Empirically-Based Staffing in Call Centers

Simple Models at the Service of Complex Realities

Sergey Zeltyn

Technion, Haifa, Israel

IBM Thomas J. Watson Research Center, October 3, 2006

Joint work with Professor Avi Mandelbaum

Outline

Subject/Flow of the Talk:
Hierarchy of operational decisions in **call centers**, with emphasis on **staffing**.

Main message I:
Simple Models can be excellent tools at the service of **Complex Realities**.

Supported by:
Erlang-A Model and the **QED** operational **regime** applied to call center **staffing**.

Main message II:
Empirically-Based Analysis is a Prerequisite for Research, Teaching and Practice of service operations.

Supported by:
DataMOCCA – Data MOdel for Call Center Analysis.

Call Centers Industry

U.S. Statistics

- Over 60% of annual business volume via the telephone
- 70,000 – 200,000 call centers
- 3 – 6.5 million employees (3% – 6% workforce)
- 20% annual growth rate
- $100 – $300 billion annual expenditures
- 1000’s agents in a “single" call center.

Quality/Efficiency Tradeoff

- 65 – 80% personnel costs
- Over 90% U.S. consumers form company image via call center experience.

Objective:

Having the right number of appropriately skilled agents when needed.

Staffing Problem

Determination of load-dependent **number of agents**.

Prevailing method:
SIPP (Stationary Independent Period-by-Period),
a constant number of agents over each period (15, 30 or 60 min).

Agents and **customers** are assumed **homogeneous**.
In particular, every agent can potentially serve every customer.

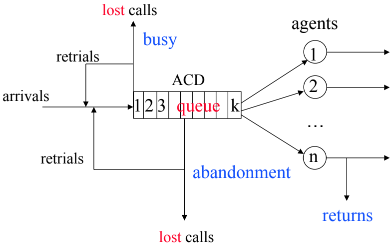
Main Approaches to Staffing

Constraint Satisfaction: find minimal number of agents n^* that satisfies some performance goal(s) (e.g. less than 3% abandonment). Prevalent in practice, Service-Level Agreements (SLA) in outsourcing.

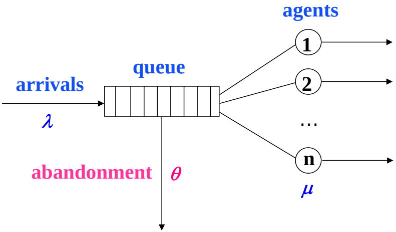
Cost/Revenue Optimization: find n^* that optimizes service revenues and costs of staffing, abandonment and waiting.

M/M/*n*+M (Erlang-A, Palm): Main Staffing Model

Call Center: Schematic Representation



Erlang-A Queue



Erlang-A Assumptions:

- λ – **Poisson** arrival rate
- μ – **Exponential** service rate
- n – number of service agents
- θ – **Exponential** individual abandonment rate

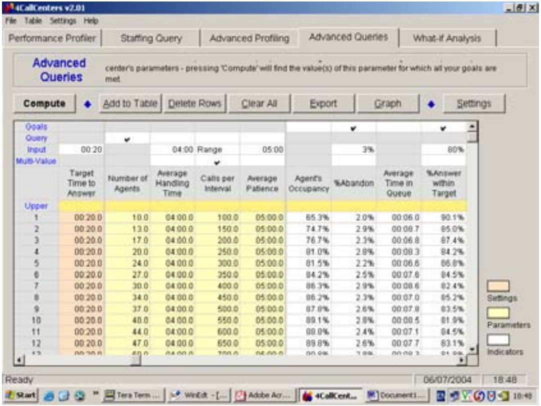
- **No busy** signals
- First Come First Served.

Erlang-A Model: Calculations

Performance measures can be calculated relatively easily.

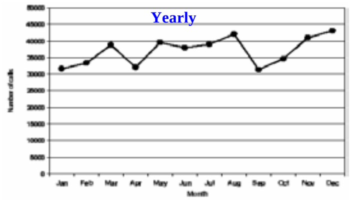
4CallCenters - A Personal Tool for Workforce Management.
Based on the M.Sc. thesis of Ofer Garnett.

http://iew3.technion.ac.il/serveng2006S/4CallCenters/Downloads.htm

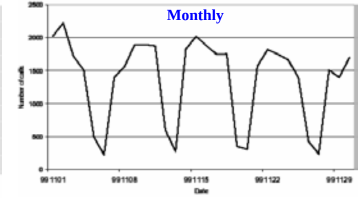


Time Scale: Arrival to a Call Center in 1999

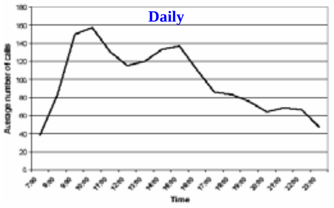
Strategic



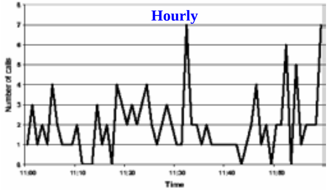
Tactical



Operational



Stochastic



Queueing Science:
Arrival to a Call Center in 1976

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

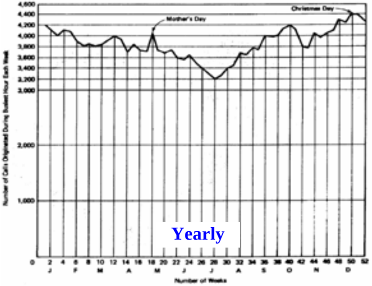


Figure 1 Typical distribution of calls during the busiest hour for each week during a year.

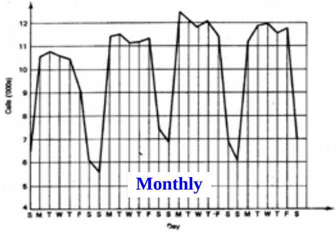


Figure 2 Daily call load for Long Beach, January 1972.

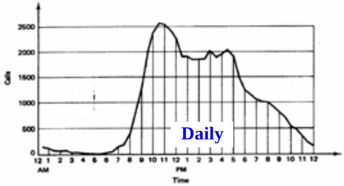


Figure 3 Typical half-hourly call distribution (Bundy D A).

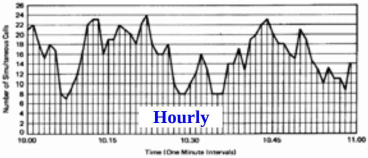
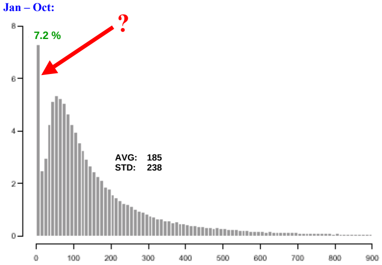


Figure 4 Typical intrahour distribution of calls, 10:00–11:00 A.M.

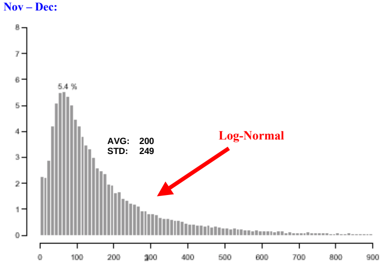
Service Times: Distribution and Psychology

Histogram of Service Times in an Israeli Call Center

January-October



November-December



- **Lognormal** service times prevalent in call centers
- **7.2% Short-Services:** Agents' "Abandon" (improve bonus, rest)
- **Distributions**, not only Averages, must be measured.

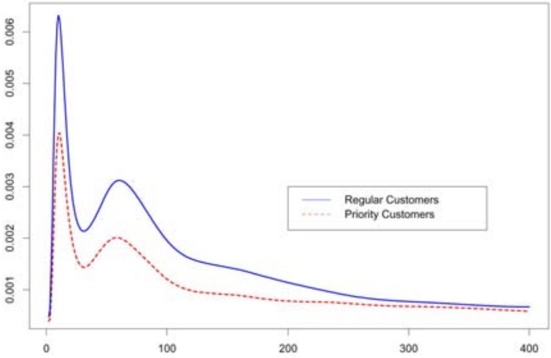
Erlang-A: Modelling (Im)Patience

- **Patience Time:** $\tau \sim \exp(\theta)$
  Time a customer is **willing to wait** for service.
- **Offered Wait:** V
  Time a customer is **required to wait** for service.
  (= Waiting time of a customer with infinite patience.)
- If $\tau \leq V$ then customer **abandons**, else **served**.
- **Actual Wait** $W_q = \min(\tau, V)$.

Patience data is **censored**: 2% abandoning implies 98% censored!

Measuring Patience

Hazard Rates of Patience in an Israeli Bank: Regular over VIP Customers

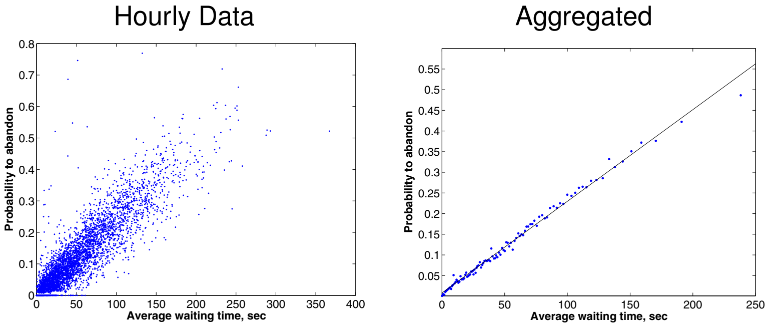


- **VIP** customers are **more patient** (needy).
- Why the peaks in abandonment? **Announcements!**
- **Call-by-call data** required to obtain this graph (+Uncensoring).

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

In queues with **exp(θ)** patience: $\boxed{\text{P\{Ab\}} = \theta \cdot \text{E}[W_q]}$.

Israeli Bank: Yearly Data

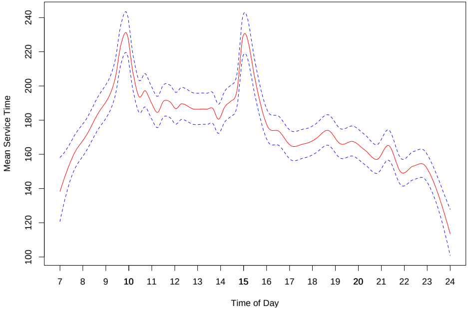


Graphs are based on 4158 hour intervals.

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

Building block: Interrelation

Average Service Time over the Day – Israeli Bank



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

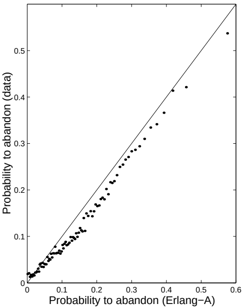
Explanations:

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

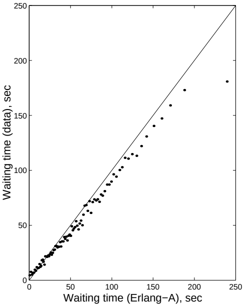
Erlang-A: Fitting a Simple Model to a Complex Reality

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

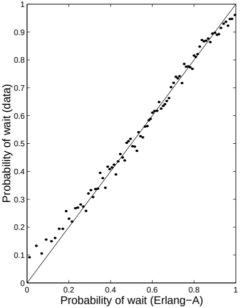
P{Ab}



E[W_q]



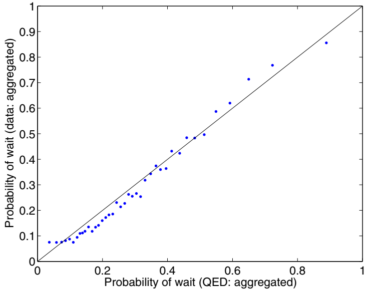
P{$W_q > 0$}



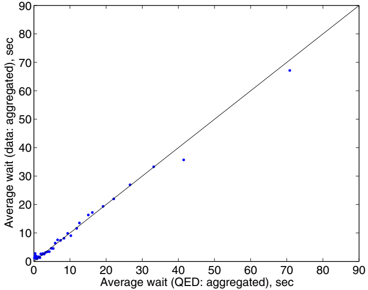
Erlang-A:
Fitting a Simple Model to a Complex Reality II

Large U.S. Bank

Retail. P$\{W_q > 0\}$



Telesales. E$[W_q]$



Partial success, in some cases Erlang-A does not work well (Networking, SBR).

Large Israeli call center – underway.

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

We have learned:

- Arrival process can by approximated by **Poisson**
- Service times are **not exponential** (typically close to lognormal)
- Patience times are **not exponential** (various patterns are observed).

Question: why Erlang-A works with non-exponential patience and service times?

Answer: via study of **operational regimes** in call centers.

Operational Regimes: Motivating Example

Health Insurance. ACD Report.

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
Total	20,577	19,860	**3.5%**	**30**	307	95.1%	
8:00	332	308	7.2%	27	302	87.1%	59.3
8:30	653	615	5.8%	58	293	96.1%	104.1
9:00	866	796	8.1%	63	308	97.1%	140.4
9:30	1,152	1,138	1.2%	28	303	90.8%	211.1
10:00	1,330	1,286	3.3%	22	307	98.4%	223.1
10:30	1,364	1,338	1.9%	33	296	99.0%	222.5
11:00	1,380	1,280	7.2%	34	306	98.2%	222.0
11:30	1,272	1,247	2.0%	44	298	94.6%	218.0
12:00	1,179	1,177	0.2%	1	306	91.6%	218.3
12:30	1,174	1,160	1.2%	10	302	95.5%	203.8
13:00	1,018	999	1.9%	9	314	95.4%	182.9
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**
14:00	1,173	1,082	7.8%	78	313	99.5%	188.9
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**
15:00	1,137	1,122	1.3%	15	320	96.9%	205.8
15:30	1,169	1,137	2.7%	17	311	97.1%	202.2
16:00	1,107	1,059	4.3%	46	315	99.2%	187.1
16:30	914	892	2.4%	22	307	95.2%	160.0
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**
17:30	420	420	0.0%	0	328	73.8%	103.5
18:00	49	49	0.0%	14	180	84.2%	5.8

Efficiency-Driven (ED) Regime

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**

- 100% occupancy
- High P{Ab}
- Considerable waiting time
- $\mathbf{P}\{W_q > 0\} \approx 1$.

Offered load:

$$R_{ED} \triangleq \frac{\lambda}{\mu} = 1061 : \frac{1800}{306} = 180.37\,.$$

Characterization:

$$n = R_{ED} \cdot (1-\gamma)\,, \qquad \gamma > 0.$$

Service grade

$$\gamma = 1 - \frac{n}{R_{ED}} = 1 - \frac{163.4}{180.37} = 0.094 \approx \mathrm{P}\{\mathrm{Ab}\}\,.$$

ED regime captured by **Fluid-Model**.

Quality-Driven (QD) Regime

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**

- Occupancy far below 100%
- Negligible P{Ab}
- Very short waiting time
- $\mathbf{P\{W_q > 0\} \approx 0}$.

Offered load:

$$R_{QD} \;=\; \frac{\lambda}{\mu} \;=\; 615 : \frac{1800}{328} \;=\; 112.07\,.$$

Characterization:

$$n \;=\; R_{QD} \cdot (1+\gamma)\,, \qquad\qquad \gamma > 0.$$

Service grade

$$\gamma \;=\; \frac{n}{R_{QD}} - 1 \;=\; \frac{135}{112.07} - 1 \;=\; 0.205\,.$$

Quality and Efficiency-Driven (QED) Regime

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**

- High occupancy, but not 100%
- Small P{Ab} and waiting
- $\mathbf{P\{W_q > 0\} \approx \alpha}$, $\quad 0 < \alpha < 1$.

Offered load: $R_{QED} \;=\; \dfrac{\lambda}{\mu} \;=\; 1212 : \dfrac{1800}{304} \;=\; 204.69\,.$

Characterization: $\boxed{n \;=\; R_{QED} + \beta\sqrt{R_{QED}}\,,}$ $\qquad -\infty < \beta < \infty\,.$

Service grade

$$\beta \;=\; \frac{n - R_{QED}}{\sqrt{R_{QED}}} = \frac{206.1 - 204.69}{\sqrt{204.69}} \;=\; 0.10\,.$$

Square-Root Staffing Rule: Described by Erlang in 1924!

Awaited the seminal formulation of **Halfin-Whitt** in 1981.

Operational Regimes: Performance.

Assume that **offered load** R is not small ($\lambda \to \infty$).

ED regime: $n \approx R - \delta R$, $0.1 \le \delta \le 0.25$.

- Essentially **all** customers are delayed
- %Abandoned $\approx \delta$ (10-25%)
- Average wait $\approx$ 30 seconds - 2 minutes.

QD regime: $n \approx R + \gamma R$, $0.1 \le \gamma \le 0.25$.
Essentially **no** delays.

QED regime: $n \approx R + \beta\sqrt{R}$, $-1 \le \beta \le 1$.

- %Delayed between 25% and 75%
- %Abandoned is 1-5%
- Average wait is one-order less than average service time (seconds vs. minutes).

Erlang-A Queue: QED Approximations

Assume that **offered load** R is not small ($\lambda \to \infty$).

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

- **Delay probability:**

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

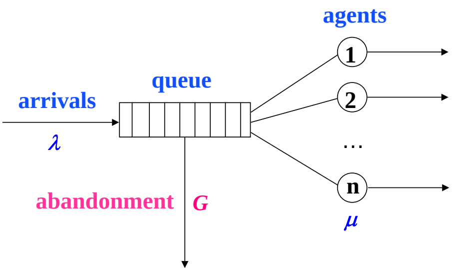
- **Probability to abandon:**

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

- **Linear relation between P{Ab} and E[W_q]:**

$$\boxed{\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W_q]} = \theta.}$$

Generally Distributed Patience: M/M/n+G Model



Back to puzzle of "**Why Erlang-A works?**"
Assume patience times are **generally distributed**.

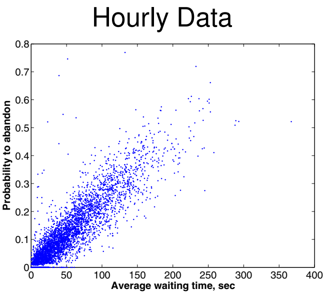
Density of patience time: $g = \{g(x), x \geq 0\}$, where $g(0) \triangleq g_0 > 0$.

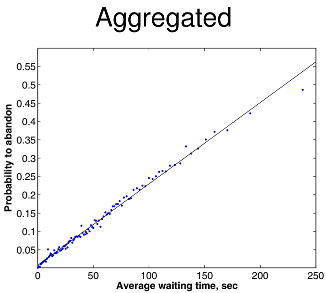
QED regime: $n \approx R + \beta\sqrt{R}$.

QED approximations: use Erlang-A, with g_0 replacing θ.

Generally Distributed Patience:
Fitting Erlang-A

Israeli Bank: Yearly Data





Theory

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

Recipe:

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

Generally Distributed Service Times

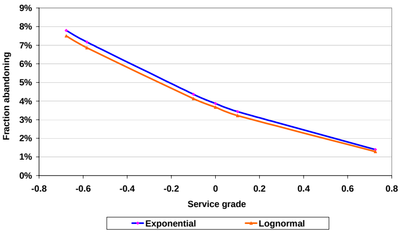
Established: M/M/n+M ≈ M/M/n+G ($\theta = g_0$).

Next: M/M/n+G ≈ M/G/n+G (same mean service).

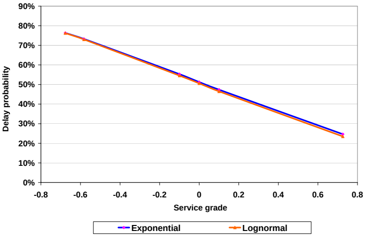
Numerical Experiments: Whitt (2004), Rosenshmidt (2006) demonstrate a **good fit** for typical call-center parameters.

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

Fraction Abandoning



Delay Probability



Simple Model for Complex Realities: More on Applications of the QED Regime

- **Simple performance approximations** that are **robust** also in the ED and QD regimes: Mandelbaum, Zeltyn.
- Optimal staffing for **cost/revenue optimization problems** (staffing, abandonment and waiting costs): Borst, Mandelbaum, Reiman, Zeltyn.
- **General service times**: Puhalskii, Reiman; Jelencović, Mandelbaum, Momčilović.
- Generalizations to **time-varying queues**: Jennings, Feldman, Mandelbaum, Massey, Whitt.
- Generalizations to system with non-homogeneous customers and/or servers (**Skills-Based Routing**): Armony, Gurvich, Mandelbaum; Atar, Shaikhet.
- **Load-balancing**: Dai, Tezcan.

The QED Regime and Stochastic-Ignorant Staffing: The Right Answer for the Wrong Reasons

If $\beta = 0$, QED staffing prescribes:

$$\boxed{n \ = \ R} \ ,$$

R = offered load (minutes of work that arrive per minute).

In word: **Assign number of agents that equals offered load**, which is common practice.

No abandonment: queue “explodes".

With abandonment, $n = 400$, reasonable (im)patience:

- %Delayed $\approx$ 50%
- %Abandoned $\approx$ 2%
- $\mathrm{E}[W_q] \ \approx \ 2\% \cdot \mathrm{E}[S]$, few seconds.

Very good service level.

Call Centers: Hierarchical Operational View

Forecasting Customers (Statistics), Agents (HRM)

Staffing: **Queueing Theory** (Erlang-A based)

Service Level, Costs

\# **FTE's** (Seats)
per unit of time

Shifts: IP, Combinatorial Optimization; LP

Union constraints, Costs

Shift structure

Rostering: Heuristics, AI (Complex)

Individual constraints

Agents Assignments

Skills-based Routing: Stochastic Control (of Q's)

DataMOCCA = Data MOdel for Call Center Analysis

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

System Components:

- Clean **Databases**: operational-data of individual calls, agents and operations.
- Friendly yet powerful **Online Interface**: enables convenient fast access to (mostly) operational and (some) administrative data (but no marketing/business data).

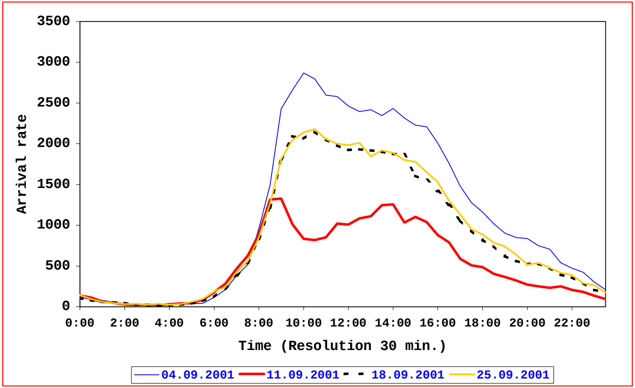
Current Databases:

- Medium-sized U.S. Bank (**2.5 years; 220M calls, 40M via agents; 800 agents at peaks**) – Completed.
- Israeli Cell-Phone Company (**2 years; 110M calls, 25M via agents; 400 agents at peaks**) – Ongoing.
- Large Israeli Bank – Pilot.

DataMOCCA will now be used to illustrate the hierarchy of operational decisions, from forecasting to SBR.

Arrivals to Service: Predictable vs. Random

Arrival Rates on Tuesdays in a September – U.S. Bank



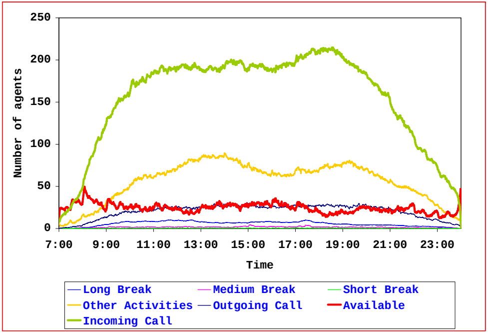
- **Tuesday**, September 4th: **Heavy**, following Labor Day
- **Tuesdays**, September 18 & 25: **Normal**
- **Tuesday, September 11th, 2001.**

Agent Status

Erlang-A Model $\Rightarrow$ optimal **Staffing Level** n.

n = overall number-of-agents that come to work? **No!**

Israeli Bank, Agent Status: Monthly Averages



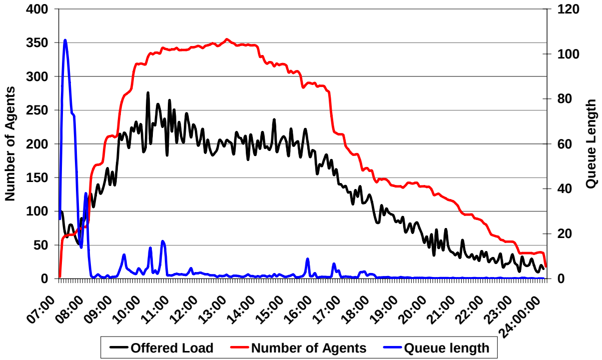
Staffing Level (FTE) = Busy with “Incoming Calls" + “Available" for service.

Shifts Scheduling

Integer Programming given interval-based Staffing Levels.

Example of a shift scheduling problem:
should we bring agents early, given a predictable arrival peak?

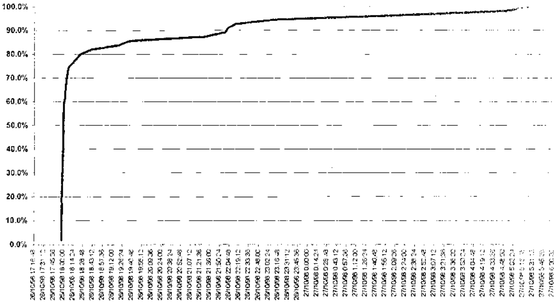
U.S. Bank: Queue-length and Staffing on May 3, 2002



Rostering

Assigning **individual agents** to **schedules**.
Typically, heuristics used to accommodate individual constraints.

Israeli Technical Support Call Center: Online Shift Bidding

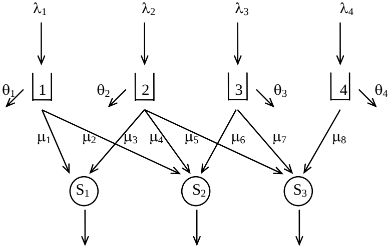


Shift-bidding starts at 18:00.

- 60% of agents are registered till **18:00**
- 80% till 18:24; 90% till 22:00; registration closed at 5:23am.

Introduction to Skills Based Routing

General Setup



Major Control Decisions

- **Customer Routing:** If an agent turns idle and there are queued customers, which customer (if any) should be routed to this agent.
- **Agent Scheduling:** If a customer arrives and there are idle agents, which agent (if any) should serve this customer.
- **Load Balancing:** Routing of customers to distributed call centers (eg. nation-wide).

Customers Relationship/Revenue Management (CRM)

NationsBank CRM relationship groups:

- RG1: high-value customers
- RG2: marginally profitable customers (with potential)
- RG3: unprofitable customer.

NationsBank's Design of the Service Encounter

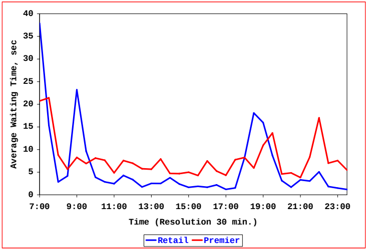
Examples of Specifications: Assignable Grade Of Service

	RG1	RG2	RG3
VRU Target	70% of calls	85% of calls	90% of calls
Abandonment rate	< 1%	< 5%	< 9%
Speed of Answer	100% in 2 rings	80% in 20 seconds	50% in 20 seconds
Average Talk Time	no limit	4 min. average	2 min. average
Rep. Training	universal	product experts	basic product
Rep. Personalization	request rep / callback	FCFS	FCFS
Trans. Confirmation	call / fax	call / mail	mail
Problem Resolution	during call	within 2 business days	within 8 business days

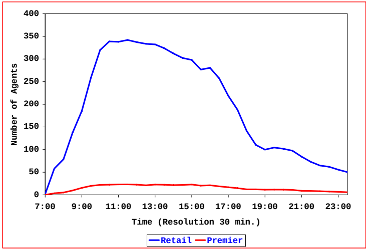
Priorities and Economies-of-Scale

U.S. Bank. Regular vs. VIP Customers. December 2002

Average Wait



Staffing Level



Premier customers do not get a better service level.

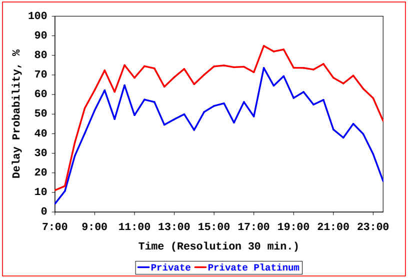
Number of agents assigned to Premier is small and they do not get enough help from regular agents.

Challenge: enable **better service level** for Premier and still serve most of them by a **small dedicated group of agents**.

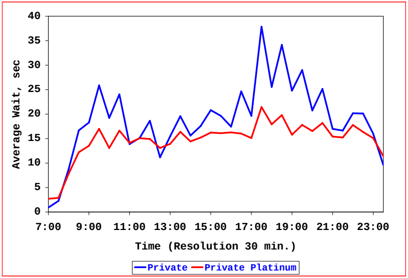
Priorities and Routing Protocols I

Israeli Bank. Regular vs. VIP Customers. October 2004

Delay Probability



Average Wait

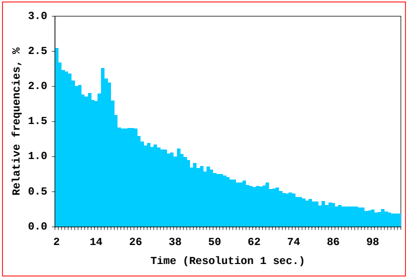


More **Platinum** customers have to wait, **but** their average wait is shorter. **How to explain?**

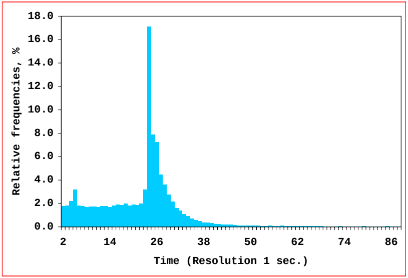
Priorities and Routing Protocols II

Histograms of Waiting Times. October 2004

Private



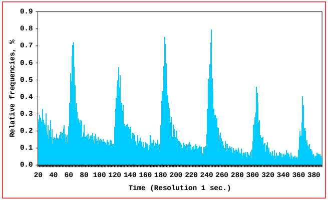
Private Platinum



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

Dynamic Priority-Upgrade

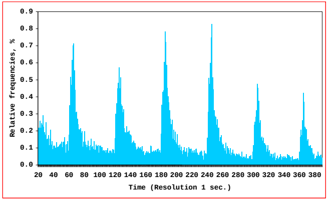
Large Israeli Bank: Histogram of Waiting Times



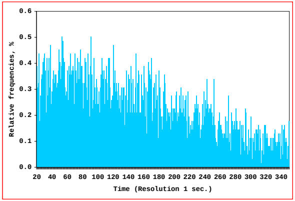
Peaks every 60 seconds. Why?

- System: **Priority-Upgrade** (unrevealed) every 60 seconds
- Human: **Voice-Announcement** every 60 seconds.

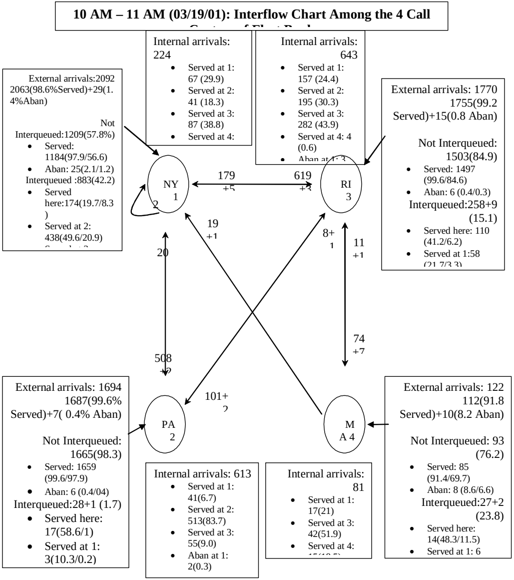
Served Customers



Abandoning Customers



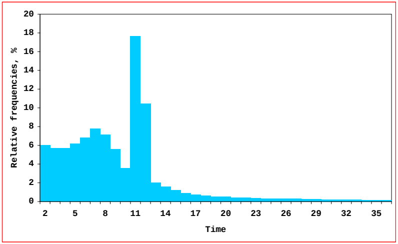
Network Balancing via Interqueue in a U.S. Bank



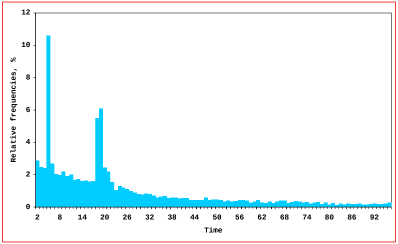
Network Balancing Protocols and Performance Level

U.S. Bank: Histograms of Waiting Times

Retail



Business



Why do we observe a peak for **Retail** service (**10 seconds**)? After 10 seconds of wait **Retail** customers were sent into the **interqueue**.

Business customers – peak at **5 seconds**, for the same reason. Second peak – unclear, maybe priority-upgrade.

Main Research and Practical Challenges

- **Skills-Based Routing:** Convergence of Practice and Theory
- **Uncertainty:** in Reality, Model Parameters, Forecasting
- **Time-Varying Queues:** Time-Stable Performance
- **General Service-Times:** Theory
- **Economic Models:** Operations (Dimensioning), Marketing
- **Human Dimensions:**
  Measure, Model, Experiment, Validate, Refine, etc.

All of the above in a **Network** of distributed call centers.

See our **Service Engineering** site for downloads:
http://iew3.technion.ac.il/serveng