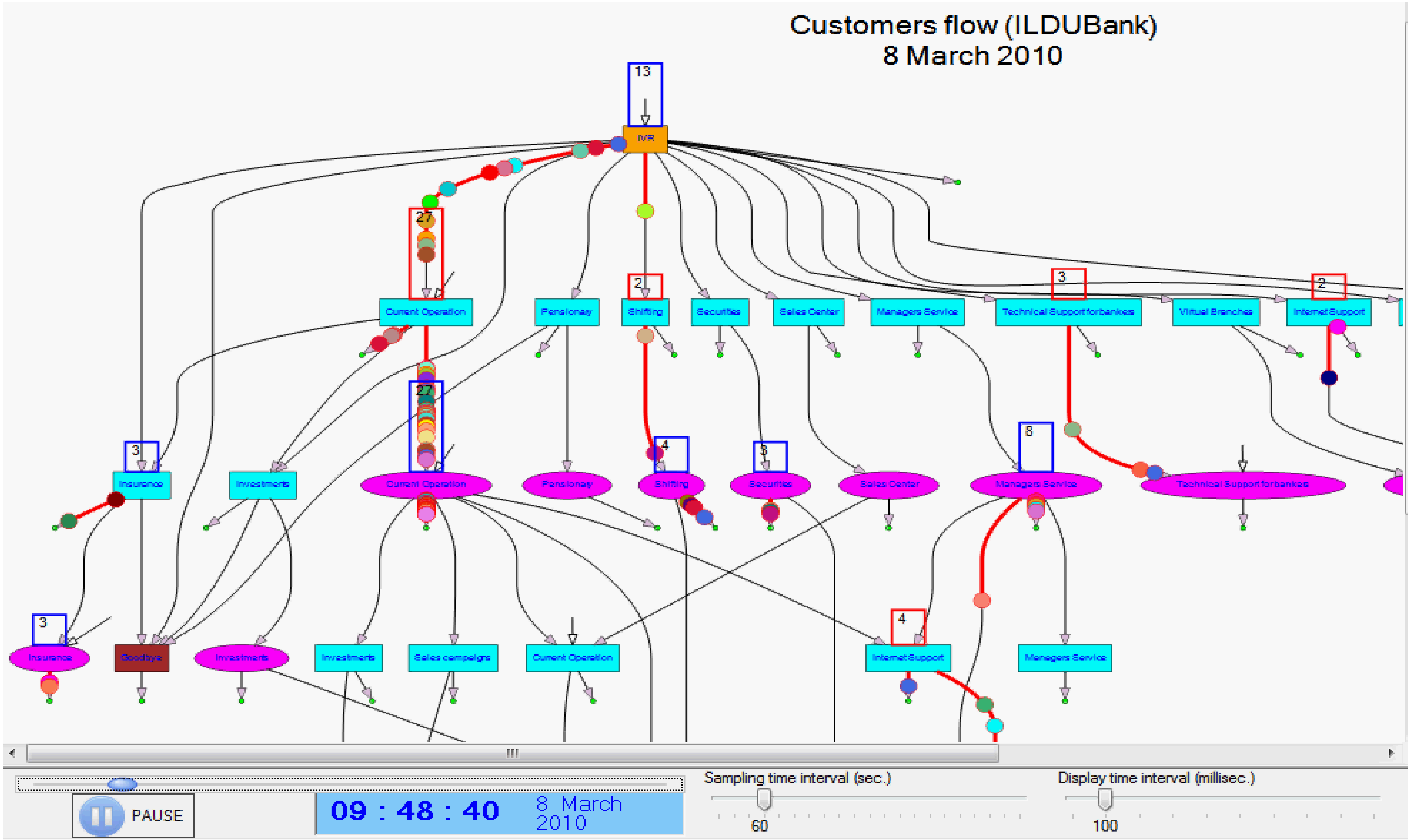
ISF: on measuring judicial workload (JW)

For the most part, … the applicant proposes to quantify the unquantifiable and solve the unsolvable, namely JW.

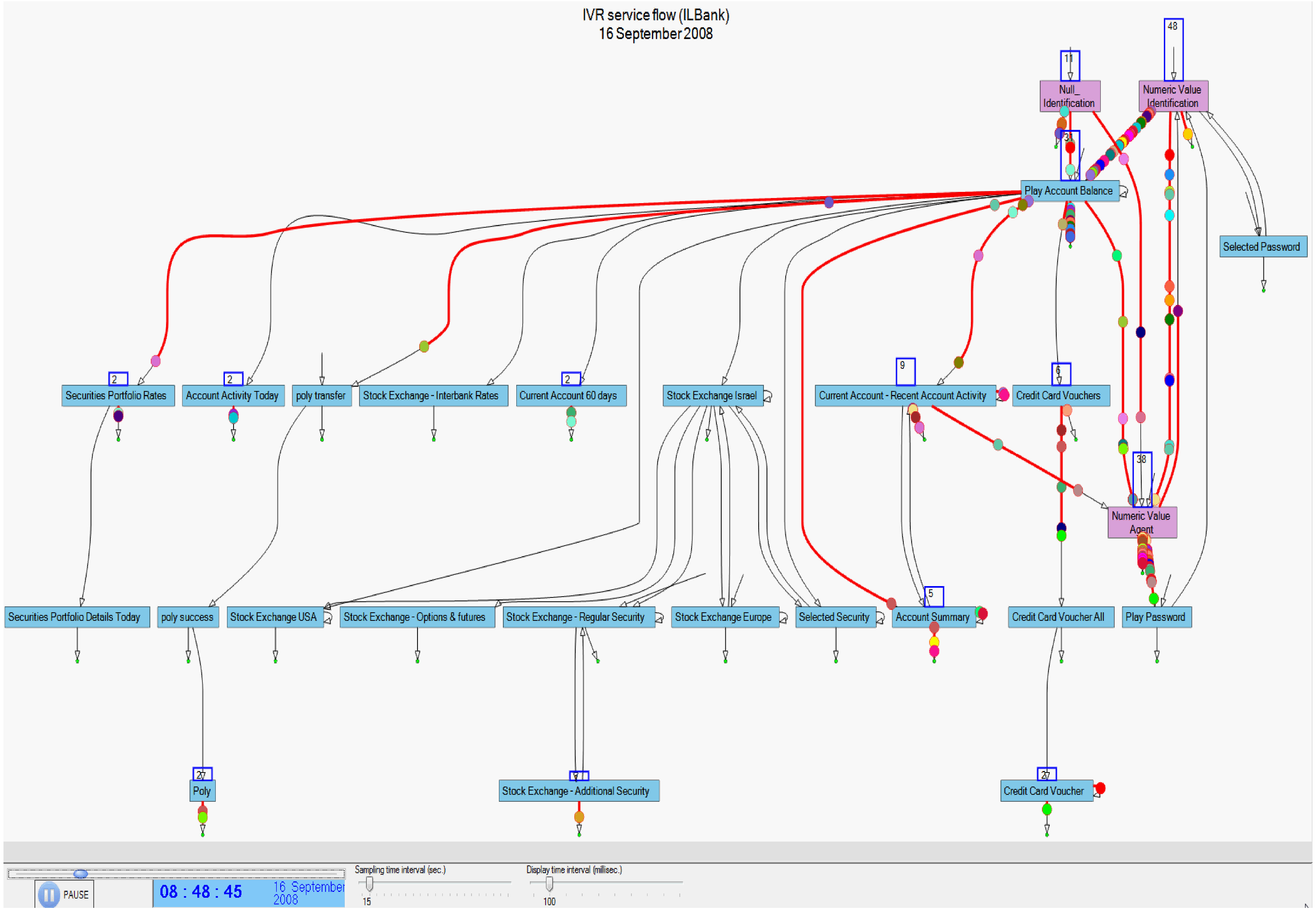
At least since the eighteenth century, there are continuous …

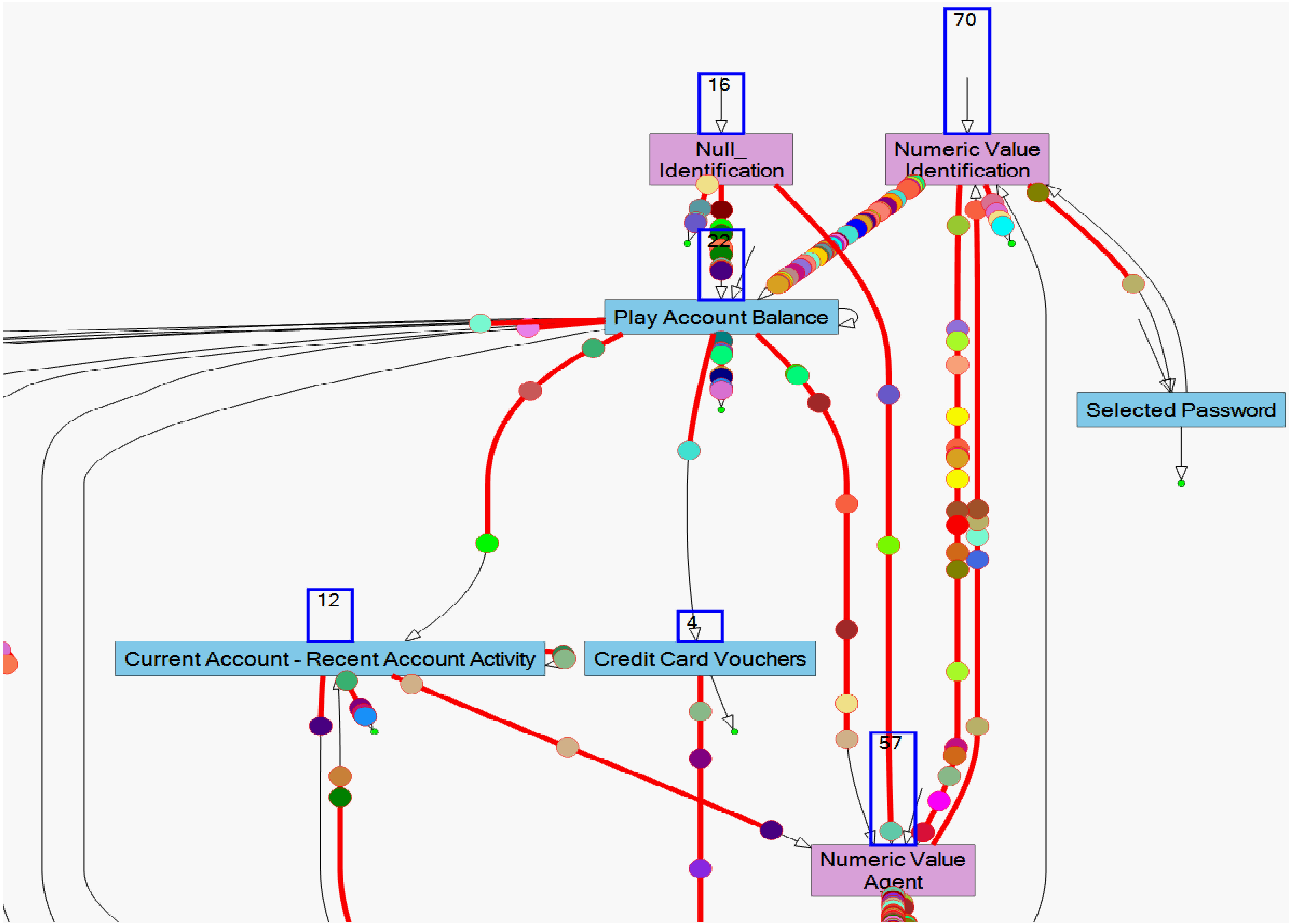
Still: Nurse-workload in maternity wards: operational+emotional(+cognitive)


Customers flow (ILDUBank)
8 March 2010
PAUSE
10 : 24 : 32
8 March
2010
Sampling time interval (sec.)
60
Display time interval (millisec.)
100



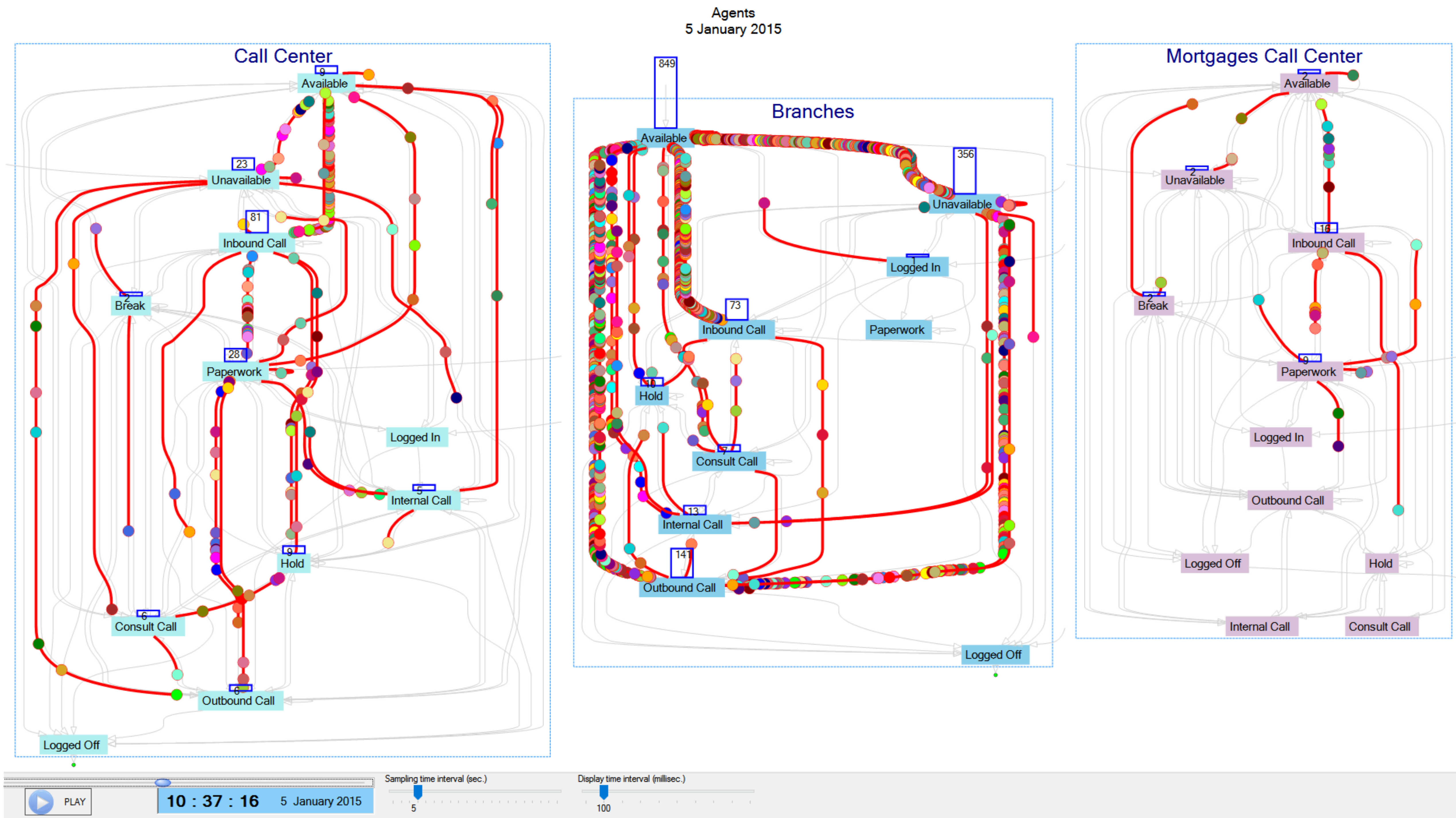
Customers flow (ILDUBank)
8 March 2010
IVR
Current Operation
Pensionay
Shifting
Securities
Sales Center
Managers Service
Technical Support for bankers
Virtual Branches
Internet Support
Insurance
Investments
Goodbye
Sales campaigns
PAUSE
09 : 48 : 40
8 March 2010
Sampling time interval (sec.)
60
Display time interval (millisec.)
100


IVR service flow (ILBank)
16 September 2008
Null_ Identification
Numeric Value Identification
Play Account Balance
Selected Password
Securities Portfolio Rates
Account Activity Today
poly transfer
Stock Exchange - Interbank Rates
Current Account 60 days
Stock Exchange Israel
Current Account - Recent Account Activity
Credit Card Vouchers
Numeric Value Agent
Securities Portfolio Details Today
poly success
Stock Exchange USA
Stock Exchange - Options & futures
Stock Exchange - Regular Security
Stock Exchange Europe
Selected Security
Account Summary
Credit Card Voucher All
Play Password
Poly
Stock Exchange - Additional Security
Credit Card Voucher
PAUSE
08 : 48 : 45
16 September 2008
Sampling time interval (sec.)
Display time interval (millisec.)

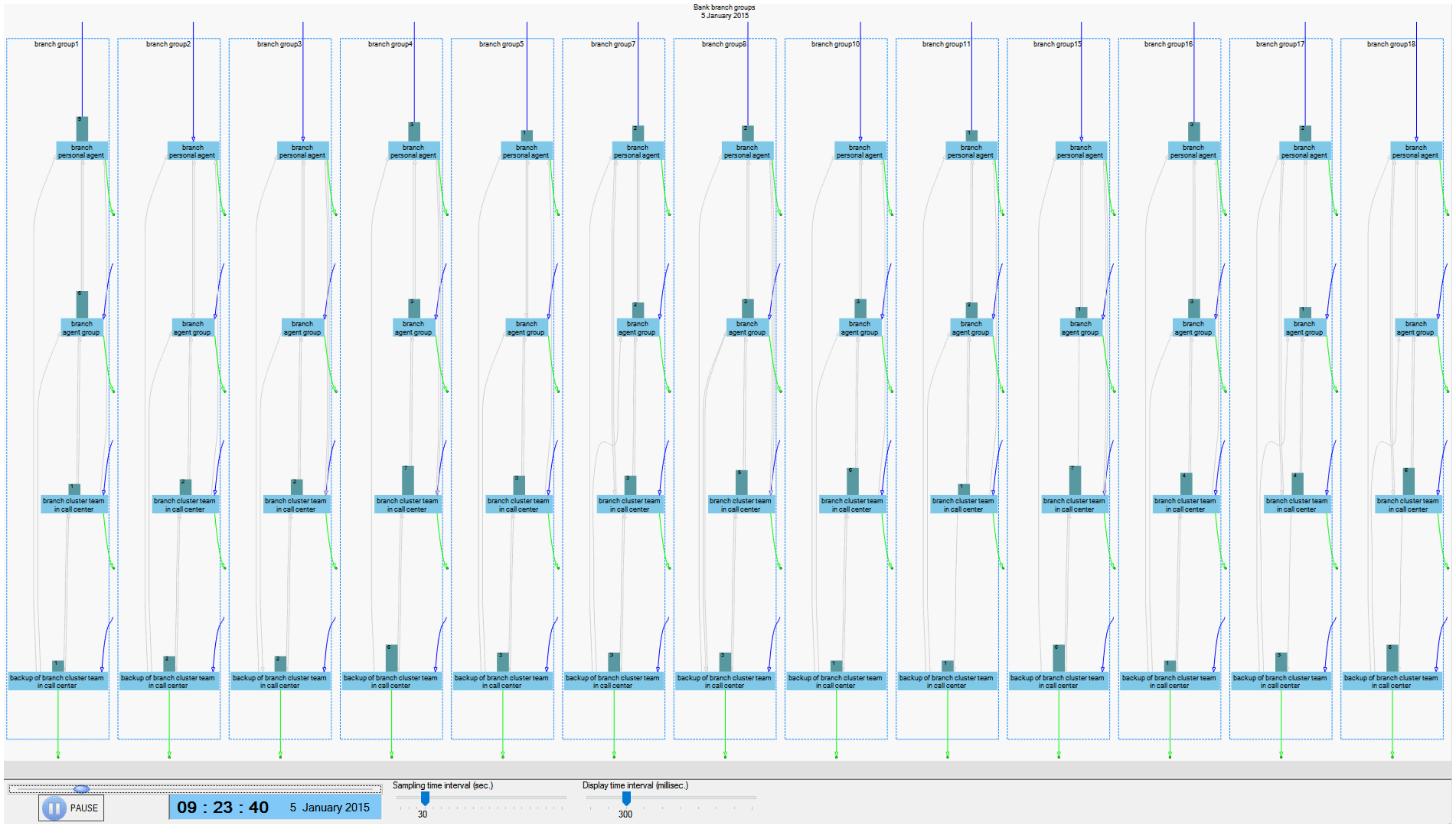


70
16
Null_ Identification
Numeric Value Identification
22
Play Account Balance
Selected Password
12
Current Account - Recent Account Activity
4
Credit Card Vouchers
57
Numeric Value Agent

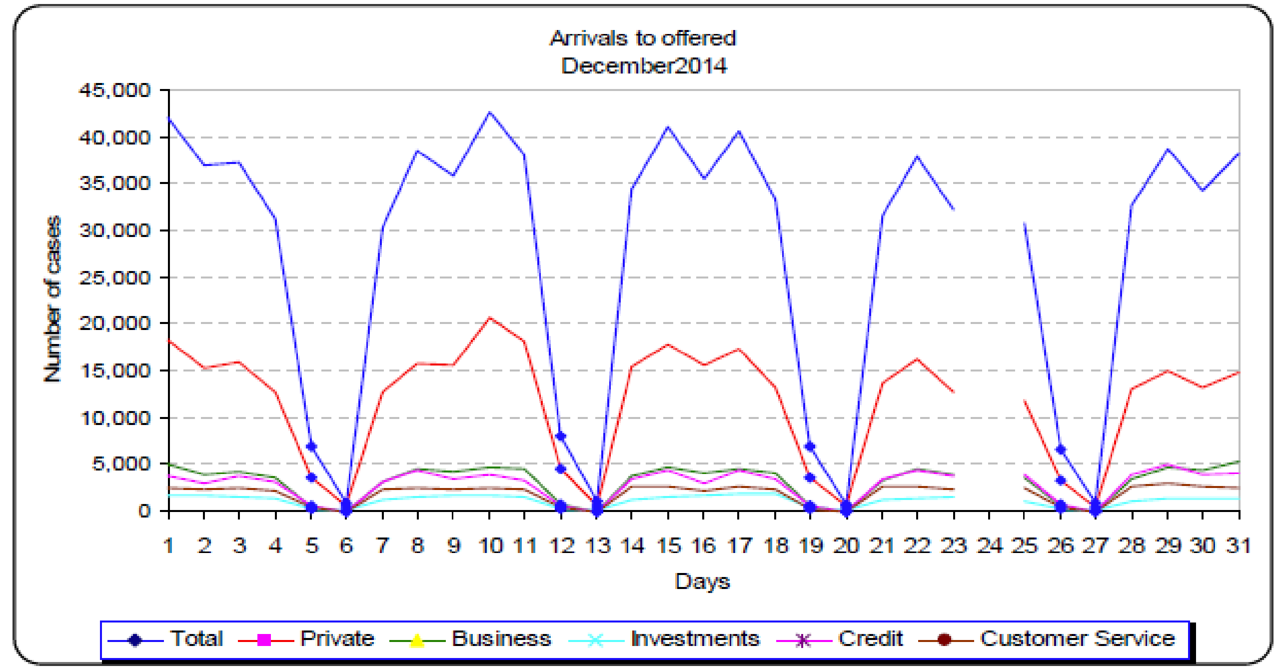
Agents (2000): Branches (1700) and Call Centers - Commercial (270), Mortgages (50)

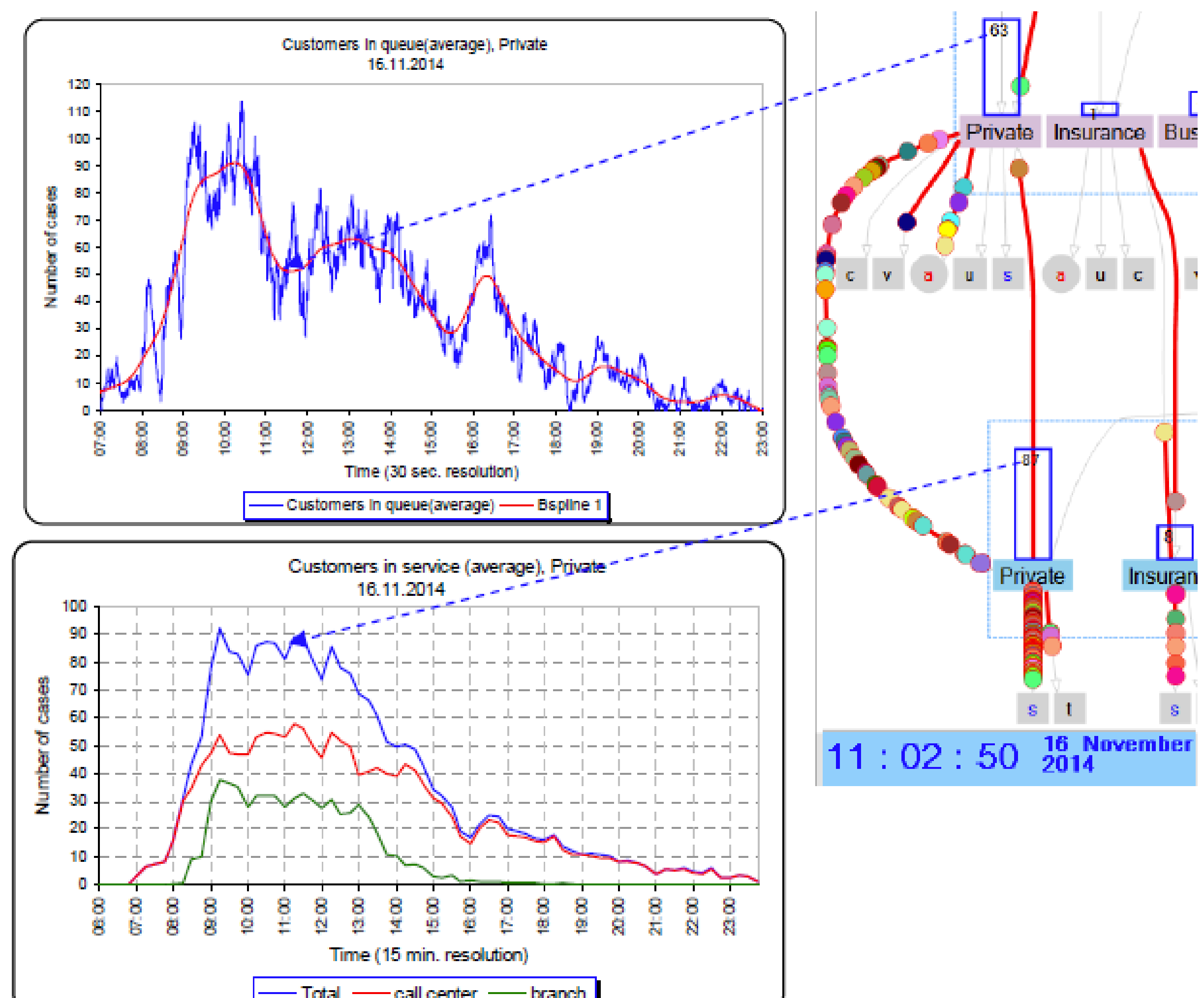


ILDU Banking: **Branch Groups Virtual Queues**



ILDU Banking:
branches and call centers
Service Types





Research Goal (within reach)

e.g. Specific Emergency-Department, with ample reliable data

- Real-time: control of patient-flow (bottlenecks); status info and prediction
- Short-term: on Monday, set Tuesday's staffing levels (or next week's); real cost of care for the individual patient (vs. mean/negotiated costs)
- Long-term: capacity allocation, facility/triage design; social network (e.g. correlated w/ outcomes); transformative change (Epic); congestion laws

All above will be enabled by parsimonious (robust) models, created in real-time **mining** (semi- or un-supervised) of ED **processes** and **models** (empirical, simulation = **SimNets**, mathematical = **QNets**, **FNets**, **DNets**,…)

Laws & Models: Data-Based Erlang A/R/S, following B/C

- **Little's Law (Steady-State, Transient), State-Space Collapse,…**

- **Erlang-B (Blocking) and Erlang-C (?)**
  - **Erlang**, Agner Krarup: Queueing Theory was born in 1909, in his paper "The Theory of Probabilities and Telephone Conversations"

- **2. Erlang-A**
  - **Abandonment**: While waiting for service, does service-value dominate residual-wait-cost

- **1. Erlang-R**
  - **Return/Feedback**: Customers often return to service (positively, negatively, just needing)

- **3. Erlang-S**
  - **Servers**: Challenging to manage, and model, no less so than customers

Above: Simple (Parsimonious/Robust) models of complex realities, yet not too simple

For the 3 Models (if time permits):

1. Motivating Phenomenon, via Data

2. Informal Description of Model

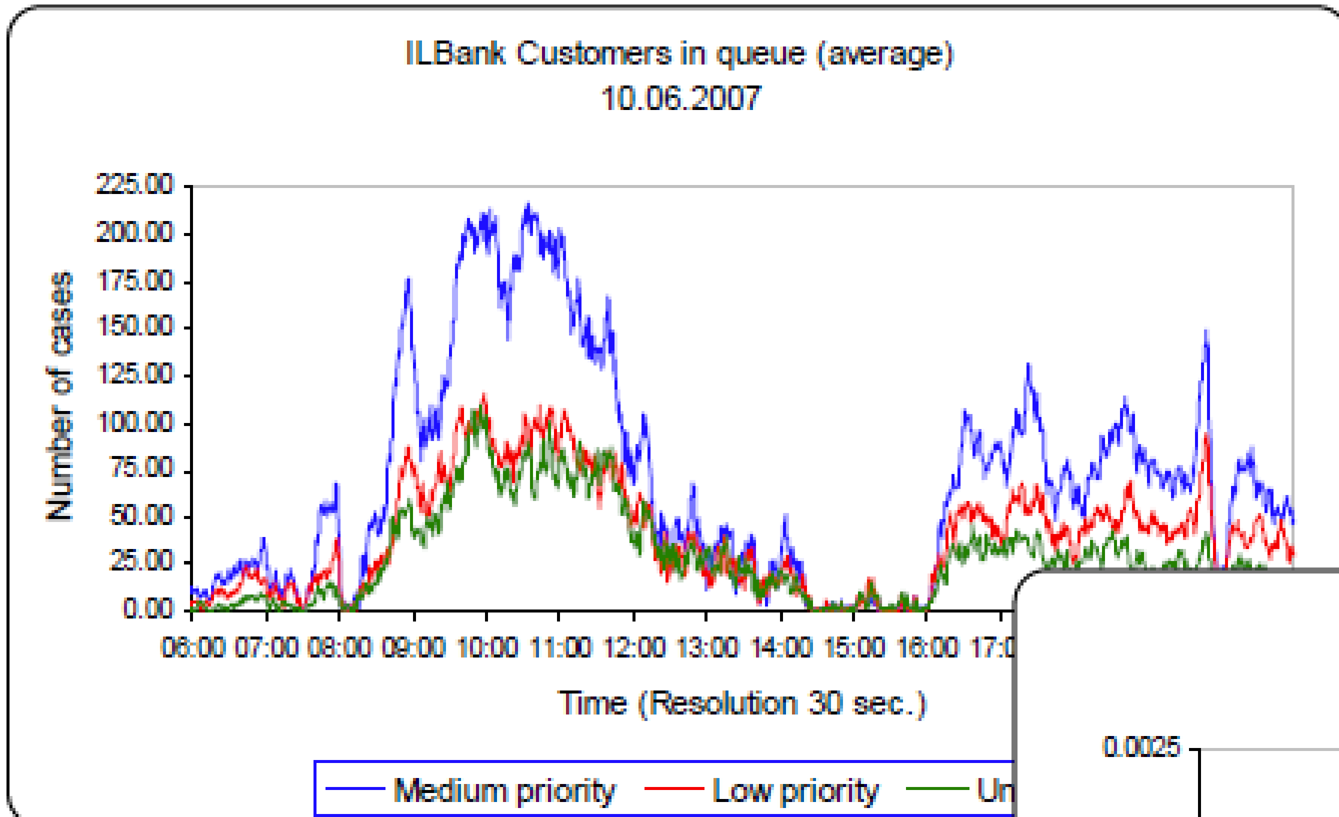
3. Example of Application(s)

4. Model Expressiveness (Strength)

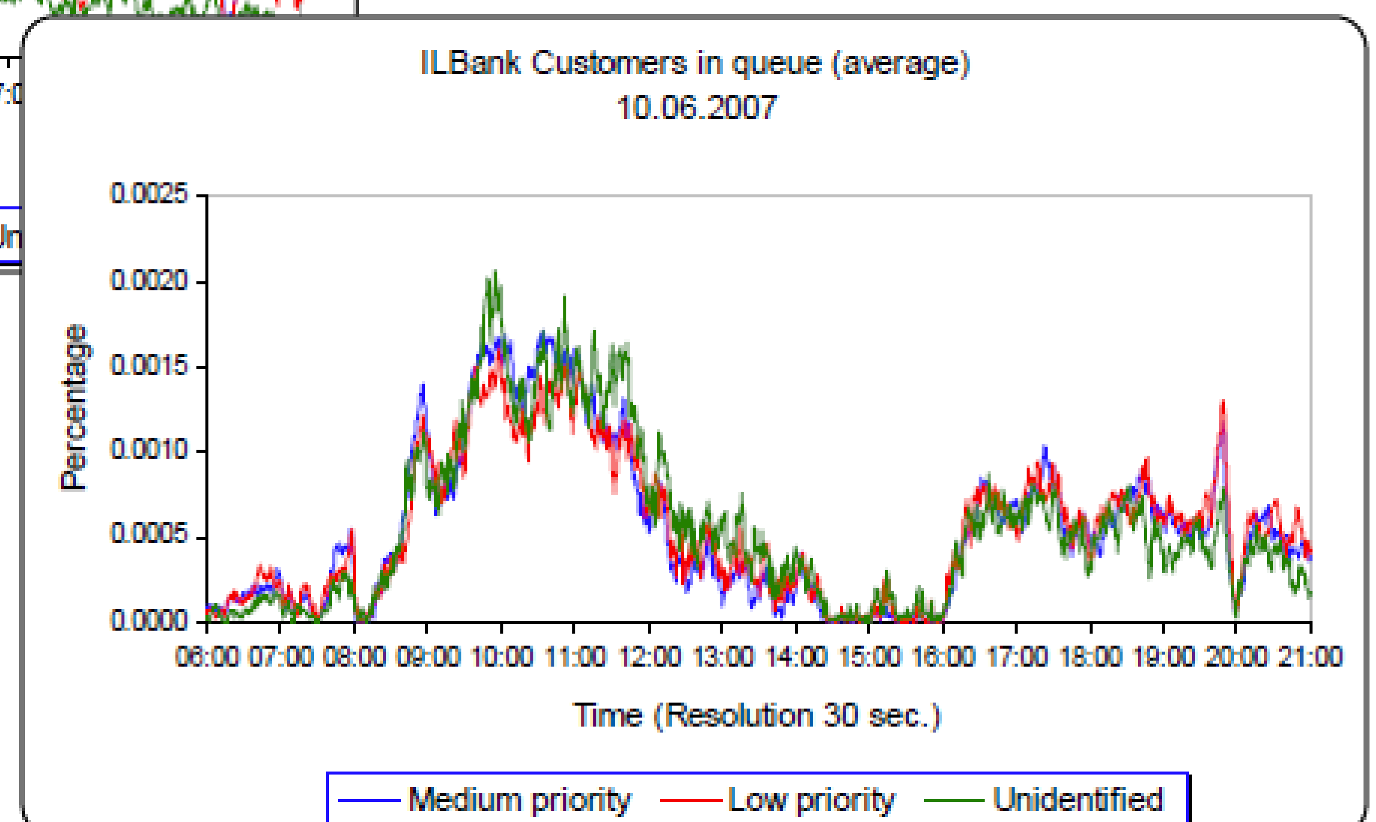
5. Insights, Extensions

Dynamics: Parsimonious Models (Congestion Laws)

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



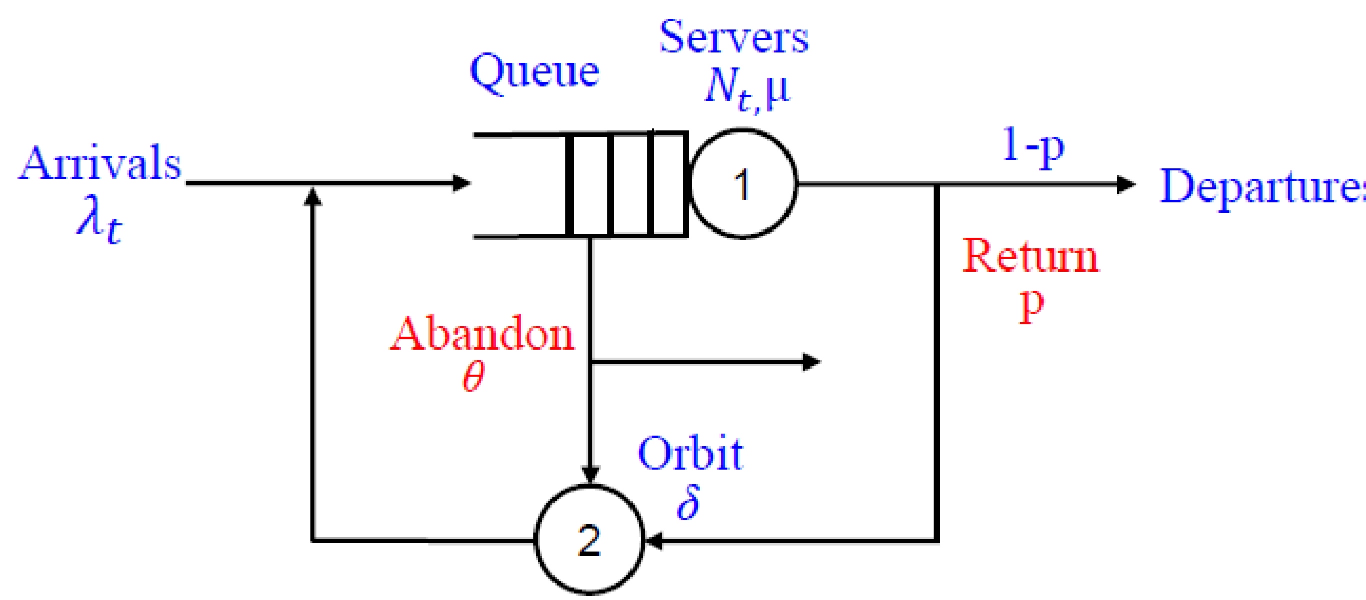
Queue "Shape"



- Area normalized to 100%
- State-Space Collapse

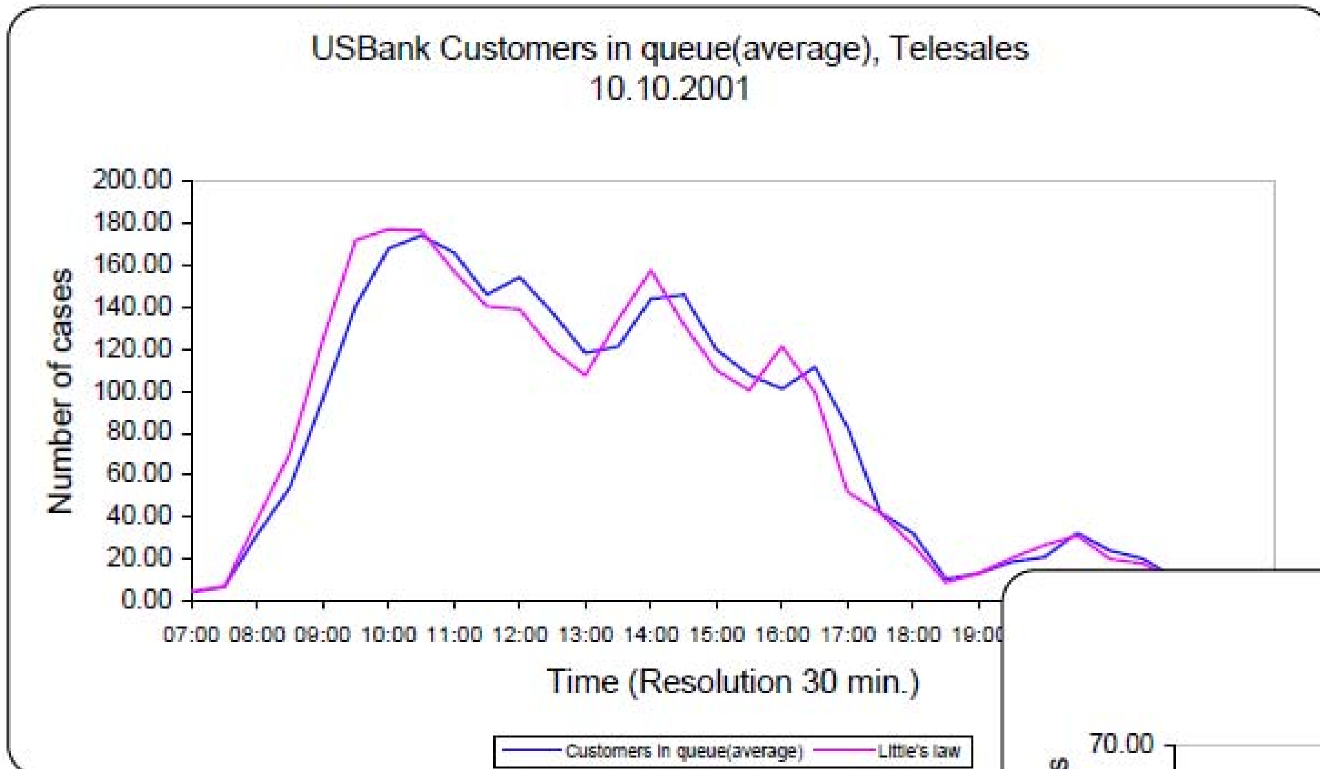
Model Selection: As Simple as Possible but Not Simpler

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

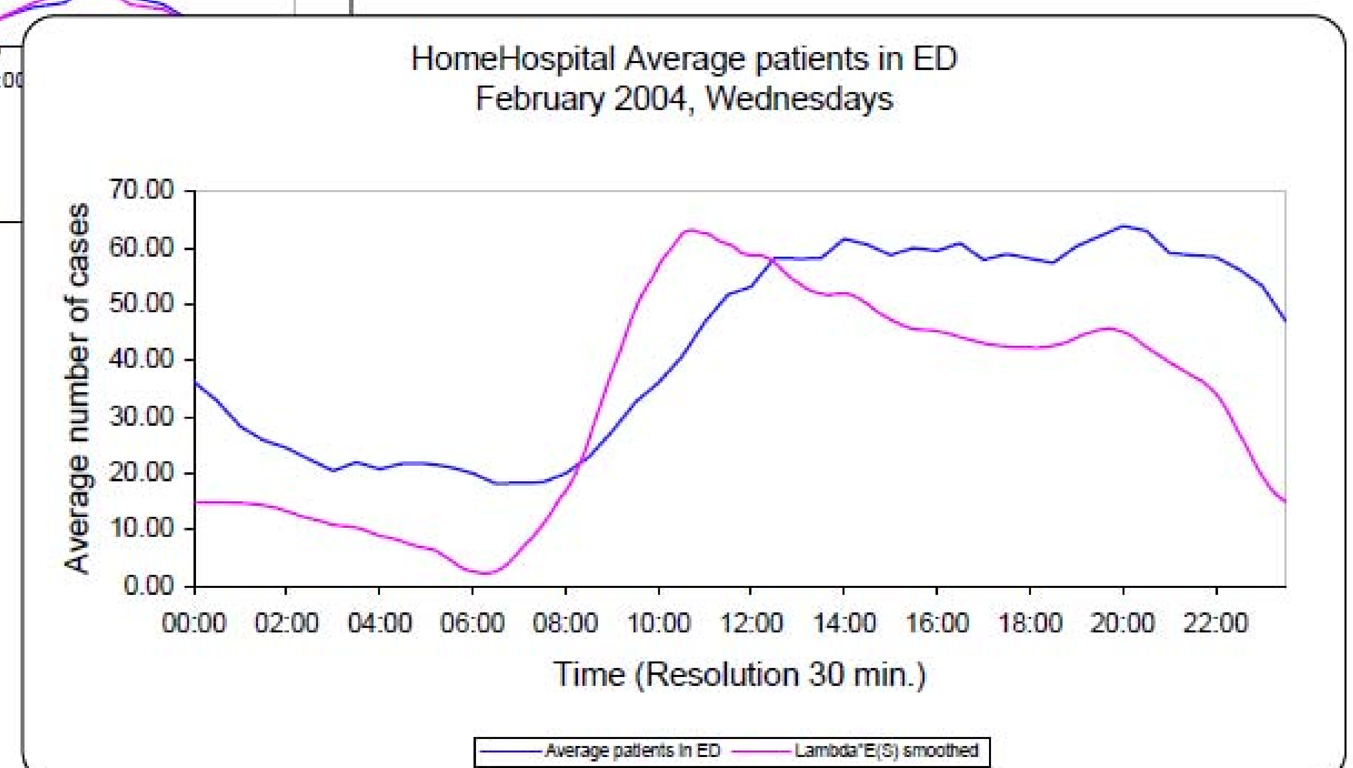


Little's Law $L = \lambda \times W$, in a Time-Varying Environment

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



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

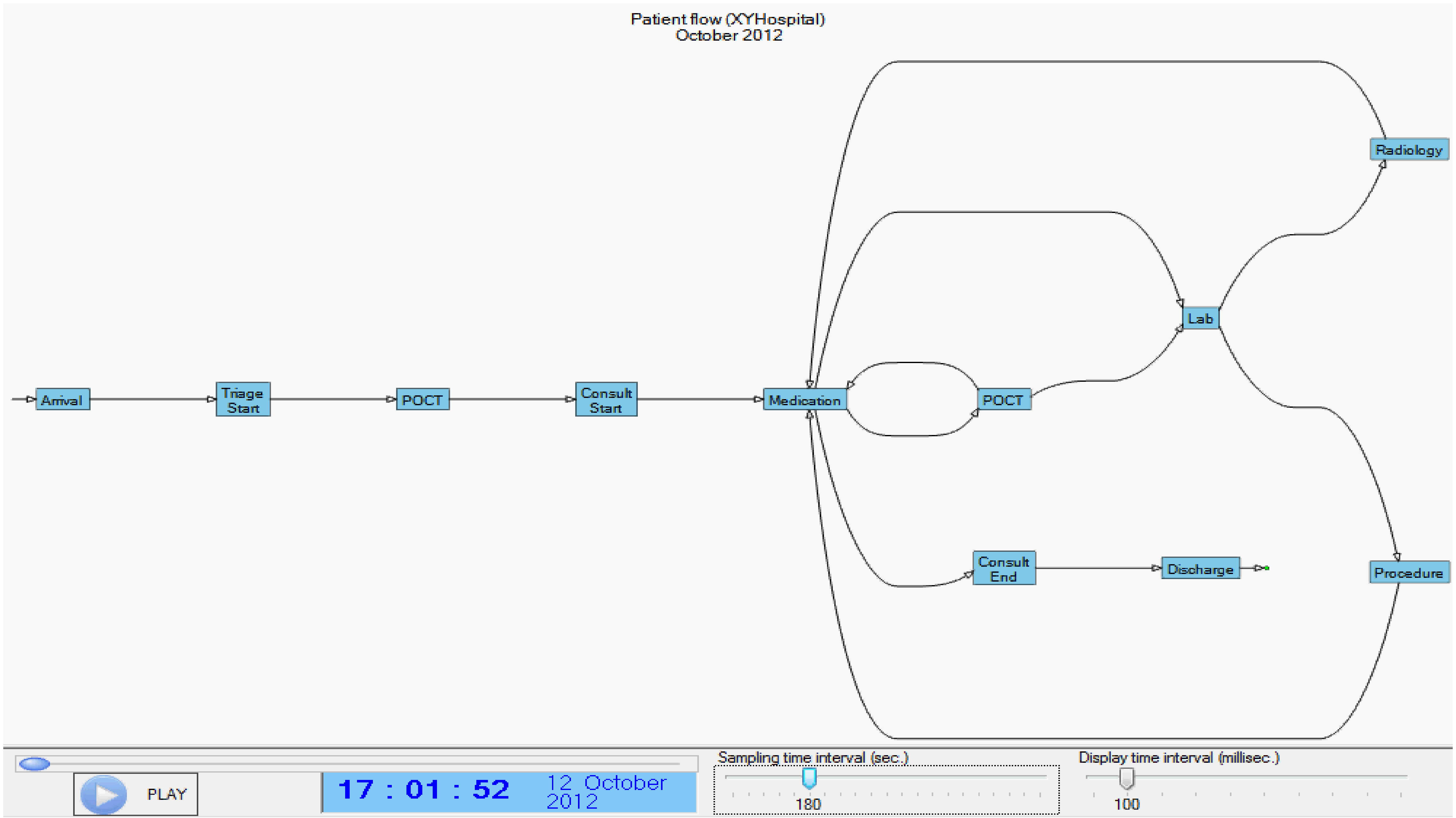


Emergency Dept:
$\Rightarrow$ **Time-Varying Transient**

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

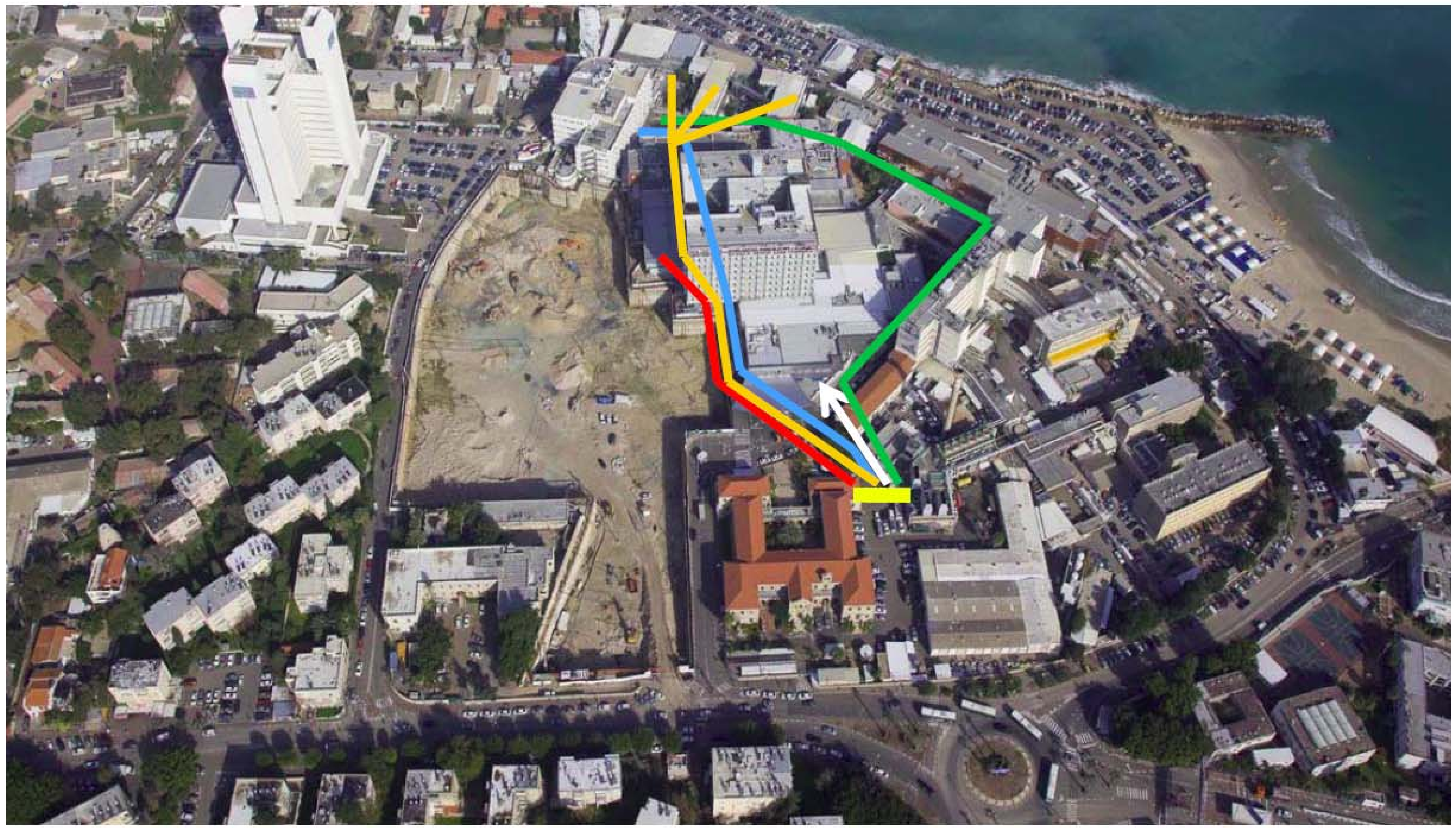
$$\tilde{\lambda}(t) = E\lambda(t - W_e).$$

(Bertsimas, Mourtzinou;
Fralix, Riano, Serfozo)


Patient flow (XYHospital)
October 2012
Arrival
Triage Start
POCT
Consult Start
Medication
POCT
Lab
Radiology
Consult End
Discharge
Procedure
PLAY
17 : 01 : 52
12 October 2012
Sampling time interval (sec.)
180
Display time interval (millisec.)
100

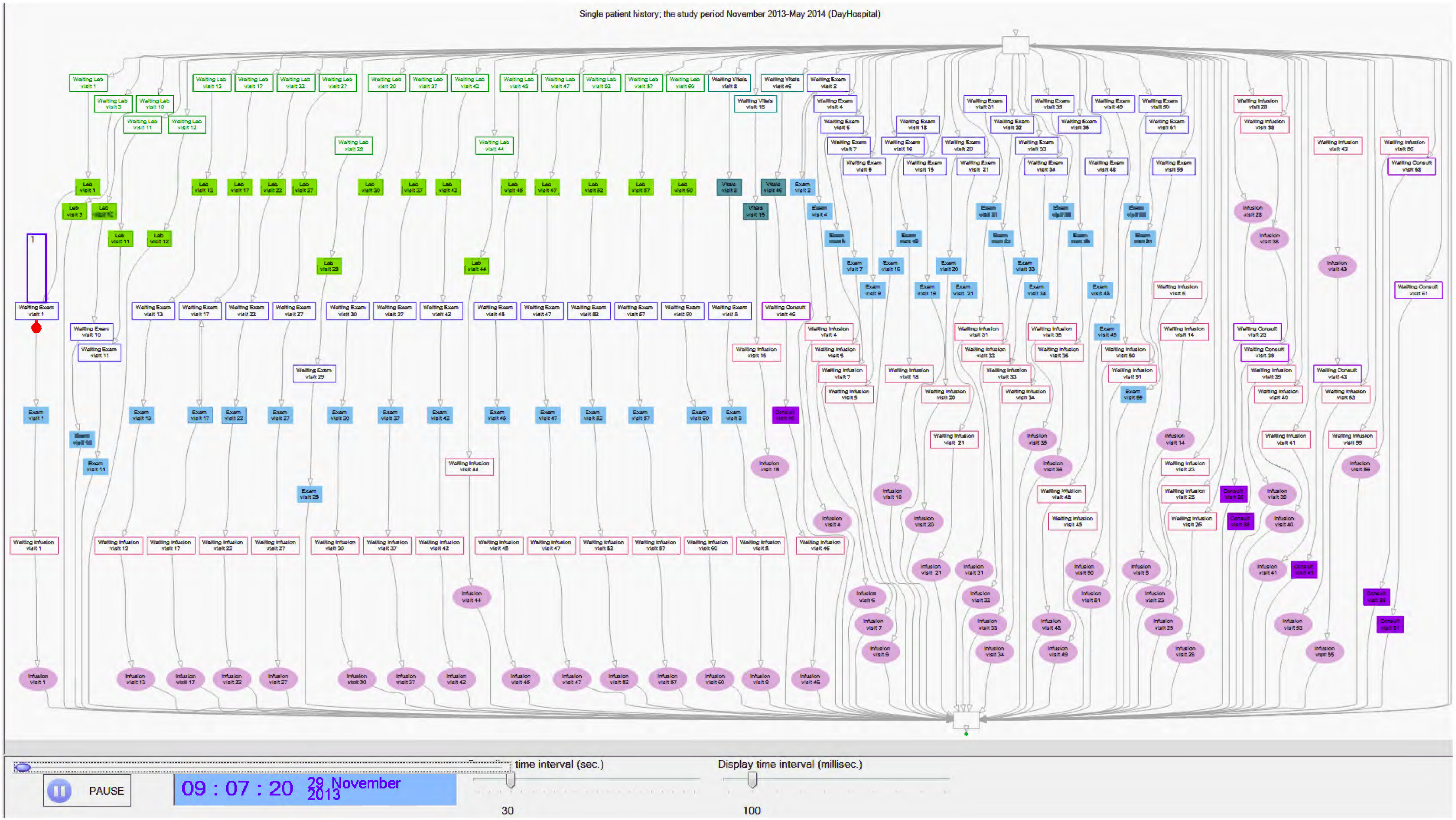
eg. RFID-Based Data: Mass Casualty Event (MCE)

Drill: Chemical MCE, Rambam Hospital, May 2010

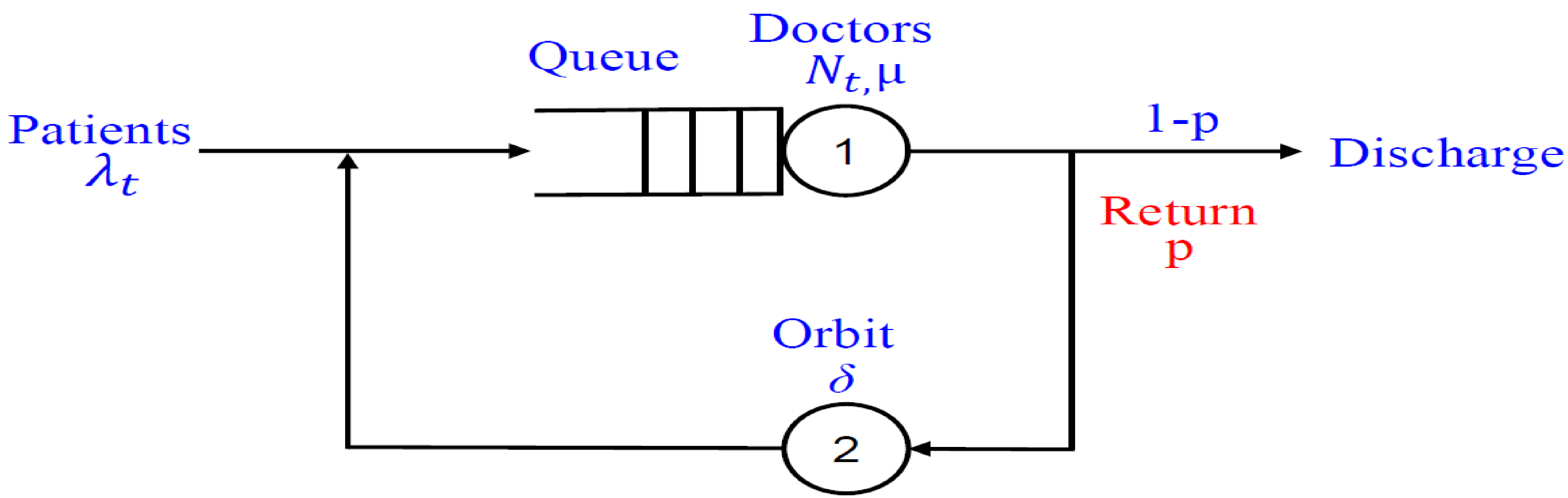


Focus on **severely wounded** casualties ($\approx$ 40 in drill)

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


Single patient history; the study period November 2013-May 2014 (DayHospital)
PAUSE
09 : 07 : 20
29 November 2013
time interval (sec.)
Display time interval (millisec.)
30
100

Erlang-R ↔ Fluid Model, w/ Galit Yom-Tov



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

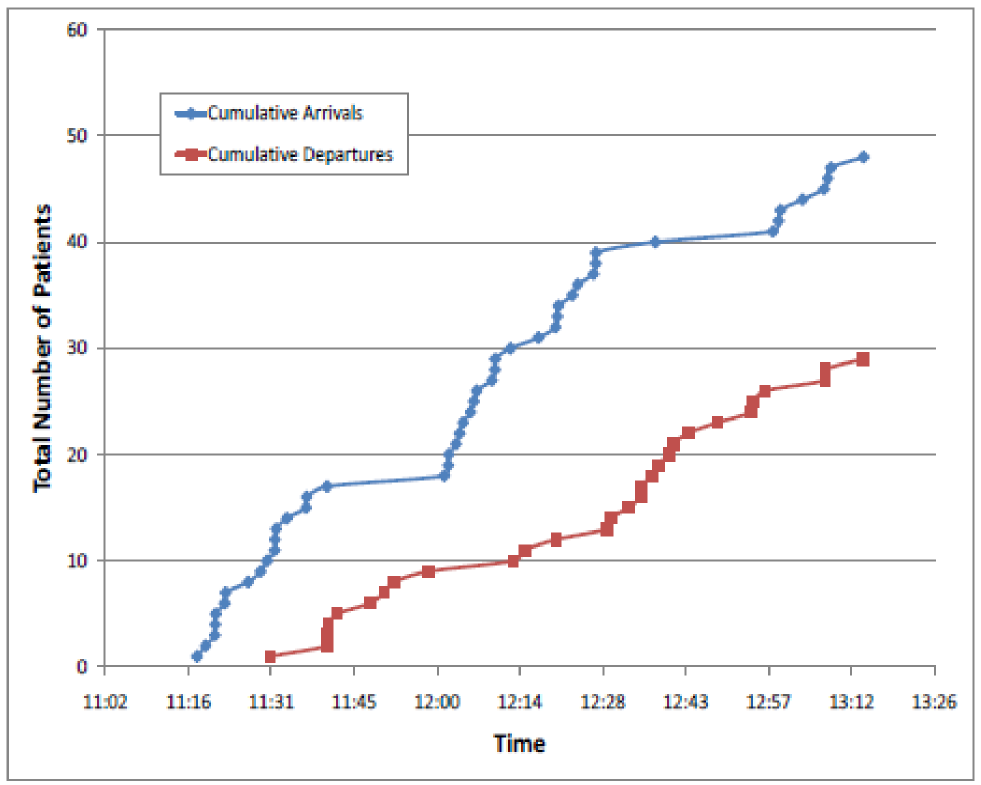
FNet = ODE: derived **directly** (no QNet), spreadsheet "solution"

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

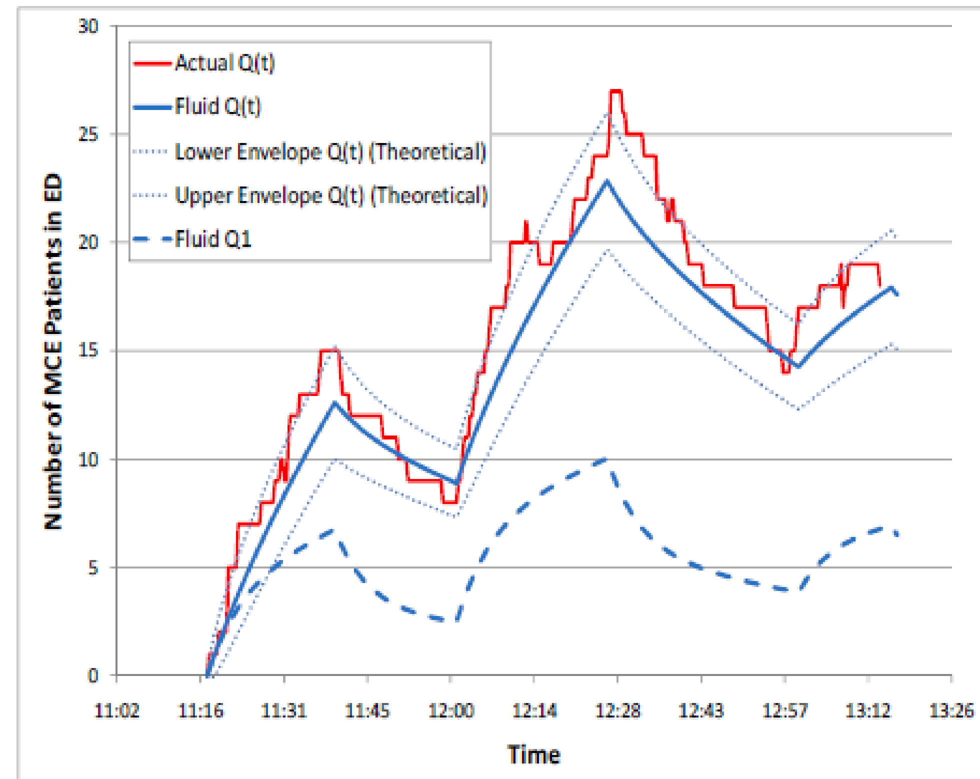
Erlang-R Value: FNet vs. Data

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

Arrivals & **Departures** (RFID)



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



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

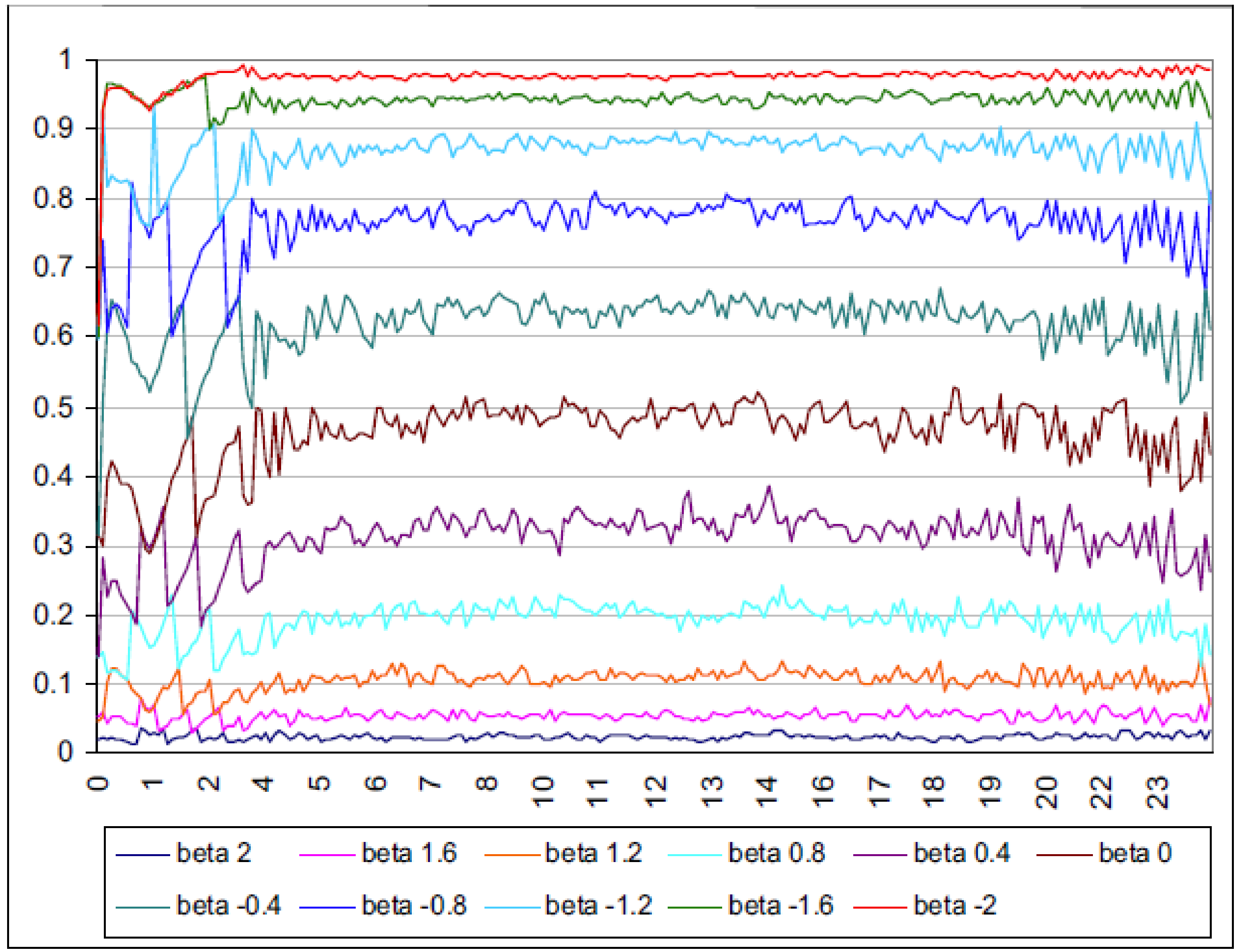
A Data-Based Framework, or "Erlang-R in the ED"

System = e.g. Emergency Department

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

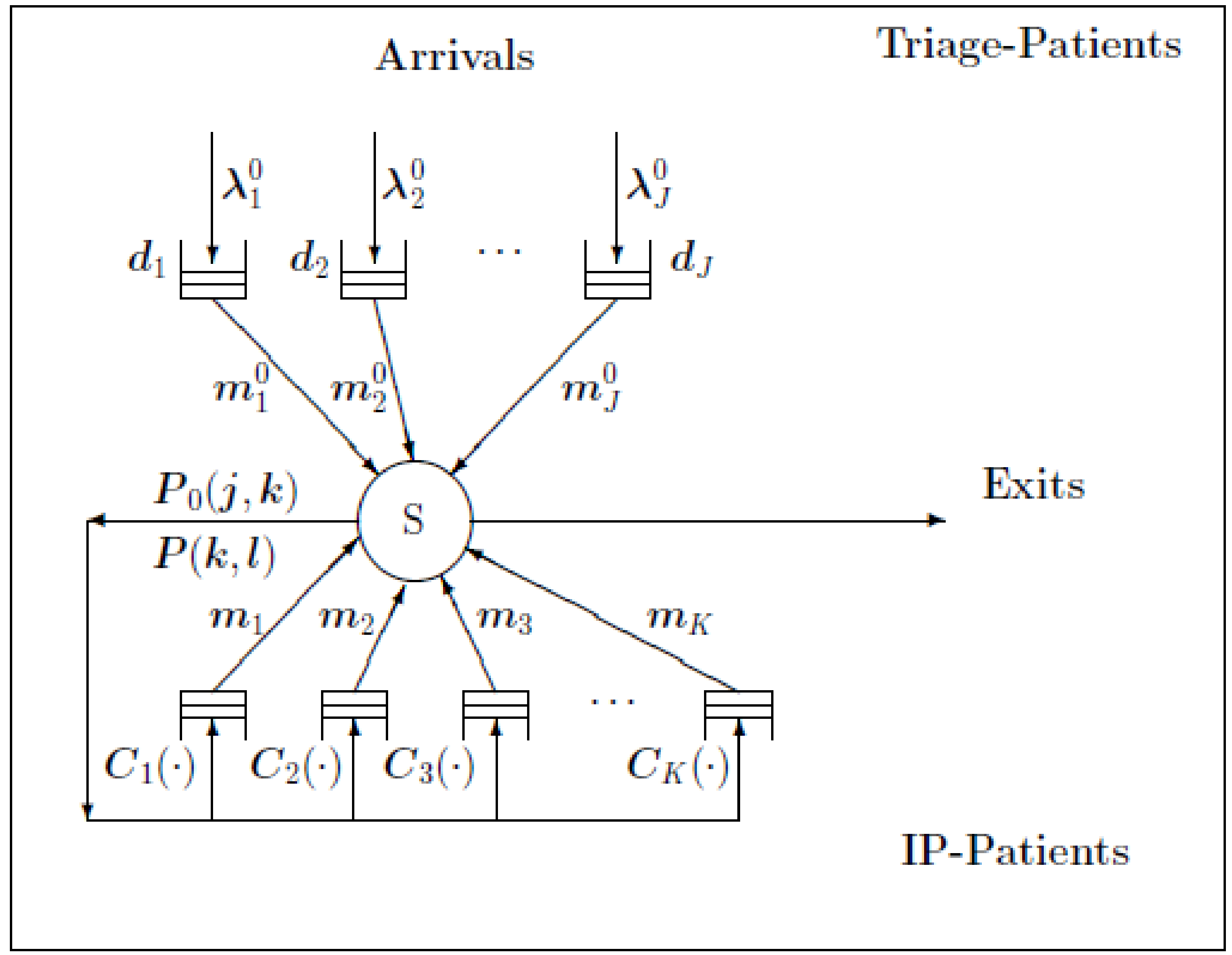
Time-Stable Performance of Time-Varying Systems

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



ED Patient Flow: The Physicians View

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

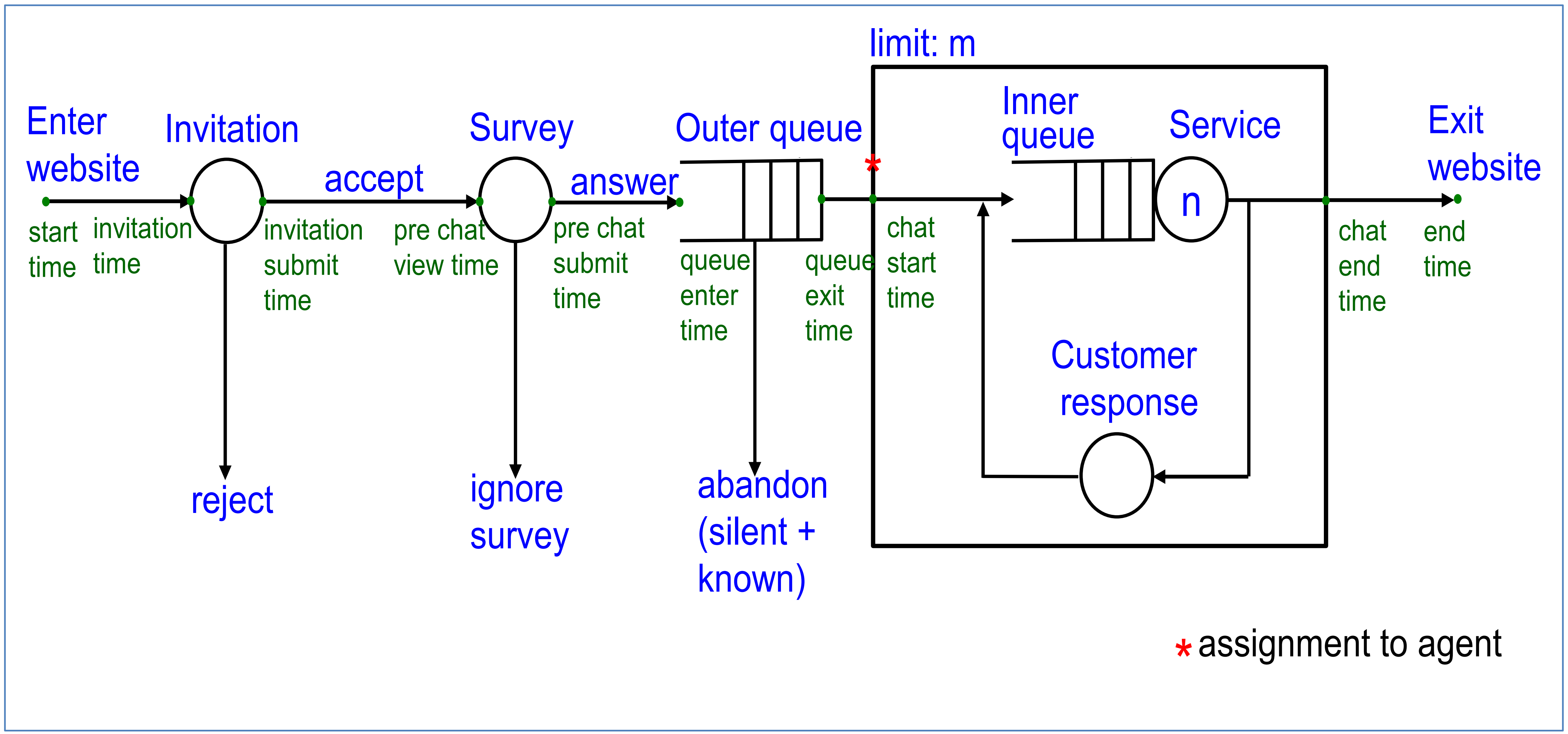


Goal: Adhere to **Triage-Constraints**, then **release In-Process** Patients

Online Chats – Returns: from Internet-Prompts to Chat-Sessions

Empirical: Galit Yom-Tov & Anat Rafaeli & graduate (OB) students

Theory: Tolga Tezcan & Jiheng Zhang, Routing & Staffing, OR 2014



Laws & Models: Data-Based Erlang A/R/S, following B/C

- **Little's Law (Steady-State, Transient), State-Space Collapse,…**

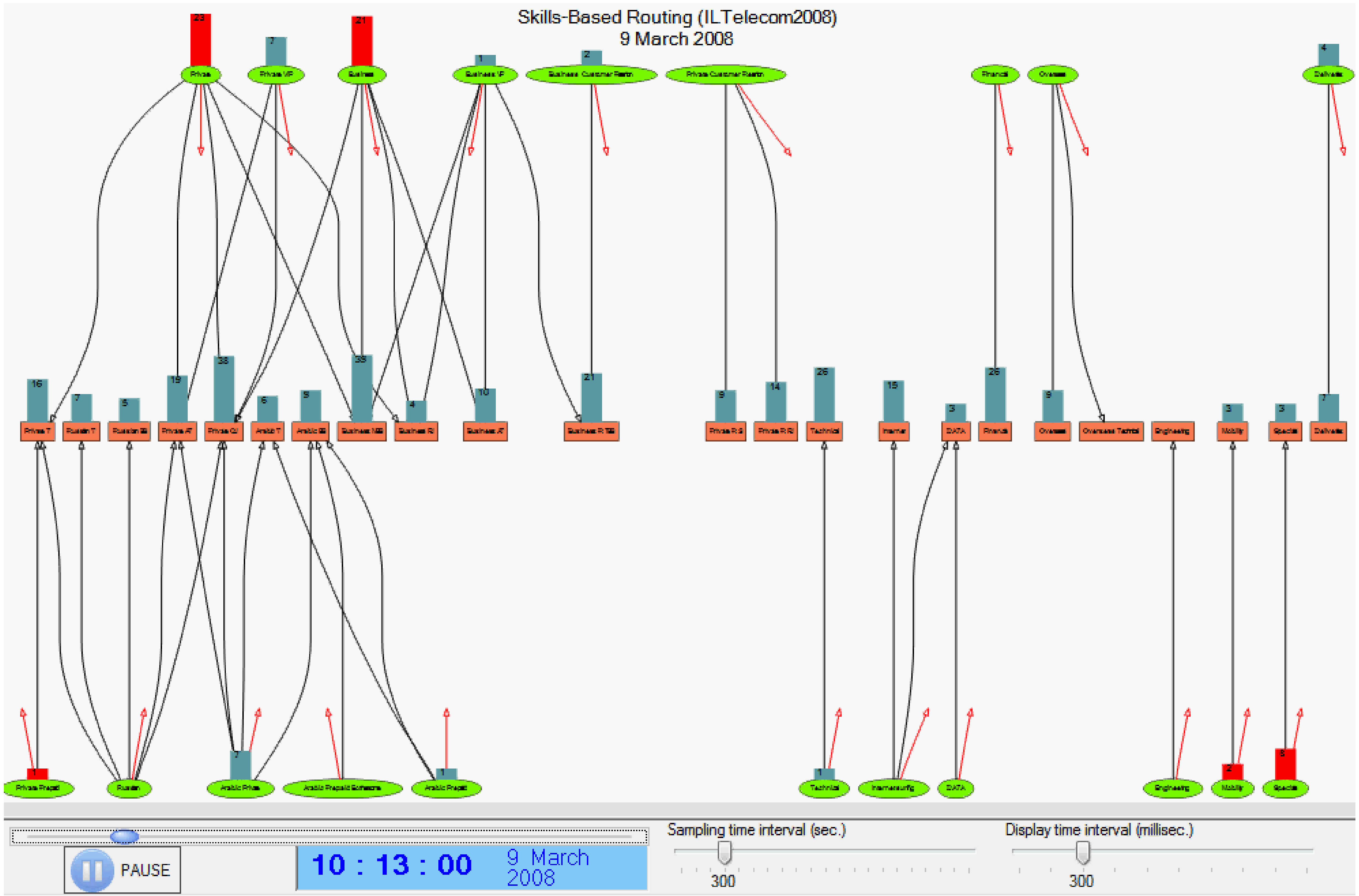
- **Erlang-B (Blocking) and Erlang-C (?)**
  - **Erlang**, Agner Krarup: Queueing Theory was born in 1909, in his paper "The Theory of Probabilities and Telephone Conversations"

- **2. Erlang-A**
  - **Abandonment**: While waiting for service, does service-value dominate residual-wait-cost

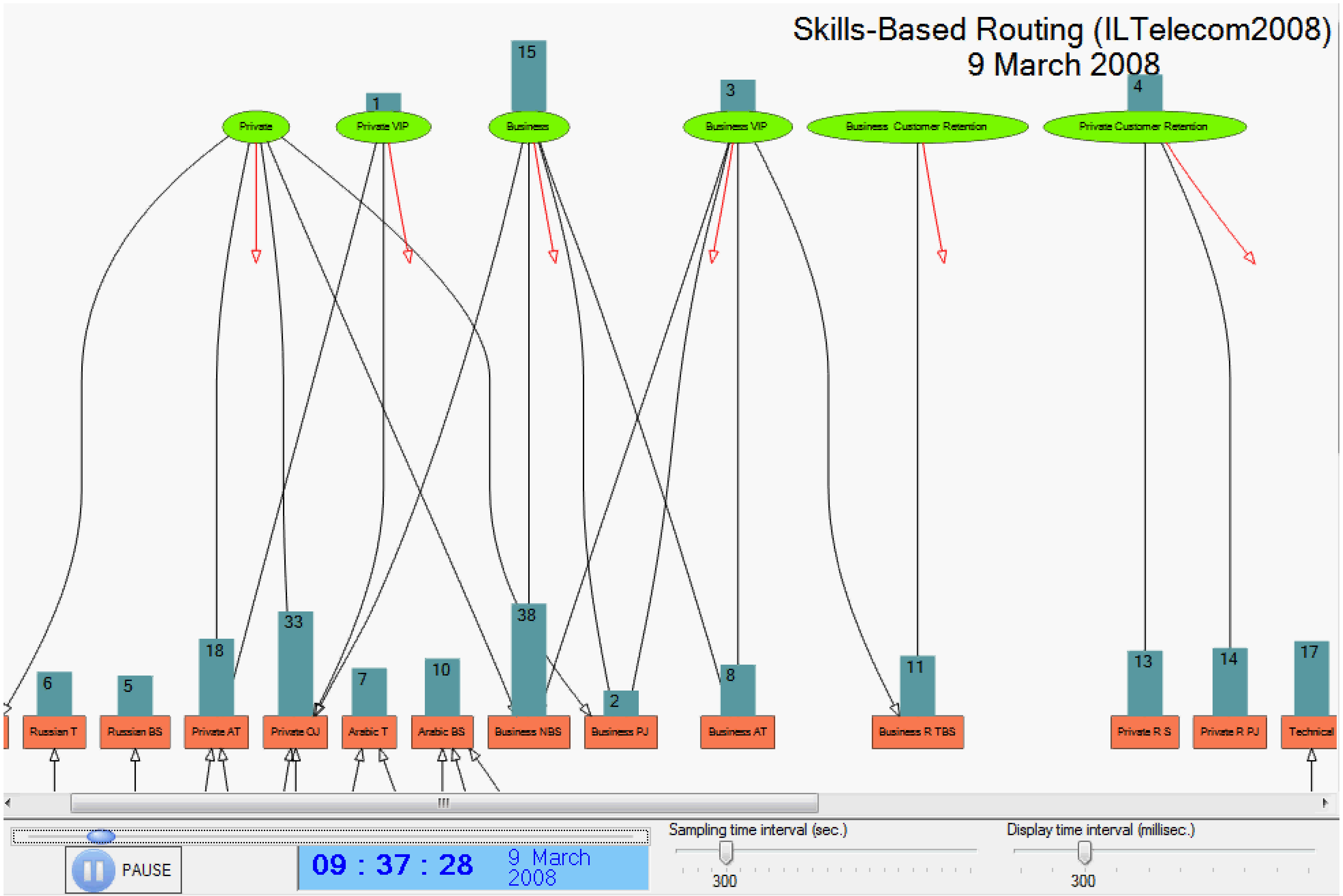
- **1. Erlang-R**
  - **Return/Feedback**: Customers often return to service (positively, negatively, just needing)

- **3. Erlang-S**
  - **Servers**: Challenging to manage, and model, no less so than customers

Above: Simple (Parsimonious) models of complex realities, yet not too simple (Robust)



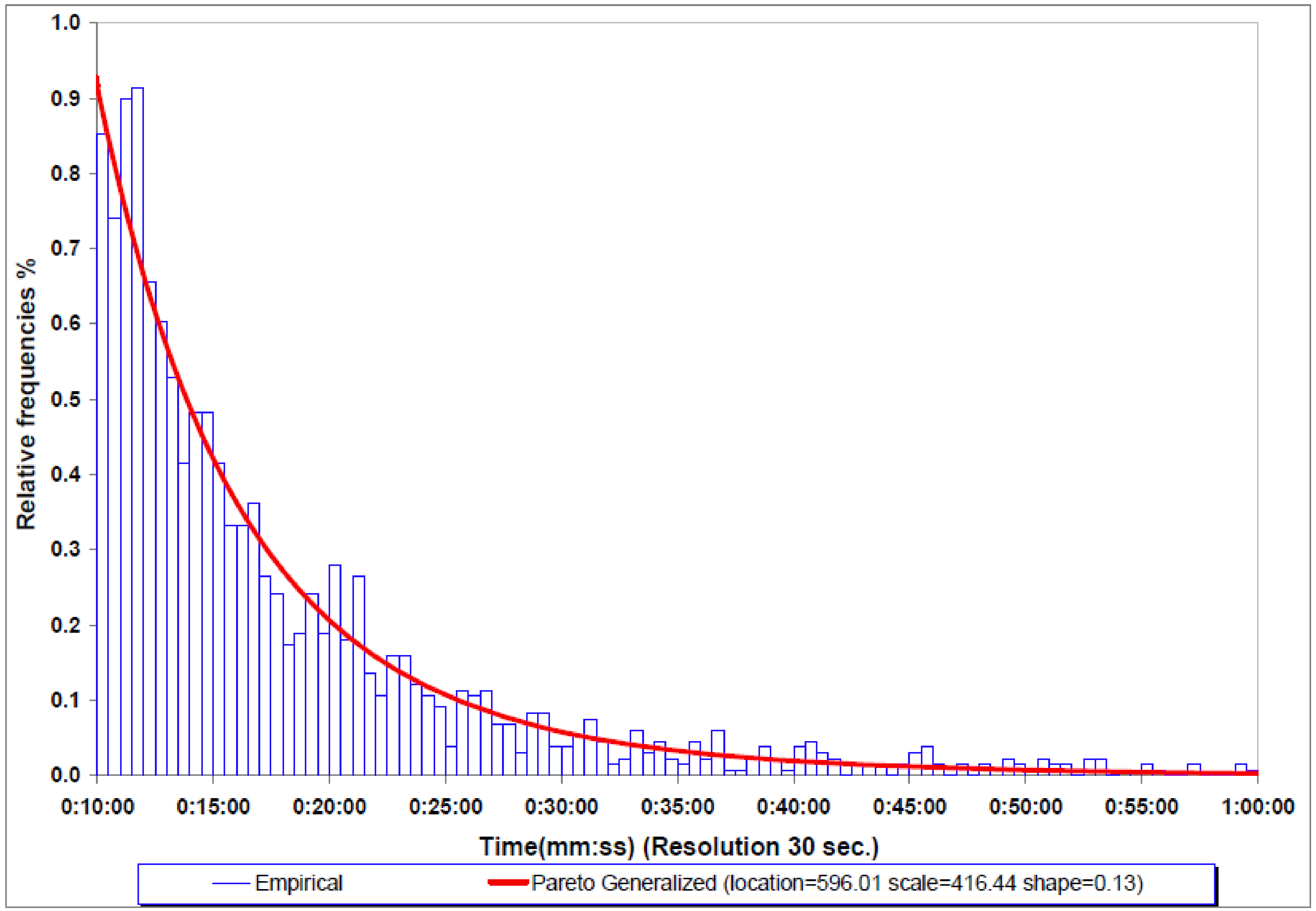
Skills-Based Routing (ILTelecom2008)
9 March 2008
PAUSE
10 : 13 : 00
9 March 2008
Sampling time interval (sec.)
300
Display time interval (millisec.)
300


Skills-Based Routing (ILTelecom2008)
9 March 2008
Private
Private VIP
Business
Business VIP
Business Customer Retention
Private Customer Retention
15
1
3
4
Russian T
Russian BS
Private AT
Private OJ
Arabic T
Arabic BS
Business NBS
Business PJ
Business AT
Business R TBS
Private R S
Private R PJ
Technical
6
5
18
33
7
10
38
2
8
11
13
14
17
PAUSE
09 : 37 : 28
9 March 2008
Sampling time interval (sec.)
300
Display time interval (millisec.)
300

Primitives: (Im)Patience

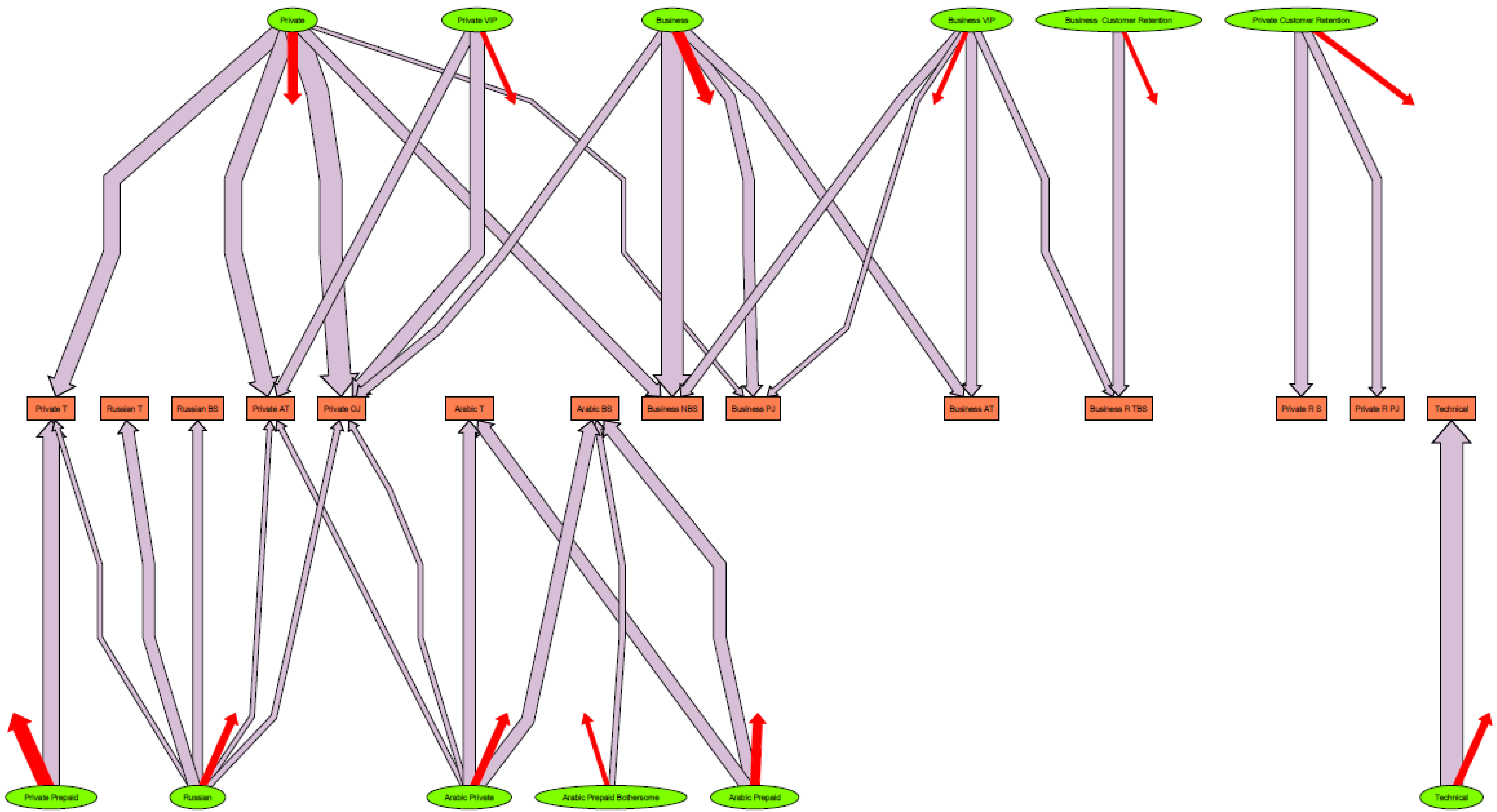
Israeli Bank: Uncensored 13,000 Customers, 24/11/2008

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



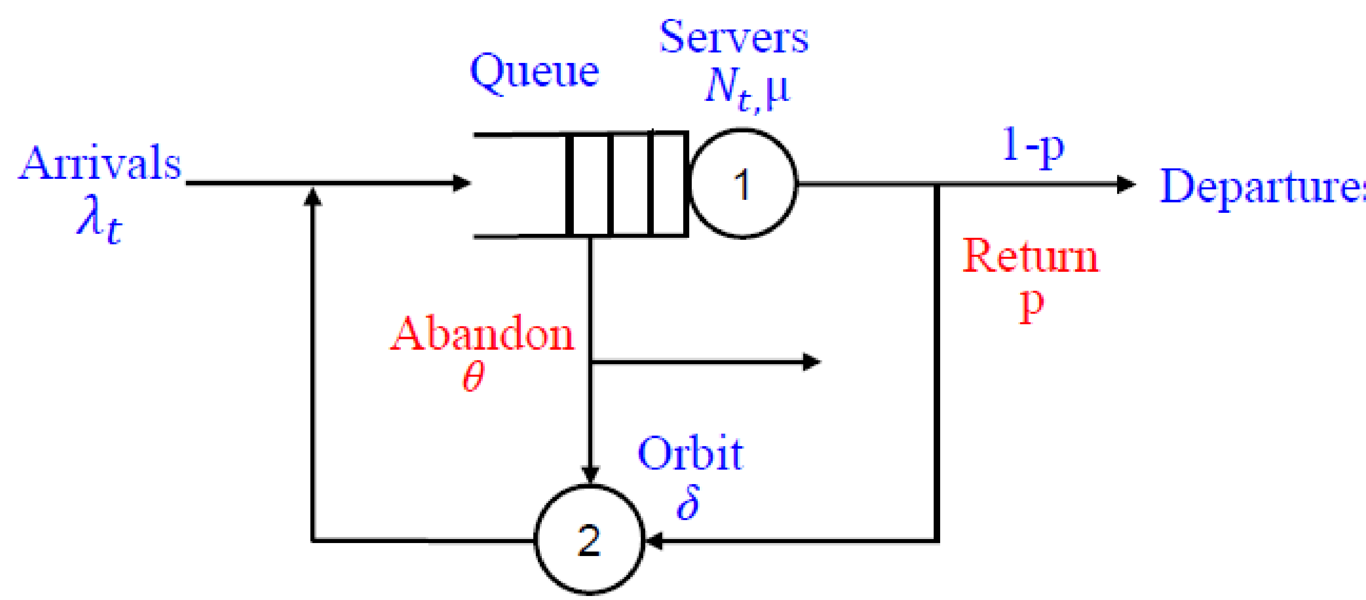
Impatient Customers - Isolate or Aggregate

ILTelecom 9/3/2008



Model Selection: As Simple as Possible but Not Simpler

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

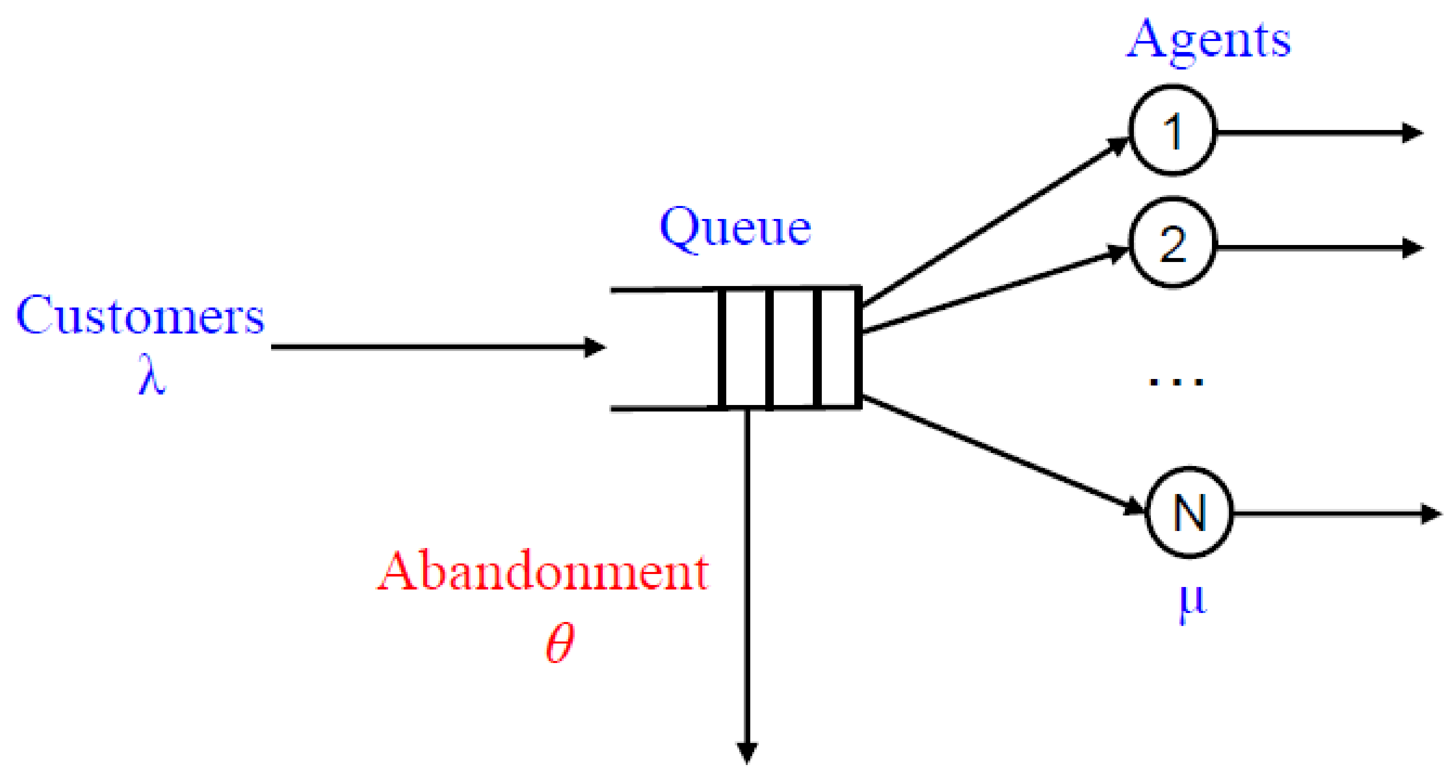


- Call centers: Visit durations naturally measured in minutes
  - Arrival rates are “constant" during visit
  - Returns occur hours after visit

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

Stationary, Abandonment

A Basic Staffing Model: Erlang-A



w/ O. Garnett

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

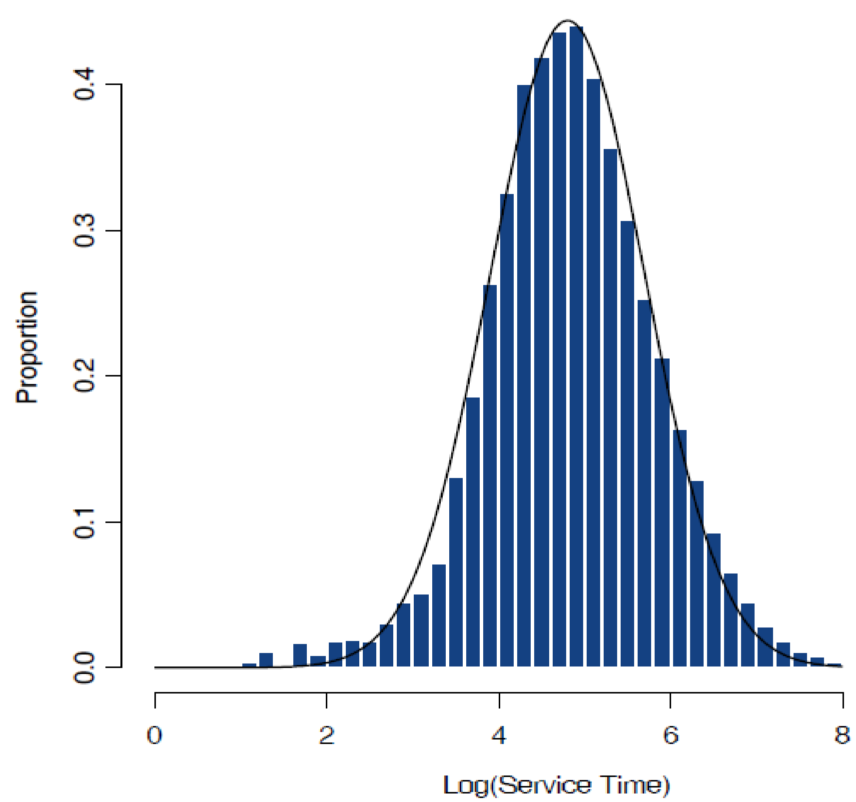
- λ – **Arrival** rate (Poisson)
- μ – **Service** rate (Exponential; $E[S] = \frac{1}{\mu}$)
- θ – **Patience** rate (Exponential, $E[\text{Patience}] = \frac{1}{\theta}$)
- N – Number of **Servers** (Agents).

Pause for a Commercial: Service Duration in a Call Center

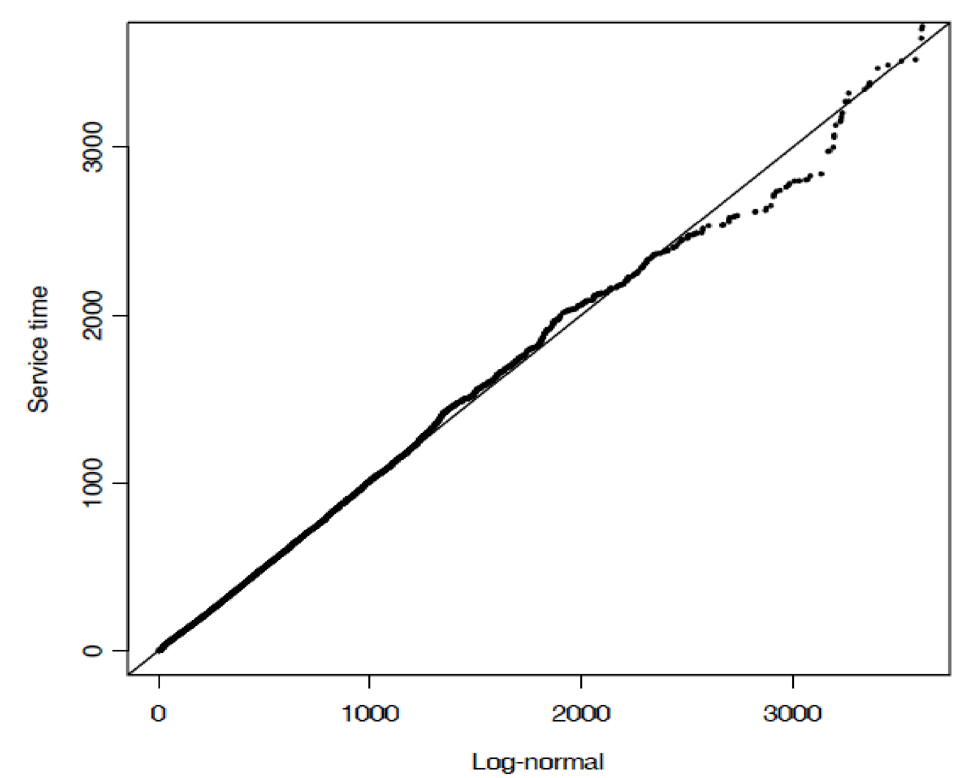
Durations: Phone Calls (Surprises)

Israeli Call Center, Nov–Dec, 1999

Log(Service Times)



LogNormal QQPlot



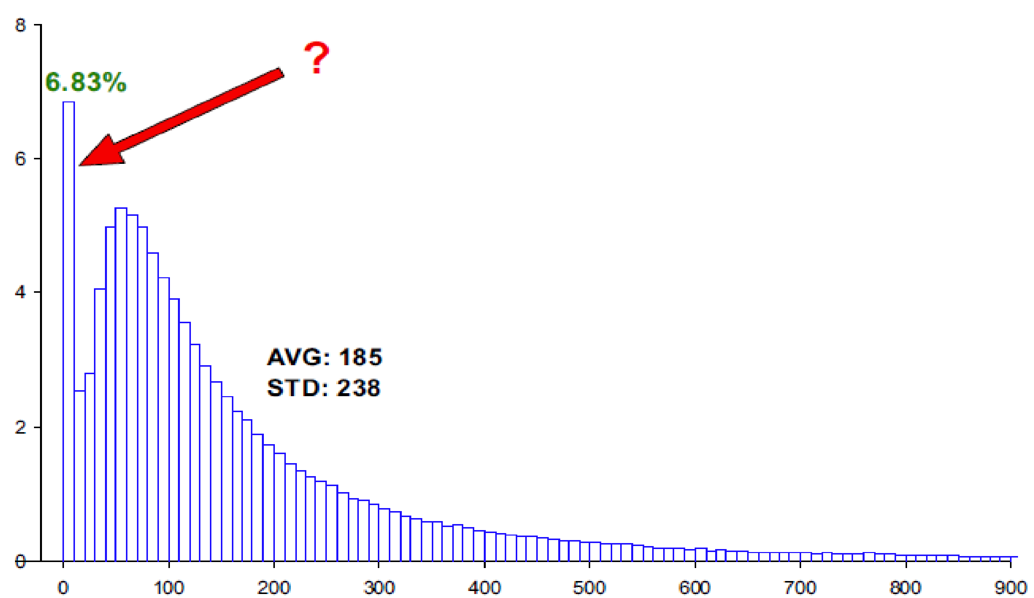
- **Practically Important**: (mean, std)(log) characterization
- **Theoretically Intriguing**: Why LogNormal ? Naturally multiplicative but, in fact, also **Infinitely-Divisible** (Generalized Gamma-Convolutions)

Building Blocks of a Model: Service Durations

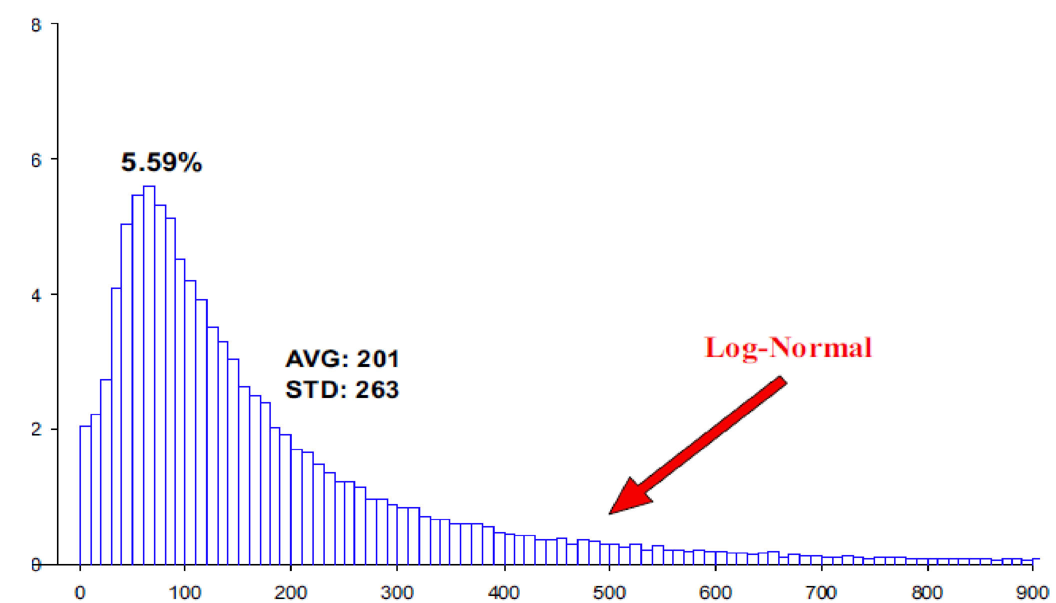
Histogram of Service-Duration in an Israeli Call Center, 1999

Why short services? Why LogNormal?

January-October



November-December



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

Erlang-A: Is it Relevant?

Experience:

- Arrival process **not pure Poisson** (time-varying, σ^2 too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (behavior-dependent).
- Building Blocks need **not be independent** (eg. long wait associated with long service; **w/ M. Reich & Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills): **w/ R. Atar, G. Shaikhet; R. Atar, I. Gurvich,** . . .
- Customers return for service (after busy, abandonment; dependently: **P. Khudiakov, R. Ghebali, M. Gorfine, P. Feigin**)
- · · · , and more.

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

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

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

Asymptotic Landscape: 9 Operational Regimes, and then some

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

Erlang-A	Conventional scaling			Many-Server scaling			NDS scaling		
μ & θ fixed	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 λ	$\frac{\mu}{1+\delta}$	$\mu-\frac{\beta}{\sqrt{n}}\mu$	$\frac{\mu}{1-\gamma}$	$\frac{n\mu}{1+\delta}$	$n\mu-\beta\mu\sqrt{n}$	$\frac{n\mu}{1-\gamma}$	$\frac{n\mu}{1+\delta}$	$n\mu-\beta\mu$	$\frac{n\mu}{1-\gamma}$
# servers	1			n			n		
Time-scale	n			1			n		
Impatience rate	θ/n			θ			θ/n		
Staffing level	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}(1+\frac{\beta}{\sqrt{n}})$	$\frac{\lambda}{\mu}(1-\gamma)$	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}$	$\frac{\lambda}{\mu}(1-\gamma)$	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}+\beta$	$\frac{\lambda}{\mu}(1-\gamma)$
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
$\mathbb{E}(Q)$	$\frac{1}{\delta(1+\delta)}$	$\sqrt{n}g(\hat{\beta})$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{1}{\delta}\varrho_n$	$\sqrt{n}g(\hat{\beta})\alpha$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$o(1)$	$ng(\hat{\beta})$	$\frac{n^2\mu\gamma}{\theta(1-\gamma)}$
$\mathbb{P}(Ab)$	$\frac{1}{n}\frac{1}{\delta}\frac{\theta}{\mu}$	$\frac{\theta}{\sqrt{n}\mu}g(\hat{\beta})$	γ	$\frac{1}{n}\frac{(1+\delta)}{\delta}\frac{\theta}{\mu}\varrho_n$	$\frac{\theta}{\sqrt{n}\mu}g(\hat{\beta})\alpha$	γ	$o(\frac{1}{n^2})$	$\frac{\theta}{n\mu}g(\hat{\beta})$	γ
$\mathbb{P}(W_q>0)$	$\frac{1}{1+\delta}$	≈ 1		ϱ_n	$\alpha\in(0,1)$	≈ 1	≈ 0	≈ 1	
$\mathbb{P}(W_q>T)$	$\frac{1}{1+\delta}e^{-\frac{\delta}{1+\delta}\mu T}$	$1+O(\frac{1}{\sqrt{n}})$	$1+O(\frac{1}{n})$	≈ 0		$f(T)$	≈ 0	$\frac{\bar{\Phi}(\hat{\beta}+\sqrt{\theta\mu}T)}{\bar{\Phi}(\hat{\beta})}$	$1+O(\frac{1}{n})$
Congestion $\frac{EW_q}{ES}$	$\frac{1}{\delta}$	$\sqrt{n}g(\hat{\beta})$	$n\mu\gamma/\theta$	$\frac{1}{n}\frac{(1+\delta)}{\delta}\varrho_n$	$\frac{\alpha}{\sqrt{n}}g(\hat{\beta})$	$\frac{\mu\gamma}{\theta}$	$o(\frac{1}{n})$	$g(\hat{\beta})$	$n\mu\gamma/\theta$

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

Pause for a Commercial: Finally a Theorem (QED)

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 ⇒ **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

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

Here $\boldsymbol{R}$ = **Offered Load**
eg. $R = 25$ 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:**

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

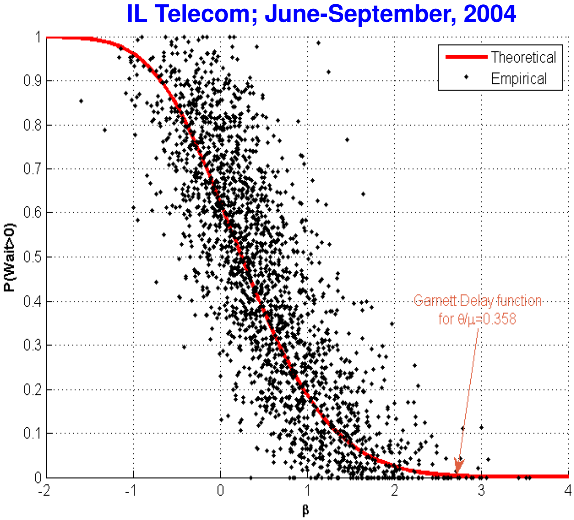
- **Probability to Abandon:**

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

- $P\{Ab\} \propto \mathrm{E}[W_q]$, 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: $\hat{Q}_N(t) \xrightarrow{d} \hat{Q}(t)$.

$$\begin{tikzcd}
\hat{Q}_N(t) \arrow[r, "t \to \infty"] \arrow[d, "N \to \infty"'] & \hat{Q}_N(\infty) \arrow[d, "N \to \infty"] \\
\hat{Q}(t) \arrow[r, "t \to \infty"'] & Q(\infty)
\end{tikzcd}$$

Approximating (Virtual) Waiting Time

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

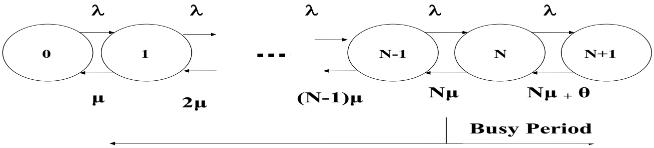
QED Intuition: Why $P\{W_q > 0\} \in (0,1)$?

1. Why **subtle**: Consider a large service system (e.g. call center).
   - Fix λ 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 λ 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 $Z \stackrel{d}{=} N(0,1)$.
   - $P\{W_q(M/M/N+M) > 0\} \stackrel{PASTA}{=} P\{L(M/M/N+M) \geq N\} \stackrel{\mu=\theta}{=}$

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

     $P\{Z \geq (N-R)/\sqrt{R}\} \stackrel{\sqrt{\cdot}\text{ staffing}}{\approx} P\{Z \geq \beta\} = 1 - \Phi(\beta)$.
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 $T_{N-1,N} = \dfrac{1}{\lambda_N E_{1,N-1}} \sim \dfrac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \dfrac{1}{\sqrt{N}} \cdot \dfrac{1/\mu}{h(-\beta)}$

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

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

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

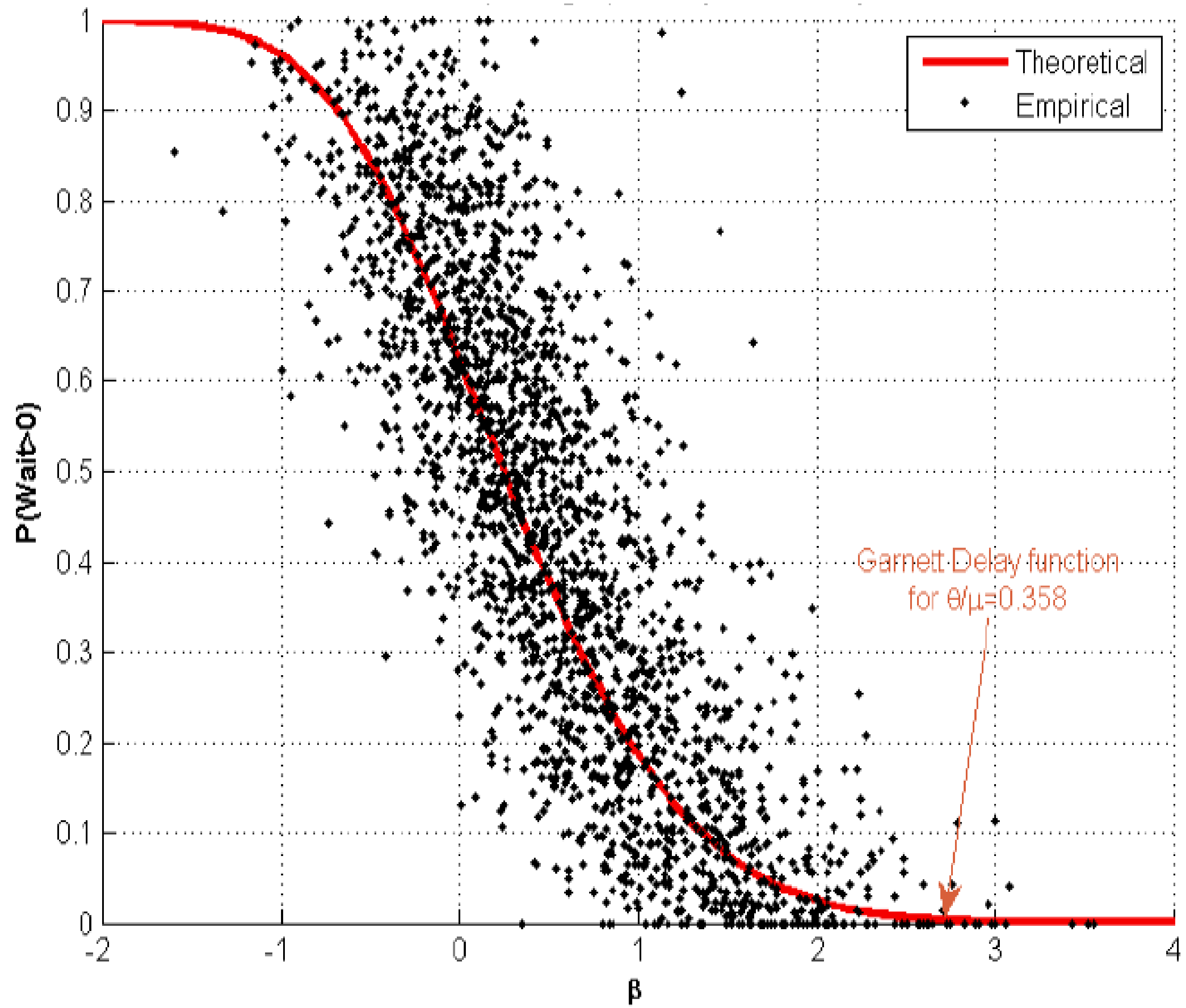
Special case: $\boldsymbol{\mu = \theta}$ (Impatient):

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

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

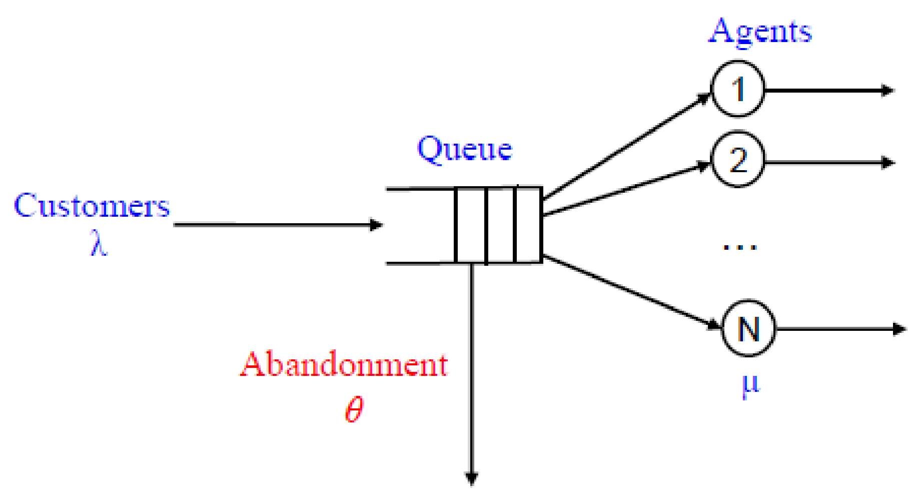
Erlang-A Value: DNet $P(W_q > 0)$ vs. Data

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



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

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



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

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

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

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

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

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

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

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

Parsimonious (Tractable, Robust), Accurate, Valuable

Abandonment: Further Applications

- **Personalized Queueing**: e.g. when Impatience has been estimated (personalized), exactly or approximately, choosing who to serve next will benefit from taking this information into account (e.g. Shortest-Patience-First); **w/ P. Momcilovic**
- **CCs:** Control of congestion via announcement / information that monitors the queue**, w/ Junfei Huang, Hanqin Zhang, Jiheng Zhang**
- **IVRs** (Self-Services): identify via Mixture Fitting, **w/ N. Carmeli, H. Kaspi**
- **EDs**: Left-Without-Being-Seen. Theory exists if All-Unknown – Current-Status Data. Otherwise, namely when having both Right-Censoring and Current-Status data, theory developed **w/ Y. Yefenof, Y. Goldberg, Y. Ritov**
- **Chats:** Before or within session (registered vs. silent abandonment)

Laws & Models: Data-Based Erlang A/R/S, following B/C

- **Little's Law (Steady-State, Transient), State-Space Collapse,…**

- **Erlang-B (Blocking) and Erlang-C (?)**
  - **Erlang**, Agner Krarup: Queueing Theory was born in 1909, in his paper "The Theory of Probabilities and Telephone Conversations"

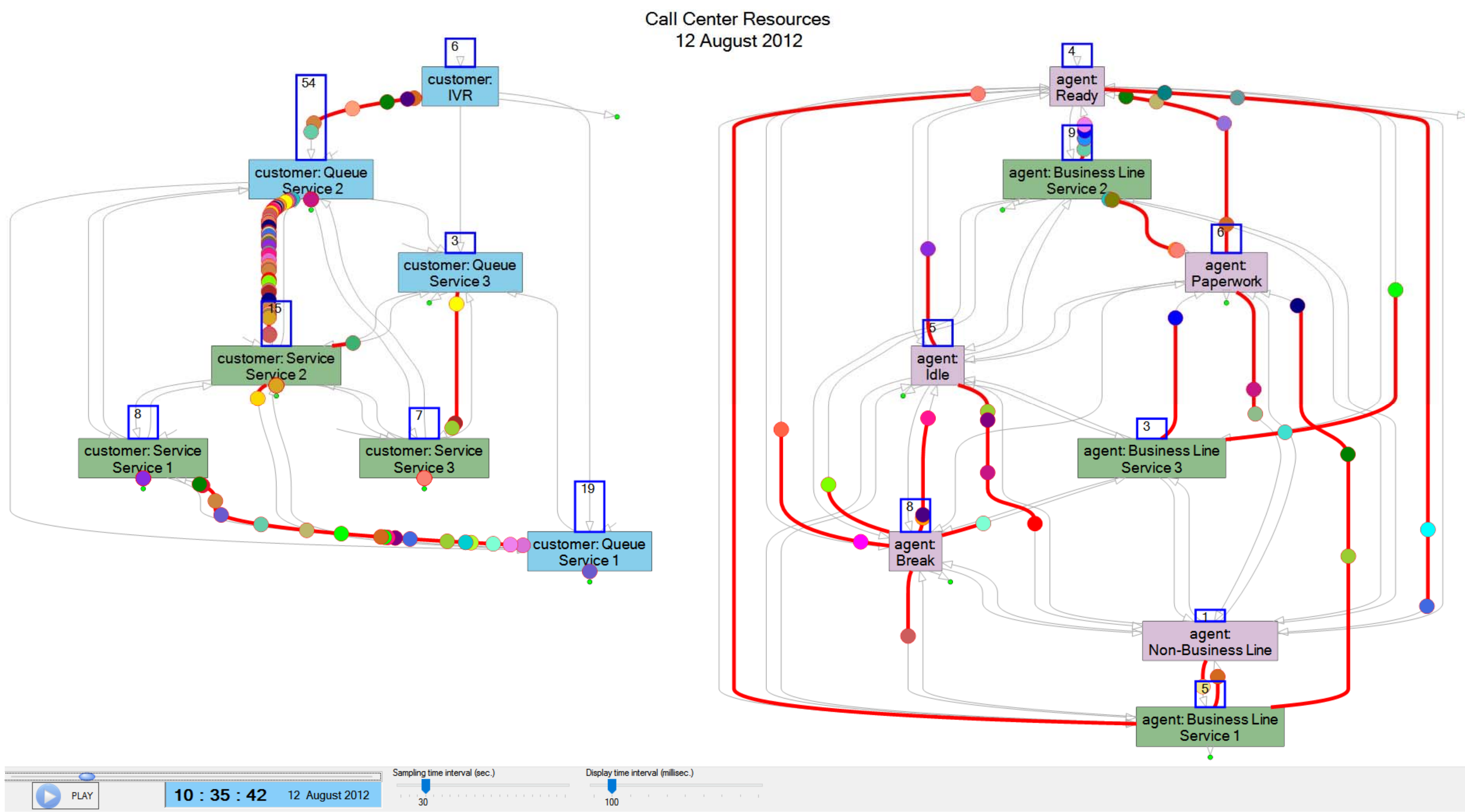
- **2. Erlang-A**
  - **Abandonment**: While waiting for service, does service-value dominate residual-wait-cost

- **1. Erlang-R**
  - **Return/Feedback**: Customers often return to service (positively, negatively, just needing)

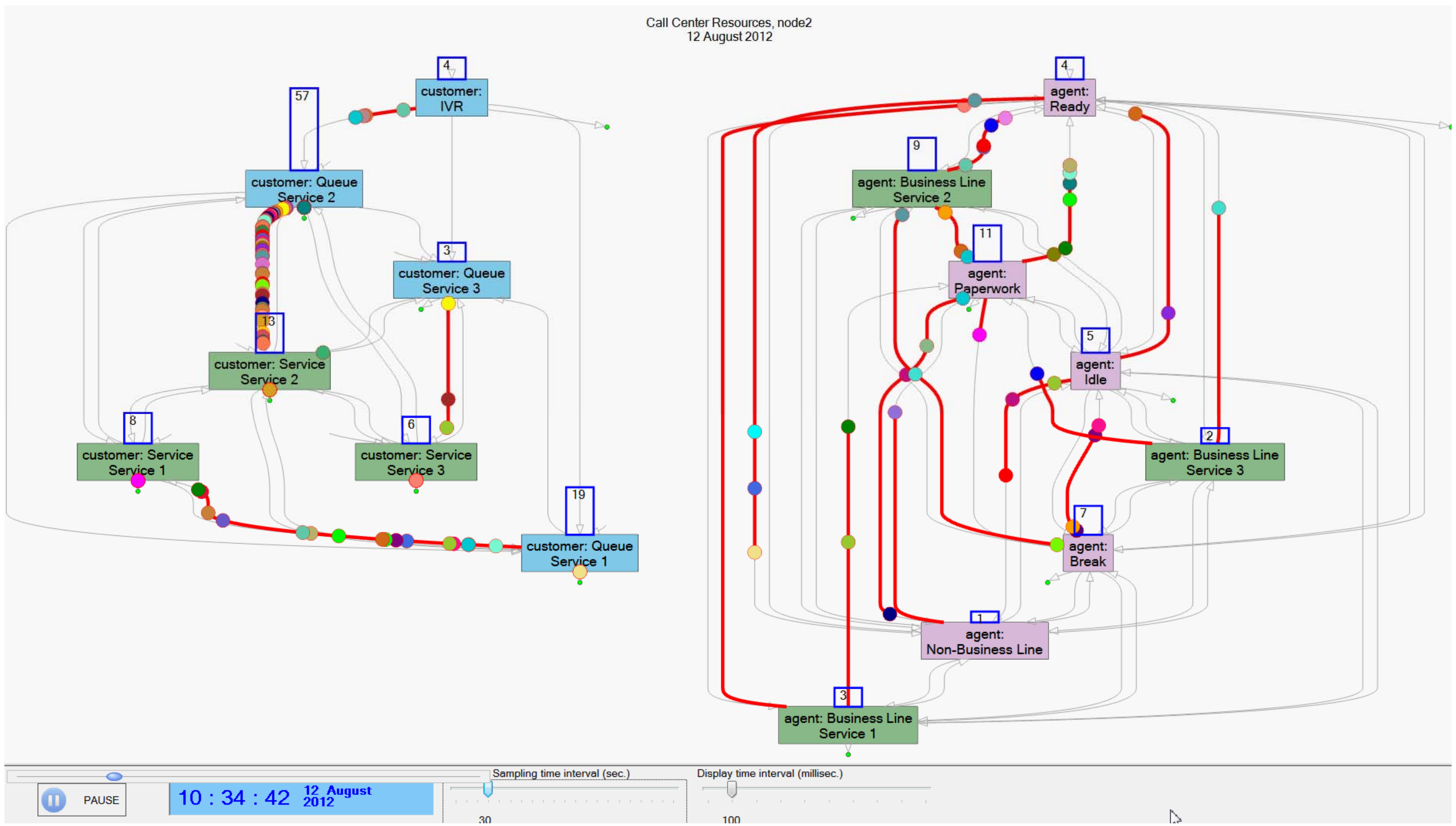
- **3. Erlang-S**
  - **Servers**: Challenging to manage, and model, no less so than customers

Above: Simple (Parsimonious) models of complex realities, yet not too simple (Robust)

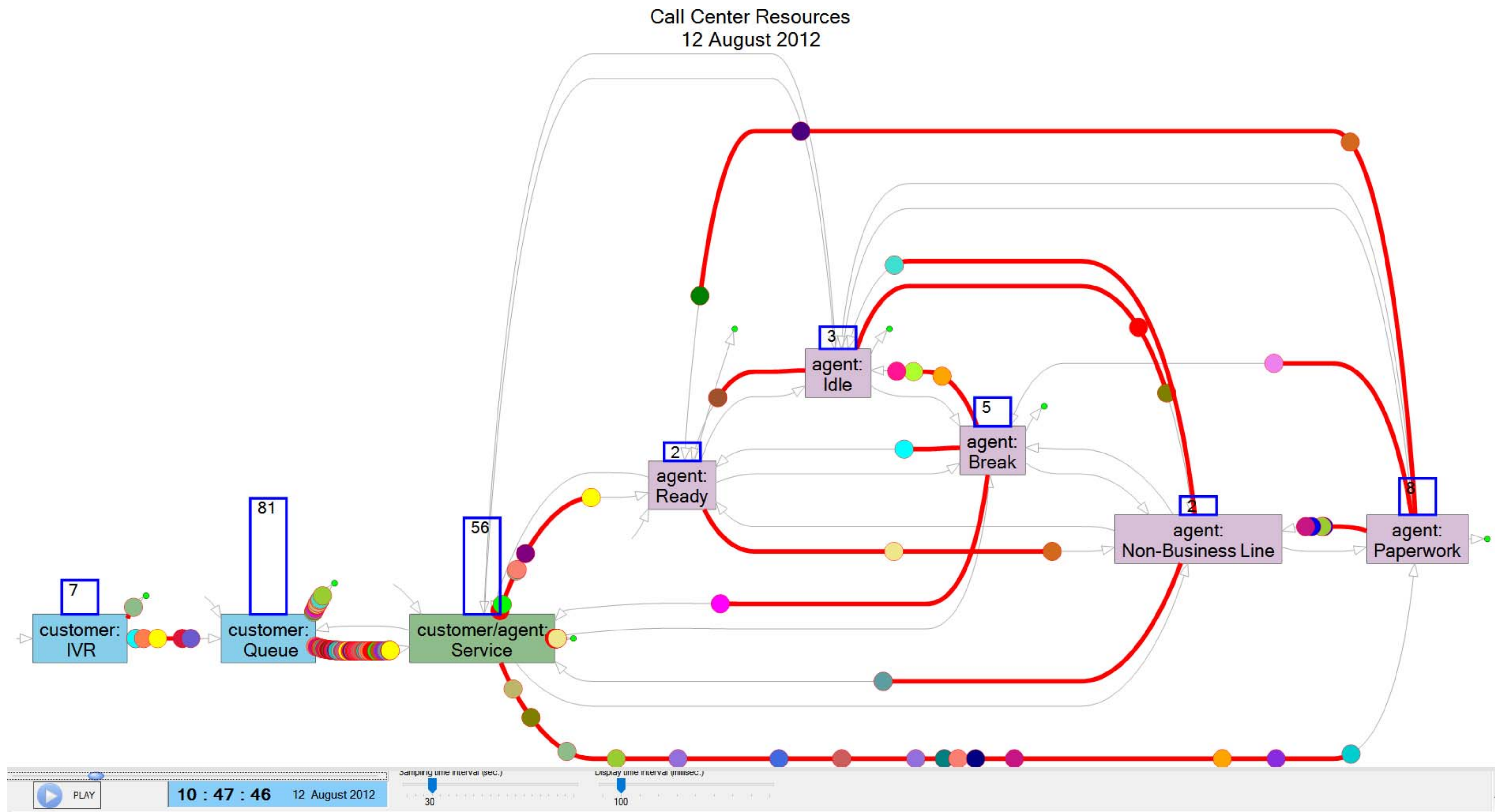
Telephone Services: Customers and Servers - Symmetric Viewpoint



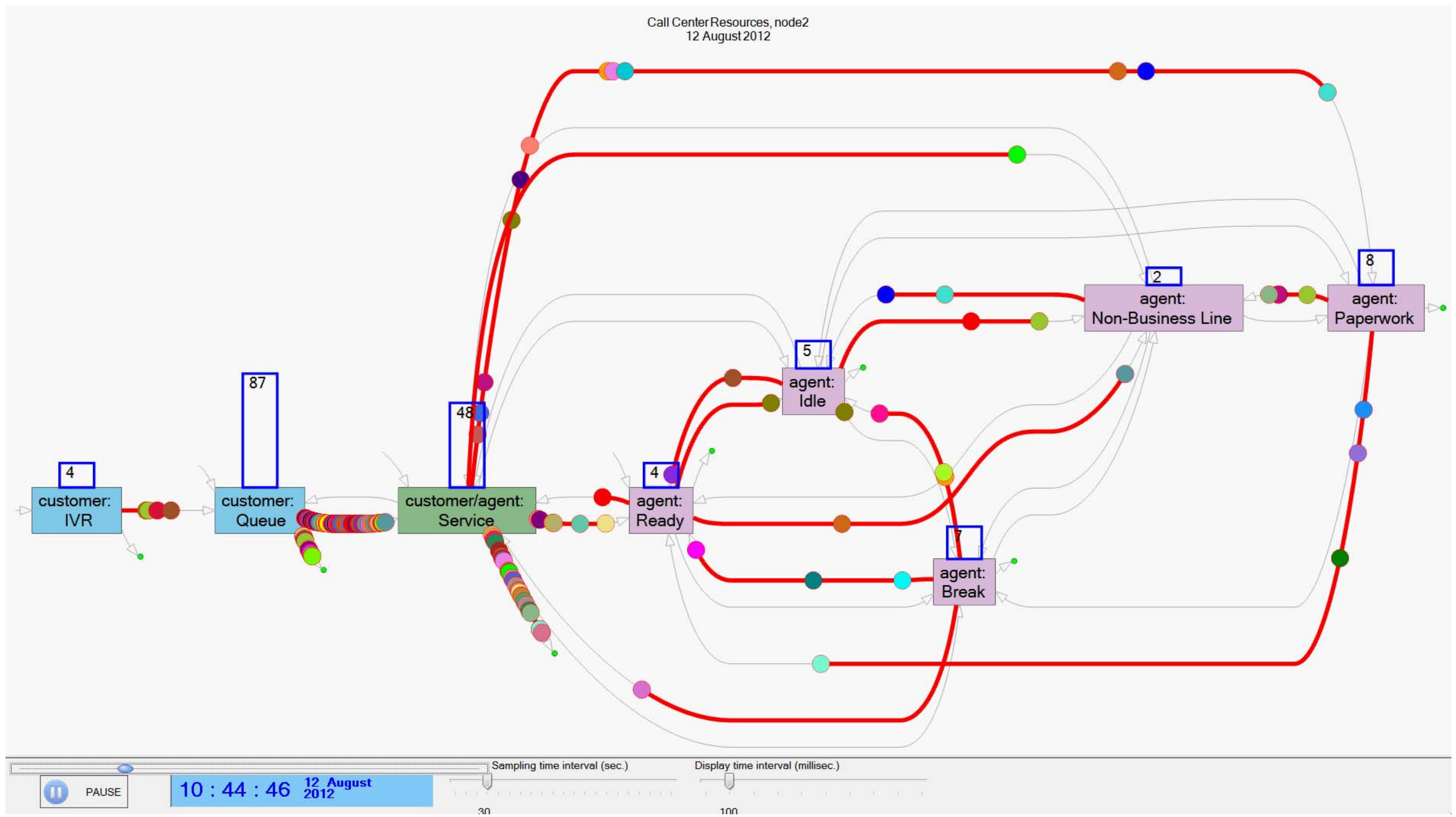
Telephone Services: Customers and Servers Symmetric Viewpoint



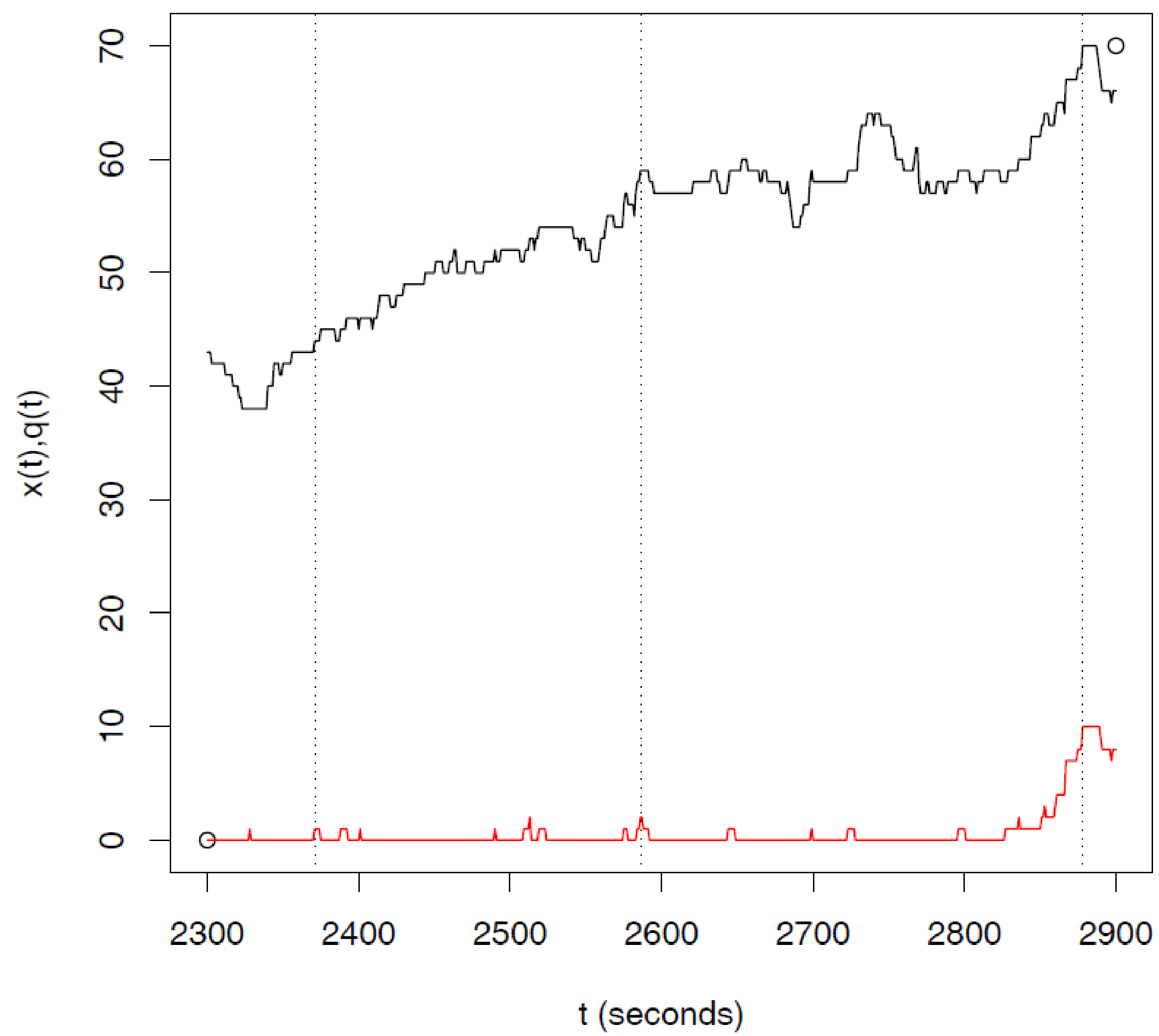
Symmetric View: **Customers & Servers**
Originates in Resource-Driven Activity-Networks
w/ **Armony, Carmeli, Momcilovic**



Symmetric View: Customers and Servers
Originates in Resource-Driven Activity-Networks:
w/ **Armony, Carmeli, Momcilovic**



Number of Servers: Present (=Constant) vs. Available (Random)

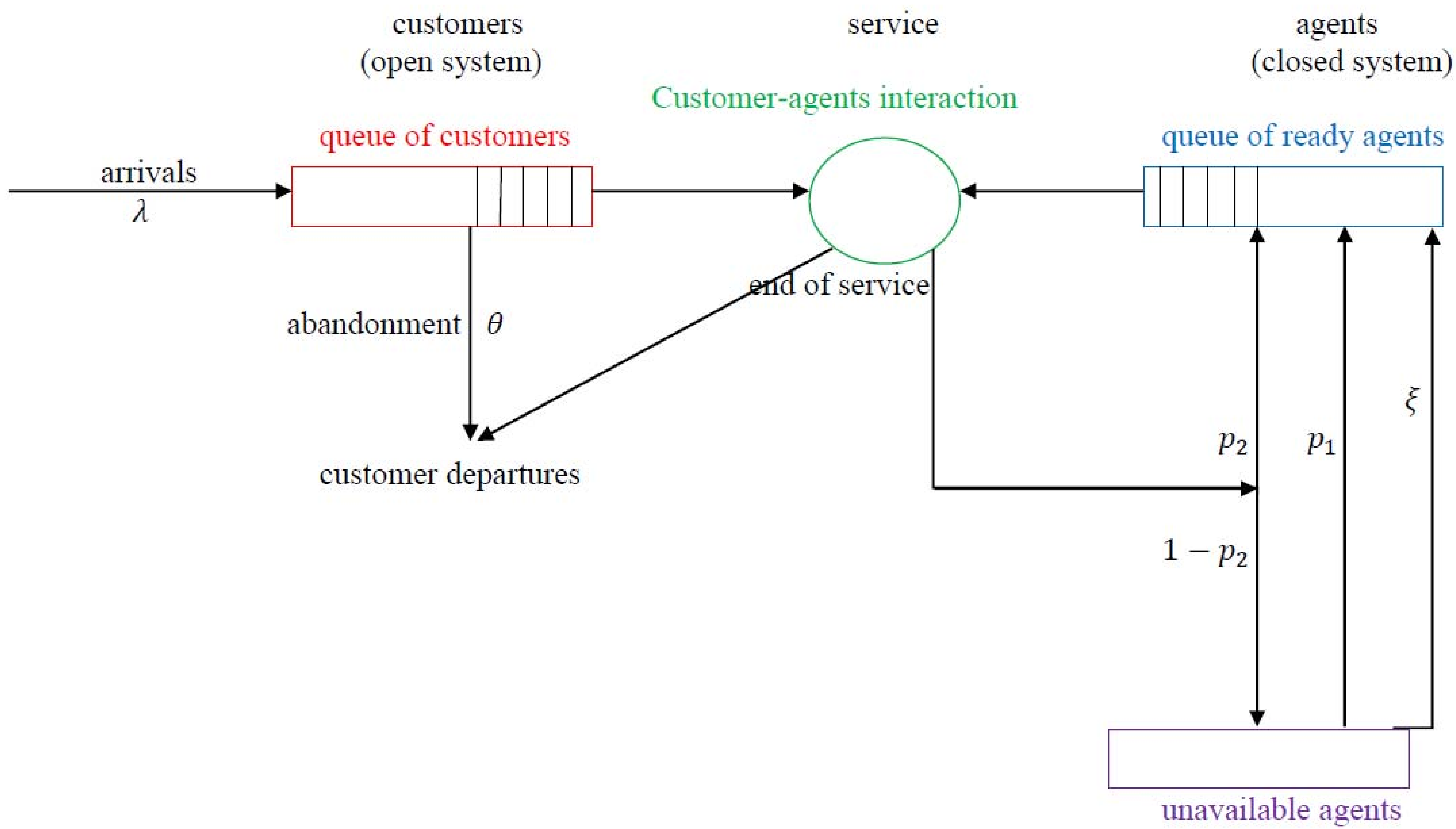


x(t), q(t) in real data (U.S. Bank; telesales; 12/1/2002)

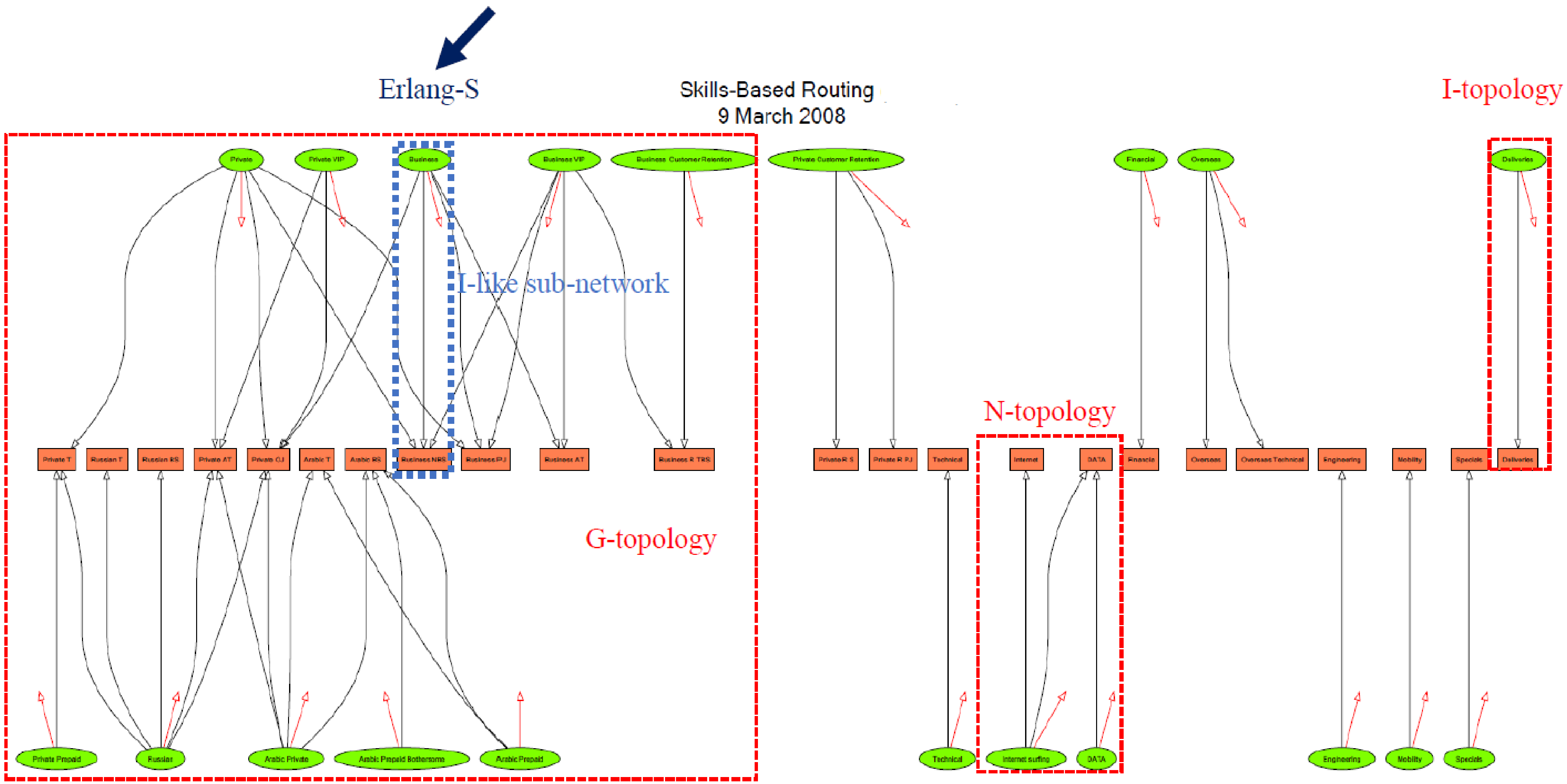
n(t) = (43,57,60)   @*t* = (10:36:54, 10:43:07, 10:47:23):

17 agents became available within **11** minutes

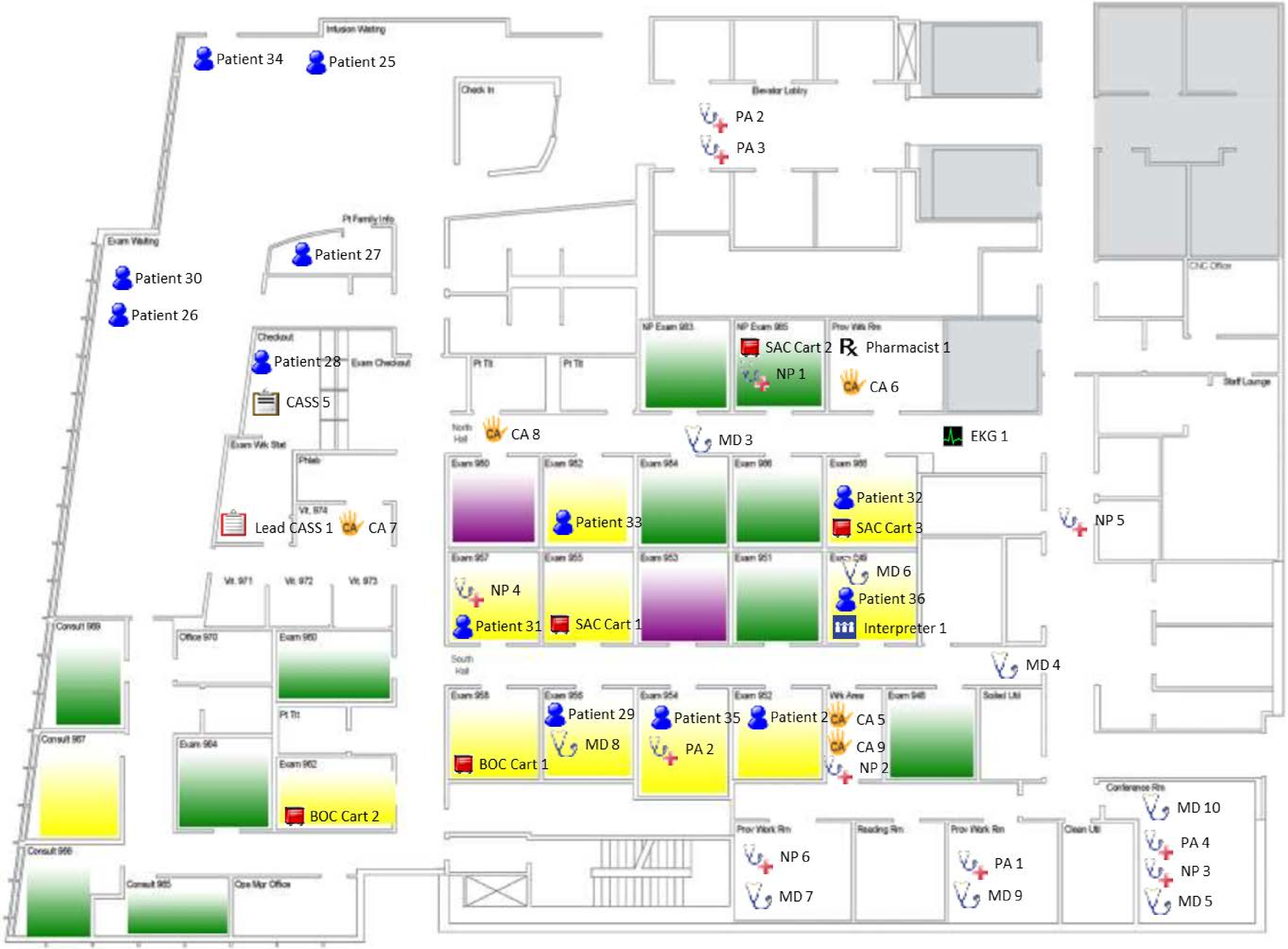
Erlang-S w/ David Azriel & Paul Feigin



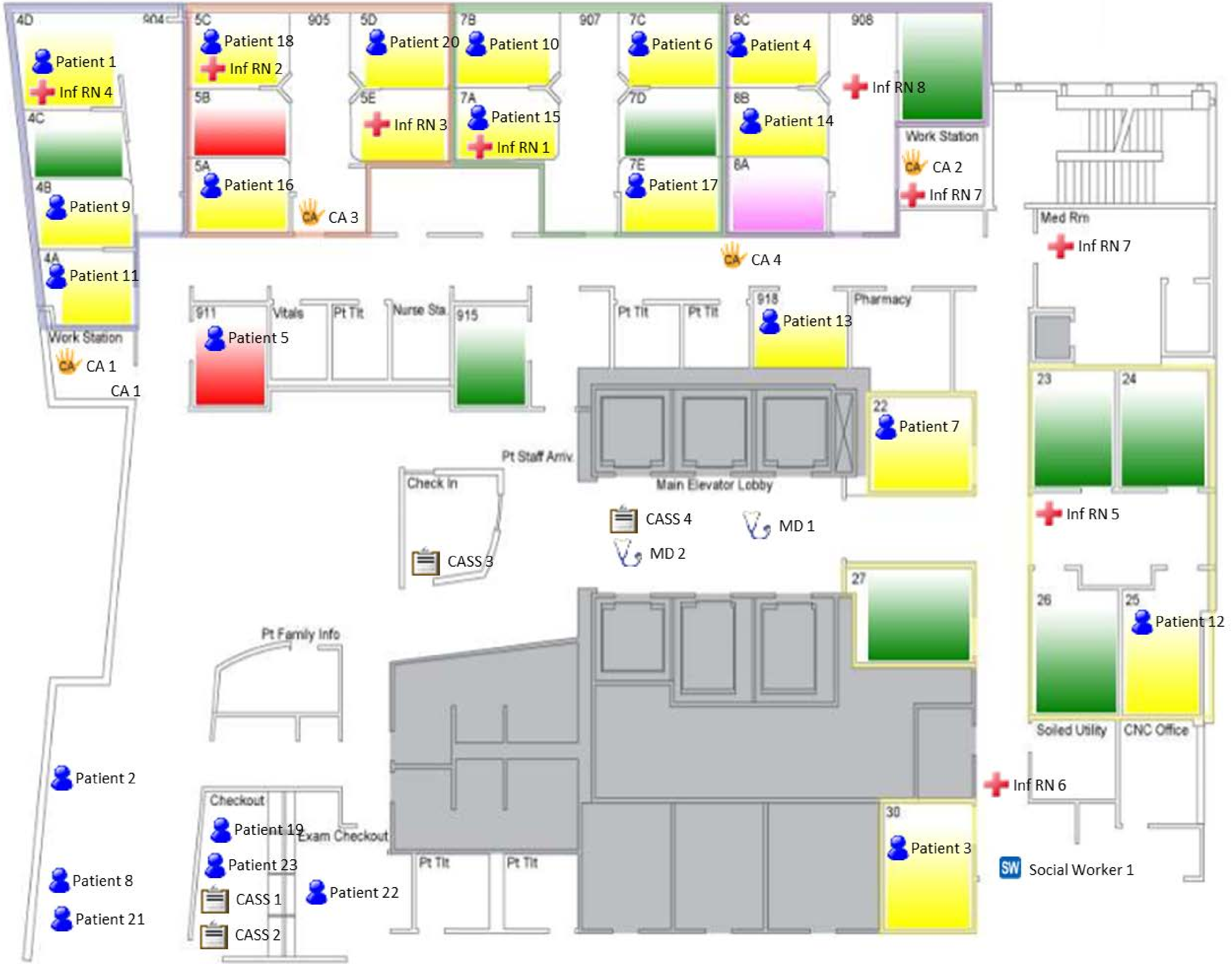
Expressiveness of Erlang-S (Zoom): Servers in Qnets

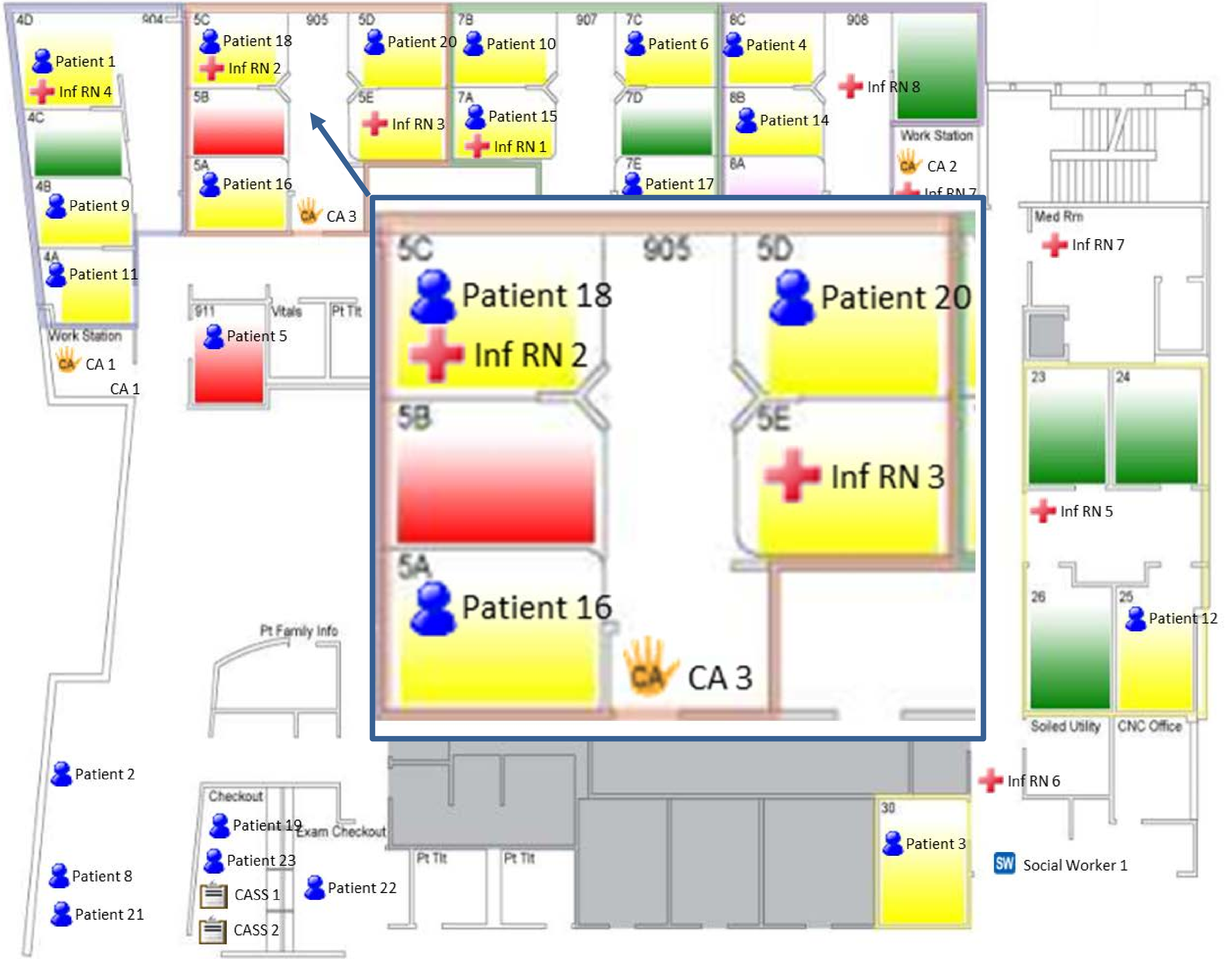


Topology of a call center:
Server-queues are in the rectangles and customer-queues are in the ovals



Infusion Waiting
Patient 34
Patient 25
Check In
Elevator Lobby
PA 2
PA 3
Pt Family Info
Patient 27
Exam Waiting
Patient 30
Patient 26
Checkout
Patient 28
Exam Checkout
CASS 5
NP Exam 983
NP Exam 985
SAC Cart 2
NP 1
Pharmacist 1
CA 6
CNC Office
Staff Lounge
CA 8
MD 3
EKG 1
Patient 32
SAC Cart 3
NP 5
Patient 33
Lead CASS 1
CA 7
NP 4
Patient 31
SAC Cart 1
MD 6
Patient 36
Interpreter 1
MD 4
Patient 29
MD 8
BOC Cart 1
Patient 35
PA 2
Patient 2
CA 5
CA 9
NP 2
BOC Cart 2
MD 10
PA 4
NP 3
MD 5
NP 6
MD 7
PA 1
MD 9


4D
904
Patient 1
Inf RN 4
4C
4B
Patient 9
4A
Patient 11
Work Station
CA 1
CA 1
5C
Patient 18
Inf RN 2
905
5B
5A
Patient 16
CA 3
5D
Patient 20
5E
Inf RN 3
7B
Patient 10
7A
Patient 15
Inf RN 1
907
7C
Patient 6
7D
7E
Patient 17
8C
Patient 4
8B
Patient 14
8A
908
Inf RN 8
Work Station
CA 2
Inf RN 7
CA 4
Med Rm
Inf RN 7
911
Patient 5
Vitals
Pt Tlt
Nurse Sta
915
Pt Tlt
Pt Tlt
918
Patient 13
Pharmacy
22
Patient 7
Pt Staff Arriv.
Main Elevator Lobby
Check In
CASS 3
CASS 4
MD 1
MD 2
23
24
Inf RN 5
27
26
25
Patient 12
Soiled Utility
CNC Office
Pt Family Info
Inf RN 6
Patient 2
Checkout
Patient 19
Exam Checkout
Patient 23
CASS 1
CASS 2
Patient 22
Patient 8
Patient 21
Pt Tlt
Pt Tlt
30
Patient 3
SW
Social Worker 1


4D
904
5C
905
5D
7B
907
7C
8C
908
Patient 1
Inf RN 4
4C
4B
Patient 9
4A
Patient 11
Work Station
CA 1
CA 1
Patient 18
Inf RN 2
5B
5A
Patient 16
CA 3
Patient 20
5E
Inf RN 3
Patient 10
7A
Patient 15
Inf RN 1
Patient 6
7D
7E
Patient 17
Patient 4
8B
Patient 14
8A
Inf RN 8
Work Station
CA 2
Inf RN 7
Med Rm
Inf RN 7
911
Vitals
Pt Tlt
Patient 5
23
24
Inf RN 5
26
25
Patient 12
Soiled Utility
CNC Office
Inf RN 6
Pt Family Info
Patient 2
Patient 8
Patient 21
Checkout
Patient 19
Exam Checkout
Patient 23
CASS 1
CASS 2
Patient 22
Pt Tlt
Pt Tlt
30
Patient 3
Social Worker 1
5C
905
5D
Patient 18
Inf RN 2
Patient 20
5B
5E
Inf RN 3
5A
Patient 16
CA 3

System Badge Type Icons

STAFF

MD

Physician

NP/PA

Nurse Practitioner / Physician Assistant

Fellow

Fellow

CA

Clinic Assistant

Infusion RN

Inf Nurse

Res RN

Research RN

Program RN

Program RN

CNC

Clinical Nurse Coordinator

PB9

Phlebotomist

Pharm

Pharmacist

Social Worker

Social Worker

CASS

CASS

- CPC

CPC

- Ambassador

Ambassador

Manager

Manager

IV RN

IV Nurse

Lead CA

Lead CA

Charge RN

Charge RN

Charge Pharm

Charge Pharm

Pharm Tech

Pharm Tech

Scheduler

Scheduler

Interpreter

Interpreter

Nutritionist

Nutritionist

Respiratory Therapist

Respiratory Therapist

Genetic Counselor

Genetic Counselor

Lead CASS

Lead CASS

CRC

CRC

Exam Practice RN

Exam Practice RN

Massage Therapist

Massage Therapist

PATIENT, FAMILY & OTHER

Patient

Patient

Family

Family

Versus

Versus Staff

SysAdmin

System Administrator

Other Staff

Other Staff

EQUIPMENT

Thermometer

Thermometer

IV Pump

IV Pump

CAS Monitor

CAS Monitor

Stretcher Chair

Stretcher Chair

O2 Sat Monitor

O2 Sat Monitor

AccuVein

AccuVein

Chart

Chart

Phone

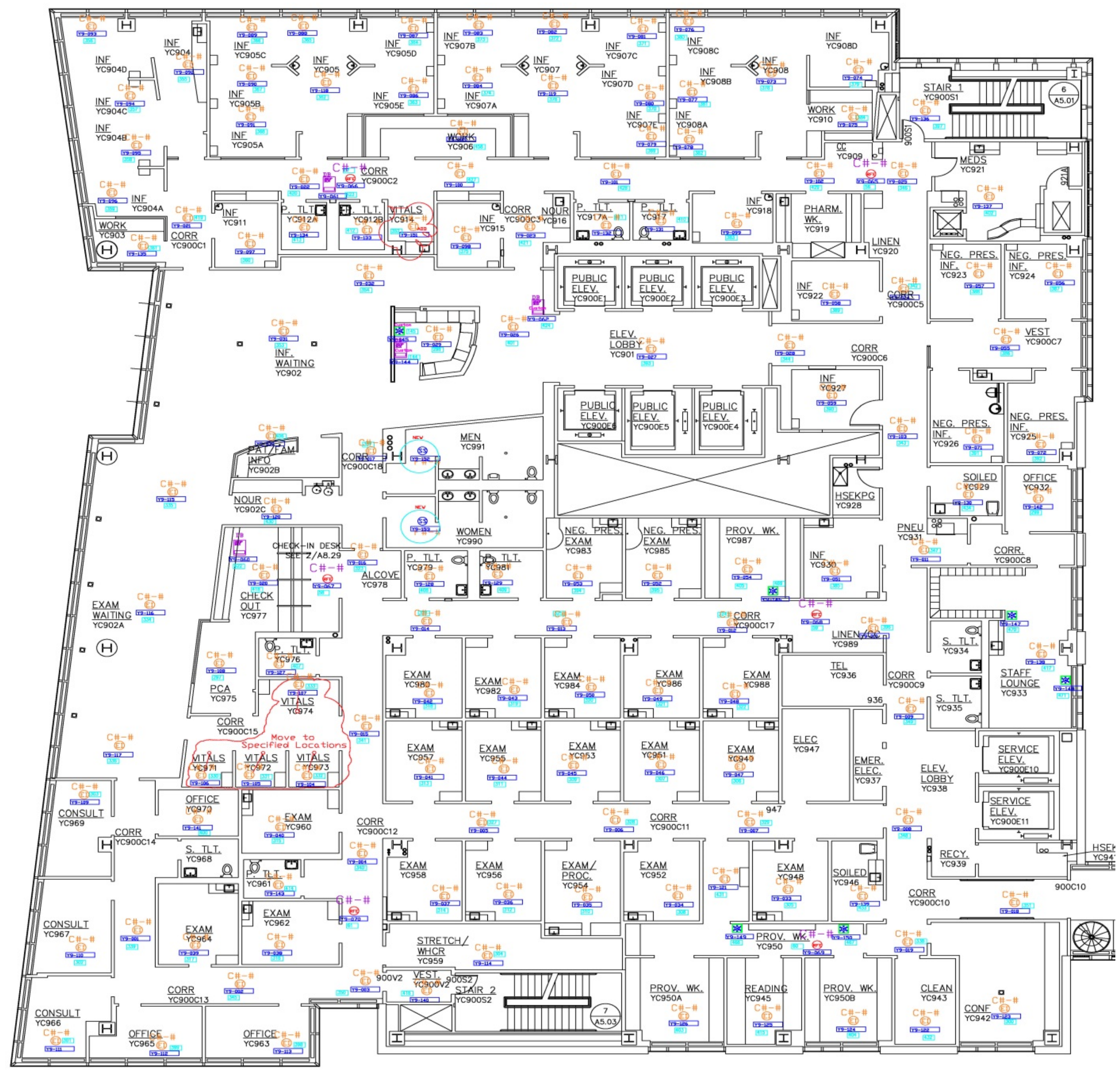
Phone

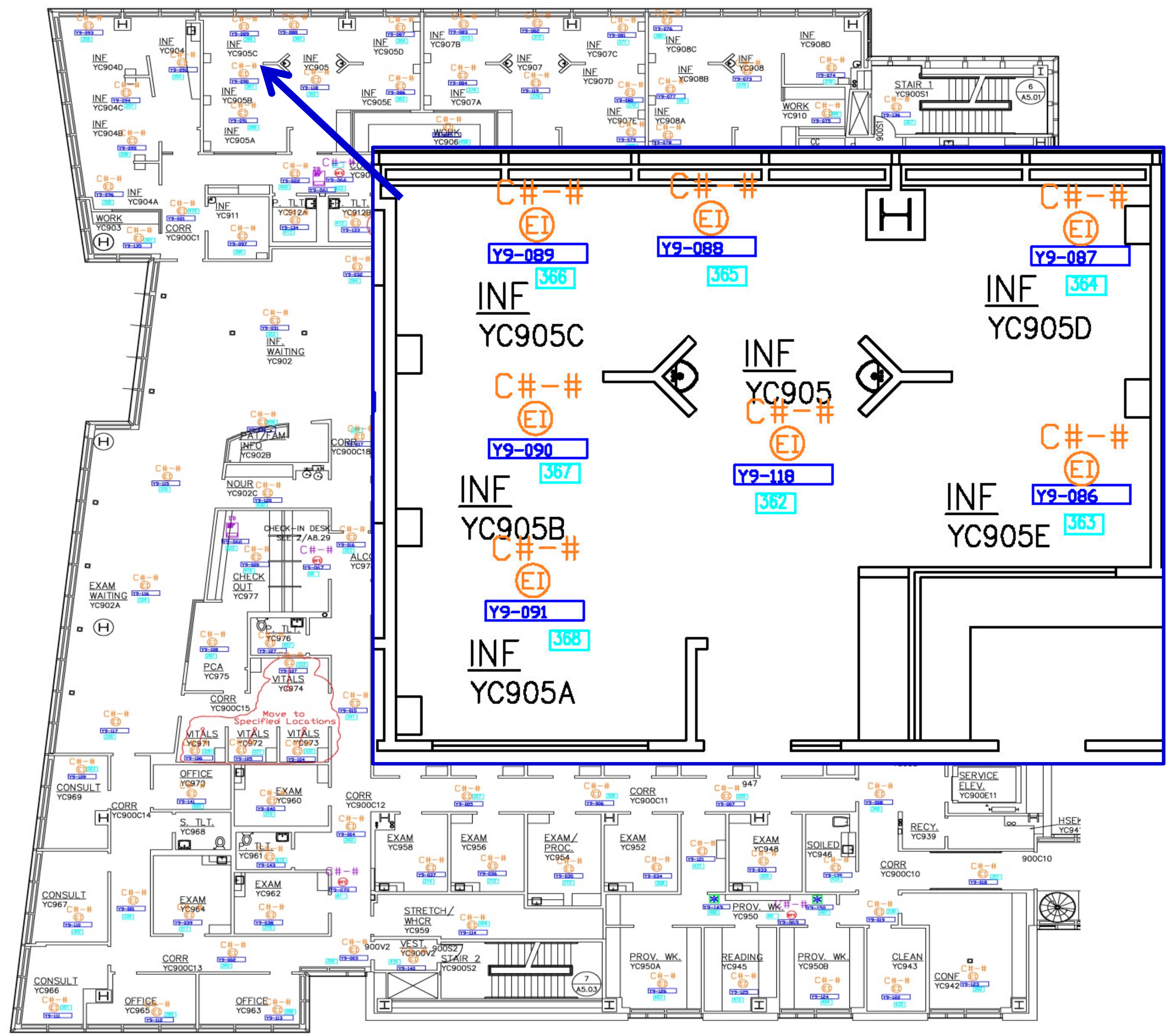
Cart

Cart

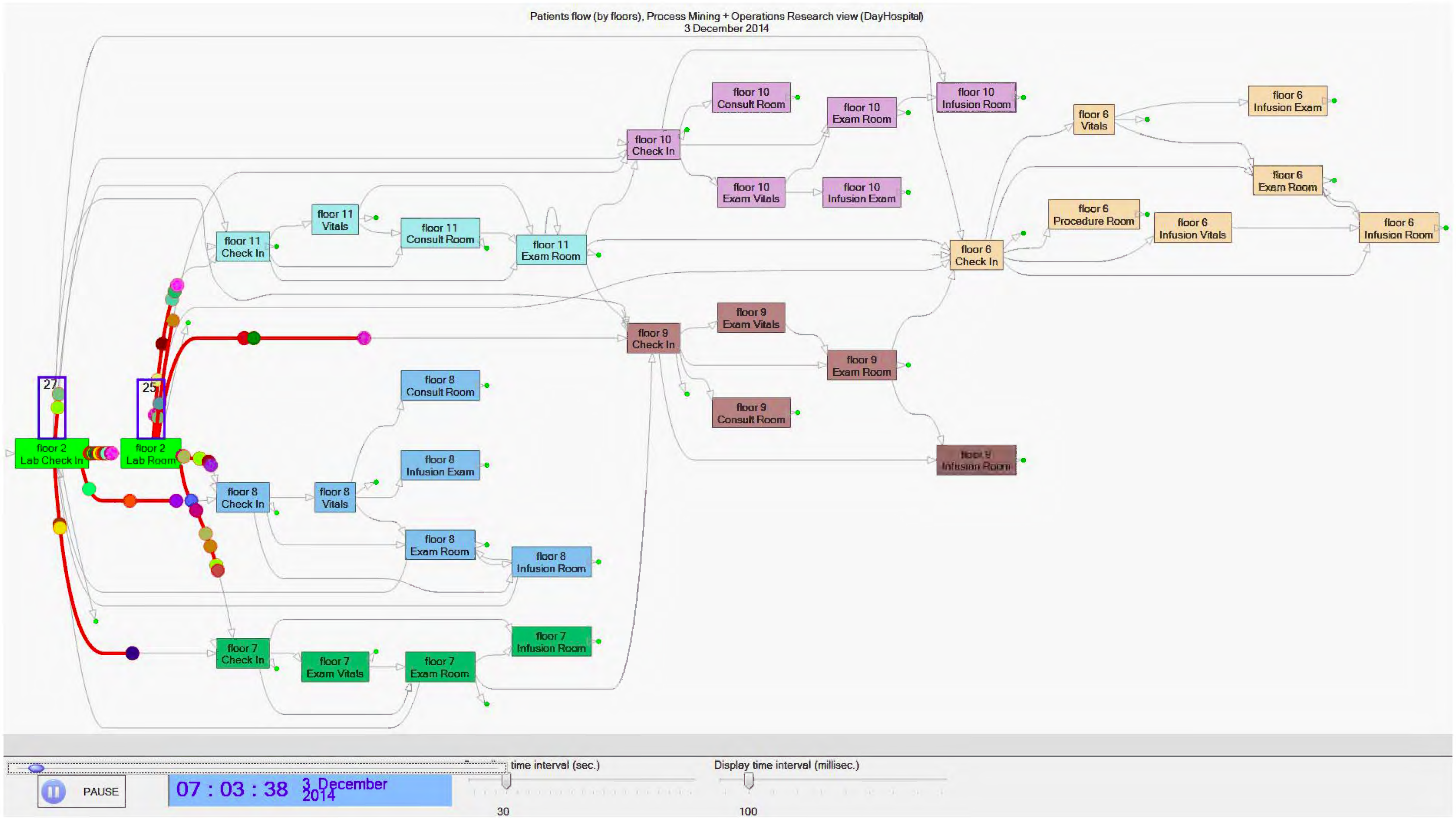
EKG

EKG

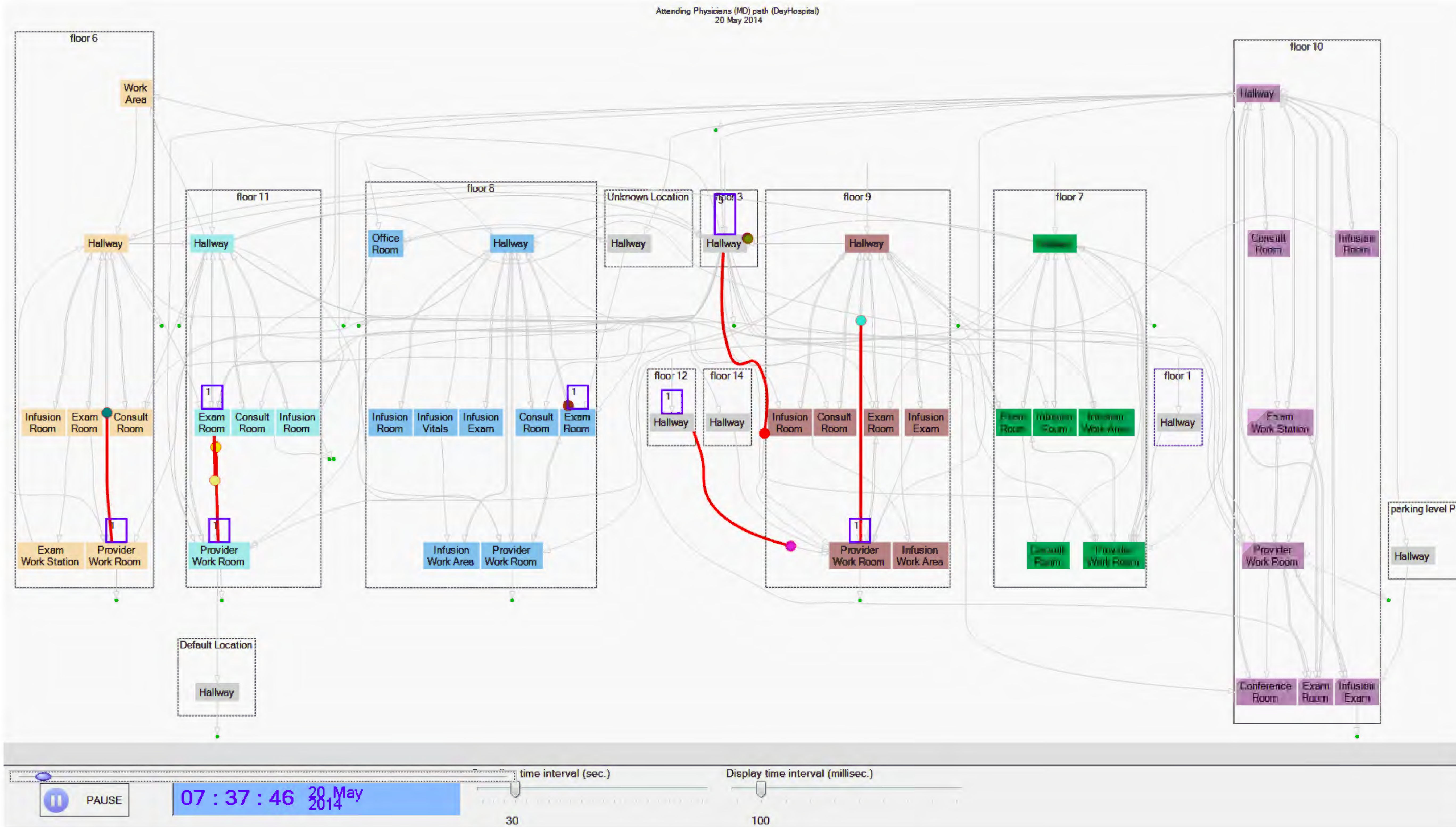

Move to Specified Locations



C#-#
EI
Y9-089
366
Y9-088
365
Y9-087
364
INF
YC905C
INF
YC905D
INF
YC905
Y9-090
367
Y9-118
362
Y9-086
363
INF
YC905B
INF
YC905E
Y9-091
368
INF
YC905A
Move to Specified Locations



Patients flow (by floors), Process Mining + Operations Research view (DayHospital)
3 December 2014
floor 10 Consult Room
floor 10 Exam Room
floor 10 Infusion Room
floor 10 Check In
floor 10 Exam Vitals
floor 10 Infusion Exam
floor 6 Vitals
floor 6 Infusion Exam
floor 6 Exam Room
floor 6 Procedure Room
floor 6 Infusion Vitals
floor 6 Infusion Room
floor 6 Check In
floor 11 Vitals
floor 11 Check In
floor 11 Consult Room
floor 11 Exam Room
floor 9 Exam Vitals
floor 9 Check In
floor 9 Exam Room
floor 9 Consult Room
floor 9 Infusion Room
floor 8 Consult Room
floor 8 Infusion Exam
floor 8 Check In
floor 8 Vitals
floor 8 Exam Room
floor 8 Infusion Room
floor 2 Lab Check In
floor 2 Lab Room
27
25
floor 7 Check In
floor 7 Exam Vitals
floor 7 Exam Room
floor 7 Infusion Room
PAUSE
07 : 03 : 38
3 December 2014
time interval (sec.)
30
Display time interval (millisec.)
100


Attending Physicians (MD) path (DayHospital)
20 May 2014
floor 6
Work Area
Hallway
Infusion Room
Exam Room
Consult Room
Exam Work Station
Provider Work Room
floor 11
Hallway
Exam Room
Consult Room
Infusion Room
Provider Work Room
floor 8
Office Room
Hallway
Infusion Room
Infusion Vitals
Infusion Exam
Consult Room
Exam Room
Infusion Work Area
Provider Work Room
Unknown Location
Hallway
floor 3
Hallway
floor 12
Hallway
floor 14
Hallway
floor 9
Hallway
Infusion Room
Consult Room
Exam Room
Infusion Exam
Provider Work Room
Infusion Work Area
floor 7
floor 1
Hallway
floor 10
Hallway
Consult Room
Infusion Room
Exam Work Station
Provider Work Room
Conference Room
Exam Room
Infusion Exam
parking level P2
Hallway
Default Location
Hallway
PAUSE
07 : 37 : 46
20 May 2014
time interval (sec.)
30
Display time interval (millisec.)
100

Applications in DFCI

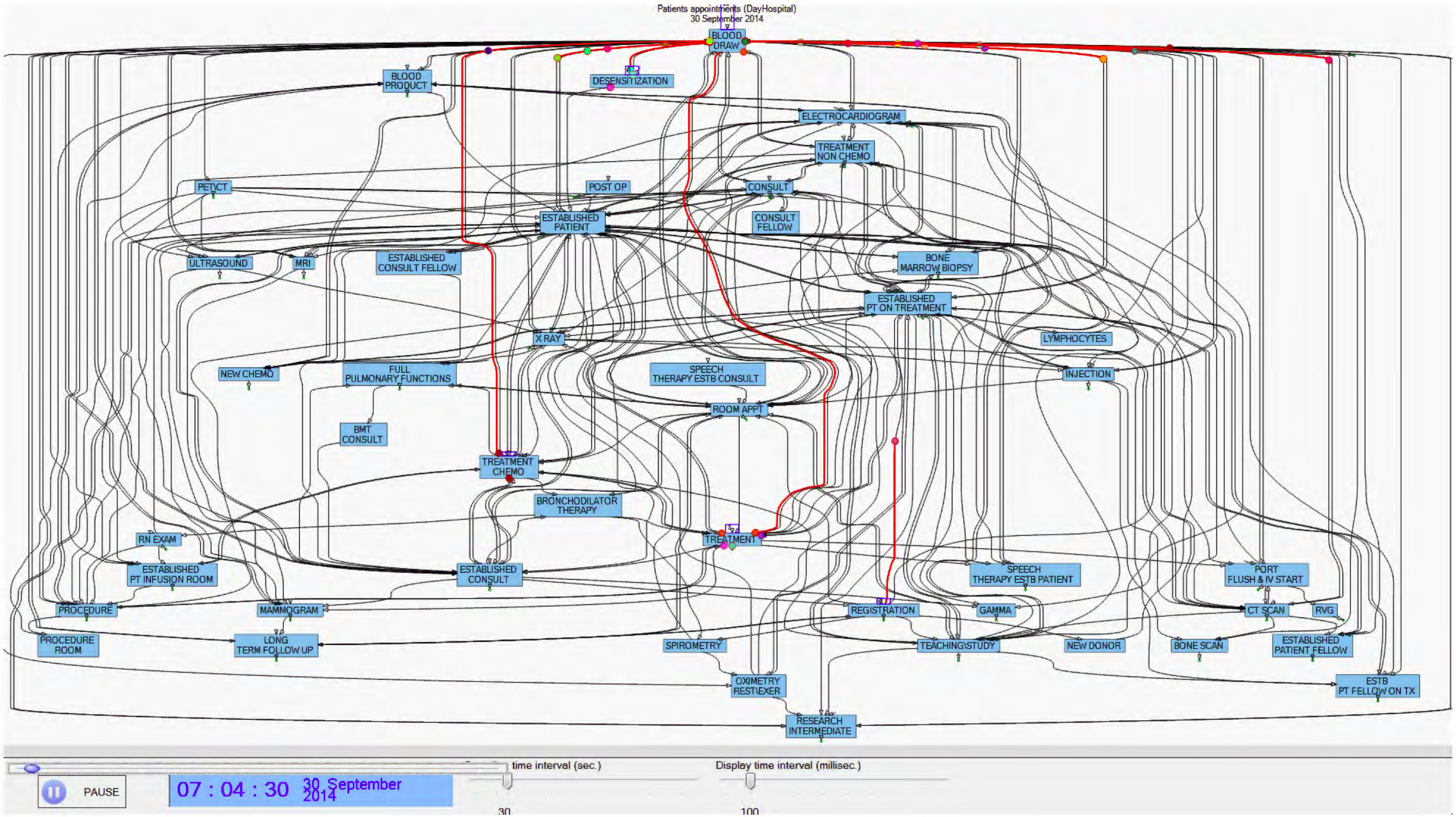
Control: rooms status, physicians location, long wait times

Planning: number infusion chairs, load-balancing among floors

Management: evidence-based

Design/Plan vs. Performance: exam durations (no 15-minutes)

Motivating improvement: room for physician vs. room for patient

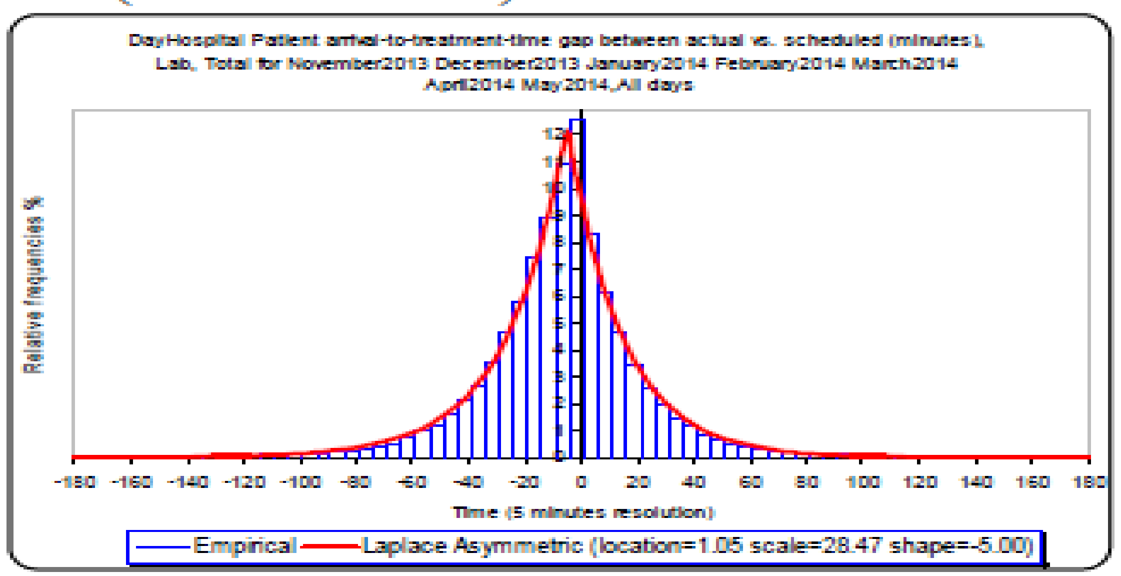


Patients appointments (DayHospital)
30 September 2014
BLOOD DRAW
BLOOD PRODUCT
DESENSITIZATION
ELECTROCARDIOGRAM
TREATMENT NON CHEMO
PET\CT
POST OP
CONSULT
ESTABLISHED PATIENT
CONSULT FELLOW
ULTRASOUND
MRI
ESTABLISHED CONSULT FELLOW
BONE MARROW BIOPSY
ESTABLISHED PT ON TREATMENT
X RAY
LYMPHOCYTES
NEW CHEMO
FULL PULMONARY FUNCTIONS
SPEECH THERAPY ESTB CONSULT
INJECTION
ROOM APPT
BMT CONSULT
TREATMENT CHEMO
BRONCHODILATOR THERAPY
RN EXAM
TREATMENT
ESTABLISHED PT INFUSION ROOM
ESTABLISHED CONSULT
SPEECH THERAPY ESTB PATIENT
PORT FLUSH & IV START
PROCEDURE
MAMMOGRAM
REGISTRATION
GAMMA
CT SCAN
RVG
PROCEDURE ROOM
LONG TERM FOLLOW UP
SPIROMETRY
TEACHING\STUDY
NEW DONOR
BONE SCAN
ESTABLISHED PATIENT FELLOW
OXIMETRY REST\EXER
ESTB PT FELLOW ON TX
RESEARCH INTERMEDIATE
time interval (sec.)
Display time interval (millisec.)
PAUSE
07 : 04 : 30
30 September 2014
30
100

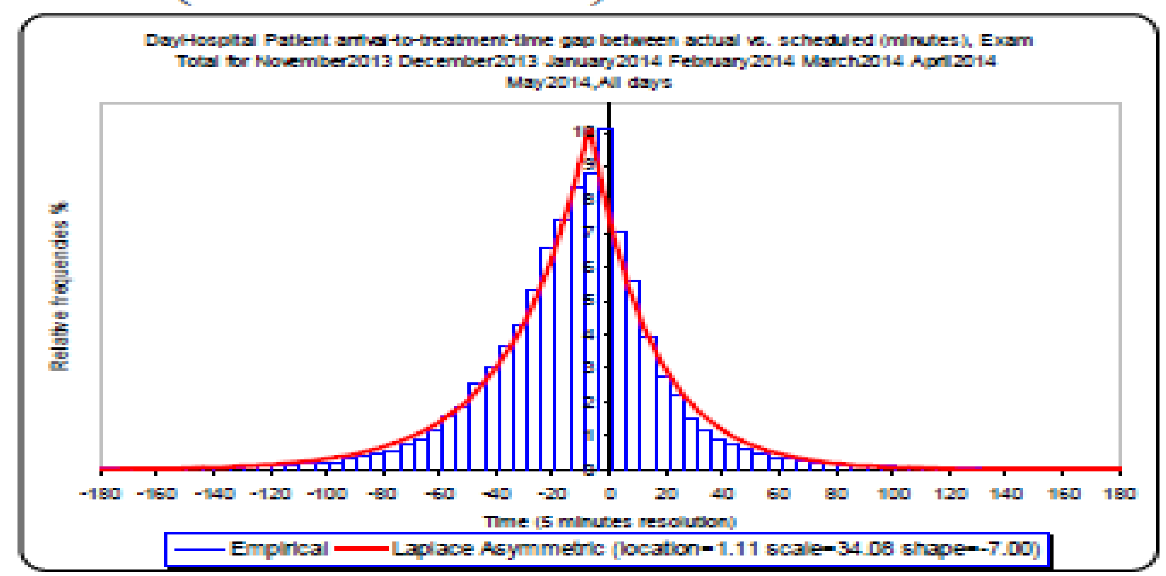
Planned vs. Actual: Improve **Plan**, Reduce **Gap**, then **Control**

w/ A. Senderovich & A. Gal

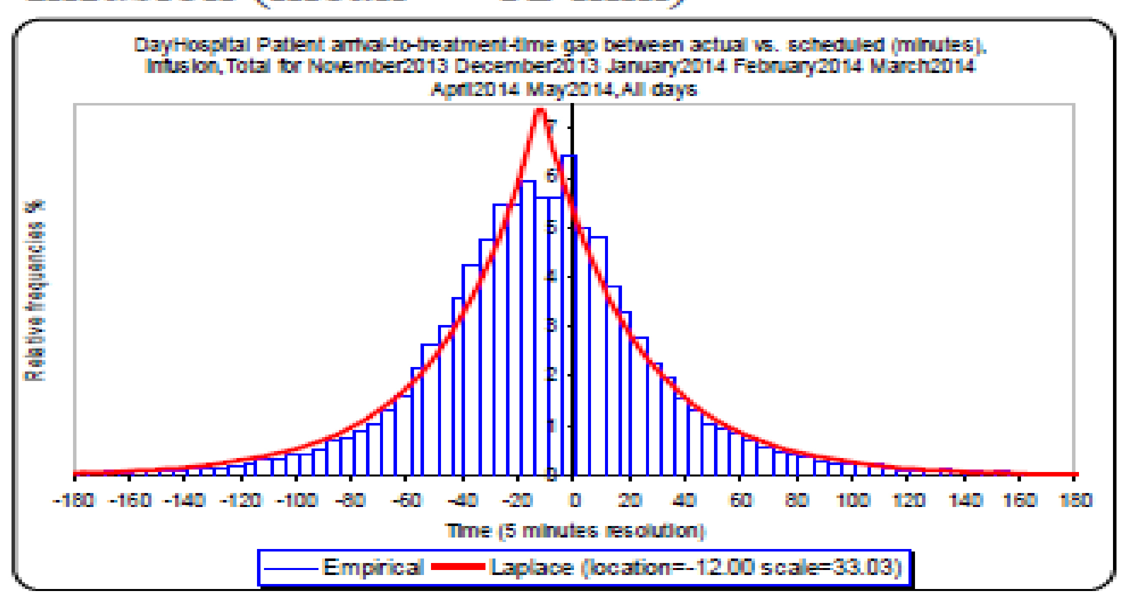
Lab (mean = -7 min)



Exam (mean = -12 min)



Infusion (mean = -12 min)



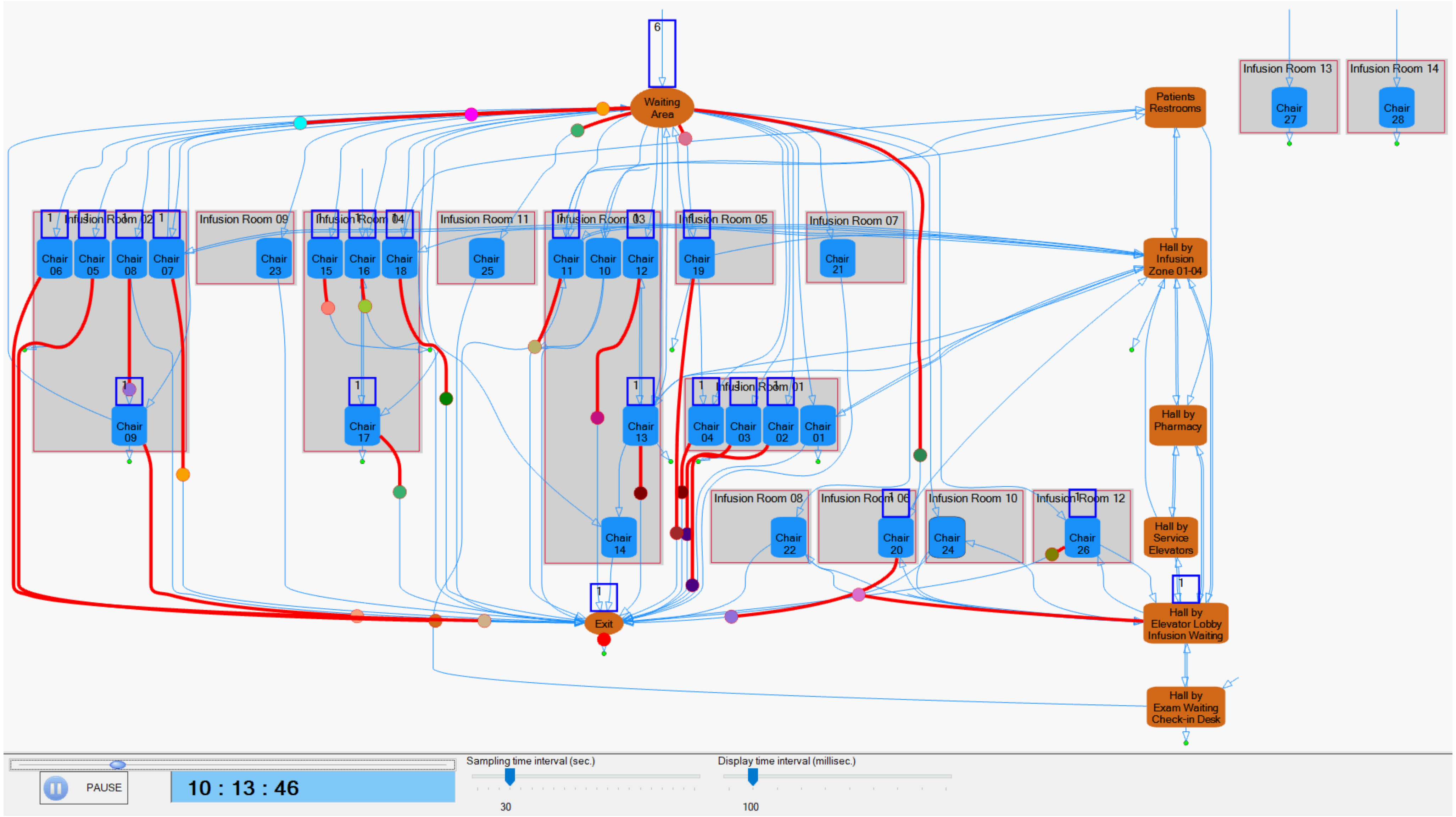
Appointments dominate healthcare systems.
Such systems are intrinsically stochastic, yet appointment systems (too) often view them deterministically:

DFCI Sample Infusion Schedule on Sep 3, 2014

time	patient_id	duration (min)	link_flag	floor_id	disease_center
15:00	01	60	unlinked	9	breast onc.
12:30	02	120	unlinked	9	breast onc.
10:30	03	180	linked	9	genitourinary onc.
12:30	04	60	linked	9	breast onc.
12:00	05	120	linked	9	genitourinary onc.
07:00	06	60	unlinked	9	breast onc.
⋮	⋮	⋮	⋮	⋮	⋮

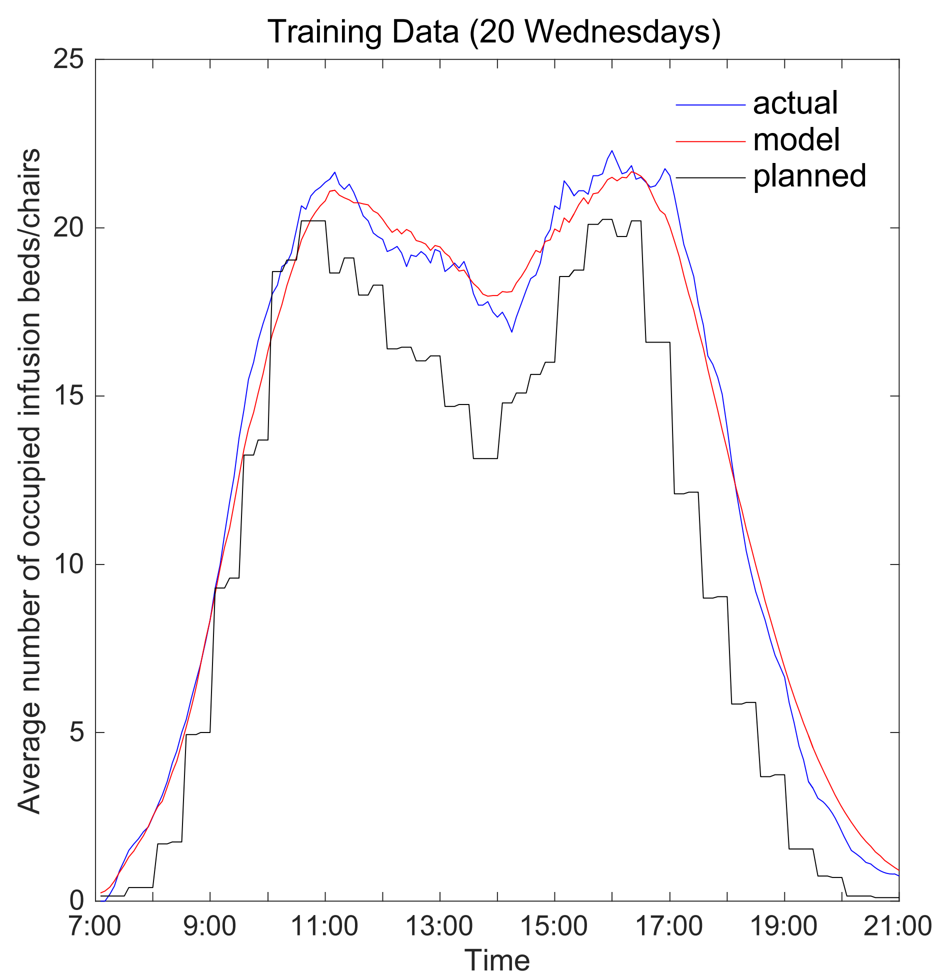
w/ P. Momcilovic & N. Trichakis: Develop methodologies for appointment scheduling, in **multi-server** environments, that take into account **stochastic** punctuality and service-durations

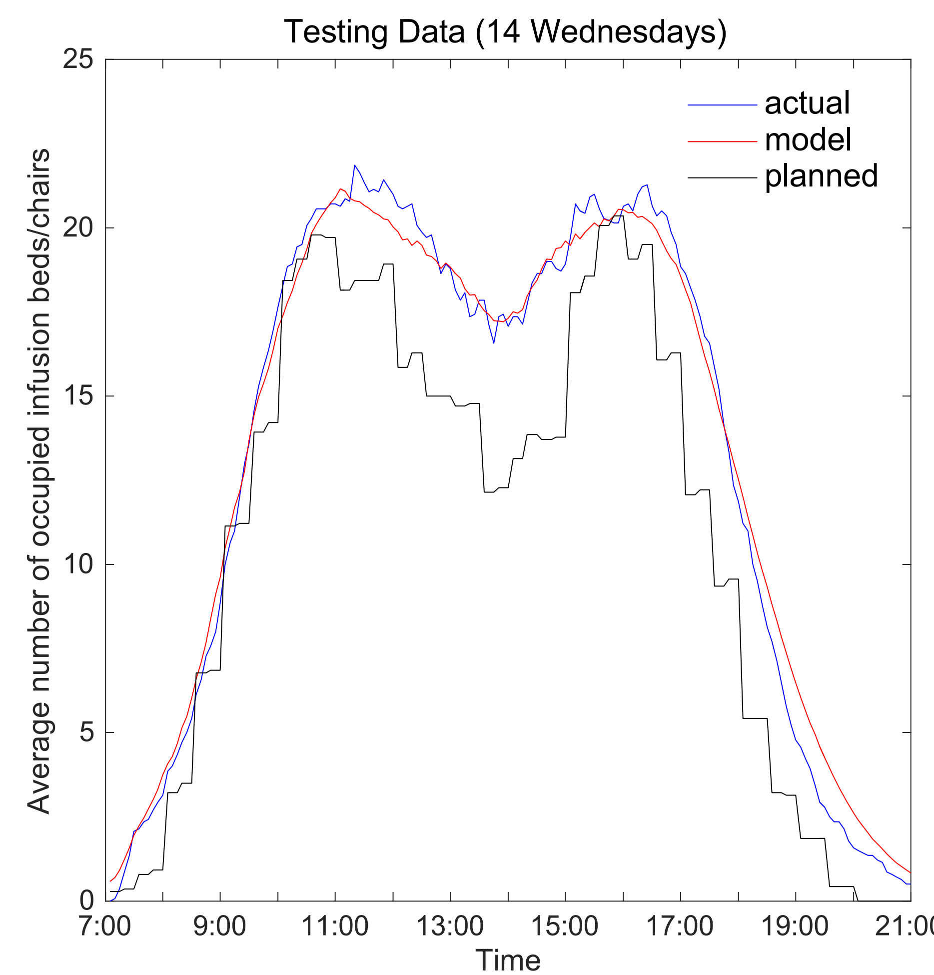
Patients Flow: Nth Floor, Infusion Unit   w/ P. Momcilovic & N. Trichakis



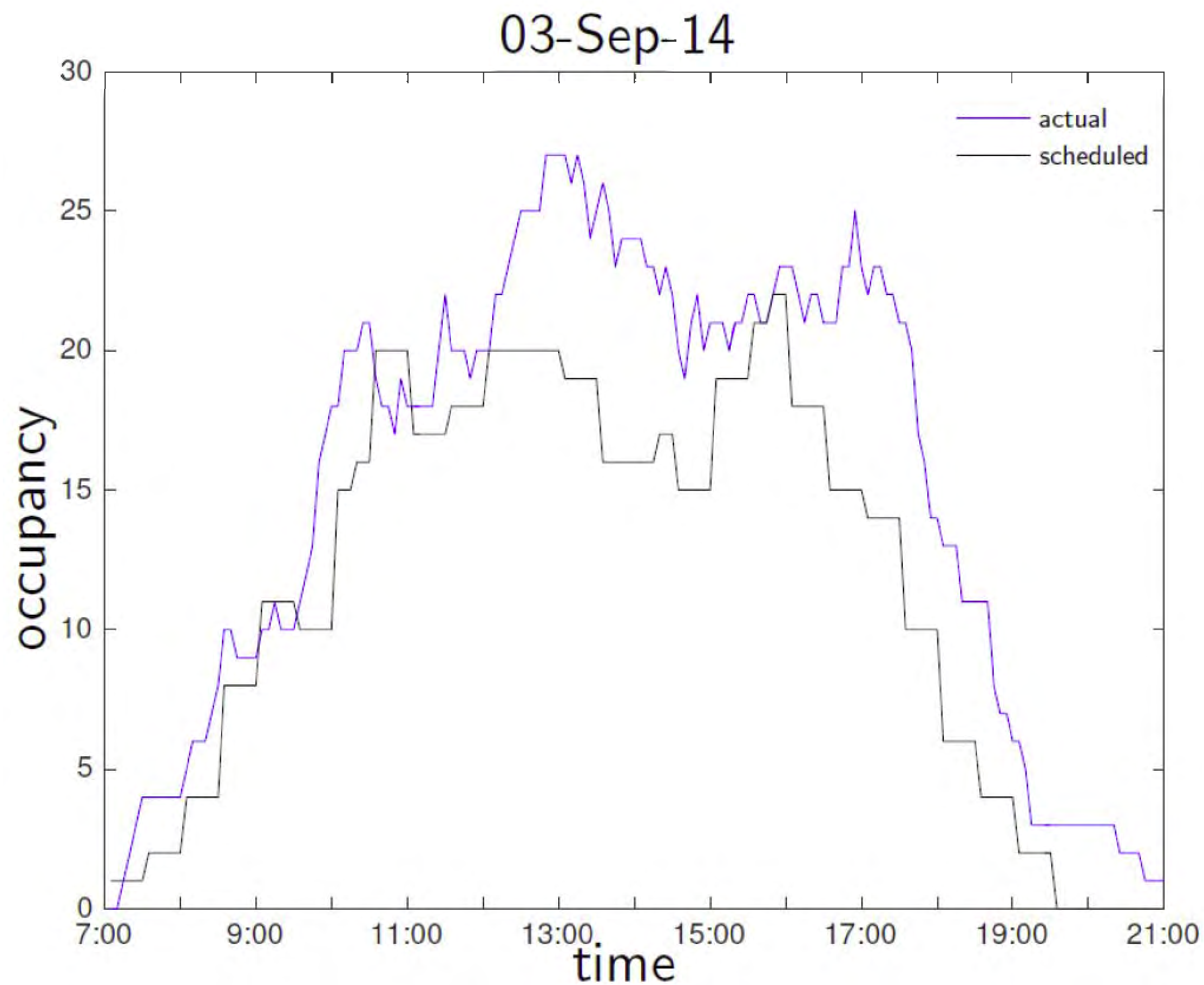
Appointment Scheduling: Approach & Validation

- Infinite-server model
- CLT approximation
- Benchmark based on Robust-Optimization framework

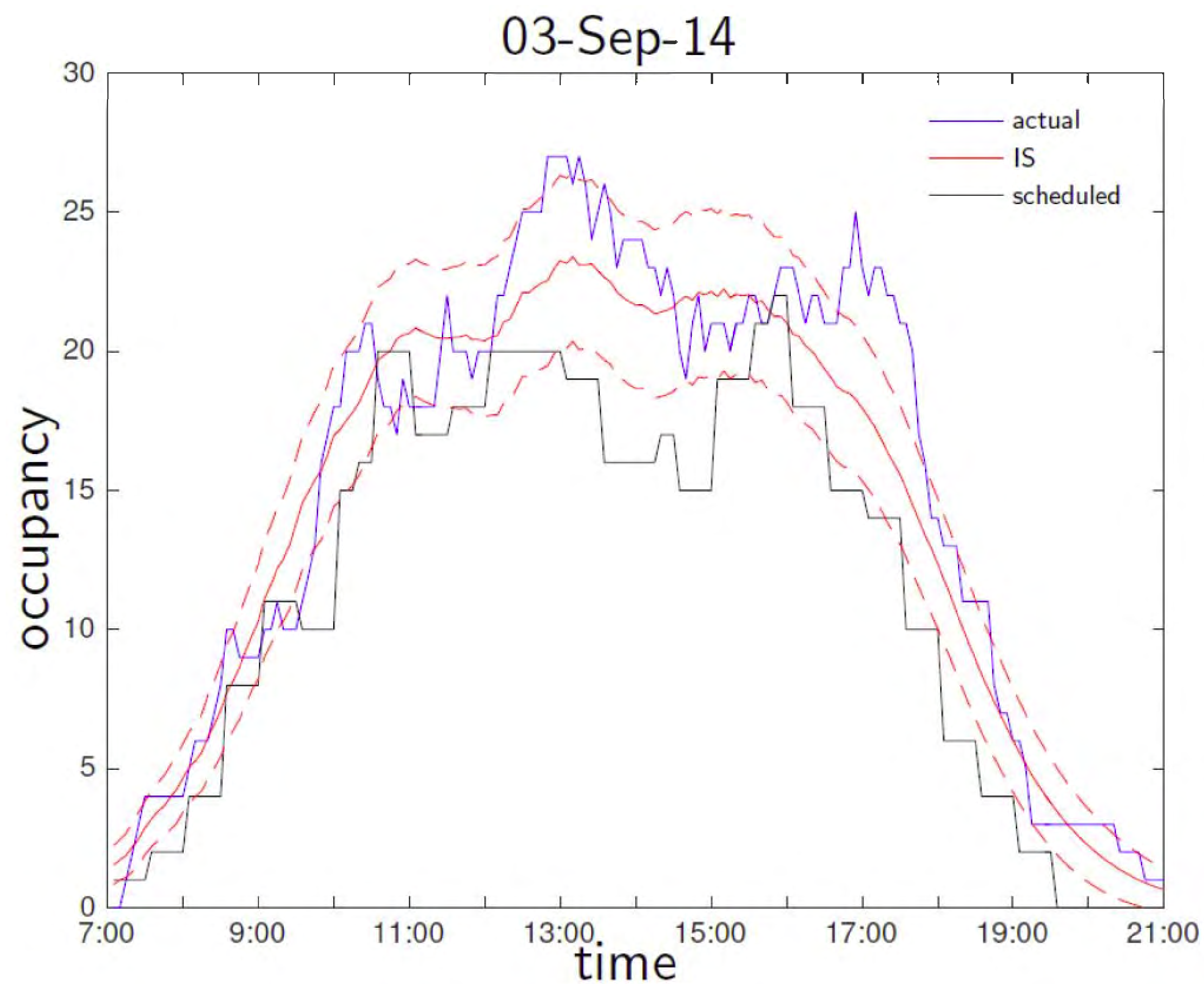




Scheduled *vs* Actual



Validation with DFCI Data



Comparison with DFCI Practice

- infusion service on floors 8, 9
- $c_t \approx 25$, $n \approx 90$
- compare IS with "means-based" scheduling
- consider 85% and 95% utilization levels

Reduce costs (waiting, overtime) by approximately 30%

Some Observations on “Academia-Industry Partnerships”
(not MSR or the Original-Bell-Labs)

- Intriguingly: RTLS successful at a **leading research** hospital?
- Intriguingly: initiative of **Physicians** at DFCI (as opposed to managers)?
- **Conjectures?**

Partnership with “Strong” partners, who appreciate Research/Evidence-Based-Management, and it is naturally based on Knowledge partnering with Data

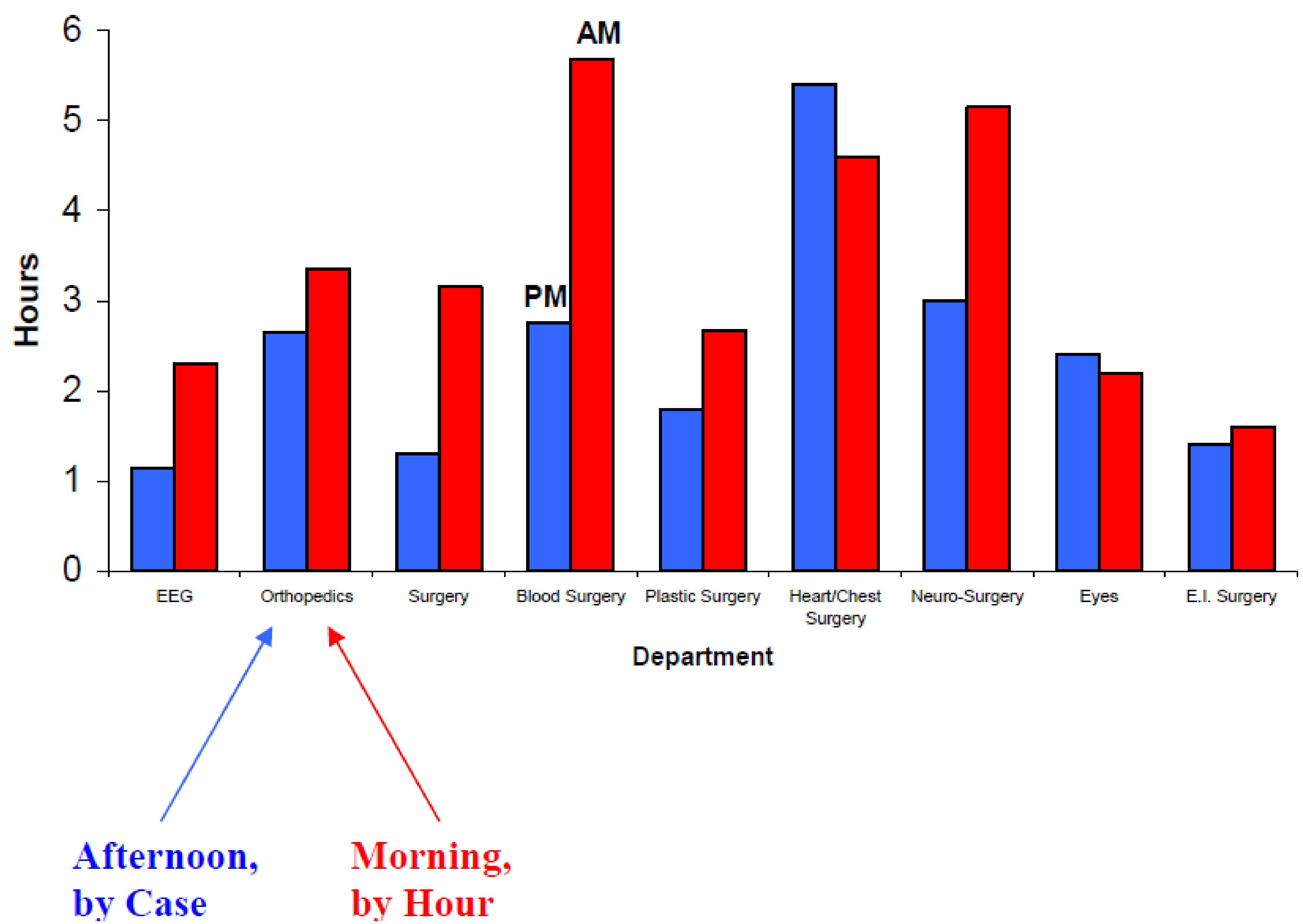
Data enables SYMMETRIC Partnerships, caters to goals of both partners (even IP)

Last Commercial: Key Missing Piece – Incentives

or: Even doctors (professors,…) can manage

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

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



Lessons from my Journeys

- Data is the language of
  - ✓ Multi-disciplinary research (e.g. OR + Psychology + NLP)
  - ✓ Academia-industry partnerships (e.g. university & bank)

- Parsimonious (robust) models at the service of complex realities

- OR needs Machine-Learning/Process-Mining/… to overcome its "curse of parsimony" (build high-stakes confidence - think "clinical-trials" via simulation);
  - ✓ The latter need OR to overcome their "curse of dimensionality" (overfitting); go beyond data-limits = predicting an existing system (think "ED counterfactuals")

- Academia & Industry must acknowledge (organize for, fund) data-professionalism

- Theompirical (data-based OR) research is feasible (1st data acquisition most challenging), important, "dangerous" before tenure, but mostly it is fun & **worth it**

The Technion SEE Center / Laboratory

Data-Based Service Science / Engineering

Technion
S.E.E
Center

Service Enterprise Engineering

Mixture Fitting

The mixture distribution function:

$$\boxed{p(x) = \sum_{j=1}^{k} w_j \varphi(x;\theta_j); \qquad \sum_{j=1}^{k} w_j = 1; \qquad w_j > 0}$$

θ_j– distribution parameters (shape, scale or location parameters)
φ – given distribution function Normal, Lognormal, Gamma, Exponential, Weibull or Invers Gaussian)
k – given number of components of the mixture
w_j– weighting coefficient or mixing proportions

Task: To find the optimal w_j and θ_j for sample X and given k and φ.

Solution:

- Specify initial values for w_j by using grid.
- For each point in a grid calculate mean, standard deviation (specify initial values for θ_j) and objective function (sum of squares of difference between empirical and fitted cumulative distribution functions).
- Select points with smallest values of objective function (best points).
- For each best point compute fitted distribution by using optimization algorithm.
- The best fit with smallest value of objective function is selected.

Distribution Fitting

Algorithms of distribution fitting (SEEStat software):

1. Maximum likelihood estimates
(11 distributions)

2. L-moments estimates (linear combination of order statistics)
(31 distributions)

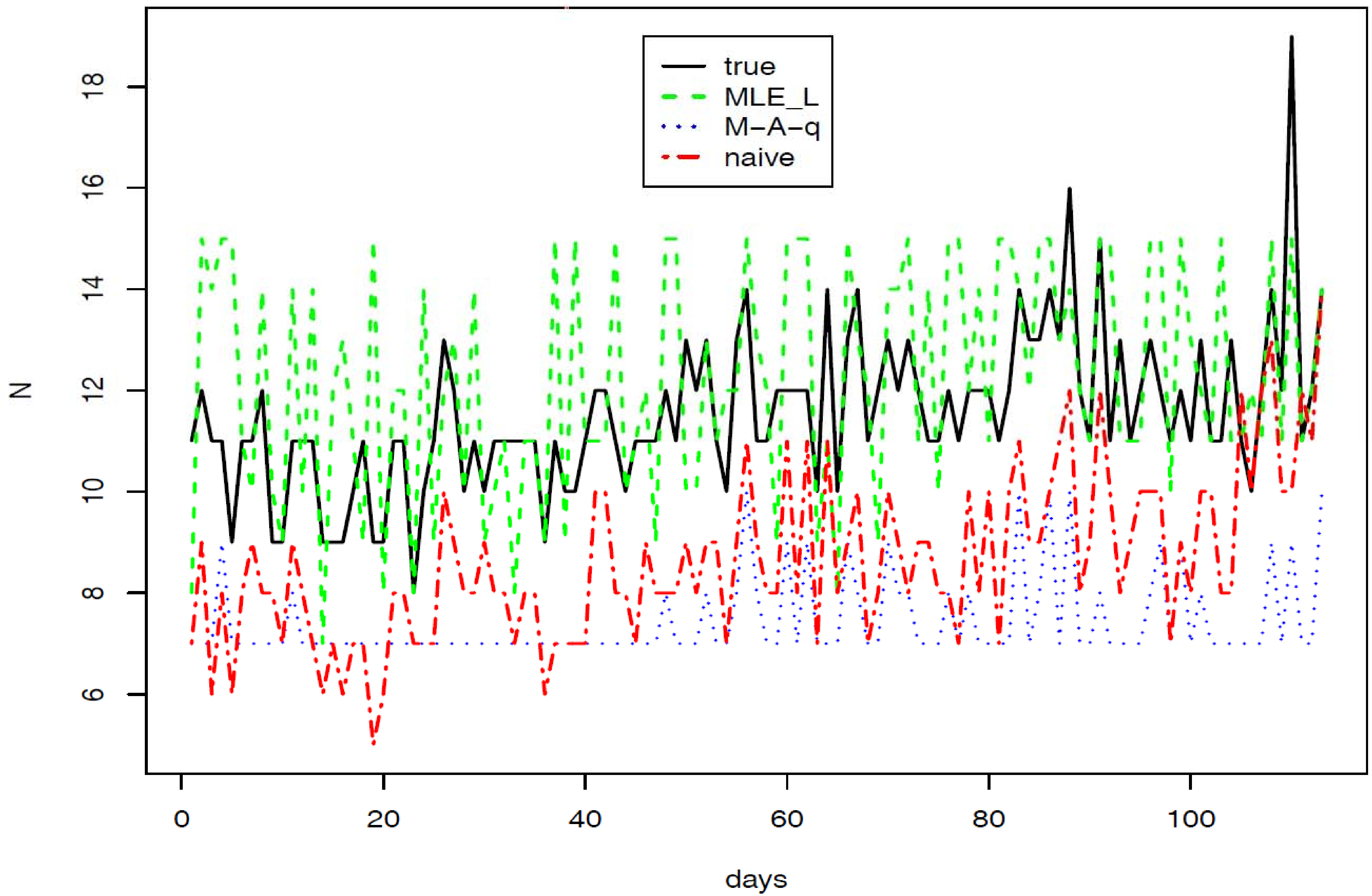
3. Maximization of likelihood function
(7 distributions)

4. Optimization of the object function (sum of squares of differences between empirical and fitted cumulative distribution functions)
(1 distribution)

Data-Based Creation of ServNets: some Technicalities

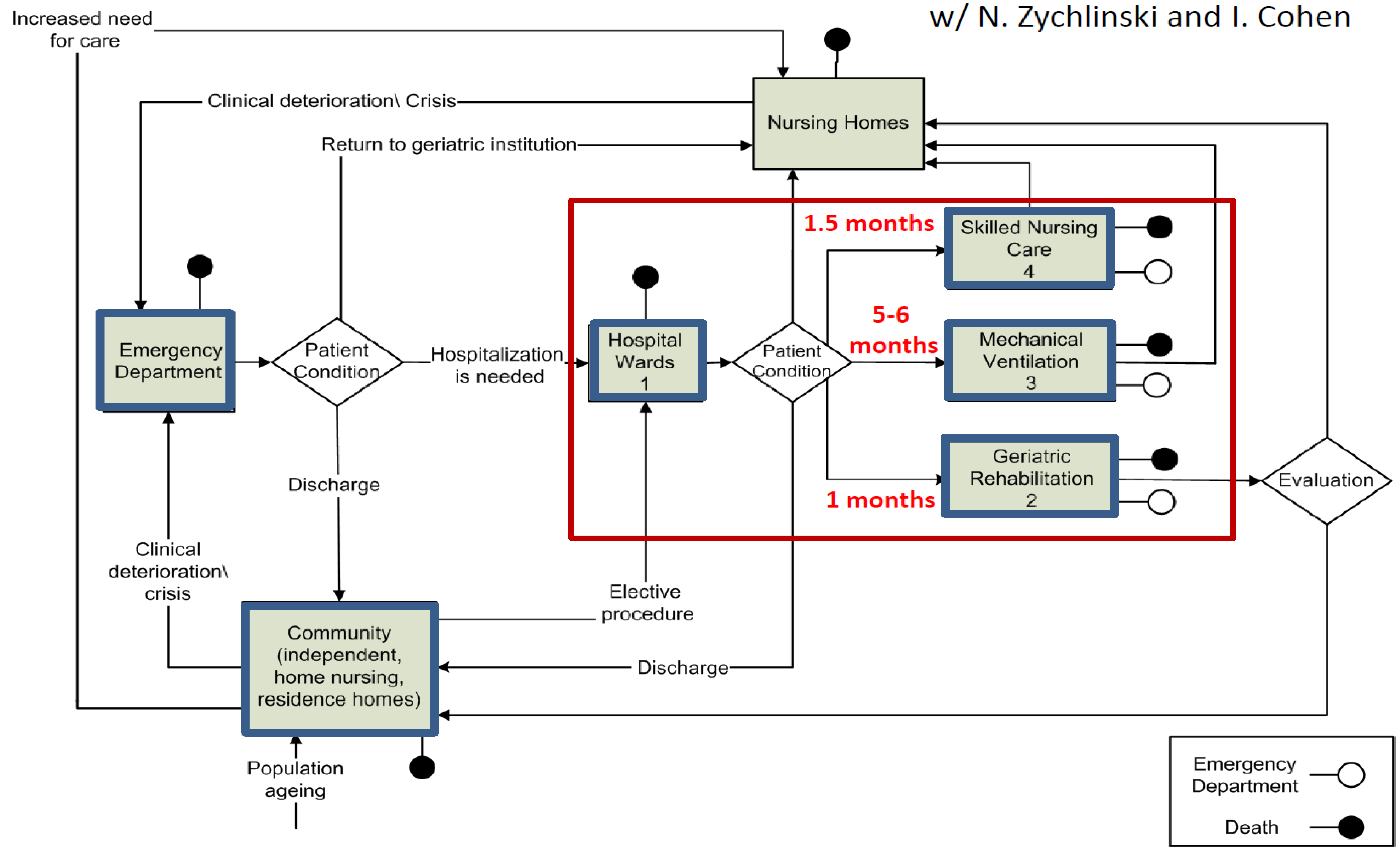
- **ServNets** = QNets, SimNets, FNets, DNets
- **Graph Layout**: Adapted from but significantly extends Graphviz (AT&T, 90's); eg. *edge-width*, which must be restricted to *poly-lines*, since there are "no parallel Bezier (Cubic) curves $(B_n(p) = E_p F[B(n,p)], 0 \leq p \leq 1)$
- **Algorithm**: Dot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

Estimating the Number of Present Servers
2/1/2005 – 27/6/2005



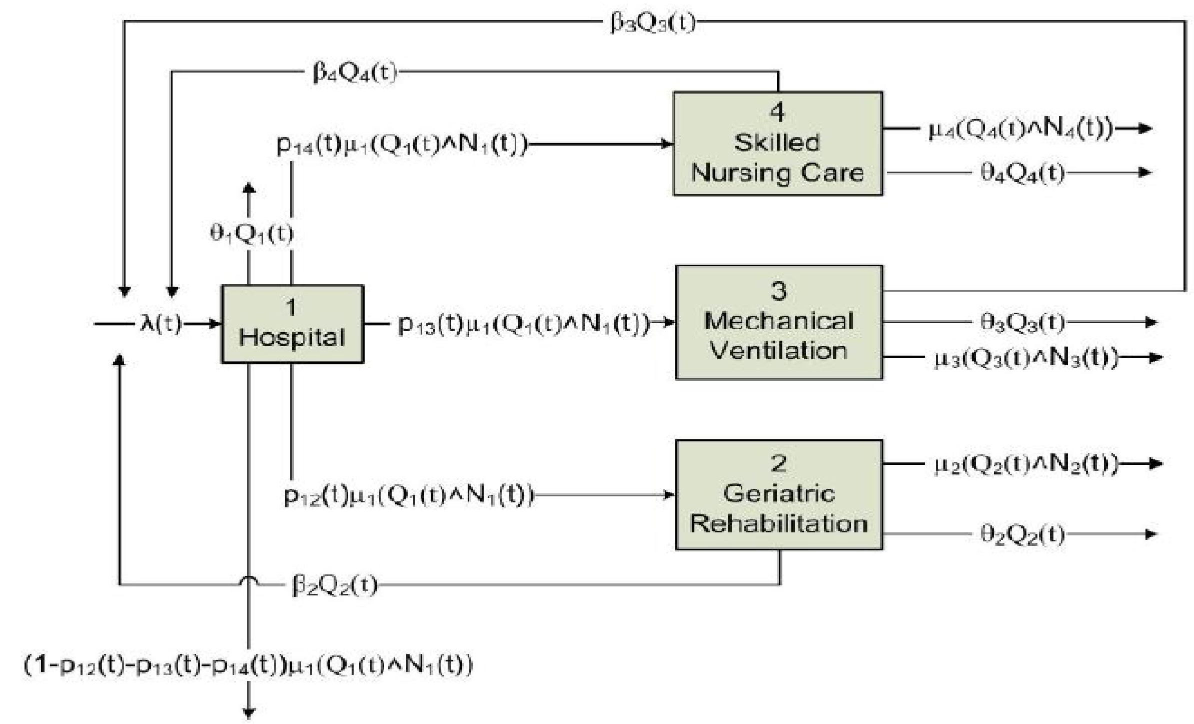
Patient Flow: ED to Wards to Nursing/Geriatric Institutions

w/ N. Zychlinski and I. Cohen



A Fluid Model

Stations 2,3 and 4



$$\dot{Q}_2(t) = (p_{12}(t)\mu_1 \cdot (Q_1(t) \wedge N_1)) \wedge (N_2 - Q_2(t))^+) -$$
$$\theta_2 Q_2(t) - \beta_2 Q_2(t) - \mu_2 \cdot (Q_2(t) \wedge N_2)$$
$$\dot{Q}_3(t) = (p_{13}(t)\mu_1 \cdot (Q_1(t) \wedge N_1)) \wedge (N_3 - Q_3(t))^+) -$$
$$\theta_3 Q_3(t) - \beta_3 Q_3(t) - \mu_3 \cdot (Q_3(t) \wedge N_3)$$
$$\dot{Q}_4(t) = (p_{14}(t)\mu_1 \cdot (Q_1(t) \wedge N_1)) \wedge (N_4 - Q_4(t))^+) -$$
$$\theta_4 Q_4(t) - \beta_4 Q_4(t) - \mu_4 \cdot (Q_4(t) \wedge N_4).$$

Technion

SEELab

Technion
S.E.E
Center
Service Enterprise Engineering

Data ⇔ Questions, Directions, Territories

- **Open Questions**
  - Why are service durations **Log-Normal**?
  - Staffing to **stabilize** time-varying performance: w/ Huang & Gurvich
  - **Fork-join** networks (Petri Nets): w/ Atar & Zviran; Zaeid & Kaspi
- **New Directions**
  - Multi-channel **design**:
    self-service (eg. IVR, w/ Carmeli & Kaspi), tele-medicine
  - **Protocol** Inference: w/ Liberman & Meilijson; w/ Senderovic, Weidlich, Gal
  - **Personalized** Queueing Theory: w/ Momcilovic (LPF)
  - Offered-Load/Capacity calculations:
    operational, cognitive, **emotional; financial**
  - All models “born" **equal**:
    eg. inference via diffusions (w/ Pang), simulating FNets
- **Uncharted Territories**
  - **Server Nets**: w/ Armony & Momcilovic; Azriel & Feigin; Senderovic & Gal
  - **RTLS**-based research (OR/Stat/IS, e.g. real-time control)
  - (Semi)Automatic **creation** of SNets (OR + Process Mining)
  - **Integrated** healthcare:
    from onset of symptoms through the hospital to recovery