Data-Based Science for Service Engineering and Management

or: Empirical Adventures in Call-Centers and Hospitals

Avi Mandelbaum

Technion, Haifa, Israel

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

▶ **Students:**

▶ **Theory:**
Armony, Atar, Gurvich, Jelenkovic, Kaspi, Massey, Momcilovic, Reiman, Shimkin, Stolyar, Wasserkrug, Whitt, Zeltyn, ... 

▶ **Industry:**
Mizrahi Bank (A. Cohen, U. Yonissi), Rambam Hospital (R. Beyar, S. Israeliit, S. Tzafrir), IBM Research (OCR Project), Hapoalim Bank (G. Maklef, T. Shlasky), Pelephone Cellular, ... 

▶ **Technion SEE Center / Laboratory:**
Feigin; Trofimov, Nadjarov, Gavako, Kutsy; Liberman, Koren, Plonsky, Senderovic; Research Assistants, ... 

▶ **Empirical/Statistical Analysis:**
Brown, Gans, Zhao; Shen; Ritov, Goldberg; Gurvich, Huang, Liberman; Armony, Marmor, Tseytlin, Yom-Tov; Zeltyn, Nardi, Gorfine, ...
History, Resources (Downloadable)


- **Teaching:** “Service-Engineering” Course (≥ 1995):
  - [http://ie.technion.ac.il/serveng](http://ie.technion.ac.il/serveng) - website
  - [http://ie.technion.ac.il/serveng/References/teaching_paper.pdf](http://ie.technion.ac.il/serveng/References/teaching_paper.pdf)

- **Call-Centers Research** (≥ 2000)
  - e.g. `<Call Centers>` in Google-Scholar

- **Healthcare Research** (≥ 2005)
  - e.g. **OCR Project**: IBM + Rambam Hospital + Technion

- **The Technion SEE Center** (≥ 2007)
The Case for Service Science / Engineering

- **Service Science / Engineering** (vs. Management) are emerging Academic Disciplines. For example, universities (world-wide), IBM (SSME, a là Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ...
The Case for Service Science / Engineering

- **Service Science / Engineering** (vs. Management) are emerging **Academic Disciplines**. For example, universities (world-wide), IBM (SSME, a là Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ... 

- Models that explain **fundamental phenomena**, which are **common** across applications:
  - Call Centers
  - Hospitals
  - Transportation
  - Justice, Fast Food, Police, Internet, ...

- **Simple models** at the Service of **Complex Realities** (Human)
  Note: Simple yet rooted in **deep analysis**.
The Case for Service Science / Engineering

- **Service Science / Engineering** (vs. Management) are emerging **Academic Disciplines**. For example, universities (world-wide), IBM (SSME, a là Computer-Science), USA NSF (SEE), Germany IAO (ServEng), ...

- Models that explain **fundamental phenomena**, which are **common** across applications:
  - Call Centers
  - Hospitals
  - Transportation
  - Justice, Fast Food, Police, Internet, ...

- **Simple models** at the Service of **Complex Realities** (Human)  
  Note: Simple yet rooted in **deep analysis**.

- Mostly **What Can Be Done** vs. **How To**
Title: Expands the Scientific Paradigm

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

Physics, Biology, ...: Measure, Model, Experiment, Validate, Refine. **Human-complexity** triggered above in Transportation, Economics. Starting with **Data**, expand to:

1. Measurements / Data
2. Modeling, Analysis
3. Validation
4. Maturity enables Deployment
5. Implementation
6. Improvement
7. Feedback
8. Novel needs, necessitating Science

**Management**

**Engineering**

**Science**

- e.g. Validate, refute or discover congestion laws (Little, PASTA, SSC, ?, ?, ...), in call centers and hospitals
Title: Expands the Scientific Paradigm

Physics, Biology, ... : Measure, Model, Experiment, Validate, Refine. **Human-complexity** triggered above in Transportation, Economics. Starting with **Data**, expand to:

1. Measurements / Data
2. Modeling, Analysis
3. Validation
4. Maturity enables Deployment
5. Implementation
6. Improvement
7. Feedback
8. Novel needs, necessitating Science

**Management** ➔ **Engineering** ➔ **Science**

- e.g. Validate, refute or discover **congestion laws** (Little, PASTA, SSC, ?, ?,...), in call centers and hospitals
Little’s Law: Call Center & Emergency Department

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

---

**USBank Customers in queue (average), Telesales 10.10.2001**

---

**HomeHospital Average patients in ED February 2004, Wednesdays**
Little’s Law: Call Center & Emergency Department

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

- USBank Customers in queue (average), Telesales 10.10.2001

- HomeHospital Average patients in ED February 2004, Wednesdays

⇒ Time-Varying Little’s Law
  - Berstemas & Mourtzinou;
  - Fralix, Riano, Serfozo; ...
Empirical Analysis of a QED Call Center

- 2205 half-hour intervals of an Israeli call center
- Almost all intervals are within $R - \beta$ and $R + 2\beta$ (i.e. $12\beta \leq \beta$), implying:
  - Very decent forecasts made by the call center
  - A very reasonable level of service (or does it?)

QED?

Recall: $GMR \times \theta \mu = $

- In this data-set, $0.35 \times \theta \mu = $
- Theoretical analysis suggests that under this condition, for $12 \beta \leq \beta$, we get $0 \leq \frac{P(Wait > 0)}{1} \leq \ldots$
QED Call Center: Performance

Large Israeli Bank

\[ P\{W_q > 0\} \text{ vs. } (R, N) \]

\[ \text{R-Slice: } P\{W_q > 0\} \text{ vs. } N \]

3 Operational Regimes:
- \textbf{QD:} \leq 25\%
- \textbf{QED:} 25\% – 75\%
- \textbf{ED:} \geq 75\%
### Operational Regimes: Scaling, Performance

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

<table>
<thead>
<tr>
<th>Erlang-A</th>
<th>Conventional scaling</th>
<th>MS scaling</th>
<th>NDS scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sub</td>
<td>Critical</td>
<td>Super</td>
</tr>
<tr>
<td><strong>Offered load per server</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arrival rate $\lambda$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Number of servers</strong></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Time-scale</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Abandonment rate</strong></td>
<td>$\frac{\theta}{n}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Staffing level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Utilization</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>E(Q)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P(\text{Ab})</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P(W_q &gt; 0)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>P(W_q &gt; T)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Congestion</strong></td>
<td>$\frac{E(W_q)}{E(S)}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $\delta > 0, \gamma \in (0, 1)$ and $\beta \in (-\infty, \infty)$;
- QD: $\phi = \frac{1 + \delta}{1 + \delta} < 1$;
- ED (ED+QED): $G(x) = \gamma$;
- QED: $\alpha_2 = \left[1 + \sqrt{\frac{\theta}{\mu} h(\beta)}\right]^{-1}$, here $\hat{\beta} = \beta \sqrt{\frac{\theta}{\mu}}$ and $h(x) = \frac{\phi(x)}{\Phi(x)}$
- ED+QED: $\alpha_3 = G(T) \Phi(\frac{\beta}{\sqrt{\frac{\theta}{\mu}}})$;
- Conventional: $P(W > T) = P\left(\frac{W}{\sqrt{n}} > \frac{T}{\sqrt{n}}\right)$, super: $P(W > T) = P\left(\frac{W}{n} > \frac{T}{n}\right)$; NDS: Super: $P(W > T) = P\left(\frac{W}{n} > \frac{T}{n}\right)$.
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.
Prerequisite I: Data

Averages Prevalent (and could be useful / interesting).

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

Sources: “Service-floor” (vs. Industry-level, Surveys, . . .)

- Administrative (Court, via “paper analysis”)
- Face-to-Face (Bank, via bar-code readers)
- Telephone (Call Centers, via ACD / CTI, IVR/VRU)
- Hospitals (Emergency Departments, . . .)
Prerequisite I: Data

Averages Prevalent (and could be useful / interesting).
But I need data at the level of the Individual Transaction:
For each service transaction (during a phone-service in a call center, or a patient’s visit in a hospital, or browsing in a website, or ...), its operational history = time-stamps of events.

Sources: “Service-floor” (vs. Industry-level, Surveys, ...)
- Administrative (Court, via “paper analysis”)
- Face-to-Face (Bank, via bar-code readers)
- Telephone (Call Centers, via ACD / CTI, IVR/VRU)
- Hospitals (Emergency Departments, ...)
- Expanding:
  - Hospitals, via RFID
  - Operational + Financial + Contents (Marketing, Clinical)
  - Internet, Chat (multi-media)
Pause for a Commercial:
Pause for a Commercial: The Technion SEE Center
SEELab: Data-repositories for research and teaching

▶ For example:

▶ Bank Anonymous: 1 years, 350K calls by 15 agents - in 2000. Brown, Gans, Sakov, Shen, Zeltyn, Zhao (JASA), paved the way for:

▶ U.S. Bank: 2.5 years, 220M calls, 40M by 1000 agents.
▶ Israeli Cellular: 2.5 years, 110M calls, 25M calls by 750 agents.
▶ Israeli Bank: from January 2010, daily-deposit at a SEESafe.
▶ Israeli Hospital: 4 years, 1000 beds; 8 ED’s- Sinreich’s data.
Technion SEE = Service Enterprise Engineering

SEELab: Data-repositories for research and teaching

- For example:
  - Bank Anonymous: 1 years, 350K calls by 15 agents - in 2000. Brown, Gans, Sakov, Shen, Zeltyn, Zhao (JASA), paved the way for:
    - U.S. Bank: 2.5 years, 220M calls, 40M by 1000 agents.
    - Israeli Cellular: 2.5 years, 110M calls, 25M calls by 750 agents.
    - Israeli Bank: from January 2010, daily-deposit at a SEESafe.
    - Israeli Hospital: 4 years, 1000 beds; 8 ED’s- Sinreich’s data.

SEEStat: Environment for graphical EDA in real-time

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 for:
    - U.S. Bank: 2.5 years, 220M calls, 40M by 1000 agents.
    - Israeli Cellular: 2.5 years, 110M calls, 25M calls by 750 agents.
    - Israeli Bank: from January 2010, daily-deposit at a SEESafe.
    - Israeli Hospital: 4 years, 1000 beds; 8 ED’s- Sinreich’s data.

SEEStat: Environment for graphical EDA in real-time


SEEServer: Free for academic use
Register, then access (presently) U.S. Bank and Bank Anonymous.

Visitor: run mstsc, seeserver.iem.technion.ac.il ; Self-Tutorial
Tutorial Cover; State-Space Collapse from Tutorial

4 overheads:

- Cover (make sure relevant to the lecture (e.g. APS, HKUST)
- Page 2 (again, make sure relevant to the lecture)
- Contents (with Stat-Space Collapse yellowed)
- The page with State-Space Collapse.
Focus on severely wounded casualties ($\approx 40$ in drill)

**Note:** 20 observers support real-time control (helps validation)
### Data Cleaning: MCE with RFID Support

<table>
<thead>
<tr>
<th>Asset id</th>
<th>order</th>
<th>Entry date</th>
<th>Exit date</th>
<th>Entry date</th>
<th>Exit date</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>1:14:07 PM</td>
<td>12:33:10 PM</td>
<td>12:02:00 PM</td>
<td>12:33:00 PM</td>
<td>exit is missing</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>12:02:02 PM</td>
<td>12:33:10 PM</td>
<td>12:02:00 PM</td>
<td>12:33:00 PM</td>
<td>entry is missing</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>11:37:15 AM</td>
<td>12:40:17 PM</td>
<td>11:37:00 AM</td>
<td>12:40:17 PM</td>
<td>entry is missing</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>12:23:32 PM</td>
<td>12:38:23 PM</td>
<td>12:23:00 PM</td>
<td>12:38:23 PM</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>12:12:47 PM</td>
<td>12:35:33 PM</td>
<td>12:12:00 PM</td>
<td>12:35:00 PM</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1:07:15 PM</td>
<td>1:07:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>11:18:19 AM</td>
<td>11:31:04 AM</td>
<td>11:18:00 AM</td>
<td>11:31:00 AM</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1:03:31 PM</td>
<td>1:07:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1:07:54 PM</td>
<td>12:01:58 PM</td>
<td>12:01:00 PM</td>
<td>12:01:58 PM</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>11:37:21 AM</td>
<td>12:57:02 PM</td>
<td>11:37:00 AM</td>
<td>12:57:00 PM</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>12:01:16 PM</td>
<td>12:37:16 PM</td>
<td>12:01:00 PM</td>
<td>12:37:16 PM</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>12:04:31 PM</td>
<td>12:20:40 PM</td>
<td></td>
<td></td>
<td>first customer is missing</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>12:27:37 PM</td>
<td>12:27:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>12:27:35 PM</td>
<td>1:07:28 PM</td>
<td>12:27:00 PM</td>
<td>1:07:00 PM</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>12:06:53 PM</td>
<td>12:06:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>11:21:34 AM</td>
<td>11:41:06 AM</td>
<td>11:21:00 AM</td>
<td>11:41:00 AM</td>
<td>exit time instead of entry time</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>12:21:06 PM</td>
<td>12:54:29 PM</td>
<td>12:21:00 PM</td>
<td>12:54:00 PM</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>11:40:54 AM</td>
<td>12:30:16 PM</td>
<td>11:40:00 AM</td>
<td>12:30:00 PM</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>12:37:57 PM</td>
<td>12:54:51 PM</td>
<td>12:37:00 PM</td>
<td>12:54:00 PM</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>11:27:11 AM</td>
<td>12:15:17 PM</td>
<td>11:27:00 AM</td>
<td>12:15:00 PM</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>12:05:50 PM</td>
<td>12:13:12 PM</td>
<td>12:05:00 PM</td>
<td>12:15:00 PM</td>
<td>wrong exit time</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td>11:31:48 AM</td>
<td>11:40:50 AM</td>
<td>11:31:00 AM</td>
<td>11:40:00 AM</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>12:06:23 PM</td>
<td>12:29:30 PM</td>
<td>12:06:00 PM</td>
<td>12:29:00 PM</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>1</td>
<td>11:31:50 AM</td>
<td>11:48:18 AM</td>
<td>11:31:00 AM</td>
<td>11:48:00 AM</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>2</td>
<td>12:59:21 PM</td>
<td>12:59:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Imagine “Cleaning" 60,000+ customers per day (call centers)!
Beyond Averages: The Human Factor

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

January-October

November-December

AVG: 185
STD: 238

AVG: 201
STD: 263

5.59%

6.83%

Log-Normal

▶ 6.8% Short-Services:
Beyond Averages: The Human Factor

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

January-October

AVG: 185
STD: 238

November-December

AVG: 201
STD: 263

6.8% Short-Services: Agents’ “Abandon” (improve bonus, rest), (mis)lead by incentives

Distributions must be measured (in seconds = natural scale)

LogNormal service times common in call centers
Validating LogNormality of Service-Duration

Israeli Call Center, Nov-Dec, 1999

Log(Service Times)  LogNormal QQPlot

[Histogram showing log(normal) distribution]

Practically Important: (mean, std)(log) characterization
Theoretically Intriguing: Why LogNormal? Naturally multiplicative but, in fact, also Infinitely-Divisible (Generalized Gamma-Convolutions)
Simple-model of a complex-reality? The Service Process:
(Telephone) Service-Process = “Phase-Type” Model

Retail Service (Israeli Bank)

Statistics OR IE
Individual Agents: Service-Duration, Variability

w/ Gans, Liu, Shen & Ye

Agent 14115

Service-Time Evolution: 6 month

Log(Service-Time)

Learning: Noticeable decreasing-trend in service-duration

LogNormal Service-Duration, individually and collectively
Individual Agents: Learning, Forgetting, Switching

Daily-Average Log(Service-Time), over 6 months
Agents 14115, 14128, 14136
Individual Agents: Learning, Forgetting, Switching

Daily-Average Log(Service-Time), over 6 months
Agents 14115, 14128, 14136

Weakly Learning-Curves for 12 Homogeneous(?!) Agents
Why Bother?

In large call centers:

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

⇒ Time and "Motion" Studies (Classical IE with New-age IT)
Why Bother?

In large call centers:

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

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

▶ **Service-Process Model**: Customer-Agent Interaction

▶ **Work Design** *(w/ Khudiakov)*
  eg. **Cross-Selling**: higher profit vs. longer (costlier) services;
  Analysis yields (congestion-dependent) cross-selling protocols

▶ **“Worker” Design** *(w/ Gans, Liu, Shen & Ye)*
  eg. **Learning, Forgetting, . . .**: Staffing & individual-performance
  prediction, in a heterogenous environment
Why Bother?

In large call centers:
+One Second to Service-Time implies +Millions in costs, annually

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

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

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

Israeli Bank: IVR/VRU Only, May 2008

IVR_only
May 2008, Week days

Relative frequencies %
mean=99
st.dev.=101

Mixture: 7 LogNormals

Fitting Mixtures of Distributions for VRU only time
May 2008, Week days

Empirical Total Lognormal Lognormal Lognormal
IVR-Process: “Phase-Type” Model
Started with Call Centers, Expanded to Hospitals

Call Centers - U.S. (Netherlands) Stat.

- $200 – $300 billion annual expenditures (0.5)
- 100,000 – 200,000 call centers (1500-2000)
- “Window" into the company, for better or worse
- Over 3 million agents = 2% – 4% workforce (100K)
Started with Call Centers, Expanded to Hospitals

Call Centers - U.S. (Netherlands) Stat.

- $200 – $300 billion annual expenditures (0.5)
- 100,000 – 200,000 call centers (1500-2000)
- “Window” into the company, for better or worse
- Over 3 million agents = 2% – 4% workforce (100K)

Healthcare - similar and unique challenges:

- Cost-figures far more staggering
- Risks much higher
- ED (initial focus) = hospital-window
- Over 3 million nurses
Call-Center Environment: Service Network
Call-Centers: “Sweat-Shops of the 21st Century”
Call-Center Network: Gallery of Models

Service Engineering: Multi-Disciplinary Process View

Call Center Design

Service Completion (75% in Banks)

Lost Calls

Lost Calls

Redial (Retrial)

Busy (Rare) Good or Bad

Arrivals (Business Frontier of the 21th Century)

Forecasting Statistics

Service Completion

VRU/IVR

Computer-Telephony Integration - CTI MIS/CS

Customers Interface Design

Human Factors Engineering

Internet Chat Email Fax

Customers Segmentation - CRM Marketing

To Avoid Starvation

To Avoid Delay

VIP Queue

Abandonment

Psychology, Statistics Lost Calls

Positive: Repeat Business

Negative: New Complaint

Information Design

Marketing, Operations Research

(→ Waiting Time ↔ Return Time)

Queue (Invisible)

Agents (Consultants)

Tele-Stress Psychology

(Turnover up to 200% per Year) (Sweat Shops of the 21th Century)

Job Enrichment Training, Incentives Human Resource Management

Skill Based Routing (SBR) Design

Marketing, Human Resources, Operations Research, MIS

Back-Office

Service Process Design

Logistics

Psychology, Operations Research, Marketing

New Services Design (R&D)

Operations, Marketing

Operations/Process Archive

Database Design

Data Mining: MIS, Statistics, Operations Research, Marketing

Psychological Process Archive

Expect 3 min Willing 8 min Perceive 15 min

Psychology, Sociology/Psychology, Operations Research

Service Engineering: Multi-Disciplinary Process View

Forecasting Statistics

Computer-Telephony Integration - CTI MIS/CS

Customers Interface Design

Human Factors Engineering

Internet Chat Email Fax

Customers Segmentation - CRM Marketing

To Avoid Starvation

To Avoid Delay

VIP Queue

Abandonment

Psychology, Statistics Lost Calls

Positive: Repeat Business

Negative: New Complaint

Information Design

Marketing, Operations Research

(→ Waiting Time ↔ Return Time)

Queue (Invisible)

Agents (Consultants)

Tele-Stress Psychology

(Turnover up to 200% per Year) (Sweat Shops of the 21th Century)

Job Enrichment Training, Incentives Human Resource Management

Skill Based Routing (SBR) Design

Marketing, Human Resources, Operations Research, MIS

Back-Office

Service Process Design

Logistics

Psychology, Operations Research, Marketing

New Services Design (R&D)

Operations, Marketing

Operations/Process Archive

Database Design

Data Mining: MIS, Statistics, Operations Research, Marketing

Psychological Process Archive

Expect 3 min Willing 8 min Perceive 15 min

Psychology, Sociology/Psychology, Operations Research

Service Engineering: Multi-Disciplinary Process View

Forecasting Statistics

Computer-Telephony Integration - CTI MIS/CS

Customers Interface Design

Human Factors Engineering

Internet Chat Email Fax

Customers Segmentation - CRM Marketing

To Avoid Starvation

To Avoid Delay

VIP Queue

Abandonment

Psychology, Statistics Lost Calls

Positive: Repeat Business

Negative: New Complaint

Information Design

Marketing, Operations Research

(→ Waiting Time ↔ Return Time)

Queue (Invisible)

Agents (Consultants)

Tele-Stress Psychology

(Turnover up to 200% per Year) (Sweat Shops of the 21th Century)

Job Enrichment Training, Incentives Human Resource Management

Skill Based Routing (SBR) Design

Marketing, Human Resources, Operations Research, MIS

Back-Office

Service Process Design

Logistics

Psychology, Operations Research, Marketing

New Services Design (R&D)

Operations, Marketing

Operations/Process Archive

Database Design

Data Mining: MIS, Statistics, Operations Research, Marketing

Psychological Process Archive

Expect 3 min Willing 8 min Perceive 15 min

Psychology, Sociology/Psychology, Operations Research
Call-Center Network: Gallery of Models

Add marks of topics to focus on
Skills-Based Routing in Call Centers
EDA and OR, with I. Gurvich and P. Liberman

Flow chart - March 2008

Mktg. ⇒

OR ⇒

HRM ⇒

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

Israeli Cellular, March 2008
SBR: Class-Dependent Services

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

PhD’s: Tezcan, Dai; Shaikhet, w/ Atar; Gurvich, Whitt
SBR: **Pool**-Dependent Services

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

PhD’s: Tezcan, Dai; Shaikhet, w/ Atar; Gurvich, Whitt
Waiting Times in a Call Center (Theory?)

Exponential in Heavy-Traffic (min.)
Small Israeli Bank

Routing via Thresholds (sec.)
Large U.S. Bank

Scheduling Priorities (sec) (later: Hospital LOS, hr.)
Medium Israeli Bank
ER / ED Environment: Service Network

Acute (Internal, Trauma)

Walking

Multi-Trauma
Queueing in a “Good” Beijing Hospital, at 6am
Emergency-Department Network: Gallery of Models

- **Service Completion** (sent to other department)
- **Blocked** (Ambulance Diversion)
- **Arrivals**
- **Emergency-Department Network: Gallery of Models**
- **Information Design**
  - MIS, HFE, Operations Research
  - (Waiting Time, Active Dashboard)
- **Internal Queue**
- **Acute, Walking**
- **Nurses**
  - Job Enrichment Training
  - HRM
- **Experts**
  - ED-Stress Psychology
  - (High turnovers Medical-Staff shortage)
- **Physicians**
  - Skill Based Routing (SBR) Design
  - Operations Research, HRM, MIS, Medicine
- **Interns**
  - Service Process Design
  - Operations Research, Medicine
- **Imaging Laboratory**
- **Home**
- **Hospital**

- **Forecasting**
  - Abandonment = LWBS, SBR \(\approx\) Flow Control
Emergency-Department Network: Gallery of Models

Add ED-to-IW routing
ED Design, with B. Golany, Y. Marmor, S. Israelit

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

e.g. Fast Track most suitable when elderly dominate

(a) Triage Model

(b) Fast-Track Model

(c) Illness-based Model

(d) Walking-Acute Model
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. In-Process / Release in EDs, w/ Carmeli, Huang, Shimkin

Workload and Offered-Load in Fork-Join Networks, w/ Kaspi, Zaeid

Synchronization Control of Fork-Join Networks, w/ Atar, Zviran

Staffing Time-Varying Q’s with Re-Entrant Customers, w/ Yom-Tov
Goal: Adhere to Triage-Constraints, then process/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, via threshold (w/ B. Carmeli, J. Huang, S. Israelit, N. Shimkin; as in Plambeck, Harrison, Kumar, who applied admission control).
Operational Fairness

1. “Punishing” fast wards in ED-to-IW Routing:
   - Parallel IWs: similar clinically, differ operationally
   - Problem: Short Length-of-Stay goes hand in hand with high bed-occupancy, bed-turnover, yet clinically apt: unfair!
   - Solution: Both nurses and managers content, w/ P. Momcilovic and Y. Tseytlin (3 time-scales: hour, day, week; “compare” with call-centers SBR)
Operational Fairness

1. “Punishing” fast wards in ED-to-IW Routing:
   - Parallel IWs: similar clinically, differ operationally
   - Problem: Short Length-of-Stay goes hand in hand with high bed-occupancy, bed-turnover, yet clinically apt: unfair!
   - Solution: Both nurses and managers content, w/ P. Momcilovic and Y. Tseytlin (3 time-scales: hour, day, week; “compare” with call-centers SBR)

2. Balancing Load across Maternity Wards:
   - 2 Maternity Wards: 1 = pre-birth, 2 = post-birth complications
   - Problem: Nurses think the “others-work-less”: unfair!
   - Goal: Balance workload, mostly via normal births
Operational Fairness

1. “Punishing” fast wards in ED-to-IW Routing:
   - Parallel IWs: similar clinically, differ operationally
   - Problem: Short Length-of-Stay goes hand in hand with high bed-occupancy, bed-turnover, yet clinically apt: unfair!
   - Solution: Both nurses and managers content, w/ P. Momcilovic and Y. Tseytlin (3 time-scales: hour, day, week; “compare” with call-centers SBR)

2. Balancing Load across Maternity Wards:
   - 2 Maternity Wards: 1 = pre-birth, 2 = post-birth complications
   - Problem: Nurses think the “others-work-less”: unfair!
   - Goal: Balance workload, mostly via normal births
   - Challenge: Workload is Operational, Cognitive, Emotional
     - Operational: Work content of a task, in time-units
     - Emotional: e.g. Mother and fetus-in-stress, suddenly fetus dies
Operational Fairness

1. “Punishing” fast wards in ED-to-IW Routing:
   - Parallel IWs: similar clinically, differ operationally
   - Problem: Short Length-of-Stay goes hand in hand with high bed-occupancy, bed-turnover, yet clinically apt: unfair!
   - Solution: Both nurses and managers content, w/ P. Momcilovic and Y. Tseytlin (3 time-scales: hour, day, week; “compare" with call-centers SBR)

2. Balancing Load across Maternity Wards:
   - 2 Maternity Wards: 1 = pre-birth, 2 = post-birth complications
   - Problem: Nurses think the “others-work-less": unfair!
   - Goal: Balance workload, mostly via normal births
   - Challenge: Workload is Operational, Cognitive, Emotional
     - Operational: Work content of a task, in time-units
     - Emotional: e.g. Mother and fetus-in-stress, suddenly fetus dies
   ⇒ Need help: A. Rafaeli & students (Psychology) - Ongoing
LogNormal & Beyond: Length-of-Stay in a Hospital

Israeli Hospital, in Days: LN

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

Why Bother?

▶ Hourly Scale: Staffing,

▶ Daily: Flow / Bed Control,

Workload at the Internal Ward (In Progress):
Arrivals, Departures, # Patients in Ward A, by Hour
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,
LogNormal & Beyond: Length-of-Stay in a Hospital

Israeli Hospital, in Days: LN

Israeli Hospital, in Hours: Mixture

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

Why Bother?
- Hourly Scale: Staffing,
- Daily: Flow / Bed Control,
Prerequisite II: Models (Fluid Q’s)

“Laws of Large Numbers” capture Predictable Variability

Deterministic Models: Scale Averages-out Stochastic Individualism

Cleaning Data – An Example:
RFID data in an MCE Drill

0
10
20
30
40
50
60

number of patients
number of patients (original)

Paths of doctors, nurses, patients (100+, 1 sec. resolution)

e.g. (could) Help predict “What if 150+ casualties severely wounded?”

▶ Transient Q’s:
Control of Mass Casualty Events (w/ I. Cohen, N. Zychlinski)

▶ Chemical MCE = Needy-Content Cycles (w/ G. Yom-Tov)
Prerequisite II: Models (Fluid Q’s)

“Laws of Large Numbers” capture Predictable Variability

Deterministic Models: Scale Averages-out Stochastic Individualism

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

Paths of doctors, nurses, patients (100+, 1 sec. resolution)

eg. (could) Help predict “What if 150+ casualties severely wounded?”

Transient Q’s:

- Control of Mass Casualty Events (w/ I. Cohen, N. Zychlinski)
- Chemical MCE = Needy-Content Cycles (w/ G. Yom-Tov)
The Basic Service-Network Model: Erlang-R

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

- $\lambda(t)$ – Time-Varying Arrival rate
- $S(\cdot)$ – Number of Servers (Nurses / Physicians).
- $\mu$ – Service rate ($E[\text{Service}] = \frac{1}{\mu}$)
- $p$ – ReEntrant (Feedback) fraction
- $\delta$ – Content-to-Needy rate ($E[\text{Content}] = \frac{1}{\delta}$)
Erlang-R: Fitting a Simple Model to a Complex Reality

Chemical MCE Drill (Israel, May 2010)

Arrivals & Departures (RFID)

Erlang-R (Fluid, Diffusion)

▶ Recurrent/Repeated services in MCE Events: eg. Injection every 15 minutes
Erlang-R: Fitting a Simple Model to a Complex Reality

Chemical MCE Drill (Israel, May 2010)

Arrivals & Departures (RFID)

Erlang-R (Fluid, Diffusion)

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

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

For example, $\text{M/M/1}$ (single-server queue): 91% server’s utilization goes with

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

and only 9% of the customers are served immediately upon arrival.
Traditional Queueing Theory predicts that Service-Quality and Servers’ Efficiency must be traded off against each other.

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

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

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

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

- **Call Centers**: Wait “seconds” for minutes service;
- **Transportation**: Search “minutes” for hours parking;
- **Hospitals**: Wait “hours” in ED for days hospitalization in IW’s;
Prerequisite II: Models (Diffusion/QED’s Q’s)

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

For example, $\text{M/M/1}$ (single-server queue): 91% server’s utilization goes with

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

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

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

- **Call Centers**: Wait “seconds” for minutes service;
- **Transportation**: Search “minutes” for hours parking;
- **Hospitals**: Wait “hours” in ED for days hospitalization in IW’s;

and, moreover, a significant fraction are not delayed in queue. (For example, in well-run call-centers, 50% served “immediately”, along with over 90% agents’ utilization, is not uncommon) ? QED
The Basic Staffing Model: Erlang-A (M/M/N + M)

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

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

- **Arrivals**: Poisson?
- **Service-durations**: Exponential?
- **(Im)Patience**: Exponential?
Testing the Erlang-A Primitives

- **Arrivals**: Poisson?
- **Service-durations**: Exponential?
- **(Im)Patience**: Exponential?
- Primitives independent (eg. Impatience and Service-Durations)?
- Customers / Servers Homogeneous?
- Service discipline FCFS?
- . . . ?

**Validation**: Support? Refute?
Arrivals to Service

Arrival-Rates to Three Call Centers

Dec. 1995 (U.S. 700 Helpdesks) vs May 1959 (England)

Arrival Process, in 1999

Random Arrivals “must be” (Axiomatically)

Time-Inhomogeneous Poisson
Arrivals to Service: only Poisson-Relatives

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


► Poisson CV (Dashed Line) = $1/\sqrt{\text{mean arrival-rate}}$

► Poisson CV’s ≪ Sampled CV’s (Solid) ⇒ Over-Dispersion
Arrivals to Service: only Poisson-Relatives

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


- Poisson CV (Dashed Line) = $1/\sqrt{\text{mean arrival-rate}}$
- Poisson CV’s ≪ Sampled CV’s (Solid) ⇒ Over-Dispersion
- Modeling (Poisson-Mixture) of and Staffing ($>\sqrt{\cdot}$) against Time-Varying Over-Dispersed Arrivals (w/ S. Maman & S. Zeltyn)
Service Durations: LogNormal Prevalent

**Israeli Bank Log-Histogram**

- **New Customers:** 2 min (NW);
- **Regulars:** 3 min (PS);

**Service-Classes Survival-Functions**

- **Stock:** 4.5 min (NE);
- **Tech-Support:** 6.5 min (IN).
Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram

Average = 2.24
St. dev. = 0.42

Service-Classes Survival-Functions

- **New** Customers: 2 min (NW);
- **Regulars**: 3 min (PS);

- **Stock**: 4.5 min (NE);
- **Tech-Support**: 6.5 min (IN).

▶ Service Durations are **LogNormal (LN)** and **Heterogeneous**
(Im)Patience while Waiting (Palm 1943-53)

Hazard Rate of (Im)Patience Distribution $\propto$ Irritation

Regular over VIP Customers – Israeli Bank

VIP Customers are more Patient (Needy)

Peaks of abandonment at times of Announcements

Challenges: Un-Censoring, Dependence (vs. KM), Smoothing - requires Call-by-Call Data

Graph showing hazard rates for Regular and Priority Customers over time.
(Im)Patience while Waiting (Palm 1943-53)

Hazard Rate of (Im)Patience Distribution $\propto$ Irritation

Regular over VIP Customers – Israeli Bank

- VIP Customers are more Patient (Needy)
- Peaks of abandonment at times of Announcements
- Challenges: Un-Censoring, Dependence (vs. KM), Smoothing
  - requires Call-by-Call Data
Dependent Primitives: Service- vs. Waiting-Time

Average Service-Time as a function of Waiting-Time
U.S. Bank, Retail, Weekdays, January-June, 2006

Introduction
- Relationship Between Service Time and Patience
- Workload and Offered-Load
- Empirical Results
- Future Research

Motivation

Focus on (Patience, Service-Time) jointly, w/ Reich and Ritov.

$E[S|\tau=W=w]$, $w \geq 0$: Service-Time of the Unserved.

53
Dependent Primitives: Service- vs. Waiting-Time

Average Service-Time as a function of Waiting-Time
U.S. Bank, Retail, Weekdays, January-June, 2006

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

Question: Is Erlang-A Relevant?  
**YES**! Fitting a Simple Model to a Complex Reality, both Theoretically and Practically.
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; with w/ M. Reich and Y. Ritov)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment; dependently; P. Khudiakov, M. Gorfine, P. Feigin)
- ···, and more.
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; with w/ M. Reich and Y. Ritov)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment; dependently; P. Khudiakov, M. Gorfine, P. Feigin)
- · · ·, and more.

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

**YES !**  **Fitting a Simple Model to a Complex Reality**, both **Theoretically** and **Practically**
Estimating (Im)Patience: via $P\{Ab\} \propto E[W_q]$ 

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

% Abandonment vs. Average Waiting-Time
Bank Anonymous (JASA): Yearly Data

Graphs based on 4158 hour intervals.
Estimate of mean (im)patience: 250/0.55 sec. $\approx 7.5$ minutes.
Erlang-A: Fitting a Simple Model to a Complex Reality

- Bank Anonymous: Small Israeli Call-Center
- (Im)Patience ($\theta$) estimated via $P\{Ab\} / E[W_q]$
- Graphs: Hourly Performance vs. Erlang-A Predictions, during 1 year (aggregating groups with 40 similar hours).

![Graphs showing comparison between Erlang-A predictions and real data for probability to abandon, waiting time, and probability of wait.]
Partial success – in some cases Erlang-A does not work well (Networking, SBR).

Ongoing Validation Project, w/ Y. Nardi, O. Plonsky, S. Zeltyn
Erlang-A: Simple, but Not Too Simple

Practical (Data-Based) questions, started in Brown et al. (JASA):
1. Fitting Erlang-A (Validation, w/ Nardi, Plonsky, Zeltyn).
2. Why does it practically work? justify robustness.
3. When does it fail? chart boundaries.
4. Generate needs for new theory.

Theoretical Framework:

Asymptotic Analysis, as load- and staffing-levels increase, which reveals model-essentials:

▶ Efficiency-Driven (ED) regime: Fluid models (deterministic)
▶ Quality- and Efficiency-Driven (QED): Diffusion refinements.

Motivation: Moderate-to-large service systems (100's - 1000's servers), notably Call-Centers.
Results turn out accurate enough to also cover <10 servers:

▶ Practically Important: Relevant to Healthcare (First: F. de Véricourt and O. Jennings; w/ G. Yom-Tov; Y. Marmor, S. Zeltyn; H. Kaspi, I. Zaeid)
Erlang-A: Simple, but Not Too Simple

Practical (Data-Based) questions, started in Brown et al. (JASA):
1. Fitting Erlang-A (Validation, w/ Nardi, Plonsky, Zeltyn).
2. Why does it practically work? justify robustness.
3. When does it fail? chart boundaries.
4. Generate needs for new theory.

Theoretical Framework: Asymptotic Analysis, as load- and staffing-levels increase, which reveals model-essentials:
  ▶ Efficiency-Driven (ED) regime: Fluid models (deterministic)
  ▶ Quality- and Efficiency-Driven (QED): Diffusion refinements.
Erlang-A: Simple, but Not Too Simple

Practical (Data-Based) questions, started in Brown et al. (JASA):
1. Fitting Erlang-A (Validation, w/ Nardi, Plonsky, Zeltyn).
2. Why does it practically work? justify robustness.
3. When does it fail? chart boundaries.
4. Generate needs for new theory.

Theoretical Framework: Asymptotic Analysis, as load- and staffing-levels increase, which reveals model-essentials:
► Efficiency-Driven (ED) regime: Fluid models (deterministic)
► Quality- and Efficiency-Driven (QED): Diffusion refinements.

Motivation: Moderate-to-large service systems (100’s - 1000’s servers), notably Call-Centers.

Results turn out accurate enough to also cover <10 servers:
► Practically Important: Relevant to Healthcare
  (First: F. de Véricourt and O. Jennings; w/ G. Yom-Tov; Y. Marmor, S. Zeltyn; H. Kaspi, I. Zaeid)
Operational Regimes: Conceptual Framework

**R:** Offered Load

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

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

$N = \#\text{Agents?}$ Intuition, as $R$ or $N$ increase unilaterally.
Operational Regimes: Conceptual Framework

**R**: Offered Load

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

eg. \( R = 25 \text{ calls/min.} \times 4 \text{ min./call} = 100 \)

\( N = \#\text{Agents} \) ? Intuition, as \( R \) or \( N \) increase unilaterally.

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

- Framework developed in O. Garnett’s MSc thesis
- Rigorously: \( (N - R)/R \rightarrow \delta \), as \( N, \lambda \uparrow \infty \), with \( \mu \) fixed.
- Performance: Delays are rare events
Operational Regimes: Conceptual Framework

**R**: Offered Load
Def. $R = \text{Arrival-rate} \times \text{Average-Service-Time} = \frac{\lambda}{\mu}$
eg  
eg eg. $R = 25 \text{ calls/min.} \times 4 \text{ min./call} = 100$

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

**QD Regime**: $N \approx R + \delta R$,  $0.1 < \delta < 0.25$  (eg. $N = 115$)
  - Framework developed in O. Garnett’s MSc thesis
  - Rigorously: $(N - R)/R \rightarrow \delta$, as $N, \lambda \uparrow \infty$, with $\mu$ fixed.
  - Performance: Delays are rare events

**ED Regime**: $N \approx R - \gamma R$,  $0.1 < \gamma < 0.25$  (eg. $N = 90$)
  - Essentially all customers are delayed
  - Wait same order as service-time; $\gamma\%$ Abandon (10-25%).
Operational Regimes: Conceptual Framework

**R**: Offered Load

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

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

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

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

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

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

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

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

- Erlang 1913-24, Halfin & Whitt 1981 (for Erlang-C)
- %Delayed between 25% and 75%
- $E[\text{Wait}] \propto \frac{1}{\sqrt{N}} \times E[\text{Service}]$ (sec vs. min); 1-5% Abandon.
## Operational Regimes: Rules-of-Thumb, w/ S. Zeltyn

<table>
<thead>
<tr>
<th>Constraint</th>
<th>P{Ab}</th>
<th>E[W]</th>
<th>P{W &gt; T}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Offered Load</strong></td>
<td>Tight 1-10%</td>
<td>Loose ≥ 10%</td>
<td>Tight ≤ 10%E[τ]</td>
</tr>
<tr>
<td><strong>Small (10’s)</strong></td>
<td>QED</td>
<td>QED</td>
<td>QED</td>
</tr>
<tr>
<td><strong>Moderate-to-Large (100’s-1000’s)</strong></td>
<td>QED</td>
<td>ED, QED</td>
<td>QED if τ (d) exp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Operational Regimes: Rules-of-Thumb, w/ S. Zeltyn

<table>
<thead>
<tr>
<th>Constraint</th>
<th>P{Ab}</th>
<th>E[W]</th>
<th>P{W &gt; T}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offered Load</td>
<td>Tight</td>
<td>Loose</td>
<td>Tight</td>
</tr>
<tr>
<td></td>
<td>1-10%</td>
<td>≥ 10%</td>
<td>≤ 10%E[τ]</td>
</tr>
<tr>
<td>Must</td>
<td>Tight</td>
<td>Loose</td>
<td>Tight</td>
</tr>
<tr>
<td></td>
<td>0 ≤ T ≤ 10%E[τ]</td>
<td>5% ≤ α ≤ 50%</td>
<td>0 ≤ T ≤ 10%E[τ]</td>
</tr>
<tr>
<td>Small (10’s)</td>
<td>QED</td>
<td>QED</td>
<td>QED</td>
</tr>
<tr>
<td>Moderate-to-Large</td>
<td>QED</td>
<td>ED, QED</td>
<td>QED</td>
</tr>
<tr>
<td>(100’s-1000’s)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ED**: \(N \approx R - \gamma R\) \((0.1 \leq \gamma \leq 0.25)\).

**QD**: \(N \approx R + \delta R\) \((0.1 \leq \delta \leq 0.25)\).

**QED**: \(N \approx R + \beta \sqrt{R}\) \((-1 \leq \beta \leq 1)\).

**ED+QED**: \(N \approx (1 - \gamma)R + \beta \sqrt{R}\) \((\gamma, \beta \text{ as above})\).
**Operational Regimes: Rules-of-Thumb, w/ S. Zeltyn**

<table>
<thead>
<tr>
<th>Constraint</th>
<th>P{Ab}</th>
<th>E[W]</th>
<th>P{W &gt; T}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offered Load</td>
<td>Tight 1-10%</td>
<td>Loose ≥ 10%</td>
<td>Tight ≤ 10%E[τ]</td>
</tr>
<tr>
<td>Small (10’s)</td>
<td>QED</td>
<td>QED</td>
<td>QED</td>
</tr>
<tr>
<td>Moderate-to-Large</td>
<td>QED</td>
<td>ED, QED</td>
<td>QED</td>
</tr>
<tr>
<td>(100’s-1000’s)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ED:** \( N \approx R - \gamma R \) \quad (0.1 \leq \gamma \leq 0.25).

**QD:** \( N \approx R + \delta R \) \quad (0.1 \leq \delta \leq 0.25).

**QED:** \( N \approx R + \beta \sqrt{R} \) \quad (-1 \leq \beta \leq 1).

**ED+QED:** \( N \approx (1 - \gamma)R + \beta \sqrt{R} \) \quad (\gamma, \beta \text{ as above}).

**WFM:** How to determine specific staffing level \( N \)? e.g. \( \beta \).
## Operational Regimes: Scaling, Performance, w/ I. Gurvich & J. Huang

<table>
<thead>
<tr>
<th>Erlang-A</th>
<th>Conventional scaling</th>
<th>MS scaling</th>
<th>NDS scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>μ fixed</strong></td>
<td>Sub</td>
<td>Critical</td>
<td>Super</td>
</tr>
<tr>
<td><strong>Offered load per server</strong></td>
<td>( \frac{1}{1 + \delta} &lt; 1 )</td>
<td>( 1 - \frac{\beta}{\sqrt{n}} \approx 1 )</td>
<td>( \frac{1}{1 + \delta} &gt; 1 )</td>
</tr>
<tr>
<td><strong>Arrival rate λ</strong></td>
<td>( \frac{\mu}{1 + \delta} )</td>
<td>( \mu - \frac{\beta}{\sqrt{n}} )</td>
<td>( \frac{\mu}{1 + \gamma} )</td>
</tr>
<tr>
<td><strong>Number of servers</strong></td>
<td>1</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td><strong>Time-scale</strong></td>
<td>( \theta/n )</td>
<td>( \theta )</td>
<td>( \theta/n )</td>
</tr>
<tr>
<td><strong>Abandonment rate</strong></td>
<td>( \theta/n )</td>
<td>( \theta )</td>
<td>( \theta/n )</td>
</tr>
<tr>
<td><strong>Staffing level</strong></td>
<td>( \frac{\lambda}{\mu} (1 + \delta) )</td>
<td>( \frac{\lambda}{\mu} (1 + \frac{\beta}{\sqrt{n}}) )</td>
<td>( \frac{\lambda}{\mu} (1 - \gamma) )</td>
</tr>
<tr>
<td><strong>Utilization</strong></td>
<td>( \frac{1}{1 + \delta} )</td>
<td>( 1 - \sqrt{\frac{\theta}{\mu}} [h(\hat{\beta}) - \hat{\beta}] )</td>
<td>( 1 )</td>
</tr>
<tr>
<td><strong>E(Q)</strong></td>
<td>( \frac{n}{\mu} \sqrt{\frac{\theta}{\mu}} [h(\hat{\beta}) - \hat{\beta}] )</td>
<td>( \frac{n\mu}{\theta(1-\gamma)} )</td>
<td>( \frac{1}{\sqrt{2\pi}} \vartheta^{1/2} \sqrt{n} \frac{1}{\sqrt{\theta}} )</td>
</tr>
<tr>
<td><strong>P(( A_b ))</strong></td>
<td>( \frac{1 + \delta}{n} \frac{\theta}{\mu} \alpha_1 )</td>
<td>( \frac{1}{\sqrt{\frac{\theta}{\mu}}} [h(\hat{\beta}) - \hat{\beta}] )</td>
<td>( \gamma )</td>
</tr>
<tr>
<td><strong>P(( W_q &gt; 0 ))</strong></td>
<td>( \alpha_1 \in (0, 1) )</td>
<td>( \approx 1 )</td>
<td>( \frac{1}{\sqrt{\frac{\theta}{\mu}}} [h(\hat{\beta}) - \hat{\beta}] \approx 0 )</td>
</tr>
<tr>
<td><strong>P(( W_q &gt; T ))</strong></td>
<td>( \alpha_1 e^{-\frac{\lambda}{\mu} \mu_0} )</td>
<td>( 1 + O(\frac{1}{\sqrt{n}}) )</td>
<td>( 1 + O(\frac{1}{\sqrt{n}}) )</td>
</tr>
<tr>
<td><strong>Congestion</strong></td>
<td>( \frac{\alpha_1}{\frac{\theta}{\mu}} )</td>
<td>( \frac{1 + \delta}{\sqrt{\frac{\theta}{\mu}}} [h(\hat{\beta}) - \hat{\beta}] )</td>
<td>( n \mu \gamma / \theta )</td>
</tr>
</tbody>
</table>
Empirical Analysis of a QED Call Center

• 2205 half-hour intervals of an Israeli call center
• Almost all intervals are within \( R - \beta \leq \leq R + 2\beta \) (i.e. \( 12\beta - \leq \leq \)), implying:
  - Very decent forecasts made by the call center
  - A very reasonable level of service (or does it?)

QED? (with Nardi, Plonski, Zeltyn)

Recall:
\[
\text{GMR} x \theta \mu = \]

In this data-set, \( 0.35 x \theta \mu = = \)

Theoretical analysis suggests that under this condition, for \( 12\beta - \leq \leq \), we get \( 0(0)1P \text{Wait} \leq \leq \) …
QED Call Center: Performance

Large Israeli Bank

\[ P\{ W_q > 0 \} \text{ vs. } (R, N) \]

R-Slice: \[ P\{ W_q > 0 \} \text{ vs. } N \]

3 Operational Regimes:

- **QD**: \( \leq 25\% \)
- **QED**: \( 25\% - 75\% \)
- **ED**: \( \geq 75\% \)
QED Theory (Erlang ’13; Halfin-Whitt ’81; Garnett MSc; Zeltyn PhD)

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

- **QED**  
  \( \{\text{Wait} > 0\} \approx \alpha \), \( 0 < \alpha < 1 \);

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

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

- **Managers**  
  \( N \approx R + \beta \sqrt{R} \), \( R = \lambda \times \text{E(S)} \) not small;
**QED Theory** (Erlang ’13; Halfin-Whitt ’81; Garnett MSc; Zeltyn PhD)

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

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

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

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

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

**QED performance**: Laplace Method (asymptotics of integrals).

**Parameters**: Arrivals and Staffing - $\beta$, Services - $\mu$,  
(Im)Patience - $g(0) = \text{patience density at the origin}$. 
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:**
  \[
  P\{W_q > 0\} \approx \left[1 + \sqrt{\frac{\theta}{\mu} \cdot \frac{h(\hat{\beta})}{h(-\beta)}}\right]^{-1}.
  \]

- **Probability to Abandon:**
  \[
  P\{\text{Ab}|W_q > 0\} \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\theta}{\mu} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right]} .
  \]

- \( P\{\text{Ab}\} \propto \mathbb{E}[W_q] \), both order \( \frac{1}{\sqrt{n}} \):
  \[
  \frac{P\{\text{Ab}\}}{\mathbb{E}[W_q]} = \theta.
  \]
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 $n \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$. 

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

3. QED

Excursions
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 $n \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$.
   - Fix $n$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$. 
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 $n \uparrow \infty$: $P\{ W_q > 0 \} \downarrow 0$.
   - Fix $n$ and let $\lambda \uparrow \infty$: $P\{ W_q > 0 \} \uparrow 1$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $n$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $n \approx R + \beta \sqrt{R}$.
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 \( n \uparrow \infty \): \( P\{ W_q > 0 \} \downarrow 0 \).
   - Fix \( n \) and let \( \lambda \uparrow \infty \): \( P\{ W_q > 0 \} \uparrow 1 \).
   - \( \Rightarrow \) Must have both \( \lambda \) and \( n \) increase simultaneously:
   - \( \Rightarrow \) (CLT) Square-root staffing: \( n \approx R + \beta \sqrt{R} \).

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 $n \uparrow \infty$: $P\{ W_q > 0 \} \downarrow 0$.
   - Fix $n$ and let $\lambda \uparrow \infty$: $P\{ W_q > 0 \} \uparrow 1$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $n$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $n \approx R + \beta \sqrt{R}$.

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: $L(M/M/n + M) \overset{d}{=} L(M/M/\infty) \overset{d}{=} \text{Poisson}(R)$, 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 $n \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$.
   - Fix $n$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$.
   - $\Rightarrow$ Must have both $\lambda$ and $n$ increase simultaneously:
   - $\Rightarrow$ (CLT) Square-root staffing: $n \approx R + \beta \sqrt{R}$.

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: $L(M/M/n + M) \overset{d}{=} L(M/M/\infty) \overset{d}{=} \text{Poisson}(R)$, with $R = \lambda/\mu$ (Offered-Load)
   - Poisson($R$) $\overset{d}{=} R + Z \sqrt{R}$, with $Z \overset{d}{=} N(0, 1)$. 
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 $n \uparrow \infty$: $P\{W_q > 0\} \downarrow 0$.
   - Fix $n$ and let $\lambda \uparrow \infty$: $P\{W_q > 0\} \uparrow 1$.
   - $\Rightarrow$ **Must** have both $\lambda$ and $n$ increase simultaneously:
   - $\Rightarrow$ (CLT) **Square-root staffing**: $n \approx R + \beta\sqrt{R}$.

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: $L(M/M/n + M) \overset{d}{=} L(M/M/\infty) \overset{d}{=} \text{Poisson}(R)$, with $R = \lambda/\mu$ (Offered-Load)
   - Poisson$(R) \overset{d}{=} R + Z\sqrt{R}$, with $Z \overset{d}{=} N(0, 1)$.
   - $P\{W_q(M/M/n + M) > 0\} \overset{\text{PASTA}}{=} P\{L(M/M/n + M) \geq n\}^{\mu=\theta}$
     \[
P\{L(M/M/\infty) \geq n\} \approx P\{R + Z\sqrt{R} \geq n\} =
     \]
     \[
P\{Z \geq (n - R)/\sqrt{R}\}^{\text{square staffing}} \approx P\{Z \geq \beta\} = 1 - \Phi(\beta).
   \]
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 \( n \uparrow \infty \): \( P\{W_q > 0\} \downarrow 0 \).
   - Fix \( n \) and let \( \lambda \uparrow \infty \): \( P\{W_q > 0\} \uparrow 1 \).
   - \( \Rightarrow \) **Must** have both \( \lambda \) and \( n \) increase simultaneously:
   - \( \Rightarrow \) (CLT) **Square-root staffing**: \( n \approx R + \beta \sqrt{R} \).

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: \( L(M/M/n + M) \overset{d}{=} L(M/M/\infty) \overset{d}{=} Poisson(R) \), with \( R = \lambda/\mu \) (Offered-Load)
   - Poisson\((R) \overset{d}{=} R + Z \sqrt{R} \), with \( Z \overset{d}{=} N(0, 1) \).
   - \( P\{W_q(M/M/n + M) > 0\} \overset{\text{PASTA}}{=} P\{L(M/M/n + M) \geq n\} \overset{\mu=\theta}{=} \)
     \( P\{L(M/M/\infty) \geq n\} \approx P\{R + Z \sqrt{R} \geq n\} = \)
     \( P\{Z \geq (n - R)/\sqrt{R}\} \overset{\text{staffing}}{=} \approx P\{Z \geq \beta\} = 1 - \Phi(\beta) \).

3. QED **Excursions**
QED Intuition via Excursions: Busy-Idle Cycles

\[ Q(0) = N : \text{ all servers busy, no queue.} \]

Let \( T_{N,N-1} = E[\text{Busy Period}] \) down-crossing  \( N \downarrow N - 1 \)
\( T_{N-1,N} = E[\text{Idle Period}] \) 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
\[ T_{N-1,N} = \frac{1}{\lambda_N E_{1,N-1}} \sim \frac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \frac{1}{\sqrt{N}} \cdot \frac{1}{h(-\beta)} \]
\[ T_{N,N-1} = \frac{1}{N\mu \pi_+(0)} \sim \frac{1}{\sqrt{N}} \cdot \frac{\beta/\mu}{h(\delta)/\delta}, \quad \delta = \beta\sqrt{\mu/\theta} \]

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

Hence,
\[ P(\text{Wait} > 0) \sim \left[ 1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta} \right]^{-1}. \]
Process Limits (Queueing, Waiting)

- $\hat{Q}_N = \{\hat{Q}_N(t), t \geq 0\}$: stochastic process obtained by centering and rescaling:
  \[
  \hat{Q}_N = \frac{Q_N - N}{\sqrt{N}}
  \]

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

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

Approximating (Virtual) Waiting Time

\[
\hat{V}_N = \sqrt{N} V_N \Rightarrow \hat{V} = \left[ \frac{1}{\mu \hat{Q}} \right]^+ \]
QED Erlang-X (Markovian Q’s: Performance Analysis)

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

Queueing Networks:
- (Semi-)Closed: Nurse Staffing (Jennings & de Vericourt), CCs with IVR (w/ Khudiakov), Erlang-R (w/ Yom-Tov)
- CCs with Abandonment and Retrials: w. Massey, Reiman, Rider, Stolyar
- Markovian Service Networks: w/ Massey & Reiman

Leaving out:
- **Non-Exponential Service Times**: M/D/n (Erlang-D), G/Ph/n, · · ·, G/GI/n+GI, Measure-Valued Diffusions
- **Dimensioning** (Staffing): M/M/n, · · ·, time-varying Q’s, V- and Reversed-V, · · ·
- **Control**: V-network, Reversed-V, · · ·, SBRNets
Back to “Why does Erlang-A Work?”

**Theoretical (Partial) Answer:**

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

- **Over-Dispersed Arrivals:** $R + \beta R^c$, $c$-Staffing ($c \geq 1/2$).

- **General Patience:** Behavior at the origin matters most (only).

- **General Services:** Empirical insensitivity beyond the mean.

- **Heterogeneous Customers / Servers:** State-Collapse.

- **Time-Varying Arrivals:** Modified Offered-Load approximations.

- **Dependent Building-Blocks:** eg. When (Im)Patience and Service-Times correlated (positively):
  - Predict performance with $E[S \mid Served]$.
  - Calculate offered-load with $E[S] = E[S \mid \text{Wait} = 0]$.
  - Note: staffing $\leftarrow$ service-times $\leftarrow$ waiting / abandonment $\leftarrow$ staffing.
"Why does Erlang-A Work?"  General Patience

Israeli Bank: Yearly Data

Hourly Data  Aggregated

Theory:
**Erlang-A:** $P\{\text{Ab}\} = \theta \cdot E[W_q];$

**M/M/N+G:** $P\{\text{Ab}\} \approx g(0) \cdot E[W_q].$

$g(0) =$ Patience-density at origin
“Why does Erlang-A Work?” General Patience

Israeli Bank: Yearly Data

Hourly Data

Aggregated

Theory:

**Erlang-A:** \(
P\{\text{Ab}\} = \theta \cdot \mathbb{E}[W_q];\)

**M/M/N+G:** \(
P\{\text{Ab}\} \approx g(0) \cdot \mathbb{E}[W_q].\)

\(g(0) = \text{Patience-density at origin}\)

Recipe:

In both cases, use Erlang-A, with \(\hat{\theta} = \frac{\widehat{P\{\text{Ab}\}}}{\widehat{\mathbb{E}[W_q]}}\) (slope above).
"Why does Erlang-A Work?" General Patience

Israeli Bank: Yearly Data

Hourly Data

Aggregated

Theory:

Erlang-A: \( P\{\text{Ab}\} = \theta \cdot E[W_q] \);

M/M/\(N+G\): \( P\{\text{Ab}\} \approx g(0) \cdot E[W_q] \).

\( g(0) = \) Patience-density at origin

Recipe:

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

References on \( g(0) \):
- Stationary M/M/N+GI, w/ Zeltyn
- Process G/GI/N+GI: w/ Momcilovic; Dai & He;
“Why does Erlang-A Work?”  Over-Dispersion

\[ \ln(\text{STD}) \text{ vs. } \ln(\text{AVG}) \]  


\text{Tue-Wed, 30 min resolution}

\[ y = 0.8027x - 0.1235 \quad R^2 = 0.9899 \]

\[ y = 0.8752x - 0.8589 \quad R^2 = 0.9882 \]

\text{Tue-Wed, 5 min resolution}

\[ y = 0.7228x - 0.0025 \quad R^2 = 0.9937 \]

\[ y = 0.7933x - 0.5727 \quad R^2 = 0.9783 \]

\text{Significant linear relations (w/ Aldor & Feigin; then w/ Maman & Zeltyn):}

\[ \ln(\text{STD}) = c \cdot \ln(\text{AVG}) + a \]

(Poisson: STD = AVG^{1/2}, hence c = 1/2, a = 0.)
Over-Dispersion: Random Arrival-Rates

Linear relation between $\ln(\text{STD})$ and $\ln(\text{AVG})$ gives rise to:

Poisson-Mixture (Doubly-Poisson, Cox) model for Arrivals:

$\text{Poisson}(\Lambda)$ with Random-Rate of the form

$$\Lambda = \lambda + \lambda^c \cdot X, \quad c \leq 1;$$

- In Call Centers: $c \approx 0.75 - 0.85$ (significant over-dispersion).
- In Emergency Departments, $c \approx 0.5$ (Poisson).

$X$ random-variable with $E[X] = 0$ ($E[\Lambda] = \lambda$), capturing the magnitude of stochastic deviation from mean arrival-rate:

under conventional Gamma prior ($\lambda$ large), $X$ can be taken Normal with std. derived from the intercept.

QED-c

Regime: Erlang-A, with Poisson$(\Lambda)$ arrivals, amenable to asymptotic analysis (with S. Maman & S. Zeltyn).
Over-Dispersion: Random Arrival-Rates

Linear relation between $\ln(\text{STD})$ and $\ln(\text{AVG})$ gives rise to:

Poisson-Mixture (Doubly-Poisson, Cox) model for Arrivals: Poisson($\Lambda$) with Random-Rate of the form

$$\Lambda = \lambda + \lambda^c \cdot X, \quad c \leq 1;$$

- $c$ determines magnitude of over-dispersion ($\lambda^c$)
  - $c = 1$, proportional to $\lambda$; $c \leq 1/2$, Poisson-level;
  - In Call Centers: $c \approx 0.75 - 0.85$ (significant over-dispersion).
  - In Emergency Departments, $c \approx 0.5$ (Poisson).

\[
\begin{align*}
\Lambda &= \lambda + \lambda^c \cdot X, \quad c \leq 1; \\
\end{align*}
\]

\[
\begin{align*}
\text{Call Centers: } &c \approx 0.75 - 0.85 \text{ (significant over-dispersion).} \\
\text{Emergency Departments, } &c \approx 0.5 \text{ (Poisson).}
\end{align*}
\]
Over-Dispersion: Random Arrival-Rates

Linear relation between $\ln(\text{STD})$ and $\ln(\text{AVG})$ gives rise to:

**Poisson-Mixture** (Doubly-Poisson, Cox) model for Arrivals: Poisson($\Lambda$) with Random-Rate of the form

$$\Lambda = \lambda + \lambda^c \cdot X, \quad c \leq 1;$$

- $c$ determines magnitude of over-dispersion ($\lambda^c$)
  - $c = 1$, proportional to $\lambda$; $c \leq 1/2$, Poisson-level;
  - In **Call Centers**: $c \approx 0.75 - 0.85$ (significant over-dispersion).
  - In **Emergency Departments**, $c \approx 0.5$ (Poisson).
- $X$ random-variable with $E[X] = 0$ ($E[\Lambda] = \lambda$), capturing the magnitude of stochastic deviation from mean arrival-rate: under conventional Gamma prior ($\lambda$ large), $X$ can be taken Normal with std. derived from the intercept.
Over-Dispersion: Random Arrival-Rates

**Linear relation** between ln(STD) and ln(AVG) gives rise to:

**Poisson-Mixture** (Doubly-Poisson, Cox) model for Arrivals:

Poisson($\Lambda$) with **Random-Rate** of the form

$$\Lambda = \lambda + \lambda^c \cdot X, \quad c \leq 1;$$

- $c$ determines magnitude of over-dispersion ($\lambda^c$)
  - $c = 1$, proportional to $\lambda$; $c \leq 1/2$, Poisson-level;
  - In **Call Centers**: $c \approx 0.75 - 0.85$ (significant over-dispersion).
  - In **Emergency Departments**, $c \approx 0.5$ (Poisson).
- $X$ random-variable with $E[X] = 0$ ($E[\Lambda] = \lambda$), capturing the magnitude of **stochastic deviation** from mean arrival-rate:
  - under conventional Gamma prior ($\lambda$ large), $X$ can be taken Normal with std. derived from the intercept.

**QED-c** Regime: Erlang-A, with Poisson($\Lambda$) arrivals, amenable to asymptotic analysis (with **S. Maman & S. Zeltyn**).
Over-Dispersion: The QED-c Regime

**QED-c Staffing:** Under offered-load $R = \lambda \cdot E[S]$,

$$N = R + \beta \cdot R^c, \quad 0.5 < c < 1$$

**Performance measures** (M/M/N + G):

- Delay probability: $P\{W_q > 0\} \sim 1 - G(\beta)$

- Abandonment probability: $P\{Ab\} \sim \frac{E[X - \beta]^+}{n^{1-c}}$

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

- Average actual wait: $E_{\lambda,N}[W] \sim E_{\lambda,N}[V]$
Why Does Erlang-A Work? Time-Varying Arrival Rates

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

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

Arrivals, Offered-Load and Staffing
**Time-Stable Performance of Time-Varying Systems**

**Delay Probability** = As in the **Stationary Erlang-A** (Garnett)

![Graph showing delay probability for different beta values]
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).
What is the Offered-Load $R(t)$?

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

- **Offered-Load Process**: $L(\cdot) = \text{Least number of servers}$ that guarantees no delay.

- **Offered-Load Function**: $R(t) = E[L(t)], \ t \geq 0$.
  
  Think $M_t/G/N_t^? + G$ vs. $M_t/G/\infty$: Ample-Servers.

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

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

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

- $\{A(t), \ t \geq 0\}$ Arrival-Process, rate $\lambda(\cdot)$;

- $S (S_e)$ generic Service-Time (Residual Service-Time).
What is the Offered-Load $R(t)$?

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

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

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

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

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

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

What is the Offered-Load $R(t)$?

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

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

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

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

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

QED-c: $N_t = R_t + \beta R_t^c, \ 1/2 \leq c < 1$; ($c = 1$ separate analysis).
The Offered-Load $R(t), t \geq 0$

- **Backbone** of time-varying staffing:
  - Practically **robust**: up to a station within a complex network (ED).
  - Theoretically **challenging**: only Erlang-A with $\mu = \theta$ tractable.

- **Process**: $L(\cdot) =$ **Least** number of servers that guarantees **no delay**.

- **Offered-Load Function** $R(\cdot) = E[L(\cdot)] \quad (M_t / G / N_t^2 + G \leftrightarrow M_t / G / \infty)$. 
The Offered-Load \( R(t), t \geq 0 \)

- **Backbone** of time-varying staffing:
  - Practically **robust**: up to a station within a complex network (ED).
  - Theoretically **challenging**: only Erlang-A with \( \mu = \theta \) tractable.
- **Process**: \( L(\cdot) = \text{Least} \) number of servers that guarantees **no delay**.
- **Offered-Load Function** \( R(\cdot) = E[L(\cdot)] \) \( (M_t/G/N_t^2 + G \leftrightarrow M_t/G/\infty) \).
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 more for longer services.

Survival Functions by Type (Small Israeli Bank)

Service times stochastic order: $S_{New} < S_{Reg} < S_{VIP}$

Patience times stochastic order: $\tau_{New} < \tau_{Reg} < \tau_{VIP}$
Dependent Primitives: Service- vs. Waiting-Time

Average Service-Time as a function of Waiting-Time
U.S. Bank, Retail, Weekdays, January-June, 2006

Introduction

Relationship Between Service Time and Patience

Workload and Offered-Load

Empirical Results

Future Research

Focus on (Patience, Service-Time) jointly, w/ Reich and Ritov.

\[ \mathbb{E}[S | \tau > W = w], w \geq 0: \text{Service-Time of the Unserved} \]
Dependent Primitives: Service- vs. Waiting-Time

Average Service-Time as a function of Waiting-Time
U.S. Bank, Retail, Weekdays, January-June, 2006

⇒ 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** \( g(w) = E[S \mid \text{Patience} > \text{Wait} = w], \ w \geq 0: \)
   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**
   \[
   E[S \mid \text{Patience} = w] = g(w) - \frac{g'(w)}{h_\tau(w)} ;
   \]
   \( h_\tau(w) = \text{hazard-rate of (im)patience (via un-censoring)}; \)

3. **Offered-load** calculations: Impute \( E[S \mid \text{Patience} = w] \)
   (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 ⇒ *emotional* load dominates

▶ ⇒
  - Offered *Operational* Load
  - Offered *Cognitive* Load
  - Offered *Emotional* Load

▶ ⇒ **Fair** Division of Load (Routing) between 2 Maternity Wards:
  One attending to complications *before* birth, the other to complications *after* birth, and both share normal birth
The Technion SEE Center / Laboratory

Data-Based Service Science / Engineering