QED Q’s

(Service-Science & -Engineering of)

Quality- and Efficiency-Driven Queues
(Call/Contact Centers)

Avishai Mandelbaum
Technion, Haifa, Israel

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

CMU, Probability in (the Service) Industry, August 2007

Based on joint work with Sergey Zeltyn, . . .

Technion SEE Center / Lab: Paul Feigin, Valery Trofimov, RA’s, . . .
Contents

▶ **Data-Based** Introduction

▶ **Simple Models at the Service of Complex Realities:**

▶ Courts, Banks, Hospitals, Call Centers, . . .
Contents

▶ **Data-Based** Introduction
  ▶ **Simple Models at the Service of Complex Realities:** Courts, Banks, Hospitals, Call Centers, . . .
  ▶ The Basic Call-Center Model: **Palm/Erlang-A** (M/M/N+M)
  ▶ Validating Erlang-A? All **Assumptions Violated**
Contents

▶ Data-Based Introduction

▶ Simple Models at the Service of Complex Realities:
   ◀ Courts, Banks, Hospitals, Call Centers, . . .

▶ The Basic Call-Center Model: Palm/Erlang-A (M/M/N+M)

▶ Validating Erlang-A? All Assumptions Violated

▶ But Erlang-A Works! Why? Robustness via asymptotic analysis
  that reveals operational regimes: QED, ED, ED+QED
Contents

- Data-Based Introduction
  - Simple Models at the Service of Complex Realities:
    - Courts, Banks, Hospitals, Call Centers, . . .

- The Basic Call-Center Model: Palm/Erlang-A (M/M/N+M)

- Validating Erlang-A? All Assumptions Violated

- But Erlang-A Works! Why? Robustness via asymptotic analysis that reveals operational regimes: QED, ED, ED+QED

- And many ("stochastically-challenged") call centers work as well - Why? “Right Answers for the Wrong Reasons"
Contents

▶ **Data-Based** Introduction
  ▶ Simple Models at the Service of Complex Realities:
    ▶ Courts, Banks, Hospitals, Call Centers, . . .

▶ The Basic Call-Center Model: **Palm/Erlang-A** (M/M/N+M)

▶ Validating Erlang-A? All **Assumptions Violated**

▶ But **Erlang-A Works! Why?** Robustness via asymptotic analysis that reveals operational regimes: **QED, ED, ED+QED**

▶ And many (“stochastically-challenged”) call centers work as well - Why? **“Right Answers for the Wrong Reasons”**

▶ “Appendix”: Demo of **DataMOCCA**
Background Material (Downloadable)


“Production of Justice” (Administrative) Network

Skills-Based-Routing at the Labor-Court in Haifa, Israel

Avg. sojourn time \(\approx\) in months / years
Processing time \(\approx\) in mins / hours / days
Operational Performance: 5 Judges, 3 Case-Types

Judges: The Best/Worst (Operational) Performer

Judges: Performance Analysis

Judges: Performance by Case-Type

Average Number of Months - \( W \)

Average Number of Cases / Month - \( \lambda \)

Judges: Operational Performance – Base Case

45 100
118
59
33

(6.2, 7.4) (13.5, 7.4) (26.3, 4.5) (12, 4.9) (7.2, 4.6) ...
Little’s Law in Court (Creative Averaging)

Judges: The Best/Worst (Operational) Performer

![Graph showing Little's Law in Court with judges as the best/worst operational performer.](image-url)
Prerequisite: Data

Averages Prevalent.
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 stay in a hospital), its operational history = time-stamps of events.
Prerequisite: Data

Averages Prevalent.
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 stay in a hospital), 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)

Future:

- Hospitals (via RFID)
- IVR (VRU), internet, chat (multi-media)
- Operational + Financial + Marketing / Clinical history
Beyond Averages: Service Times in a Call Center

Histogram of Service Times in an Israeli Call Center

January-October

Jan – Oct:

AVG: 185
STD: 238

7.2 %

November-December

Nov – Dec:

AVG: 200
STD: 249

Log-Normal

7.2% Short-Services:
Beyond Averages: Service Times in a Call Center

Histogram of Service Times in an Israeli Call Center

January-October  
Nov - Dec:

Jan – Oct:  
Nov – Dec:

- **7.2% Short-Services:** Agents’ “Abandon” (improve bonus, rest)
- **Distributions,** not only Averages, must be measured.
- **Lognormal** service times prevalent in call centers
Measurements: Face-to-Face Services

23 Bar-Code Readers at a Bank Branch

Bank – 2nd Floor Measurements
“Face-to-Face Services” Network

Bank Branch = Jackson Network

Entrance

Manager

Teller

Xerox

Tourism

Bottleneck!
Present Focus: Call Centers

U.S. Statistics (Relevant Elsewhere)

- Over 60% of annual business volume via the telephone
- 100,000 – 200,000 call centers
- 3 – 6 million employees (2% – 4% workforce)
- 1000’s agents in a “single” call center = 70 % costs.
- 20% annual growth rate
- $200 – $300 billion annual expenditures
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

Information Design
- Marketing,
- Operations Research

Organizational Design:
- Parallel (Flat)
- Sequential (Hierarchical)

Sociology/Psychology,
- Operations Research

Forecasting

Statistics

New Services
- Design (R&D)
- Operations,
- Marketing

Operations/
- Business
- Process
- Archive

Database
- Design
- Data Mining:
- MIS, Statistics,
- Operations
- Research,
- Marketing

Service Completion

(75% in Banks)

Lost Calls

Redial
- (Retrial)

Busy
- (Rare)
- Good or
- Bad

Arrivals
- (Business Frontier
- of the 21th Century)

Forecasting

Statistics

Computer-Telephony
Integration - CTI
MIS/CS

VRU/
IVR

Agents
- (CSRs)

Agents
- (Experts)

Queue
- (Invisible)

Job Enrichment
- Training, Incentives
- Human Resource
- Management

Skill Based Routing
- (SBR) Design
- Marketing,
- Human Resources,
- Operations Research,
- MIS

VIP Queue

Back-Office

Service Process
- Design

Logistics

VIP
- (Training)

Psychological
- Process
- Archive

Expect 3 min
- Willing 8 min
- Perceive 15 min

Customers
- Segmentation - CRM
Marketing

Customers
- Interface Design

Human Factors
- Engineering

Human Resource
- Management

Service Completion

Positive: Repeat Business
Negative: New Complaint

Index
- Function
- Scientific Discipline
- Multi-Disciplinary

Operations/
- Business
- Process
- Archive

Database
- Design
- Data Mining:
- MIS, Statistics,
- Operations
- Research,
- Marketing

Service Completion

Psychological
- Process
- Archive

Expect 3 min
- Willing 8 min
- Perceive 15 min

Customers
- Segmentation - CRM
Marketing

Customers
- Interface Design

Human Factors
- Engineering

Human Resource
- Management

Service Completion

Positive: Repeat Business
Negative: New Complaint
Beyond Averages: Waiting Times in a Call Center

Small Israeli Bank

Large U.S. Bank

Mean = 98
SD = 105

Medium Israeli Bank
The “Anatomy of Waiting" for Service

Common Experience:

- Expected to wait 5 minutes, Required to 10,
- Felt like 20, Actually waited 10,
- ... etc.
The “Anatomy of Waiting" for Service

Common Experience:
  ▶ Expected to wait 5 minutes, Required to 10,
  ▶ Felt like 20, Actually waited 10,
  ▶ . . . etc.

An attempt at “Modeling the Experience”:

1. Time that a customer expects to wait
2. willing to wait                \((\text{Im})\text{Patience: } \tau\)\n3. required to wait              \((\text{Offered Wait: } V)\)
4. actually waits                \(W_q = \min(\tau, V)\)
5. perceives waiting.
The “Anatomy of Waiting” for Service

Common Experience:

▶ Expected to wait 5 minutes, Required to 10,
▶ Felt like 20, Actually waited 10,
▶ . . . etc.

An attempt at “Modeling the Experience”:

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

Experienced customers \(\Rightarrow\) Expected = Required
“Rational” customers \(\Rightarrow\) Perceived = Actual.

Then left with \((\tau, V)\).
Call Center Data: Hazard Rates (Un-Censored)

(Call)Patience Time $\tau$

Israel

U.S.
Call Center Data: Hazard Rates (Un-Censored)

(Im)Patience Time $\tau$

Required/Offered Wait $\checkmark$

Israel

U.S.
Call Center Data: Hazard Rates (Un-Censored)

(Im)Patience Time $\tau$

Israel

Required/Offered Wait $\nu$

U.S.

Note: 5% abandoning $\Rightarrow$ 95% (im)patience-observations censored!
Peaks Every 60 Seconds. Why?

- Human: **Voice-announcement** every 60 seconds.
A “Waiting-Times” Puzzle at a Large Israeli Bank

Peaks Every 60 Seconds. Why?
- Human: Voice-announcement every 60 seconds.
- System: Priority-upgrade (unrevealed) every 60 sec’s (Theory?)

Served Customers

Abandoning Customers
Models for Performance Analysis

- **(Im)Patience**: r.v. $\tau$ = Time a customer is **willing to wait**

- **Offered-Wait**: r.v. $V$ = Time a customer is **required to wait**
  (= Waiting time of a customer with infinite patience).

- **Abandonment** = $\{\tau \leq V\}$

- **Service** = $\{\tau > V\}$

- **Actual Wait** $W_q = \min\{\tau, V\}$. 

- **Modeling**:
  $\tau$ = input to the model, $V$ = output.

- **Operational Performance-Measure calculable in terms of $(\tau, V)$**:

  - **e.g.** Avg. Wait $= E[\min\{\tau, V\}]$
    
  - **e.g.** % Abandon $= P\{\tau \leq V\}$
    $P\{5 \text{ sec} < \tau \leq V\}$

- **Application**:
  Staffing – How Many Agents? (then: When? Who?)
Models for Performance Analysis

- **(Im)Patience**: \( r.v. \ \tau = \) Time a customer is **willing to wait**
- **Offered-Wait**: \( r.v. \ V = \) Time a customer is **required to wait**
  (= Waiting time of a customer with infinite patience).
- **Abandonment** = \( \{ \tau \leq V \} \)
- **Service** = \( \{ \tau > V \} \)
- **Actual Wait** \( W_q = \min\{\tau, V\} \).

Modeling: \( \tau = \) **input** to the model, \( V = \) **output**.

Operational Performance-Measure calculable in terms of \((\tau, V)\):

- eg. **Avg. Wait** = \( E[\min\{\tau, V\}] \) \( \quad (E[W_q|\text{Served}] = E[V|\tau > V]) \)
- eg. **% Abandon** = \( P\{\tau \leq V\} \) \( \quad (P\{5 \text{ sec} < \tau \leq V\}) \)

Application: **Staffing – How Many Agents?** (then: When? Who?)
**The Basic Staffing Model: Erlang-A (M/M/N + M)**

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

- $\lambda$ – **Arrival** rate (Poisson)
- $\mu$ – **Service** rate (Exponential)
- $\theta$ – **Impatience** rate (Exponential)
- $n$ – Number of **Service-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?
- Customers / Servers Heterogeneous?
- Service discipline FCFS?
- . . . ?

**Validation**: Support? Refute?
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers


- Dec. 1995: % Arrivals
  - Hourly arrivals peak at 10:00 and 15:00
- May 1959: Arrival Rate
  - Hourly rate of input
    - Observation: Peak loads at 10:00 & 15:00
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers


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

November 1999 (Israel)
Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram

Survival-Functions by Service-Class

- **New** Customers: 2 min (NW);
- **Regulars**: 3 min (PS);
- **Stock**: 4.5 min (NE);
- **Tech-Support**: 6.5 min (IN).

Observation: VIP require longer service times.
(Im)Patience while Waiting (Palm 1943-53)

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

**Regular** over **VIP** Customers – Israeli Bank
(Im)Patience while Waiting (Palm 1943-53)

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

Regular over VIP Customers – Israeli Bank

- Peaks of abandonment at times of Announcements
- Call-by-Call Data (DataMOCCA) required (& Un-Censoring).

Observation: VIP are more patient (Needy)
A “Service-Time" Puzzle at an Israeli Bank
Inter-related Primitives

Average Service Time over the Day – Israeli Bank

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

**Explanations:**

- **Common:** Service protocol different (longer) during peak times.
A “Service-Time” Puzzle at an Israeli Bank
Inter-related Primitives

Average Service Time over the Day – Israeli Bank

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

**Explanations:**
- Common: Service protocol different (longer) during peak times.
- Operational: The needy abandon less during peak times; hence the **VIP remain** on line, with their **long service** times.
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 possibly implies long service)
- Customers and Servers not homogeneous (classes, skills)
- Customers return for service (after busy, abandonment)
- ... 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 possibly implies long service)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment)
- · · · , and more.

Question: **Is Erlang-A Practically Relevant?**
Estimating (Im)Patience: via $P\{\text{Ab}\} \propto E[W_q]$

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

### Israeli Bank: Yearly Data

Graphs based on 4158 hour intervals.

Estimate of mean (im)patience: $250/0.55 \approx 450 \text{ seconds}$. 
Erlang-A: Fitting a Simple Model to a Complex Reality

- Small Israeli Banking Call-Center (10 agents)
- (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).
Erlang-A: Simple, but Not Too Simple

Further Natural Questions:

2. When does it fail? chart boundaries.
3. Generalize: time-variation, SBR, networks, uncertainty, ...
Erlang-A: Simple, but Not Too Simple

Further Natural Questions:

2. When does it fail? chart boundaries.

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

- Efficiency-Driven (ED) regime: Fluid models (deterministic)
Erlang-A: Simple, but Not Too Simple

Further Natural Questions:
2. When does it fail? chart boundaries.
3. Generalize: time-variation, SBR, networks, uncertainty, ...

Answers via Asymptotic Analysis, as load- and staffing-levels increase, which reveals model-essentials:
- Efficiency-Driven (ED) regime: Fluid models (deterministic)

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

Results turn out accurate enough to also cover 10-20 servers. Important – relevant to hospitals (nurse-staffing: de Véricourt & Jennings, 2006), ...
Operational Regimes: Conceptual Framework

Assume: Offered Load \( R = \frac{\lambda}{\mu} (= \lambda \times \text{E}[S]) \) not too small.

**QD Regime:** \( N \approx R + \delta R \)  
- Essentially no delays: \( (N - R)/R \rightarrow \delta, \text{ as } N, \lambda \uparrow \infty \) 
- \([P\{W_q > 0\} \rightarrow 0].\)

**ED Regime:** \( N \approx R - \gamma R \)  
- Garnett, M. & Reiman 2003
- Essentially all customers are delayed
- Wait same order as service-time; \( \gamma \% \) Abandon (10-25%).

**QED Regime:** \( N \approx R + \beta \sqrt{R} \)  
- Erlang 1924, Halfin & Whitt 1981
- %Delayed between 25% and 75%
- Wait one-order below service-time (sec vs. min); 1-5% Abandon.

**QED+ED:** \( N \approx (1 - \gamma)R + \beta \sqrt{R} \)  
- Zeltyn & M. 2006
- QED refining ED to accommodate "timely-delays": \( P\{W_q > T\} \).
Operational Regimes: Conceptual Framework

Assume: Offered Load $R = \frac{\lambda}{\mu}$ ($= \lambda \times E[S]$) not too small.

QD Regime: $N \approx R + \delta R$  
[($N - R)$/$R$ $\rightarrow$ $\delta$, as $N, \lambda$ $\uparrow$ $\infty$]  
► Essentially no delays: $[P\{W_q > 0]\} \rightarrow 0$.

ED Regime: $N \approx R - \gamma R$  
► Garnett, M. & Reiman 2003  
► Essentially all customers are delayed  
► Wait same order as service-time; $\gamma$% Abandon (10-25%).

QED Regime: $N \approx R + \beta \sqrt{R}$  
► Erlang 1924, Halfin & Whitt 1981  
► %Delayed between 25% and 75%  
► Wait one-order below service-time (sec vs. min); 1-5% Abandon.
Operational Regimes: Conceptual Framework

Assume: Offered Load \( R = \frac{\lambda}{\mu} (= \lambda \times \text{E}[S]) \) not too small.

**QD Regime:** \( N \approx R + \delta R \)  
\[ (N - R)/R \to \delta, \text{ as } N, \lambda \uparrow \infty \]  
- Essentially no delays:  
\[ P\{W_q > 0\} \to 0 \].

**ED Regime:** \( N \approx R - \gamma R \)  
- Garnett, M. & Reiman 2003  
- Essentially all customers are delayed  
- Wait same order as service-time; \( \gamma \)% Abandon (10-25%).

**QED Regime:** \( N \approx R + \beta \sqrt{R} \)  
- Erlang 1924, Halfin & Whitt 1981  
- %Delayed between 25% and 75%  
- Wait one-order below service-time (sec vs. min); 1-5% Abandon.

**QED+ED:** \( N \approx (1 - \gamma)R + \beta \sqrt{R} \)  
- Zeltyn & M. 2006  
- QED refining ED to accommodate “timely-delays”: \( P\{W_q > T\} \).
QED: Practical Support

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

Empirical Service Grade (Beta) American data. Beta vs ASA

-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0% 1% 2% 3% 4% 5% 6% 7% 8%

probability to abandon, %

beta

1.5
1.0
0.5
0.0
-0.5
-1.0

0% 1% 2% 3% 4% 5% 6% 7% 8%
QED: Theoretical Support (Garnett, M., Reiman ‘02; Zeltyn ‘03)

Consider a sequence of $M/M/N+G$ models, $N=1,2,3,...$

Then the following points of view are equivalent:

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

- **Customers** $\%\{\text{Abandon}\} \approx \gamma \sqrt{\frac{1}{N}}$, $0 < \gamma$;

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

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

QED performance (ASA, ...) is easily computable, all in terms of $\beta$ (the square-root safety staffing level) – see later.
QED Approximations (Zeltyn, M. ‘06)

$G$ – patience distribution,

$g_0$ – patience density at origin  \((g_0 = \theta, \text{ if } \exp(\theta))\).

\[ N = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}) , \quad -\infty < \beta < \infty . \]

\[
P\{\text{Ab}\} \approx \frac{1}{\sqrt{N}} \cdot \left[ h(\tilde{\beta}) - \tilde{\beta} \right] \cdot \left[ \sqrt{\frac{\mu}{g_0}} + \frac{h(\tilde{\beta})}{h(-\beta)} \right]^{-1},
\]

\[
P \left\{ W > \frac{T}{\sqrt{N}} \right\} \approx \left[ 1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\beta)}{h(-\beta)} \right]^{-1} \cdot \frac{\Phi(\tilde{\beta} + \sqrt{g_0\mu} \cdot T)}{\Phi(\tilde{\beta})},
\]

\[
P \left\{ \text{Ab} \bigg| W > \frac{T}{\sqrt{N}} \right\} \approx \frac{1}{\sqrt{N}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot [ h(\tilde{\beta} + \sqrt{g_0\mu} \cdot T) - \tilde{\beta} ] .
\]

Here

\[
\tilde{\beta} = \beta \sqrt{\frac{\mu}{g_0}}
\]

\[
\bar{\Phi}(x) = 1 - \Phi(x),
\]

\[
h(x) = \phi(x)/\bar{\Phi}(x) , \quad \text{hazard rate of } N(0, 1).
\]
QED Intuition via Excursions: Busy/Idle Periods

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

Let \( T_{N,N-1} = \text{Busy Period (down-crossing } N \downarrow N - 1 \text{)} \)

\( T_{N-1,N} = \text{Idle Period (up-crossing } N - 1 \uparrow N \text{)} \)

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,

\[
P(\text{Wait} > 0) \sim \left[1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta}\right]^{-1}.
\]
QED Intuition via Excursions: Asymptotics

Calculate

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

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

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

Hence,

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

Special cases:

- **\( \lambda = \theta \):** \( Q \overset{d}{=} M/M/\infty \), since sojourn-time always \( \exp(\mu = \theta) \).
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 \( \sqrt{N} \left(1 - \rho_N\right) \rightarrow \beta, -\infty < \beta < \infty \).

Hence,

\[ P(Wait > 0) \sim \left[ 1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta} \right]^{-1}. \]

Special cases:

- \( \mu = \theta \): \( Q \overset{d}{=} M/M/\infty \), since sojourn-time always \( \exp(\mu = \theta) \).

- \( \beta = 0 (N \approx R) \): \( P\{Wait > 0\} \approx [1 + \sqrt{\theta/\mu}]^{-1} \).
QED Intuition via Excursions: Asymptotics

Calculate

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

\[ T_{N,N-1} = \frac{1}{N\mu \pi_+(0)} \sim \frac{1}{\sqrt{N}} \times \frac{\beta/\mu}{h(\delta)/\delta}, \quad \delta = \beta\sqrt{\mu/\theta} \]

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

Hence,

\[ P(Wait > 0) \sim \left[ 1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta} \right]^{-1}. \]

Special cases:

- \( \mu = \theta \): \( Q \overset{d}{=} M/M/\infty \), since sojourn-time always \( \exp(\mu = \theta) \).

- \( \beta = 0 \) \( (N \approx R) \): \( P\{Wait > 0\} \approx [1 + \sqrt{\theta/\mu}]^{-1} \).

- **Both** of the above: \( P\{Wait > 0\} \approx 1/2 \).
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: $\hat{Q}_N(t) \overset{d}{\rightarrow} \hat{Q}(t)$.

Approximating (Virtual) Waiting Time

- $\hat{V}_N = \sqrt{N} V_N \Rightarrow \hat{V} = \left[ \frac{1}{\mu} \hat{Q} \right]^+$ (Puhalskii, 1994)
Dimensioning a Service System

Operational Regimes provide a conceptual framework.

Questions:
1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Answers, via many-server Asymptotic Analysis (w/ Borst & Reiman, 2004; Zeltyn, 2006):
1. Approximations are extremely accurate.
2. Dimensioning:
   ▶ Cost / Profit Optimization: eg. Min costs of Staffing + Congestion.
   ▶ Constraint Satisfaction: eg. Min. N, s.t. QOS constraints.
3. Robustness depends:
   ▶ Without Abandonment: QED covers all, at amazing accuracy.
   ▶ With Abandonment: ED, QED, ED+QED all have a role.
Dimensioning a Service System

Operational Regimes provide a conceptual framework.

Questions:
1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Answers, via many-server Asymptotic Analysis (w/ Borst & Reiman, 2004; Zeltyn, 2006):
1. Approximations are extremely accurate.
2. Dimensioning:
   ▶ Cost / Profit Optimization: eg. Min costs of Staffing + Congestion.
   ▶ Constraint Satisfaction: eg. Min. N, s.t. QOS constraints.
3. Robustness depends:
   ▶ Without Abandonment: QED covers all, at amazing accuracy.
   ▶ With Abandonment: ED, QED, ED+QED all have a role.
Dimensioning a Service System

Operational Regimes provide a **conceptual framework**.

Questions:

1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Answers, via many-server **Asymptotic Analysis** (w/ Borst & Reiman, 2004; Zeltyn, 2006):

1. Approximations are **extremely accurate**.
Operational Regimes provide a **conceptual framework**.

**Questions:**

1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

**Answers**, via many-server **Asymptotic Analysis** (w/ Borst & Reiman, 2004; Zeltyn, 2006):

1. Approximations are **extremely accurate**.
2. **Dimensioning**:
   - **Cost / Profit Optimization**: eg. Min costs of Staffing + Congestion.
   - **Constraint Satisfaction**: eg. Min. N , s.t. QOS constraints.
Dimensioning a Service System

Operational Regimes provide a conceptual framework.

Questions:

1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Answers, via many-server Asymptotic Analysis (w/ Borst & Reiman, 2004; Zeltyn, 2006):

1. Approximations are extremely accurate.
2. Dimensioning:
   ▶ Cost / Profit Optimization: eg. Min costs of Staffing + Congestion.
   ▶ Constraint Satisfaction: eg. Min. N , s.t. QOS constraints.
3. Robustness depends:
   ▶ Without Abandonment: QED covers all, at amazing accuracy.
   ▶ With Abandonment: ED, QED, ED+QED all have a role.
Operational Regimes: Rules-of- Thumb

<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>1-10%</td>
<td>$\geq 10%$</td>
<td>$\leq 10%E[\tau]$</td>
<td>$\geq 10%E[\tau]$</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>QED</td>
<td>QED</td>
<td>QED if $\tau \overset{d}{=} \text{exp}$</td>
</tr>
</tbody>
</table>

**ED:** $N \approx R - \gamma R$ \hspace{1cm} ($0.1 \leq \gamma \leq 0.25$).

**QD:** $N \approx R + \delta R$ \hspace{1cm} ($0.1 \leq \delta \leq 0.25$).

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

**ED+QED:** $N \approx (1 - \gamma)R + \beta \sqrt{R}$ \hspace{1cm} ($\gamma, \beta$ as above).
Back to “Why does Erlang-A Work?"

**Theoretical Answer:** \( M_t^J / G / N_t + G \overset{d}{\approx} (M/M/N + M)_t , \ t \geq 0. \)

- **General Patience**: Behavior at the origin is all that matters.
- **General Services**: Empirical insensitivity beyond the mean.
- **Time-Varying Arrivals**: Modified Offered-Load approximations.
- **Heterogeneous Customers**: 1-D state collapse.
Back to “Why does Erlang-A Work?”

**Theoretical Answer:** \( M_t^J / G / N_t + G \overset{d}{\approx} (M/M/N + M)_t, \ t \geq 0. \)

- **General Patience:** Behavior at the origin is all that matters.
- **General Services:** Empirical insensitivity beyond the mean.
- **Time-Varying Arrivals:** Modified Offered-Load approximations.
- **Heterogeneous Customers:** 1-D state collapse.

**Practically:** Why do (stochastically-challenged) Call Centers work?

“The right answer for the wrong reason"
“Why does Erlang-A Work?” General Patience

(Im)Patience times **Generally Distributed**: $M/M/n+G$

**Exact** analysis in steady-state (Baccelli & Hebuterne, 1981): solve Kolmogorov’s PDE’s (semi-Markov) for the offered-wait $V$. 

$\lambda$ is the arrival rate, $\mu$ is the service rate, $G$ represents generally distributed patience times.
“Why does Erlang-A Work?” General Patience

(Im)Patience times Generally Distributed: $M/M/n+G$

Exact analysis in steady-state (Baccelli & Hebuterne, 1981): solve Kolmogorov’s PDE’s (semi-Markov) for the offered-wait $V$.

QED analysis (w/ Zeltyn, 2006): $n \approx R + \beta \sqrt{R}$.
- Assume (Im)Patience density $g(0) > 0$.
- $V$ asymptotics ($\lambda \uparrow \infty$): Laplace Method, leading to
- QED Approximations: Use Erlang-A as is, with $\theta \leftrightarrow g(0)$. 
**General Patience: Fitting Erlang-A**

**Israeli Bank: Yearly Data**

**Hourly Data**

**Aggregated**

---

**Theory:**

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

**M/M/N+G:** \( P\{\text{Ab}\} \approx g(0) \cdot E[W_q]. \)
General Patience: Fitting Erlang-A

Israel Bank: Yearly Data

Hourly Data

Aggregated

Theory:

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

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

Recipe:

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

Established:  \( M/M/N+G \approx M/M/N+M \)  \((\theta = g(0))\).
Why Does Erlang-A Work?  General Services

Established: \[ M/M/N+G \approx M/M/N+M \quad (\theta = g(0)). \]

Now: \[ M/G/N+G \approx M/M/N+G \quad (E[S] \text{ same in both}). \]
Why Does Erlang-A Work?  General Services

Established: \( M/M/N+G \approx M/M/N+M \)  \((\theta = g(0))\).

Now: \( M/G/N+G \approx M/M/N+G \)  \((E[S] \text{ same in both})\).


Lognormal (CV=1) vs. Exponential Service Times, QED Regime;
100 agents, average patience = average service

Fraction Abandoning

Delay Probability
Why Does Erlang-A Work? General Services

Established: \( M/M/N+G \approx M/M/N+M \) \( (\theta = g(0)) \).

Now: \( M/G/N+G \approx M/M/N+G \) \( (E[S] \text{ same in both}) \).


Lognormal (CV=1) vs. Exponential Service Times, QED Regime; 100 agents, average patience = average service

Fraction Abandoning

Delay Probability

QED G-Services: \( G/D_K/N+G \) (w/ Momčilović, ongoing).
Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: \( M/G/N+G \approx M/M/N+M \) \((\theta = g(0))\).
Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: \( M/G/N+G \approx M/M/N+M \quad (\theta = g(0)). \)

Now: \( M_t/G/N_t + G \approx (M/G/N + G)_t \quad (N_t, \lambda \text{ well chosen}). \)
Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: \( M/G/N + G \approx M/M/N + M \) \((\theta = g(0))\).

Now: \( M_t/G/N_t + G \approx (M/G/N + G)_t \) \((N_t, \lambda \text{ well chosen})\).

Two steps (Feldman, M., Massey & Whitt, 2006):

1. **Modified Offered-Load: \( \lambda \)**
   - Consider \( M_t/G/N_t + G \) with arrival rate \( \lambda(t), t \geq 0 \).
   - Approximate its **time-varying** performance **at time** \( t \) with a stationary \( M/G/N_t + G \), in which \( \lambda = E\lambda(t - S_e) \).
     \((S_e \overset{d}{=} \text{residual-service: congestion-lag behind peak-load.})\)
Why Does Erlang-A Work? **Time-Varying Arrival Rates**

Established: \( M/G/N+G \approx M/M/N+M \) \((\theta = g(0))\).

Now: \( M_t/G/N_t + G \approx (M/G/N + G)_t \) \((N_t, \lambda \text{ well chosen})\).

Two steps (Feldman, M., Massey & Whitt, 2006):

1. **Modified Offered-Load: \( \lambda \)**
   - Consider \( M_t/G/N_t + G \) with arrival rate \( \lambda(t), t \geq 0 \).
   - Approximate its **time-varying** performance at time \( t \) with a stationary \( M/G/N_t + G \), in which \( \lambda = E\lambda(t - S_e) \).
     \((S_e \triangleq \text{residual-service: congestion-lag behind peak-load.})\)

2. **Square-Root Staffing: \( N_t \)**
   - Let \( R_t = E\lambda(t - S_e) \times ES \) be the **Offered-Load** at time \( t \)
     \((R_t = \text{Number-in-system in a corresponding } M_t/G/\infty)\).
   - Staff \( N_t = R_t + \beta \sqrt{R_t} \).
Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: \( M/G/N+G \approx M/M/N+M \quad (\theta = g(0)) \).

Now: \( M_t/G/N_t + G \approx (M/G/N + G)_t \quad (N_t, \lambda \text{ well chosen}) \).

Two steps (Feldman, M., Massey & Whitt, 2006):

1. Modified Offered-Load: \( \lambda \)
   - Consider \( M_t/G/N_t + G \) with arrival rate \( \lambda(t), t \geq 0 \).
   - Approximate its time-varying performance at time \( t \) with a stationary \( M/G/N_t + G \), in which \( \lambda = E\lambda(t - S_e) \).
     \((S_e \triangleq \text{residual-service: congestion-lag behind peak-load})\)

2. Square-Root Staffing: \( N_t \)
   - Let \( R_t = E\lambda(t - S_e) \times ES \) be the Offered-Load at time \( t \)
     \((R_t = \text{Number-in-system in a corresponding } M_t/G/\infty)\)
   - Staff \( N_t = R_t + \beta \sqrt{R_t} \).

Serendipity: Time-stable performance, supported by ISA = Iterative Staffing Algorithm, and QED diffusion limits \( (M_t/M/N + M, \mu = \theta) \).
Example: "Real" Call Center
(The "Right Answer" for the "Wrong Reasons")

Time-Varying (two-hump) arrival functions common
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)

Assume: Service and abandonment times are both Exponential, with mean 0.1 (6 min.)
Garnett / Halfin-Whitt Functions: $P\{W_q > 0\}$

- Halfin-Whitt
- QED Erlang-A
**Real Call Center:** Empirical waiting time, given positive wait

(1) $\alpha=0.1$ (QD)  
(2) $\alpha=0.5$ (QED)  
(3) $\alpha=0.9$ (ED)
QED Staffing \((\beta=0\text{ iff }\alpha=0.5)\)
The "Right Answer" (for the "Wrong Reasons")

Prevalent Practice \[ N_t = \left[ \lambda(t) \cdot E(S) \right] \] (PSA)

"Right Answer" \[ N_t \approx R_t + \beta \cdot \sqrt{R_t} \] (MOL)

\[ R_t = E\lambda(t - S) \cdot E(S) \]
The "Right Answer" (for the "Wrong Reasons")

Prevalent Practice \[ N_t = \left[ \lambda(t) \cdot E(S) \right] \] (PSA)

"Right Answer" \[ N_t \approx R_t + \beta \cdot \sqrt{R_t} \] (MOL)

\[ R_t = E \lambda(t - S) \cdot E(S) \]

Practice \approx "Right" \[ \beta \approx 0 \] (QED)

and \[ \lambda(t) \approx \text{stable over service-durations} \]

Practice Improved \[ N_t = \left[ \lambda[t - E(S)] \cdot E(S) \right] \]
The "Right Answer" (for the "Wrong Reasons")

Prevalent Practice \[ N_t = \left[ \lambda(t) \cdot E(S) \right] \] (PSA)

"Right Answer" \[ N_t \approx R_t + \beta \cdot \sqrt{R_t} \] (MOL)

\[ R_t = E\lambda(t - S) \cdot E(S) \]

Practice \approx "Right" \[ \beta \approx 0 \] (QED)

and \[ \lambda(t) \approx \text{stable over service-durations} \]

Practice Improved \[ N_t = \left[ \lambda[t - E(S)] \cdot E(S) \right] \]

When Optimal? for moderately-patient customers:

1. Satisfization \[ \Leftrightarrow \] At least 50% to be serve immediately
2. Optimization \[ \Leftrightarrow \] Customer-Time = 2 x Agent-Salary
Why Does Erlang-A Work? **Multi-Class Customers**

Now: \[ M^J_t / G / N_t + G \approx (M^J / G / N + G)_t \] (well staffed & controlled).

**Service Levels:** Class 1 = VIP, . . ., Class J = best-effort.
Why Does Erlang-A Work? **Multi-Class Customers**

Now: \( M_t^J / G / N_t + G \approx (M_t^J / G / N + G)_t \) (well staffed & controlled).

**Service Levels:** Class 1 = **VIP**, . . ., Class \( J \) = **best-effort**.

**Staffing, Control** (w/ Gurvich & Armony 2005; Feldman & Gurvich):

- Consider \( M_t^J / G / N_t + G \) with arrival rates \( \lambda_j(t), t \geq 0 \).
- Assume **i.i.d. servers**.
- Let \( R_t = E \sum_j \lambda_j(t - S_e) \times ES \) be the **Offered-Load** at time \( t \).
Why Does Erlang-A Work? Multi-Class Customers

Now: \( M_t^J / G / N_t + G \approx (M_t^J / G / N + G)_t \) (well staffed & controlled).

Service Levels: Class 1 = VIP, ..., Class \( J \) = best-effort.

Staffing, Control (w/ Gurvich & Armony 2005; Feldman & Gurvich):

- Consider \( M_t^J / G / N_t + G \) with arrival rates \( \lambda_j(t), t \geq 0 \).
- Assume i.i.d. servers.
- Let \( R_t = \mathbb{E} \sum_j \lambda_j(t - S_e) \times ES \) be the Offered-Load at time \( t \).
- Staff \( N_t = R_t + \beta \sqrt{R_t} \), with \( \beta \) determined by a desired QED performance for the lowest-priority class \( J \).
- Control via threshold priorities, where the thresholds are determined by ISA according to desired service levels.
- Approximate time-varying performance at time \( t \) with a stationary threshold-controlled \( M_t^J / G / N_t + G \), in which \( \lambda_j = \mathbb{E} \lambda_j(t - S_e) \).
Why Does Erlang-A Work? **Multi-Class Customers**

Now: \[ M^J_t / G / N_t + G \approx (M^J_t / G / N + G)_t \] (well staffed & controlled).

**Service Levels:** Class 1 = **VIP**, ..., Class \( J \) = **best-effort**.

**Staffing, Control** (w/ Gurvich & Armony 2005; Feldman & Gurvich):
- Consider \( M^J_t / G / N_t + G \) with arrival rates \( \lambda_j(t), t \geq 0 \).
- Assume i.i.d. servers.
- Let \( R_t = E \sum_j \lambda_j(t - S_e) \times ES \) be the **Offered-Load** at time \( t \).
- **Staff** \( N_t = R_t + \beta \sqrt{R_t} \), with \( \beta \) determined by a desired QED performance for the **lowest-priority** class \( J \).
- **Control** via threshold priorities, where the thresholds are determined by ISA according to desired service levels.
- Approximate **time-varying** performance at time \( t \) with a **stationary** threshold-controlled \( M^J / G / N_t + G \), in which \( \lambda_j = E\lambda_j(t - S_e) \).

**Serendipity:** **Multi-Class Multi-Skill**, w/ **class-dependent** services. Support: ISA, QED diffusion limits (Atar, M. & Shaikhet, 2007).
Additional Simple (QED) Models of Complex Realities:
Exponential Services; i.i.d. Customers, i.i.d. Servers

- **Performance Analysis:**
  - Khudiakova, Feigin, M. (Semi-Open): Call-Center + IVR/VRU;
  - De Véricourt, Jennings (Closed + Delay), then w/ Yom-Tov (Semi-Open): Nurse staffing (ratios), bed sizing;
  - Randhawa, Kumar (Closed + Loss): Subscriber queues.

- **Optimal Staffing:** Accurate to within 1, even with very small $n$’s, for both constraint-satisfaction and cost/revenue optimization (staffing, abandonment and waiting costs).
  - Armony, Maglaras: ($M_x/M/N$) Delay information (Equilibrium);
  - Borst, M., Reiman (M/M/N): Asymptotic framework;

- **Time-Varying Queues**, via 2 approaches:
  - Jennings, M., Massey, Whitt, then w/ Feldman: Time-Stable Performance (ISA, leading to Modified Offered Load);
Less-Simple (QED) Models: General Service-Times

The Challenge: Must keep track of the state of \( n \) individual servers, as \( n \uparrow \infty \). (Recall Kiefer & Wolfowitz).

- Shwartz, M. (M/G/N), Rosenshmidt, M. (M/G/N+G): Simulations; \textbf{LogNormal better then Exp}, 2-valued same as D.
- Whitt (GI/M+0/N): Covering \( CV \geq 1 \);
- Puhalskii, Reiman (GI/PH/N): Markovian process-limits (no steady-state); also priorities;
- Kaspi, Ramanan (G/G/N): Fluid, next Diffusion (measure-valued ages, following Kiefer & Wolfowitz);
- Reed (GI/GI/N): Fluid, Diffusion (\textbf{Skorohod-Like Mapping}).
Complex (QED) Models: Skills-Based Routing
(Heterogeneous Customers or/and Servers - Theory)

- **V-Model**: Harrison, Zeevi; Atar, M., Reiman; Gurvich, M., Armony;
  then **Class-dependent** services: Atar, M., Shaikhet;

- **Reversed-V**: Armony, M.;
  then **Pool-dependent** services: Dai, Tezcan; Gurvich, Whitt
  ($G-c\mu$); Atar, M., Shaikhet (Abandonment);

- **General**: Atar, then w/ Shaikhet (Null-controllability, Throughput-suboptimality); Gurvich, Whitt (FQR);

- **Distributed Networks**: Tezcan;

- **Random Service Rates**: Atar (Fastest or longest-idle server).
The Technion SEE Center / Laboratory
DataMOCCA = Data MOdels for Call Center Analysis

- **Technion**: P. Feigin, V. Trofimov, Statistics / SEE Laboratory.
- **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
- **industry**:
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents.**
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents; ongoing.**
DataMOCCA = Data MOdels for Call Center Analysis

- **Technion**: P. Feigin, V. Trofimov, Statistics / SEE Laboratory.
- **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
- **industry**:
  - U.S. Bank: 2.5 years, 220M calls, 40M by 1000 agents.
  - Israeli Cellular: 2.5 years, 110M calls, 25M calls by 750 agents; ongoing.

**Project Goal**: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing and displaying Call-by-Call-based Data / Information.
DataMOCCA = Data MOdels for Call Center Analysis

- **Technion**: P. Feigin, V. Trofimov, Statistics / SEE Laboratory.
- **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
- **industry**:  
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents.**  
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents; ongoing.**

**Project Goal**: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing and displaying **Call-by-Call-based Data / Information**.

**System Components**:  
- **Clean Databases**: operational-data of individual calls / agents.  
- **Graphical Online Interface**: easily generates graphs and tables, at varying resolutions (seconds, minutes, hours, days, months).
DataMOCCA = Data MOdels for Call Center Analysis

- **Technion**: P. Feigin, V. Trofimov, Statistics / SEE Laboratory.
- **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
- **industry**:
  - U.S. Bank: 2.5 years, 220M calls, 40M by 1000 agents.
  - Israeli Cellular: 2.5 years, 110M calls, 25M calls by 750 agents; ongoing.

**Project Goal**: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing and displaying *Call-by-Call-based Data / Information*.

**System Components**:
- **Clean Databases**: operational-data of individual calls / agents.
- **Graphical Online Interface**: easily generates graphs and tables, at varying resolutions (seconds, minutes, hours, days, months).

**Free for academic adoption**: ask for a DVD (3GB).