Telephone Call/Contact Centers

Research in State-of-the-Art

Service Engineering

ניהול ותפעול מוקדי שירות (טלפוניים)

המי"ל

e.mail : avim@tx.technion.ac.il

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

1. Supporting Material (Downloadable)

M. "Call Centers: Research Bibliography with Abstracts." Version 5, July, 2003.

Gans (U.S.A.), Koole (Europe), and M. (Israel): "Telephone Call Centers: Tutorial, Review and Research Prospects." *MSOM, 2003.*

M., Sakov, **Zeltyn**: "Empirical Analysis of a Telephone Call Center: A Queueing-Science Perspective."

Brown, Gans, M., Sakov, Shen, **Zeltyn**, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." Submitted, 2003.

Borst, M. and Reiman: "Dimensioning Large Telephone Call Centers." *OR, 2004.*

Erlang: "On the rational determination of the number of circuits." Written in 1923, Used in **1913**, Published in "*The life and works of A.K. Erlang*," 1948. Amazingly, still the most prevalent model in support of WFM.

Contents

Service Engineering – a Subjective View

- Contrast with the traditional and prevalent

  Service Management (Business Schools; U.S.A.)

  Industrial Engineering (Engineering Schools; Europe)

- Goal: Develop scientifically-based design principles (rules-of-thumb) and tools (software) that support the balance of service quality, process efficiency and business profitability, from the (often conflicting) views of customers, servers and managers.

  - Theoretical Framework: Queueing Networks
  - Applications focus: Call (Contact) Centers

Example: Staffing - How many agents required for balancing service-quality with operational-efficiency.

Example: Skills-Based Routing (SBR) – Platinum and Gold and Silver customers, all seeking Support or Purchase, via the Telephone or IVR or e.mail or Chat.

Example: Service Process Design + Staffing + SBR

Multi-Disciplinary: Typical (IE/OR, Marketing, CS, HRM)

BONUS SUPPLEMENT: E-TAILING'S FUTURE

e.biz GEN

www.businessweek.com

BusinessWeek

OCTOBER 23, 2000

A PUBLICATION OF THE McGRAW-HILL COMPANIES

Mutual Funds

How to avoid a big tax bill

Wall Street

Will tech's slide keep spreading?

Dot-coms

The search for new business models

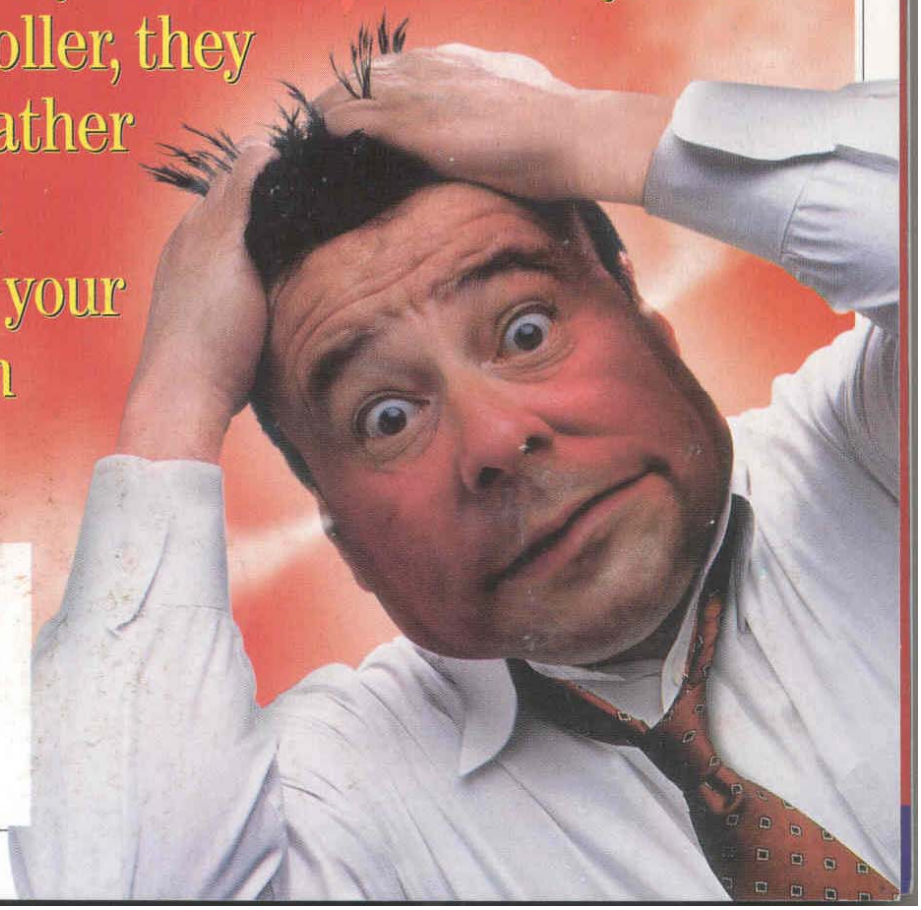
Managed Care

Employers seek a new solution

WHY SERVICE STINKS

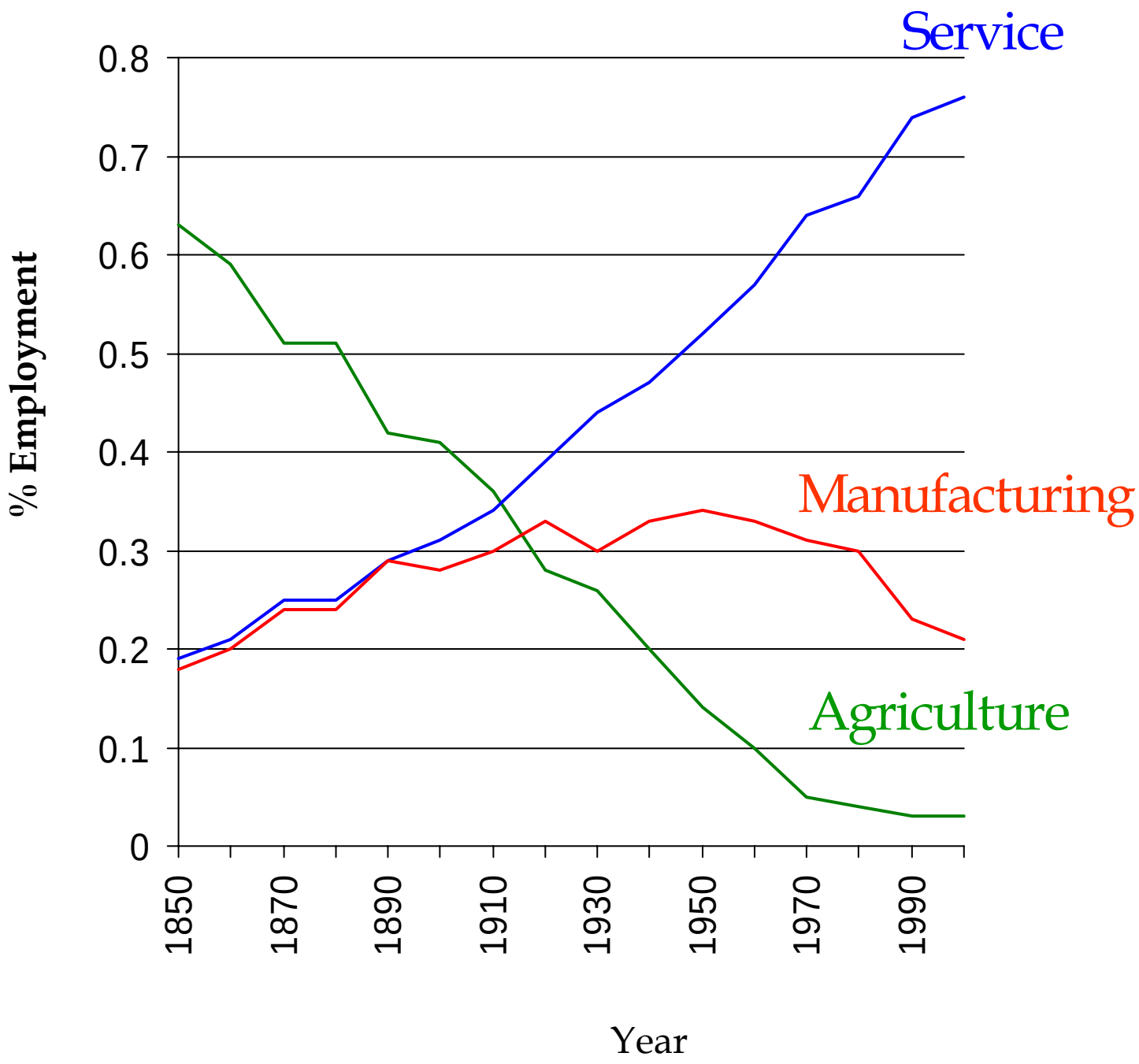
Companies know just how good a customer you are—and unless you're a high roller, they would rather lose you than fix your problem

PAGE 118



AOL Keyword: BW

Employment History: 1850 – 2000+



Services: Some Current Trends

"Everything is Service"

Rather than buying a product, why not buy its service only? For example, car leasing; or, why setup and run a help-desk for technical support, with its costly fast-to-obsolete hardware, growing-sophisticated software, high-skilled peopleware and ever-expanding infoware, rather than let **outsourcing** do it all for you?

Data; Technology and Human Interaction

Far too little reliance on **data, the language of nature**, in formulating models for the systems and processes of the deepest importance to human beings, namely those **in which we are actors**. Systems with fixed rules, such as physical systems, are relatively simple, whereas systems involving human beings expressing their microgoals … can exhibit incredible complexity; there is yet the hope to devise tractable models through remarkable collective effects…
(Robert Herman, in "Reflection on Vehicular Traffi c **Science**".)

Fusion of Disciplines: POM/IE, Marketing, IT, HRM,

The highest challenge facing banks with respect to efficient and effective innovation lies in the "**New Age Industrial Engineer**" that must combine technological knowledge with process design in order to create the delivery system of the future.
(Frei, Harker & Hunter, in "Innovation in Retail Banking").

Service Networks = Queueing Networks

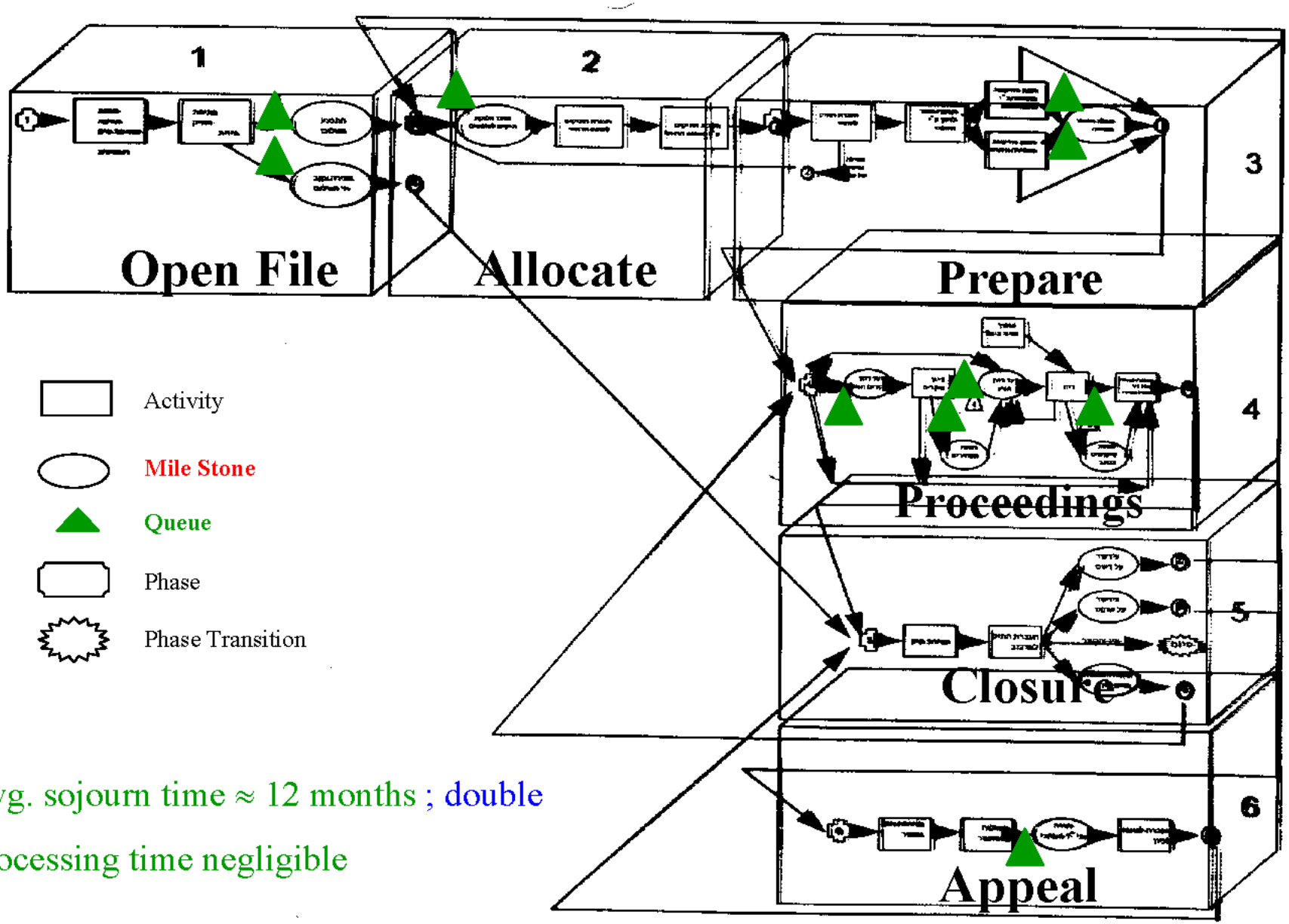
- People, waiting for service: teller, repairman, ATM
- Telephone-calls, to be answered: busy, music, info.
- Forms, to be sent, processed, printed; for a partner
- Projects, to be developed, approved, implemented
- Justice, to be made: pre-trial, hearing, retrial
- Ships, for a pilot, berth, unloading crew
- Patients, for an ambulance, emergency room, operation
- Cars, in rush hour, for parking
- Checks, waiting to be processed, cashed

- Queues Scarce Resources, Synchronization Gaps

  Costly, but here to stay

  - Face-to-face Nets (Chat) (min.)
  - Tele-to-tele Nets (Telephone) (sec.)
  - Administrative Nets (Letter-to-Letter) (days)
  - Fax, e.mail (hours)
  - Face-to-ATM, Tele-to-IVR
  - Mixed Networks (Contact Centers)

“Production” Of Justice



Avg. sojourn time ≈ 12 months ; double

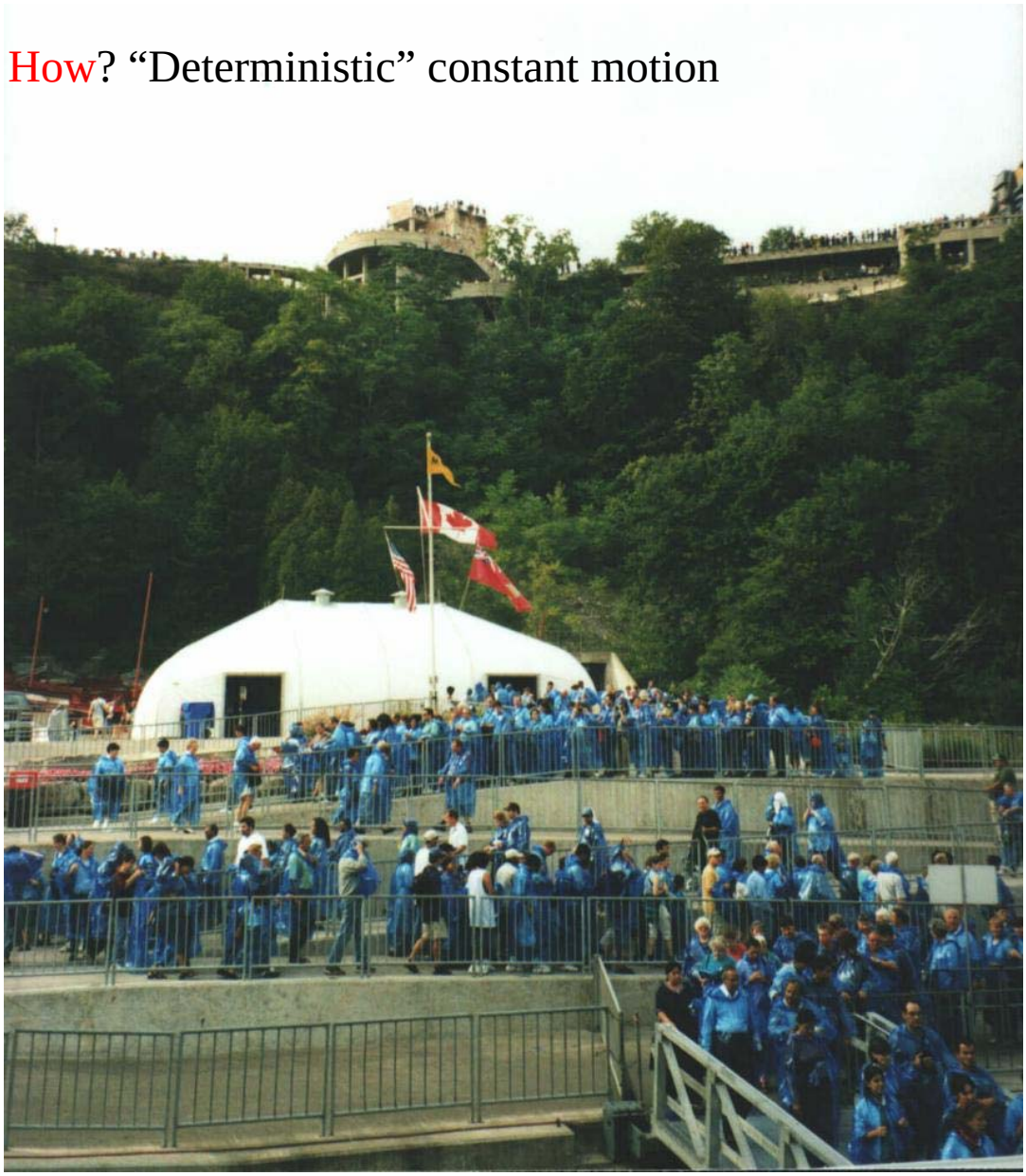
Processing time negligible

Labor-Day Queueing in Niagara Falls

Three-station Tandem Network:
Elevators, Coats, Boats

Total wait of 15 minutes
from upper-right corner to boat

How? “Deterministic” constant motion



Tele-Nets: Call/Contact Centers

Scope	Examples	Perf. Meas.
Information (uni, bi-dir)	#411, Tele-pay, Help Desks	Avg. Delay > 0
Business	Tele-Banks, #800-Retail	Abandons, Econ % Wait > T
Emergency	Police #911	% Wait > 0
Mixed		Weighted
Info + Emerg.	Utility, City Halls	
Info + Bus.	Airlines	

Scale

- 10s to 1000s of agents in a "single" Call Center
- 3% of U.S. work force in call centers (several millions)
- 70% of total business transactions in call centers
- 20% growth rate of the call center industry
- Leading-edge technology, but 70% costs for "people"

Trends: THE interface for/with customers

- Beyond the classical quality vs. efficiency paradigm (Scale)
- Contact Centers (E-Commerce/Multimedia), outsourcing,…
- Retails outlets of 21-Century
- but also the Sweat-shops of the21-Century

(Typical) History of Call Centers at "Tele-Phone"

(http://ie.technion.ac.il/serveng/Lectures/MultiServer.pdf, on pages 28-31.)

- "Tele-Phone" "born" in the late 80's;
  Phone services by regular employees;

- First Call Center "born" in 1990, with 6 agents, via a single phone-number;

- Grew to 60 agents (10-15 per shift) in 1994, with 3 phone numbers for Service, Technical Support and Business.

- In 2003, employs well over 1000 agents, who work in about 10 interconnected call centers (few large, several small), with 300-400 agents handling calls at a peak hour.

דוגמא: תכן מוקד (ראשית שנות ה-90)

* מצב קיים: 3 מספרים (טכני, חשבונות, כללי)

? מספר יחיד: השלכות

* לקוחות: עדיפות, סבלנות... פסיכולוגיה

  פילוח שיווק, סטטיס.

* מוקדנים: גמיש לעומת מומחה הנדסת תעשיה

  הכשרה, מיון, תמריצים... משאבי אנוש

* מערכת: איוש, משמרות, זימון... חקר ביצועים

  ממשק לקוח-מוקדן-מערכת מערכות מידע

אולי חזית+עורף רב תחומי

מהי רמת שירות נאותה רב תחומי

Staffing (+SBR): How Many Agents?

- Fundamental problem in service operations / call centers:
  - People = 70% operating costs; 3% U.S. workforce.
  - HiTec Retail Outlets but also Sweat Shops of 21-Century

Reality

- Workforce Management (WFM) is Erlang-based (1913!)
- Reality is complex and becoming even more so
- Solutions are urgently needed
- Technology enables smart systems
- Theory lags significantly behind needs

$\Rightarrow$ Ad-hoc methods: heuristics, simulation-based

Research Progress is based on

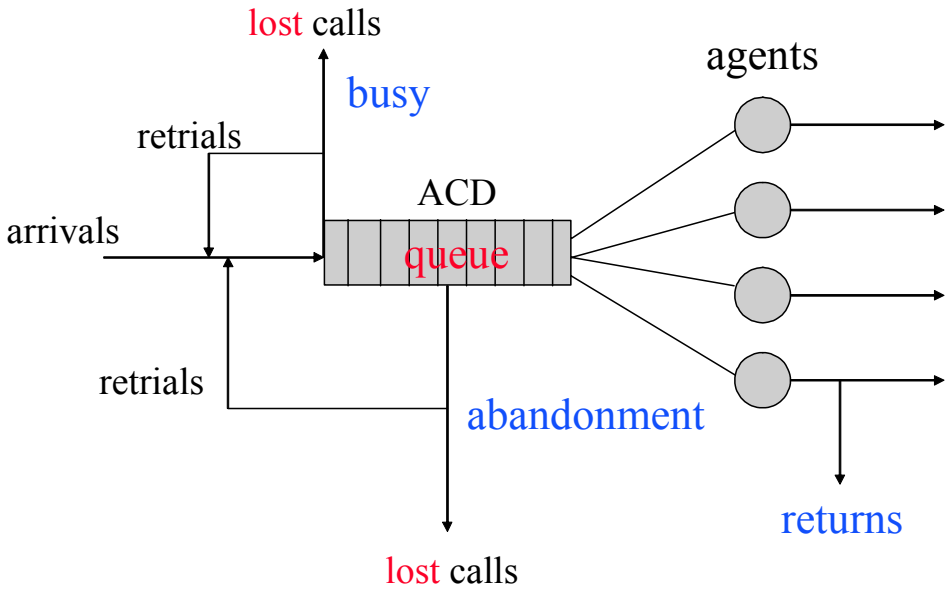
- Small yet significant models for theoretical insight

the analysis of which gives rise to

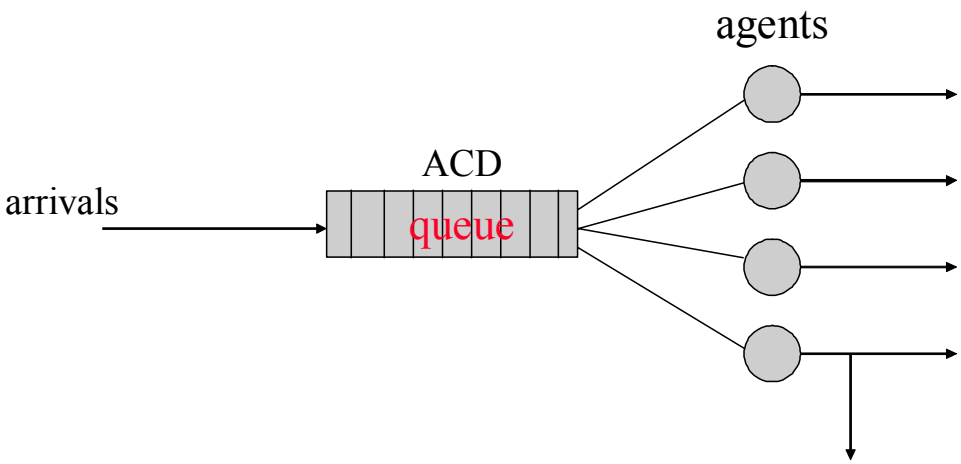
- Principles, Guidelines, Tools: Service Engineering

S

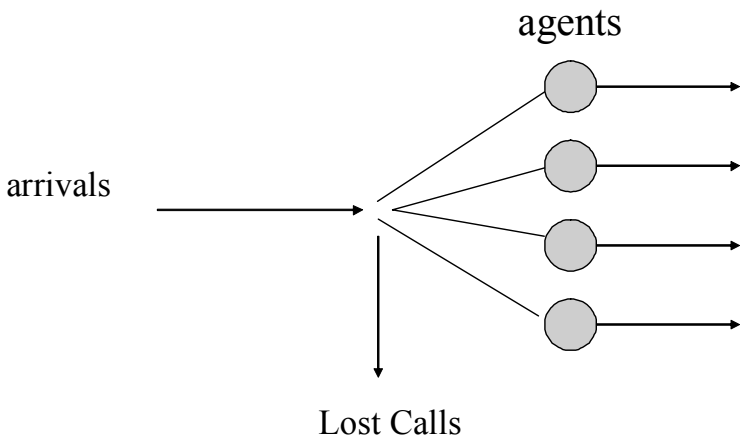
The Basic Call Center



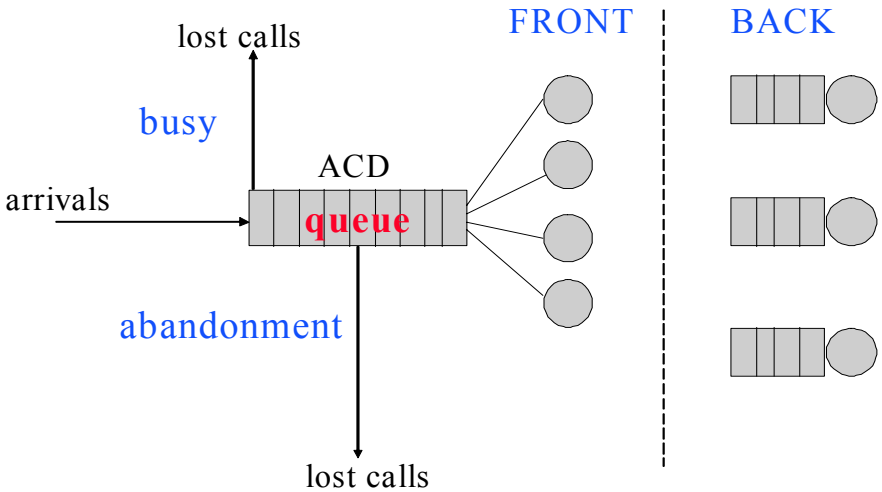
Erlang-C = M/M/N



Erlang-B

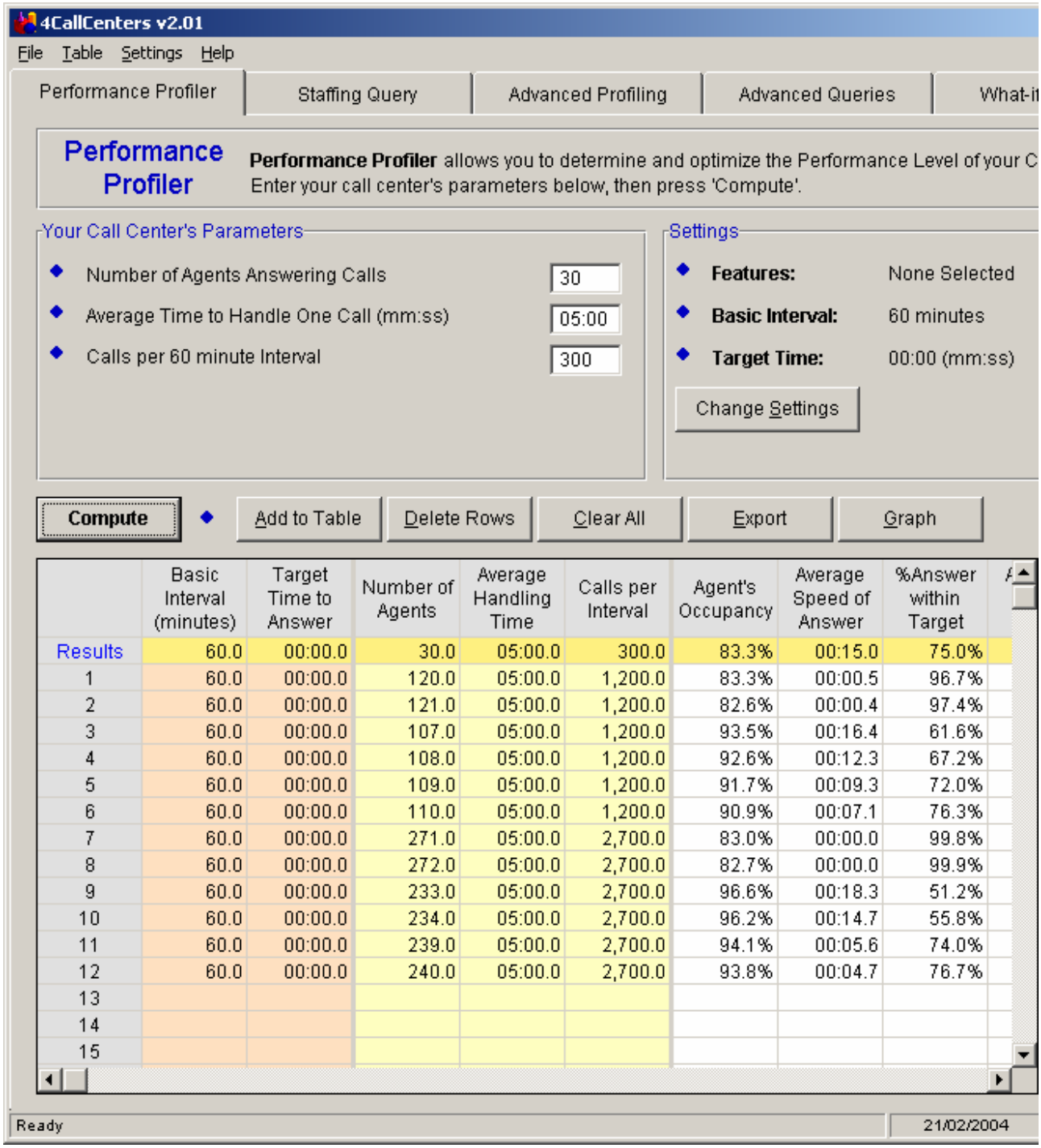


Erlang-A <4CallCenters>



4CallCenters

(Demo: Operational Regimes while Pooling)



	Basic Interval (minutes)	Target Time to Answer	Number of Agents	Average Handling Time	Calls per Interval	Agent's Occupancy	Average Speed of Answer	%Answer within Target
Results	60.0	00:00.0	30.0	05:00.0	300.0	83.3%	00:15.0	75.0%
1	60.0	00:00.0	120.0	05:00.0	1,200.0	83.3%	00:00.5	96.7%
2	60.0	00:00.0	121.0	05:00.0	1,200.0	82.6%	00:00.4	97.4%
3	60.0	00:00.0	107.0	05:00.0	1,200.0	93.5%	00:16.4	61.6%
4	60.0	00:00.0	108.0	05:00.0	1,200.0	92.6%	00:12.3	67.2%
5	60.0	00:00.0	109.0	05:00.0	1,200.0	91.7%	00:09.3	72.0%
6	60.0	00:00.0	110.0	05:00.0	1,200.0	90.9%	00:07.1	76.3%
7	60.0	00:00.0	271.0	05:00.0	2,700.0	83.0%	00:00.0	99.8%
8	60.0	00:00.0	272.0	05:00.0	2,700.0	82.7%	00:00.0	99.9%
9	60.0	00:00.0	233.0	05:00.0	2,700.0	96.6%	00:18.3	51.2%
10	60.0	00:00.0	234.0	05:00.0	2,700.0	96.2%	00:14.7	55.8%
11	60.0	00:00.0	239.0	05:00.0	2,700.0	94.1%	00:05.6	74.0%
12	60.0	00:00.0	240.0	05:00.0	2,700.0	93.8%	00:04.7	76.7%
13								
14								
15								

Workforce Management (WFM): Hierarchical Operational View

Forecasting Customers (Statistics), Agents (HRM)

Staffing: Queueing Theory (M/M/N 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)

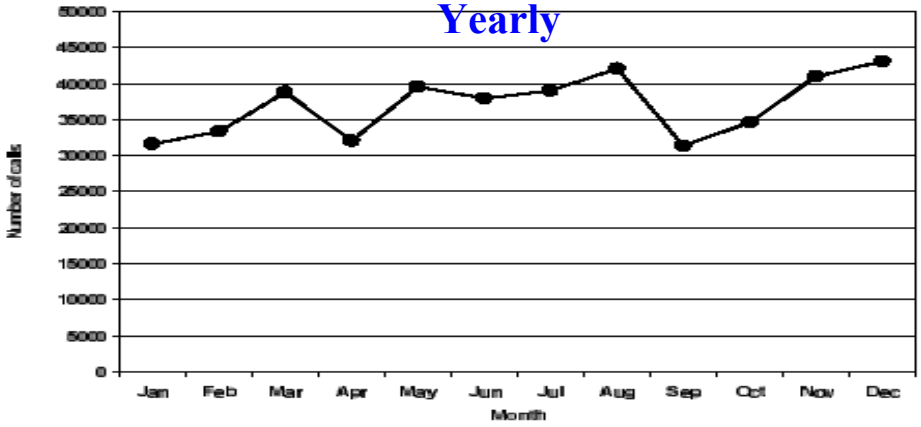
Individual constraints

Agents Assignments

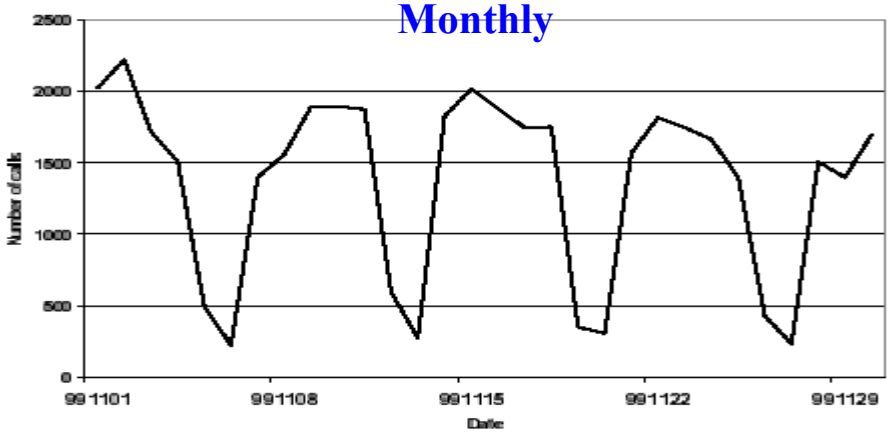
Skills-based Routing: Stochastic Control (of Q's)

Arrival Process

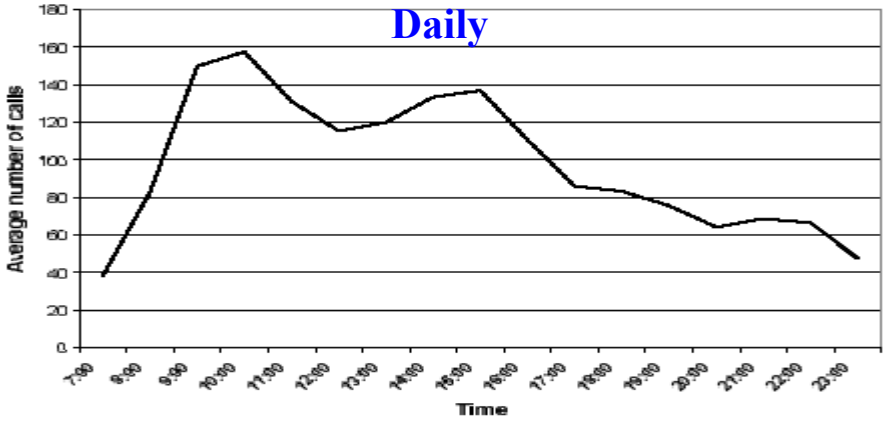
Yearly



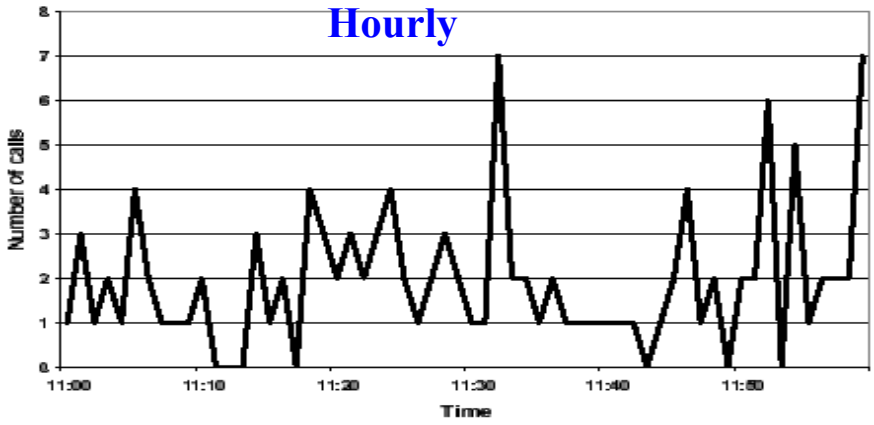
Monthly



Daily

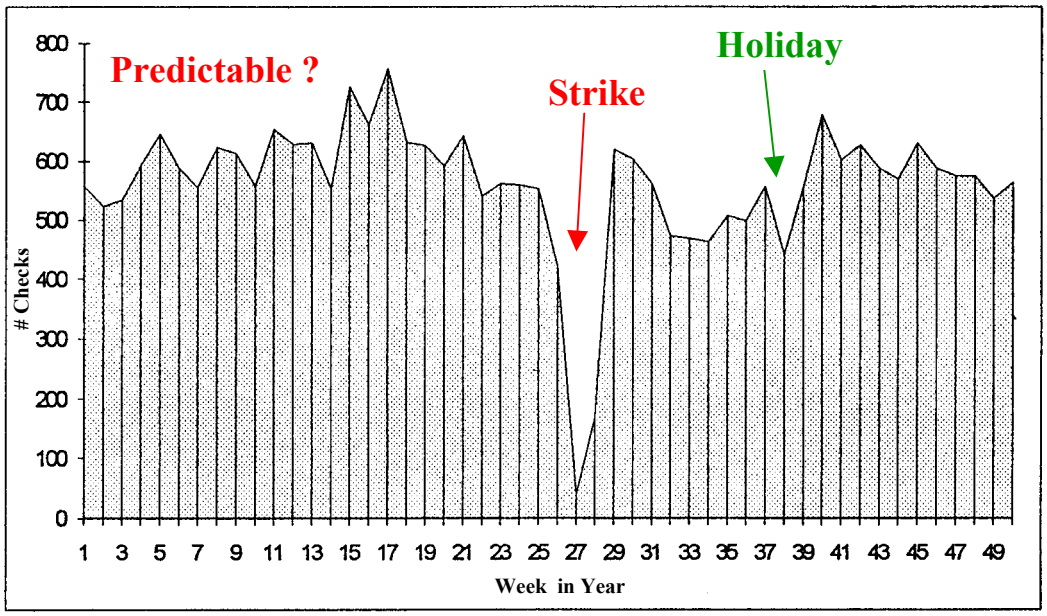


Hourly



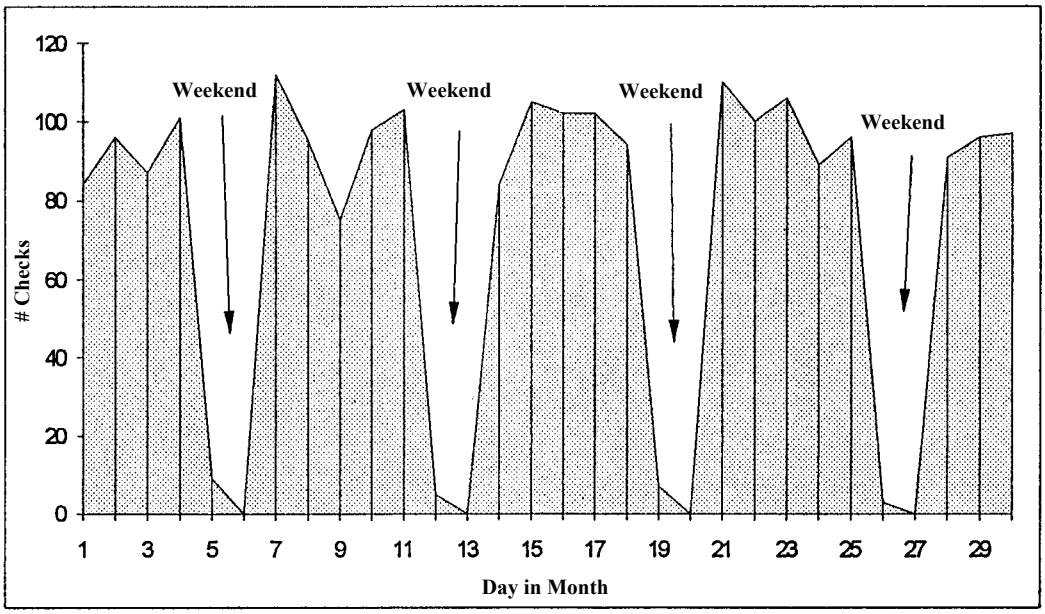
Custom Checks at an Airport

Number of Checks Made During 1993:



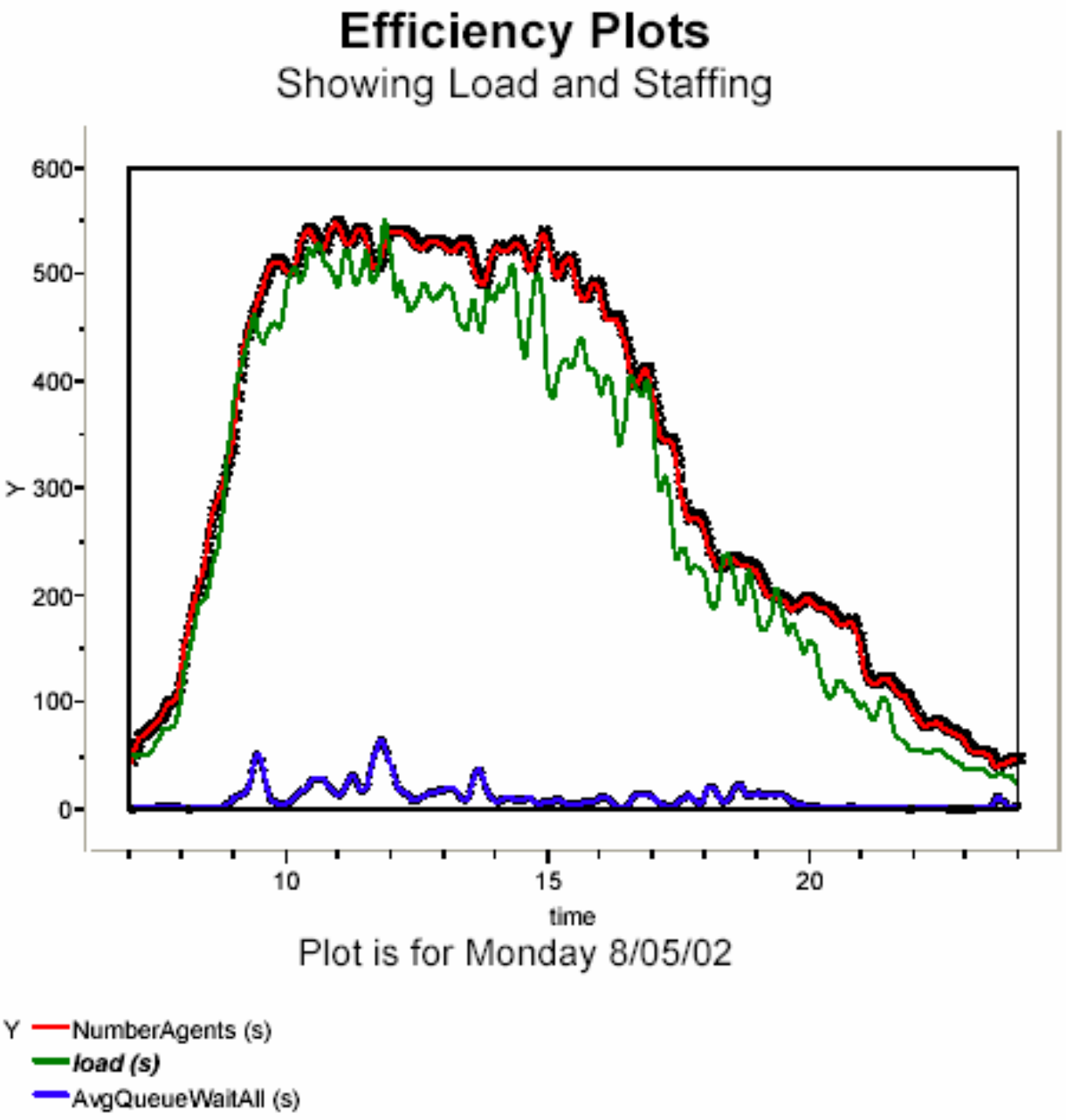
Source: Ben-Gurion Airport Custom Inspectors Division

Number of Checks Made in November 1993:



Source: Ben-Gurion Airport Custom Inspectors Division

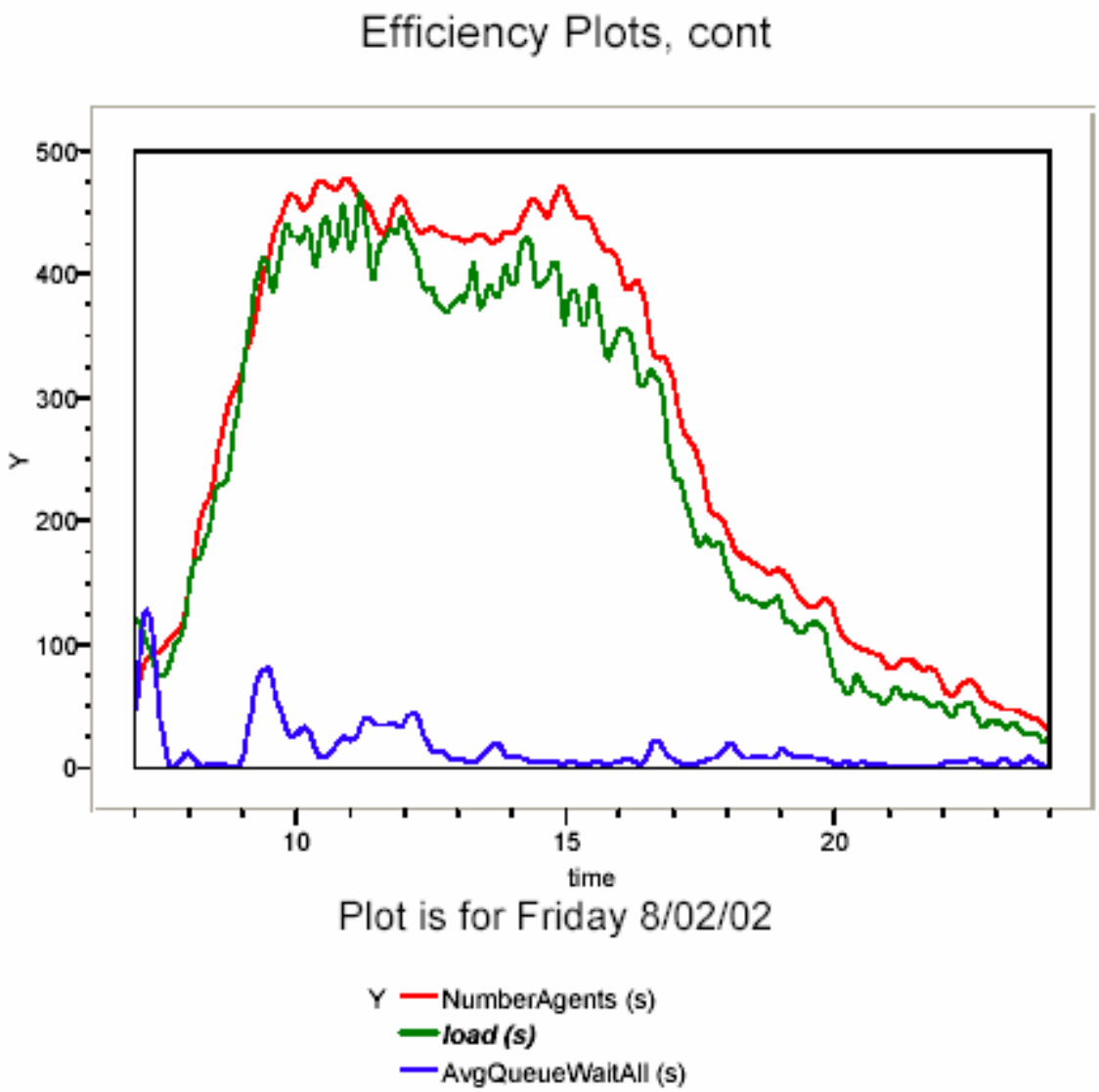
Time-Varying Workload (U.S. Bank)



"Agents" = *Estimate* of number of agents on-duty at that time.
[In each 150 second interval an agent is estimated to be on-active-duty for the entire interval if (s)he is on the phone sometime in that interval.]

In Practice: Time-Buckets (15, 30, 60 min, by volume)

Staffing Matters (on Fridays, 7:00 am)



Note increased usage from 7-7:30 am (typical of Fridays).
Note increased average Queue-Wait during this time.
(Accompanied by a rise in abandonments to about 10%.)

Overall Utilization: 8/02/02 = 88%
8/05/02 = 89%

What is a "Service-Time" (Duration) ?

(http://ie.technion.ac.il/serveng/Lectures/ServiceFull.pdf)

Does it include after-call work?

Does it include whisper-time?

Does is account for Repeat Calls (Redials) ?

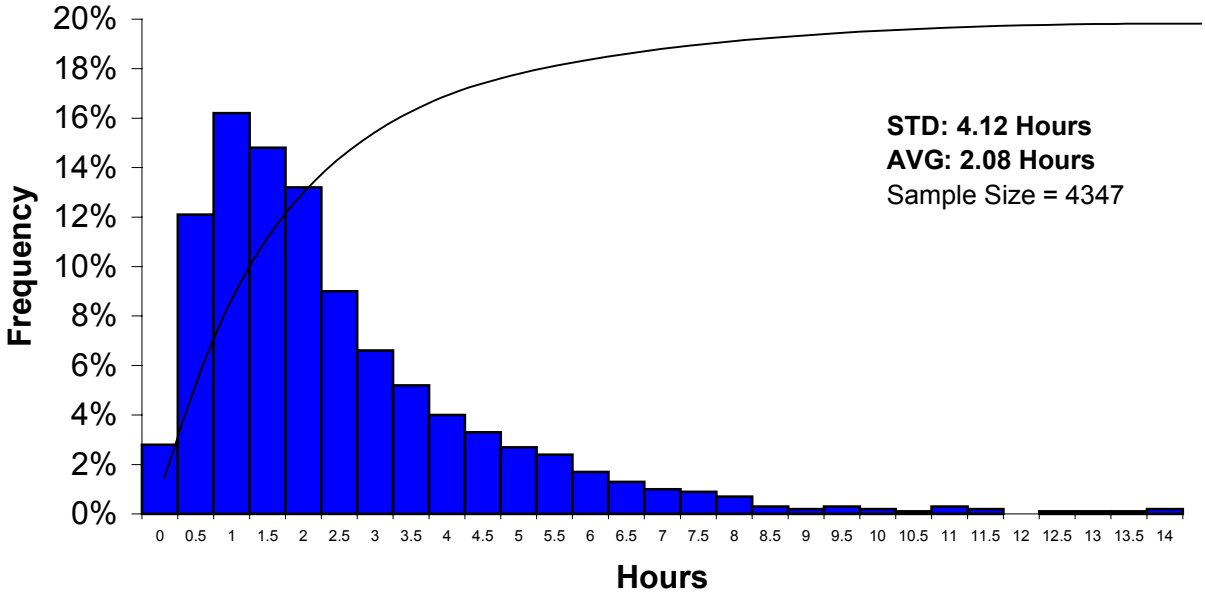
How to accurately estimate, predict; analyze, model ?

How affected by covariates:

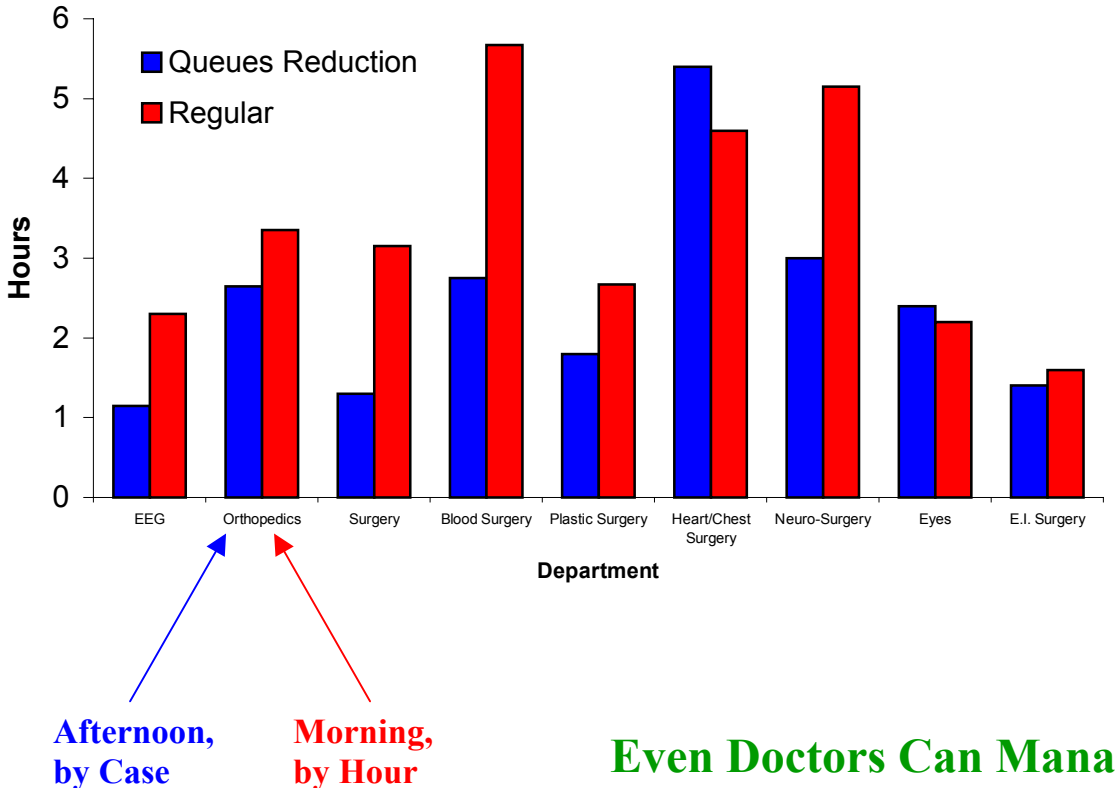
- Experience and Skill of Agents (Learning Curve)
- Type of Customer (Service Type, VIP Status)
- Time-of-Day (or Congestion-Level)
- Environment: Incentives, Workload
- …

Operations Time In a Hospital

Operations Time Histogram:



Operations Time - Morning vs. Afternoon:



Afternoon, by Case

Morning, by Hour

Even Doctors Can Manage!

Bank Classification of "Continued – Calls"

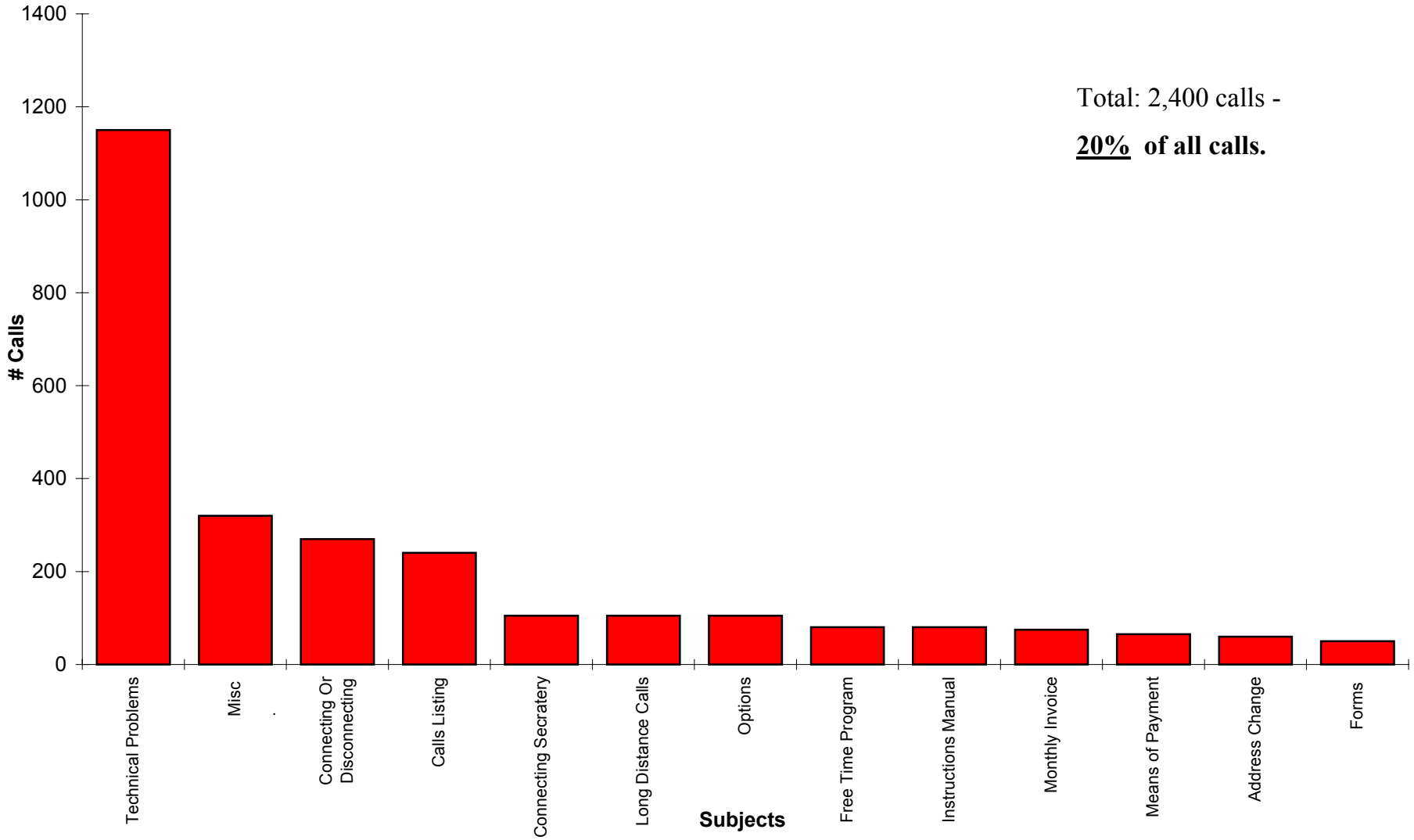
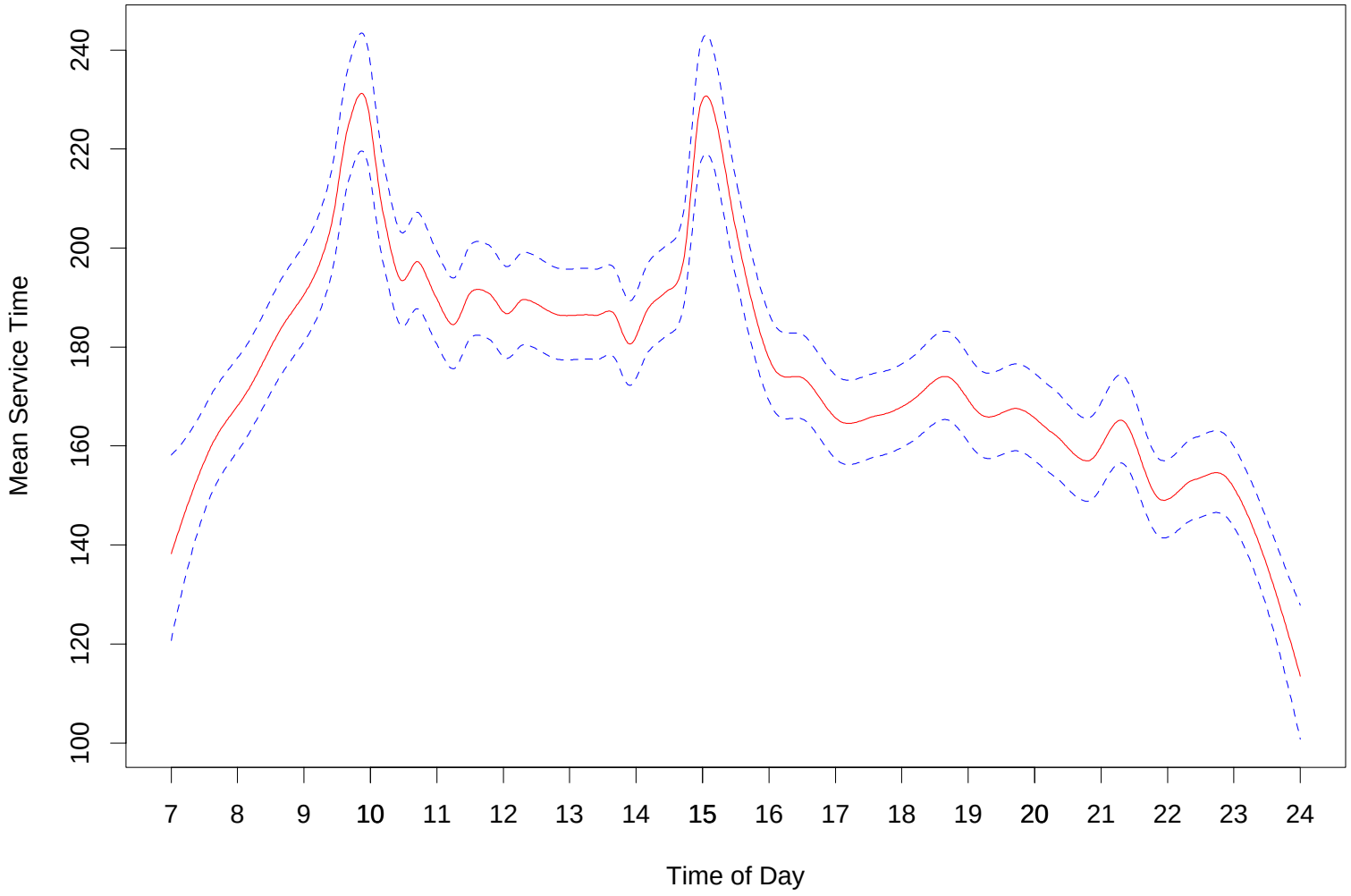
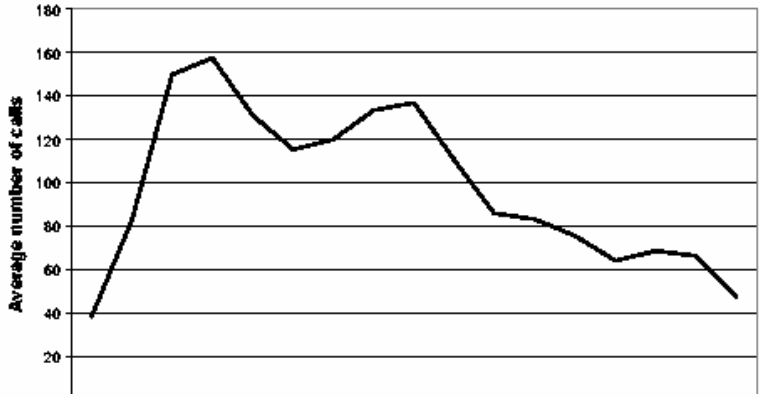


Figure 12: Mean Service Time (Regular) vs. Time-of-day (95% CI) ($n = 42613$)

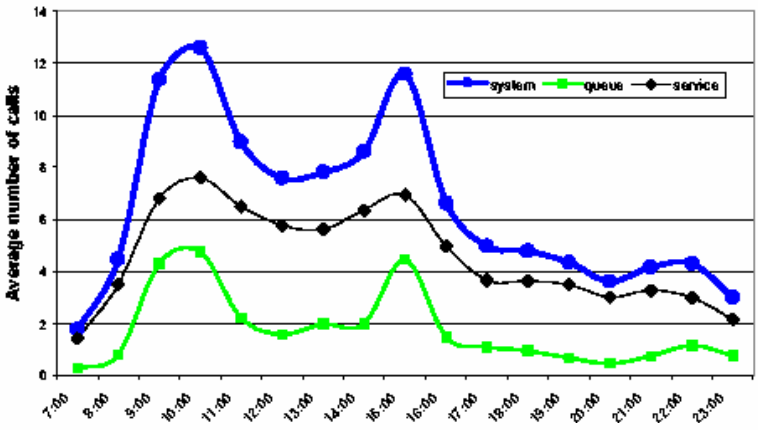


Predictable Variability

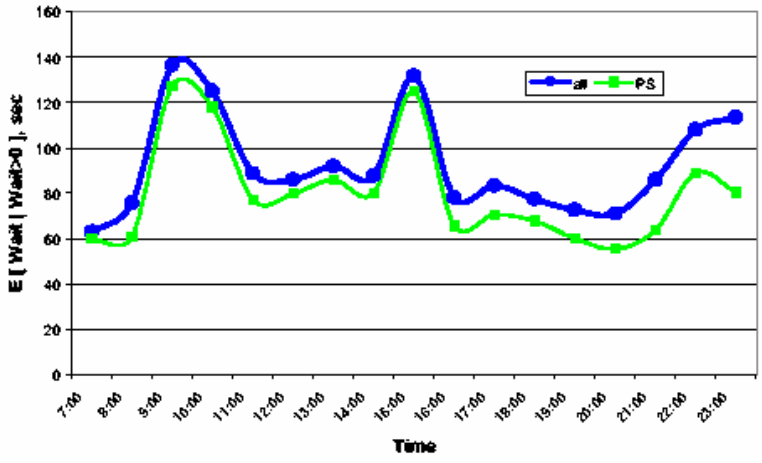
Arrivals

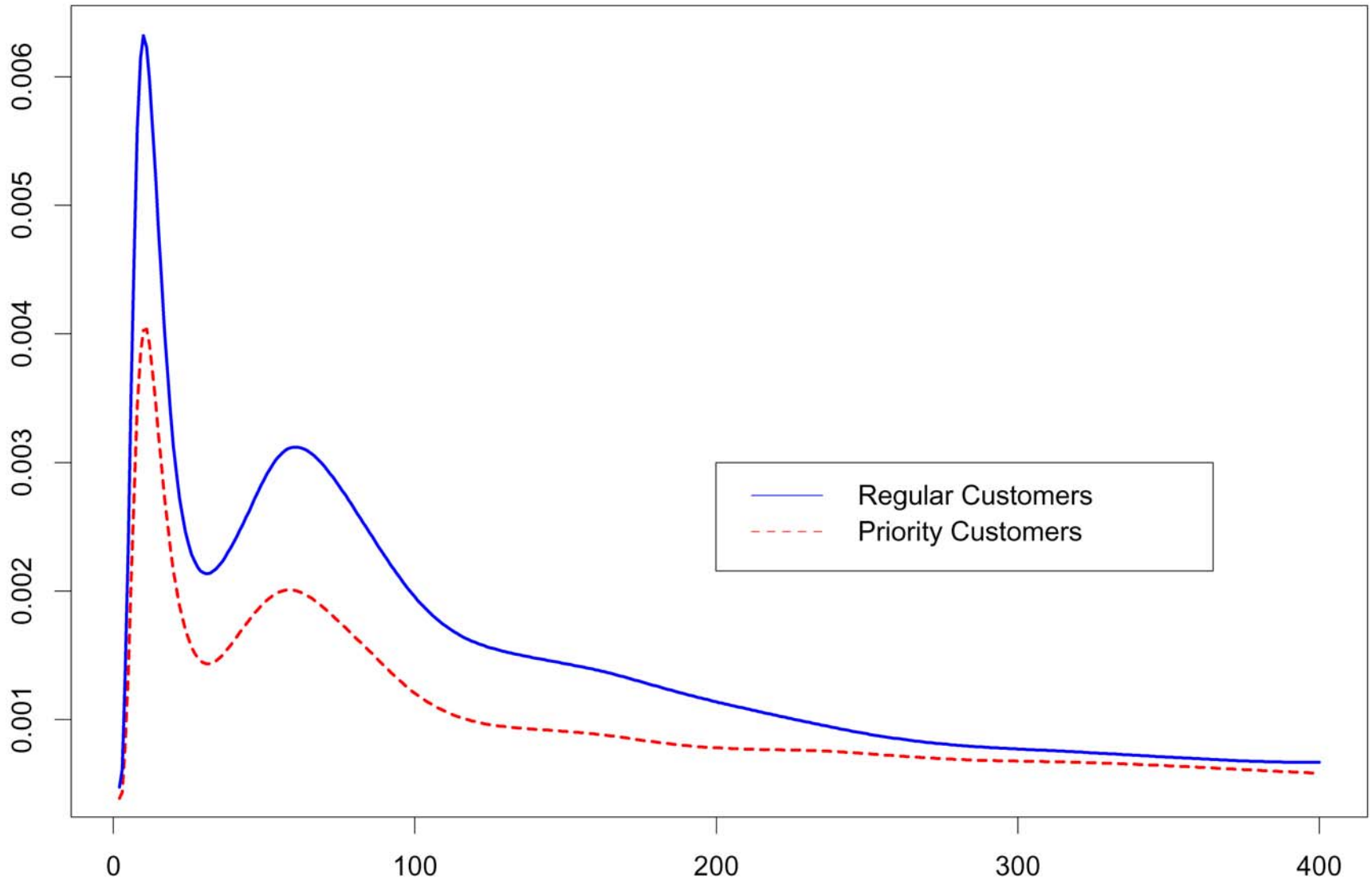


Queues



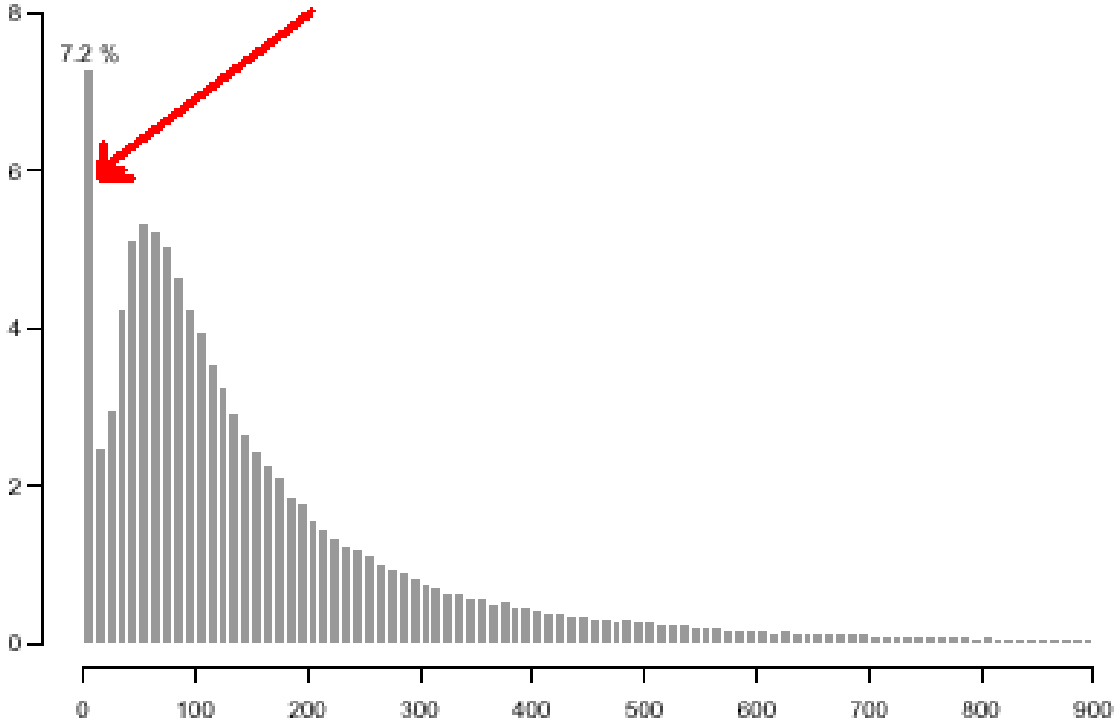
Waiting



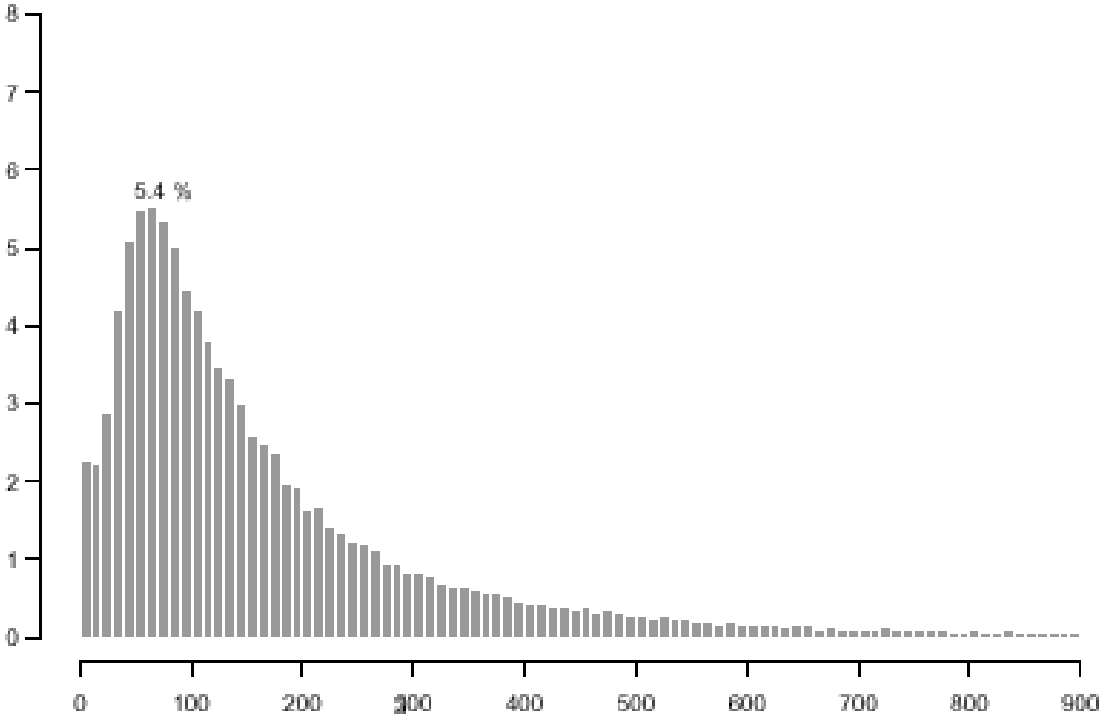

Hazard Rate: Empirical (Im)Patience
0.001
0.002
0.003
0.004
0.005
0.006
0
100
200
300
400
Regular Customers
Priority Customers

Beyond Data Averages:
Short Service Times

Jan - Oct:



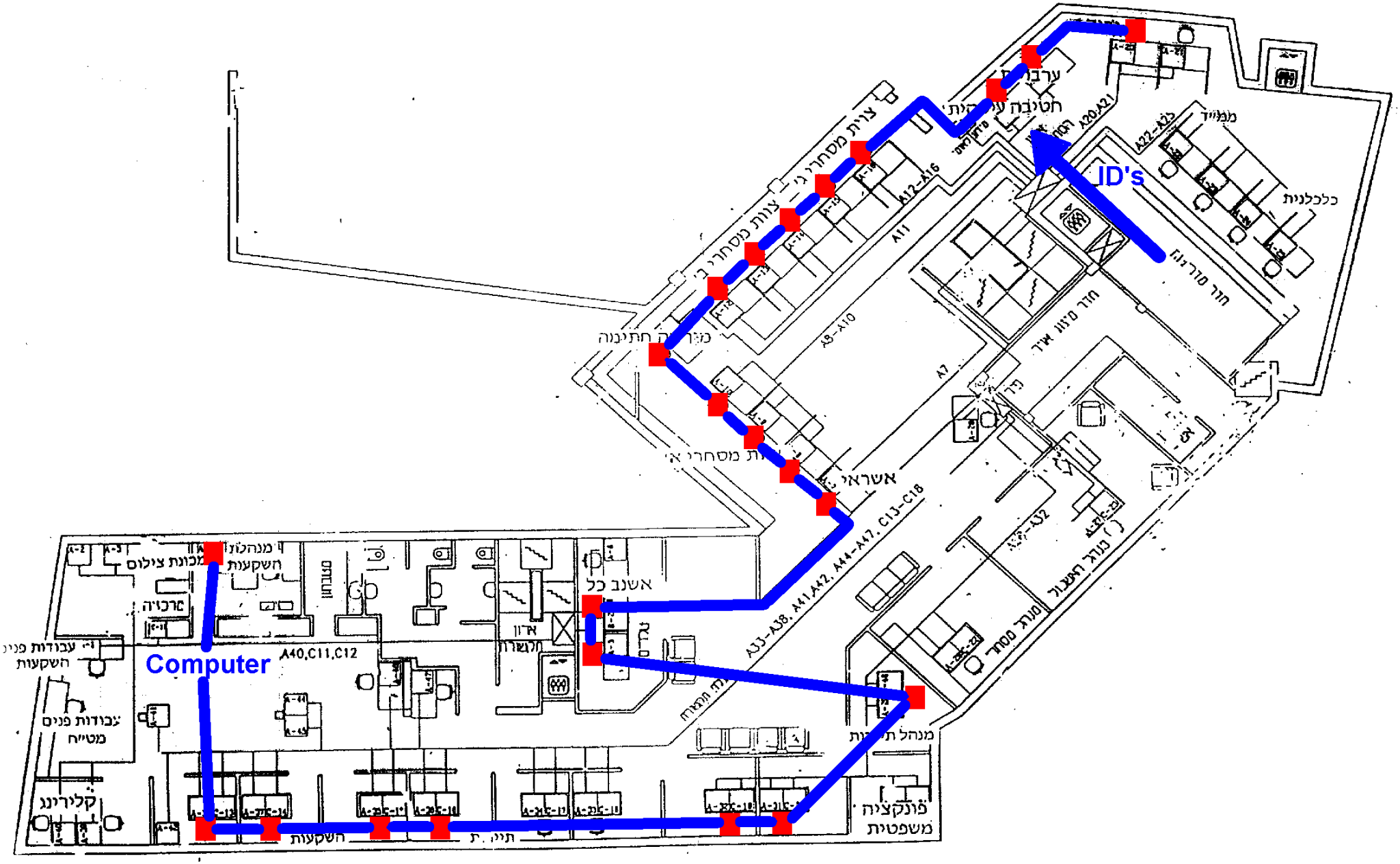
Nov - Dec:



Measurements

- Data is the Language of Nature
- Prerequisite for Science, Engineering and Management, yet
- Empirical "Axiom" = Problems with Historical Records
  - The data you need is not there for you to use:
    - Not collected or erased, contaminated, …
  - If there is data, it has 'frequencies" but no "times":
    - Fires, Courts, Hospitals, Projects, …
  - If "times", typically aggregated means but no std's:
    - Let alone histograms / distributions,
    - Typically small samples, too short time-preiods
    - Often paper-archives, not computerized
- Challenges – not Technological
  - Too little: "Complete" Data (QIE, Abandons)
  - Too much: Transaction-mgt., Big-Brother, Data Mining
- Scope
  - Face-to-face services
  - Tele-services
  - Administrative processes

Bank – 2nd Floor Measurements



Telephone Service: Call-by-Call Data

vru+line	call_id	customer_id	priority	type	date	vru_entry	vru_exit	vru_time	q_start	q_exit	q_time	outcome	ser_start	ser_exit	ser_time	server
AA0101	44749	27644400	2	PS	990901	11:45:33	11:45:39	6	11:45:39	11:46:58	79	AGENT	11:46:57	11:51:00	243	DORIT
AA0101	44750	12887816	1	PS	990905	14:49:00	14:49:06	6	14:49:06	14:53:00	234	AGENT	14:52:59	14:54:29	90	ROTH
AA0101	44967	58660291	2	PS	990905	14:58:42	14:58:48	6	14:58:48	15:02:31	223	AGENT	15:02:31	15:04:10	99	ROTH
AA0101	44968	0	0	NW	990905	15:10:17	15:10:26	9	15:10:26	15:13:19	173	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44969	63193346	2	PS	990905	15:22:07	15:22:13	6	15:22:13	15:23:21	68	AGENT	15:23:20	15:25:25	125	STEREN
AA0101	44970	0	0	NW	990905	15:31:33	15:31:47	14	00:00:00	00:00:00	0	AGENT	15:31:45	15:34:16	151	STEREN
AA0101	44971	41630443	2	PS	990905	15:37:29	15:37:34	5	15:37:34	15:38:20	46	AGENT	15:38:18	15:40:56	158	TOVA
AA0101	44972	64185333	2	PS	990905	15:44:32	15:44:37	5	15:44:37	15:47:57	200	AGENT	15:47:56	15:49:02	66	TOVA
AA0101	44973	3.06E+08	1	PS	990905	15:53:05	15:53:11	6	15:53:11	15:56:39	208	AGENT	15:56:38	15:56:47	9	MORIAH
AA0101	44974	74780917	2	NE	990905	15:59:34	15:59:40	6	15:59:40	16:02:33	173	AGENT	16:02:33	16:26:04	1411	ELI
AA0101	44975	55920755	2	PS	990905	16:07:46	16:07:51	5	16:07:51	16:08:01	10	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44976	0	0	NW	990905	16:11:38	16:11:48	10	16:11:48	16:11:50	2	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44977	33689787	2	PS	990905	16:14:27	16:14:33	6	16:14:33	16:14:54	21	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44978	23817067	2	PS	990905	16:19:11	16:19:17	6	16:19:17	16:19:39	22	AGENT	16:19:38	16:21:57	139	TOVA
AA0101	44764	0	0	PS	990901	15:03:26	15:03:36	10	00:00:00	00:00:00	0	AGENT	15:03:35	15:06:36	181	ZOHARI
AA0101	44765	25219700	2	PS	990901	15:14:46	15:14:51	5	15:14:51	15:15:10	19	AGENT	15:15:09	15:17:00	111	SHARON
AA0101	44766	0	0	PS	990901	15:25:48	15:26:00	12	00:00:00	00:00:00	0	AGENT	15:25:59	15:28:15	136	ANAT
AA0101	44767	58859752	2	PS	990901	15:34:57	15:35:03	6	15:35:03	15:35:14	11	AGENT	15:35:13	15:35:15	2	MORIAH
AA0101	44768	0	0	PS	990901	15:46:30	15:46:39	9	00:00:00	00:00:00	0	AGENT	15:46:38	15:51:51	313	ANAT
AA0101	44769	78191137	2	PS	990901	15:56:03	15:56:09	6	15:56:09	15:56:28	19	AGENT	15:56:28	15:59:02	154	MORIAH
AA0101	44770	0	0	PS	990901	16:14:31	16:14:46	15	00:00:00	00:00:00	0	AGENT	16:14:44	16:16:02	78	BENSION
AA0101	44771	0	0	PS	990901	16:38:59	16:39:12	13	00:00:00	00:00:00	0	AGENT	16:39:11	16:43:35	264	VICKY
AA0101	44772	0	0	PS	990901	16:51:40	16:51:50	10	00:00:00	00:00:00	0	AGENT	16:51:49	16:53:52	123	ANAT
AA0101	44773	0	0	PS	990901	17:02:19	17:02:28	9	00:00:00	00:00:00	0	AGENT	17:02:28	17:07:42	314	VICKY
AA0101	44774	32387482	1	PS	990901	17:18:18	17:18:24	6	17:18:24	17:19:01	37	AGENT	17:19:00	17:19:35	35	VICKY
AA0101	44775	0	0	PS	990901	17:38:53	17:39:05	12	00:00:00	00:00:00	0	AGENT	17:39:04	17:40:43	99	TOVA
AA0101	44776	0	0	PS	990901	17:52:59	17:53:09	10	00:00:00	00:00:00	0	AGENT	17:53:08	17:53:09	1	NO_SERVER
AA0101	44777	37635950	2	PS	990901	18:15:47	18:15:52	5	18:15:52	18:16:57	65	AGENT	18:16:56	18:18:48	112	ANAT
AA0101	44778	0	0	NE	990901	18:30:43	18:30:52	9	00:00:00	00:00:00	0	AGENT	18:30:51	18:30:54	3	MORIAH
AA0101	44779	0	0	PS	990901	18:51:47	18:52:02	15	00:00:00	00:00:00	0	AGENT	18:52:02	18:55:30	208	TOVA
AA0101	44780	0	0	PS	990901	19:19:04	19:19:17	13	00:00:00	00:00:00	0	AGENT	19:19:15	19:20:20	65	MEIR
AA0101	44781	0	0	PS	990901	19:39:19	19:39:30	11	00:00:00	00:00:00	0	AGENT	19:39:29	19:41:42	133	BENSION
AA0101	44782	0	0	NW	990901	20:08:13	20:08:25	12	00:00:00	00:00:00	0	AGENT	20:08:28	20:08:41	13	NO_SERVER
AA0101	44783	0	0	PS	990901	20:23:51	20:24:05	14	00:00:00	00:00:00	0	AGENT	20:24:04	20:24:33	29	BENSION
AA0101	44784	0	0	NW	990901	20:36:54	20:37:14	20	00:00:00	00:00:00	0	AGENT	20:37:13	20:38:07	54	BENSION
AA0101	44785	0	0	PS	990901	20:50:07	20:50:16	9	00:00:00	00:00:00	0	AGENT	20:50:15	20:51:32	77	BENSION
AA0101	44786	0	0	PS	990901	21:04:41	21:04:51	10	00:00:00	00:00:00	0	AGENT	21:04:50	21:05:59	69	TOVA
AA0101	44787	0	0	PS	990901	21:25:00	21:25:13	13	00:00:00	00:00:00	0	AGENT	21:25:13	21:28:03	170	AVI
AA0101	44788	0	0	PS	990901	21:50:40	21:50:54	14	00:00:00	00:00:00	0	AGENT	21:50:54	21:51:55	61	AVI
AA0101	44789	9103060	2	NE	990901	22:05:40	22:05:46	6	22:05:46	22:09:52	246	AGENT	22:09:51	22:13:41	230	AVI
AA0101	44790	14558621	2	PS	990901	22:24:11	22:24:17	6	22:24:17	22:26:16	119	AGENT	22:26:15	22:27:28	73	VICKY
AA0101	44791	0	0	PS	990901	22:46:27	22:46:37	10	00:00:00	00:00:00	0	AGENT	22:46:36	22:47:03	27	AVI
AA0101	44792	67158097	2	PS	990901	23:05:07	23:05:13	6	23:05:13	23:05:30	17	AGENT	23:05:29	23:06:49	80	VICKY
AA0101	44793	15317126	2	PS	990901	23:28:52	23:28:58	6	23:28:58	23:30:08	70	AGENT	23:30:07	23:35:03	296	DARMON
AA0101	44794	0	0	PS	990902	00:10:47	00:12:05	78	00:00:00	00:00:00	0	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44795	0	0	PS	990902	07:16:52	07:17:01	9	00:00:00	00:00:00	0	AGENT	07:17:01	07:17:44	43	ANAT
AA0101	44796	0	0	PS	990902	07:50:05	07:50:16	11	00:00:00	00:00:00	0	AGENT	07:50:16	07:53:03	167	STEREN

DATA-MOCCA

DATA MOdel for Call Center Analysis

Dr Valery Trofimov
Professor Paul D. Feigin
Professor Avishai Mandelbaum
Ms Eva Ishay

December 2003

Contents

What can be achieved

At what cost

Copy of Summary Interval - Order PK

Date: 7/7/97

Spli/Skill: Order PK

Time	Avg Speed Ans (W)	Avg Aban Time	ACD Calls (Δ)	Avg ACD Time (1/μ)	Avg ACW Time	Aban Calls (#Aban)	% ACD Time	% Ans Calls	Avg Pos Staf (N)	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time (P)
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	98	70	149		8		
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	51	7	4	51	2	16	61
12:30 AM*	:00:03	:04:10	14	:07:27	:00:33	1	89	52	5	3	48	1	26	63
1:00 AM*	:00:00		9	:04:54	:11:29	0	91	90	1	7	90	0	26	65
5:30 AM*			0			0	0		0	0		33	0	0
6:00 AM*	:00:00		12	:03:21	:00:19	0	21	100	7	2	100	9	2	19
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	29
7:00 AM*	:00:00		62	:03:34	:00:15	0	38	100	21	3	100	13	4	34
7:30 AM*	:00:00		93	:03:11	:00:34	0	36	100	30	3	100	7	4	32
8:00 AM*	:00:00		120	:03:37	:00:40	0	39	100	47	3	100	8	6	33
8:30 AM*	:00:00		193	:03:04	:00:14	0	44	100	61	3	100	10	7	37
9:00 AM*	:00:01		293	:03:25	:00:25	0	54	99	75	4	97	9	7	47
9:30 AM*	:00:02	:00:06	381	:03:45	:00:22	2	60	97	91	4	93	8	8	52
Peak → 10:00 AM*	:00:02	:00:01	416	:03:49	:00:26	1	63	97	94	4	96	5	8	55
10:30 AM*	:00:00		349	:03:35	:00:33	0	52	99	96	4	99	6	8	44
11:00 AM*	:00:00		352	:03:50	:00:27	0	51	100	102	3	100	7	8	45
11:30 AM*	:00:00		349	:03:44	:00:18	0	49	100	97	4	100	8	5	45
12:00 PM*	:00:01		354	:03:59	:00:18	0	52	95	95	4	95	8	5	47
12:30 PM*	:00:00		336	:03:38	:00:21	0	52	99	97	3	99	9	8	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	8	44
1:30 PM*	:00:00		368	:03:52	:00:14	0	56	99	99	4	99	11	7	50
2:00 PM*	:00:01		393	:03:55	:00:17	0	51	100	106	4	100	10	5	46
2:30 PM*	:00:00		403	:03:58	:00:13	0	54	100	112	4	100	10	4	50
3:00 PM*	:00:00	:00:04	410	:04:02	:00:16	1	57	98	110	4	98	8	5	51
3:30 PM*	:00:00		347	:03:59	:00:14	0	50	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:48	:01:37	0	54	100	98	4	100	8	7	47
4:30 PM*	:00:00		378	:03:41	:00:19	0	55	99	97	4	99	8	5	50
5:00 PM*	:00:00		411	:03:53	:00:19	0	53	100	109	4	100	9	5	48
5:30 PM*	:00:01		387	:03:58	:00:19	0	58	99	98	4	99	10	6	51
6:00 PM*	:00:01	:00:21	371	:03:28	:00:25	1	53	98	91	4	98	9	8	47
6:30 PM*	:00:00		280	:03:26	:00:13	0	41	100	90	3	100	8	4	37
7:00 PM*	:00:00		289	:03:24	:00:17	0	42	100	78	3	100	9	5	38

Rough Performance Analysis

Peak 10:00 – 10:30 a.m., with 100 agents

400 calls

3:45 minutes average service time

2 seconds ASA

Offered load $R = \lambda \times E(S)$

$= 400 \times 3{:}45 = 1500$ min./30 min.

= 50 Erlangs

Occupancy $\rho = R/N$

$= 50/100 = 50\%$

$\Rightarrow$ **Quality-Driven Operation** (Light-Traffic)

$\Rightarrow$ Classical Queueing Theory (M/G/N approximations)

Above: $R = 50$, $N = R + 50$, $\approx$ all served immediately.

Rule of Thumb: $N = \lceil R + \delta R \rceil$, $\delta > 0$ service-grade.

Quality-driven: 100 agents, 50% utilization

⇒ **Can** increase offered load - but **by how much?**

Erlang-C **N=100** **E(S) = 3:45 min.**

λ/hr	ρ	$E(W_q)$ = ASA	% Wait = 0
800	50%	0	100%
1000	62.5%	0	100%
1200	75%	0	99.7%
1400	87.5%	0:02 min.	88%
1500	93.8%	0:15 min.	60%
1550	96.9%	0:48 min.	35%
1580	98.8%	2:34 min.	15%
1585	**99.1%**	**3:34 min.**	12%

⇒ **Efficiency-driven Operation** **(Heavy Traffic)**

Above: R = 99, N = R + 1, ≈ all delayed.

Rule of Thumb: $N = \lceil R + \gamma \rceil$, $\gamma > 0$ **service grade**.

Changing N (Staffing) in Erlang-C

E(S) = 3:45

λ/hr	N	OCC	ASA	% Wait = 0
1585	100	99.1%	3:34	12%
1599	**100**	99.9%	**59:33**	0%
1599	**100+1**	98.9%	**3:06**	13%
1599	102	98.0%	1:24	24%
1599	105	**95.2%**	0:23	**50%**

⇒ **New Rationalized Operation**

Efficiently driven, in the sense that OCC > 95%;

Quality-Driven, 50% answered **immediately**

QED Regime = Quality- and Efficiency-Driven Regime

Economies of Scale in a Frictionless Environment

Above: R = 100, N = R + 5, 50% delayed.

√· **Safety-Staffing** $N = \lceil R + \beta\sqrt{R} \rceil$, $\beta > 0$.

√· Safety-Staffing: Performance

$R = \lambda \times E(S)$ Offered load (Erlangs)

$N = R + \underbrace{\beta\sqrt{R}}$ β = "service-grade" > 0

$= R + \Delta$ √· safety-staffing

Expected Performance:

$$\%\ \text{Delayed} \approx P(\beta) = \left[1 + \frac{\beta\phi(\beta)}{\varphi(\beta)}\right]^{-1}, \quad \beta > 0$$ Erlang-C

$$\text{Congestion index} = E\left[\frac{\text{Wait}}{E(S)} \,\middle|\, \text{Wait} > 0\right] = \frac{1}{\Delta}$$ ASA

$$\%\left\{\frac{\text{Wait}}{E(S)} > T \,\middle|\, \text{Wait} > 0\right\} = e^{-T\Delta}$$ TSF

$$\text{Servers' Utilization} = \frac{R}{N} \approx 1 - \frac{\beta}{\sqrt{N}}$$ Occupancy

Rules-of-Thumb in an "Erlang-C World"

R = Offered Load (not small)

Efficiency-Driven: $N = R + 2$ (or 3, or…);

Expect that essentially **all** customers are delayed in queue, that average delay is about 1/2 (or 1/3, or…) average service-time, and that agents utilization is extremely high (close to 100%).

Quality-Driven: $N = R + (10\% - 20\%)\ R$

Expect essentially **no** delays of customers.

QED: $\mathbf{N = R + 0.5\sqrt{R}}$

Expect that about **half** of the customers are not delayed in queue, that average delay is about one-order less than average service-time (seconds vs. minutes), and that agents utilization is high (90-95%).

Can determine regime scientifically:

Strategy: Retain performance levels under Pooling (4CC demo)

Economics: Minimize agent salaries + congestion cost, or

Satisfization: Least Number of Agents s.t. Constraints

Strategy: Sustain Regime under Pooling

Base: $\lambda = 300$/hr, AHT = 5 min, N = 30 agents

OCC = 83.3%, ASA = 15 sec, %{Wait>0} = 25%

4 CC: $\lambda = 1200$, AHT = 5 min, N=?

Quality-Driven: maintain OCC at 83.3%.

N = 120, ASA = .5 sec

Efficiency-Driven: maintain ASA at 15 sec.

N = 107, OCC = 93.5%

QED: maintain %{Wait>0}) at 25%

$N = 100 + 1 \cdot \sqrt{100} = 110$, OCC = 91%, ASA = 7 sec

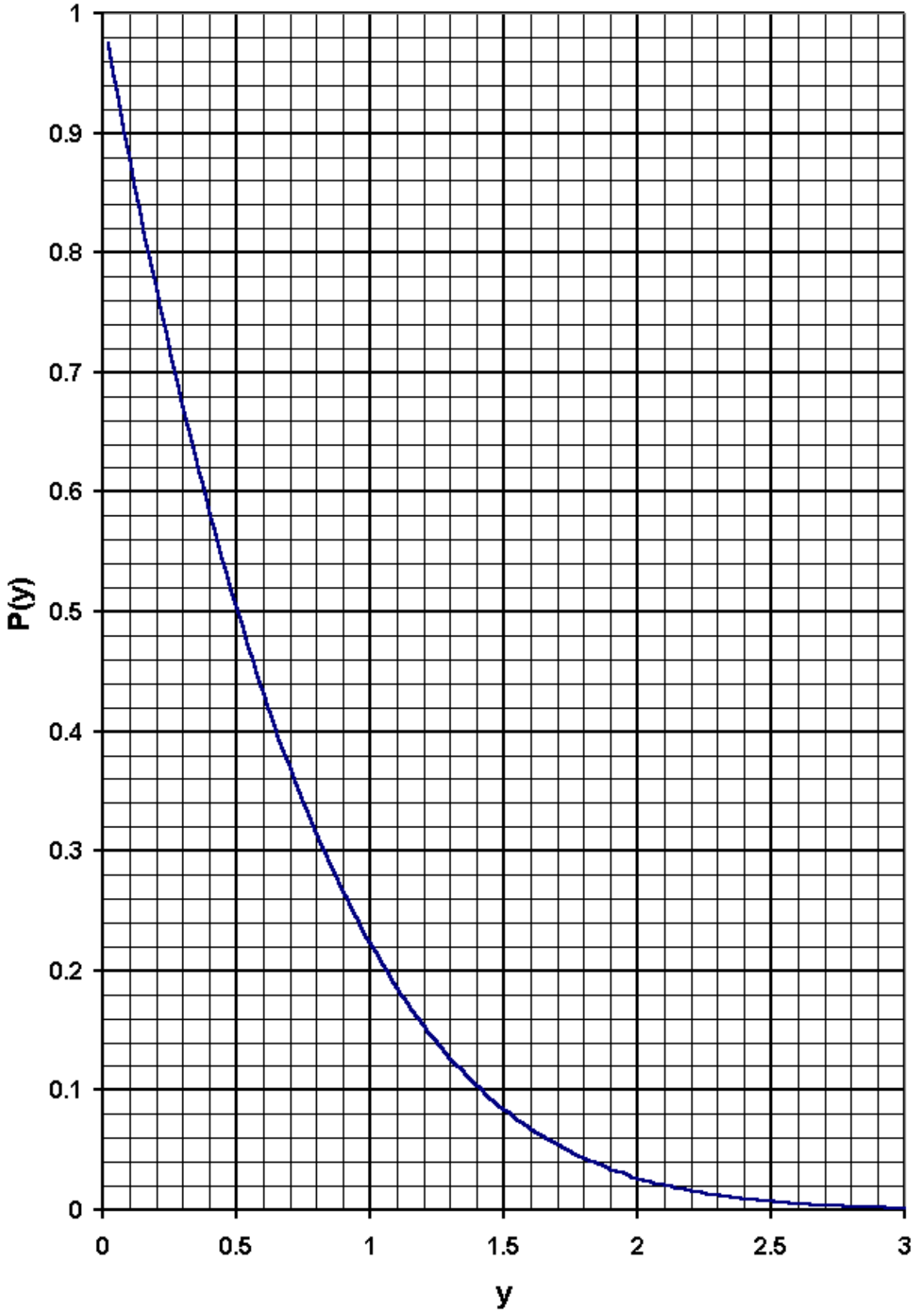
9 CC: $\lambda = 2700$, AHT = 5, R = 225

Q: N = 270

E: N = 233

QED: $N = 225\ 1 \cdot \sqrt{225} = 240$, OCC = 94%, ASA = 4.7 sec

The Halfin -Whitt Delay Function $P(\beta)$



Economics: √· Safety-Staffing:

Optimal $$\mathbf{N^* \approx R + y^*\left(\frac{d}{c}\right)\sqrt{R}}$$

where **d** = delay/waiting costs

c = staffing costs

in which $y^*(\mathbf{r})$ easily calculable (see next page)

Robust: covers **also** efficiency- and quality-driven

Accurate: to within **1 agent** (from few to many 100's) typically

Relevant: Medium-Large well-run CC do perform as above.

Instructive: In medium-large call centers, high resource utilization and service levels could **coexist**, which is enabled by **economies of scale** that dominate stochastic variability. BUT --

Example: 100 calls per minute, at 4 min. per call

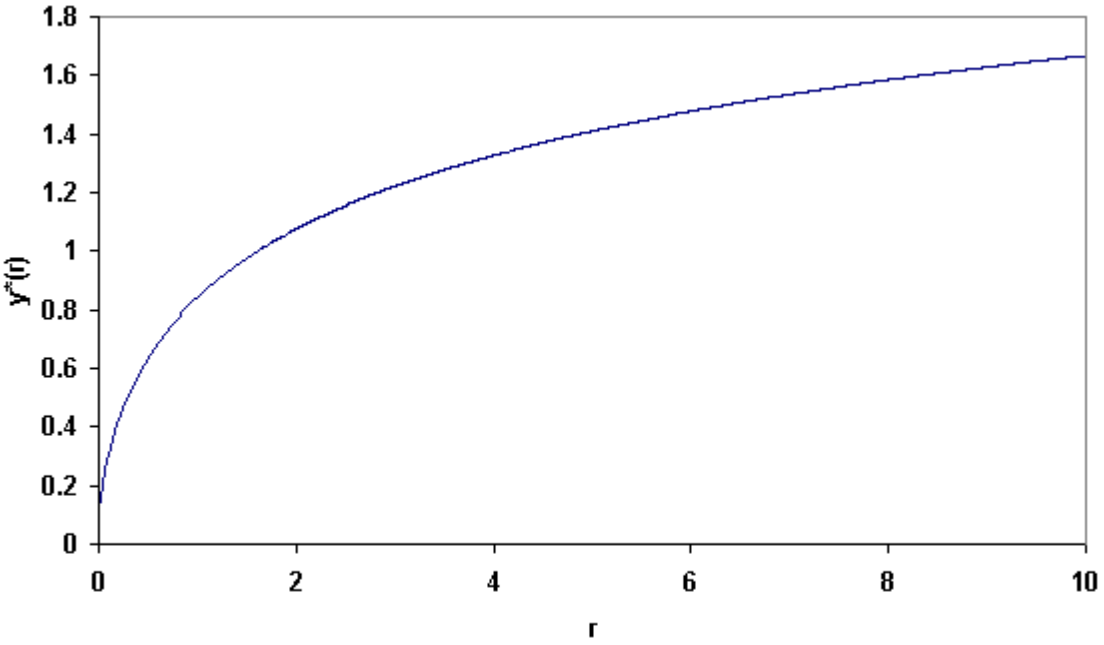
$\Rightarrow$ R = 400, least number of agents

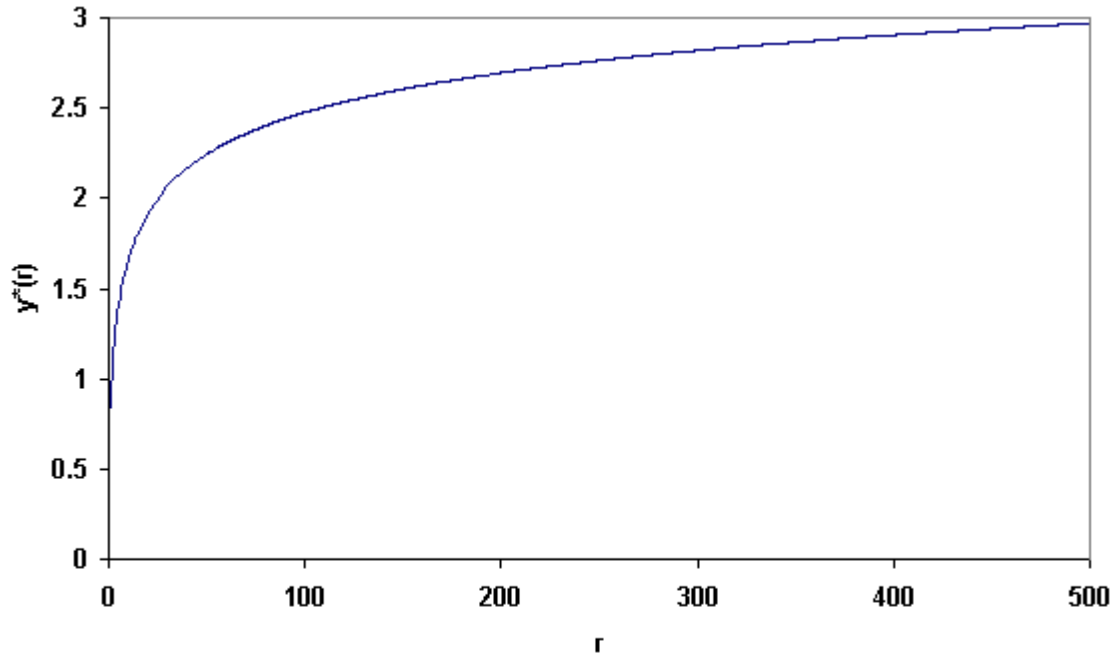
$$\frac{\Delta}{R} \approx \frac{y^*(r)}{\sqrt{R}} = \frac{y^*}{20}, \quad \text{with } y^*: 0.5\text{–}1.5 \ ;$$

Safety staffing: 2.5%–7.5% of R=Min ! $\Rightarrow$ **"Real" Challenge?**

Square-Root Safety Staffing: $N = R + y^*(r)\sqrt{R}$

r = cost of delay / cost of staffing





√· Safety-Staffing: Overview

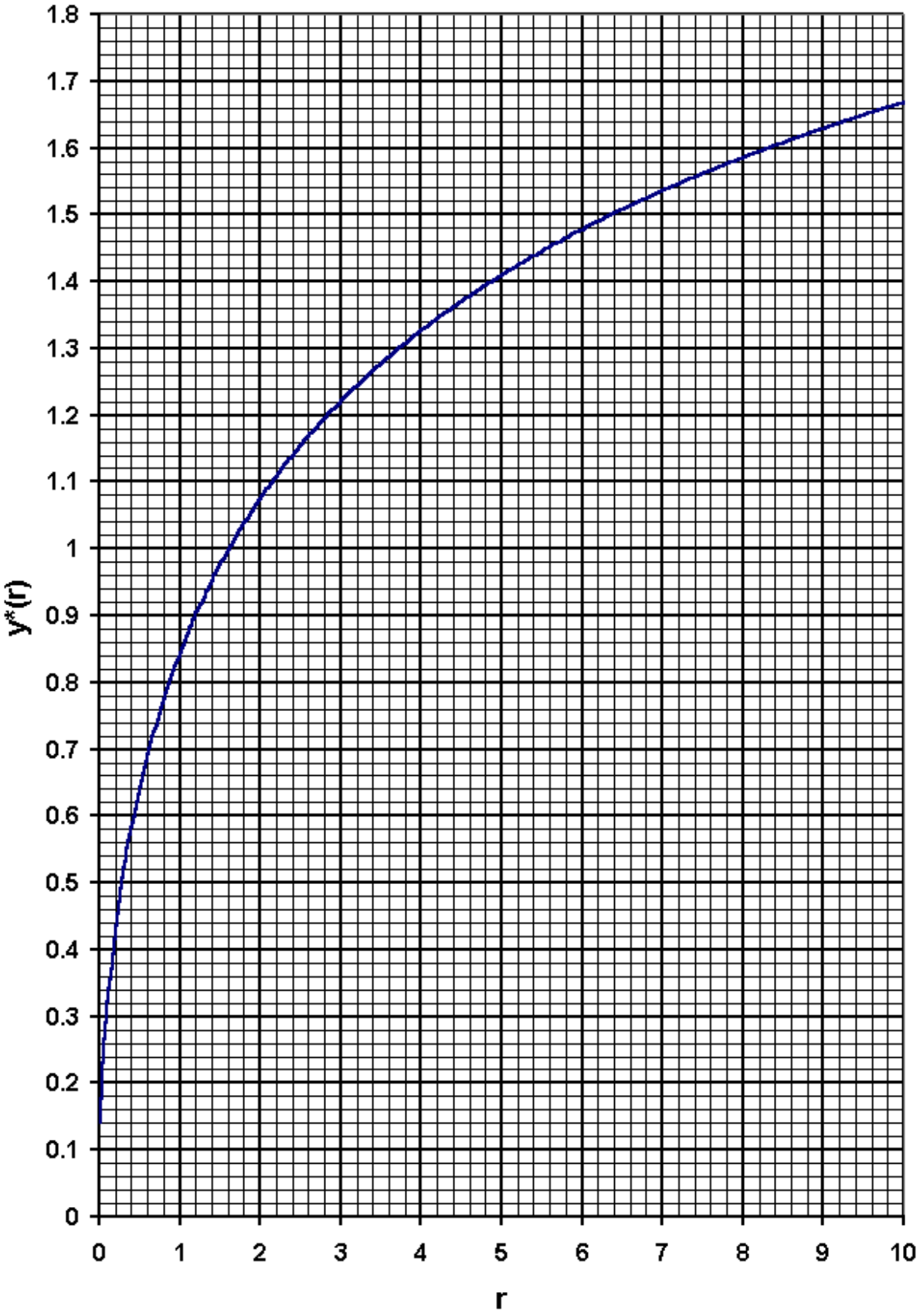
- Parameters

Average service time	M seconds	
Arrival rate	λ calls per hour	
$\Rightarrow$ Offered load	$\mathbf{R} = \lambda \times M / 3600$	hours work per hour
Delay costs	d	per customer-min
Staffing costs	c	per agent-hour
$\Rightarrow \frac{\text{Waiting}}{\text{Staffing}}$	$\mathbf{r} = \frac{d}{c} \times 60$	

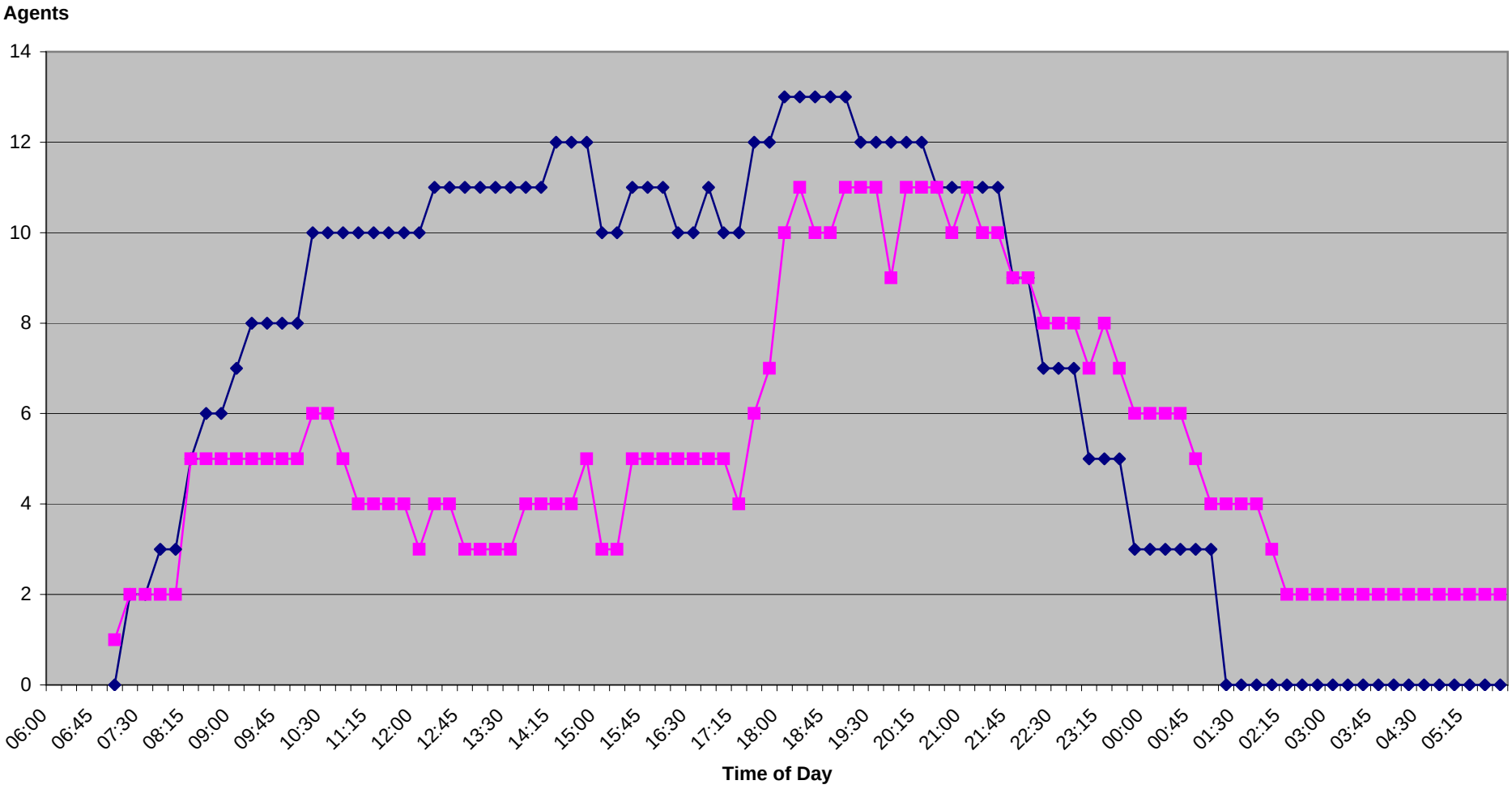
- **Optimal** $\mathbf{N^* \approx R + \underbrace{y^*(r)\sqrt{R}}}$

Δ : **safety-staffing**

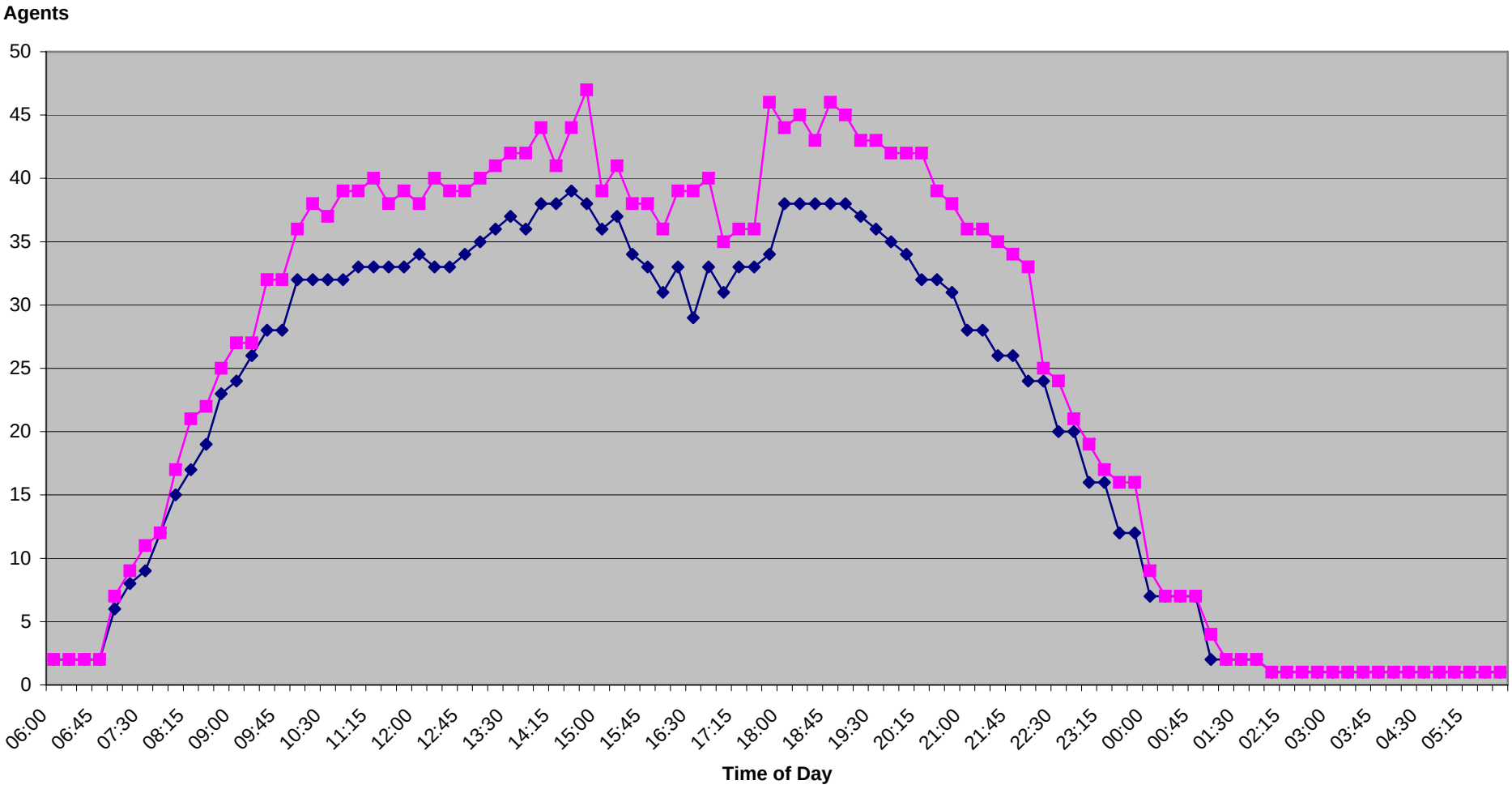
$y^*(r)$, **r = cost of delay / cost of staffing**



Staffing: "Actual" vs. "Planned" (where is "Needed"?)



Staffing: "Actual" vs. "Planned" (where is "Needed"?)



Scenario Analysis: 80:20 Rule (Large Call Center)

Prevalent std: at least 80% customers wait less than 20 sec.

Formally: %(Wait > 20 sec.) < 0.2

- Base Case: $\lambda = 100$ calls per min (avg)
  M = 4 min. service time (avg)
  R = 400 Erlangs offered load (large)

$y^*(\frac{d}{c}) = 0.53$, by $\%\{\text{Wait} > 20 \text{ sec.}\} = P(y^*)\, e^{-1.67y^*} = 0.2$

Hence: $N^* = 400 + 0.53\sqrt{400} = 411$, by $\sqrt{\cdot}$ safety-staffing

And $\frac{d}{c} = (y^*)^{-1}(0.53) = 0.32$, by inverting y^*

Low valuation of customers' time, at $\frac{1}{3}$ of servers' time, yet reasonable 80:20 performance? enabled by scale!

- What if $\frac{d}{c} = 5$?

N* = 429 agents (vs. 411 before)

Agents' accessibility (idelness) = 7% (vs. 3% before)

Hence, 1 out of 100 waits over 20 sec. (vs. 1 out of 5)

Scenario Analysis: A "Best-Practice" Call Center

- 15,000 callers per day, with 1,800 calls at peak hour (avg);
- 4 min. service time (avg);
- Significant service variability: 5% served over 12 min. (avg);
- 90% servers' utilization (avg).
- No "busy" signals, mere seconds waits, no abandonment.

Peak hour analysis:

$R = \lambda \times M = 1800 \times 4/60 = 120$ Erlangs offered-load

$N = R/\rho = 120/0.9 = 133.3$ agents

$\Delta = N - R = 13.3$ safety staffing

$y^*(d/c) = \Delta / \sqrt{R} = 13.3 / \sqrt{120} = \mathbf{1.22}$

$\frac{d}{c} = (y^*)^{-1}(1.22) \approx \mathbf{3}$, **service index**

1 hr of customers' wait is valued at 3 times hr wage of agents

Performance (via Erlang-C):

- %(Wait > 0) = P(1.22)	= 15%	delayed
- %(Wait > 20 sec)	= 5%	delayed over 20 sec.
- ASA = E[Wait]	= **2.7** sec	average wait
- ASA | Wait > 0	= **18** sec	average wait of delayed

Scenario Analysis: 80:20 Rule (Large Call Center)

Prevalent std: at least 80% customers wait less than 20 sec.

Formally: %(Wait > 20 sec.) < 0.2

- Base Case: $\lambda = 100$ calls per min (avg)
  M = 4 min. service time (avg)
  R = 400 Erlangs offered load (large)

$y^*(\frac{d}{c}) = 0.53$, by $\%\{\text{Wait} > 20 \text{ sec.}\} = P(y^*)\, e^{-1.67y^*} = 0.2$

Hence: $N^* = 400 + 0.53\sqrt{400} = 411$, by $\sqrt{\cdot}$ safety-staffing

And $\frac{d}{c} = (y^*)^{-1}(0.53) = 0.32$, by inverting y^*

Low valuation of customers' time, at $\frac{1}{3}$ of servers' time, yet reasonable 80:20 performance? enabled by scale!

- What if $\frac{d}{c} = 5$?

N* = 429 agents (vs. 411 before)

Agents' accessibility (idelness) = 7% (vs. 3% before)

Hence, 1 out of 100 waits over 20 sec. (vs. 1 out of 5)

Scenario Analysis: on Economies of Scale

"Best Practice" call center

λ = 1,800 calls per peak hour; **M = 4 min.**

$$R = 1800 \times \frac{4}{60} = 120 \quad \text{Erlangs offered-load}$$

- Base Case: How many agents are required so that, on avg, only 1 out of 100 wait more than 20 sec.? $N^* = 140$

$\Delta = 140 - 120 = 20$ (safety staffing)

$$y^*\left(\frac{d}{c}\right) = \frac{\Delta}{\sqrt{R}} = \frac{20}{\sqrt{120}} = 1.75$$

$\frac{d}{c} = 12.5$, namely customers' wait is highly valued.

- What if M = **30 sec.** (as in 411 services), $N^* = 126$ suffices for the above performance, which implies

$$\Rightarrow \; y^*\left(\frac{d}{c}\right) \approx \frac{6}{\sqrt{120}} = 0.53, \quad \text{or} \quad \frac{d}{c} = 0.32.$$

This equals the performance of a **large** call center (R = 400), **but** with E(S) = 4 min. (vs. only 30 sec. here).

Scenario Analysis: "Satisfization" vs. Optimization

Theory: The least N that guarantees $\%\{\text{Wait} > 0\} < \varepsilon$ is close to $N^* = R + P^{-1}(\varepsilon)\sqrt{R}$ (again $\sqrt{\cdot}$ safety-staffing).

(Folklore: $N^* = R + \bar{\phi}^{-1}(\varepsilon)\sqrt{R}$, $\bar{\phi} = 1 - \phi$; based on "classical" normal approximations to infinite-servers models. The two essentially coincide for small ε.)

Example: λ = 1,800 calls at peak hour (avg)

M = 4 min. service time (avg)

$R = 1800 \times \frac{4}{60} = 120$ Erlangs offered-load

Service level constraint: less than 15% delayed, equivalently at least 85% answered immediately.

$\Rightarrow N^* = R + P^{-1}(0.15)\sqrt{R} = 120 + 1.22\sqrt{120} = 133$ agents

$\Rightarrow$ $\%\{\text{Wait} > 20 \text{ sec.}\} = 5\%$ delayed over 20 sec.

ASA = E[Wait] = **2.7** sec. average wait

ASA | Wait > 0 = **18** sec. average wait of delayed

Scenario Analysis: "Reasonable" Service Level ?

Theory: The least N that guarantees $\%\{\text{Wait} > 0\} < \varepsilon$ is close to $N^* = R + P^{-1}(\varepsilon)\sqrt{R}$ (again $\sqrt{\cdot}$ safety-staffing).

Example: $\lambda = 1{,}800$ calls at peak hour (avg)

$M = 4$ min. service time (avg)

$R = 1800 \times \frac{4}{60} = 120$ Erlangs offered-load

Service level constraint: 1 out of 100 delayed (avg), namely 99% answered immediately.

$\Rightarrow N^* = R + P^{-1}(0.01)\sqrt{R} = 120 + 2.38\sqrt{120} = 146$ agents

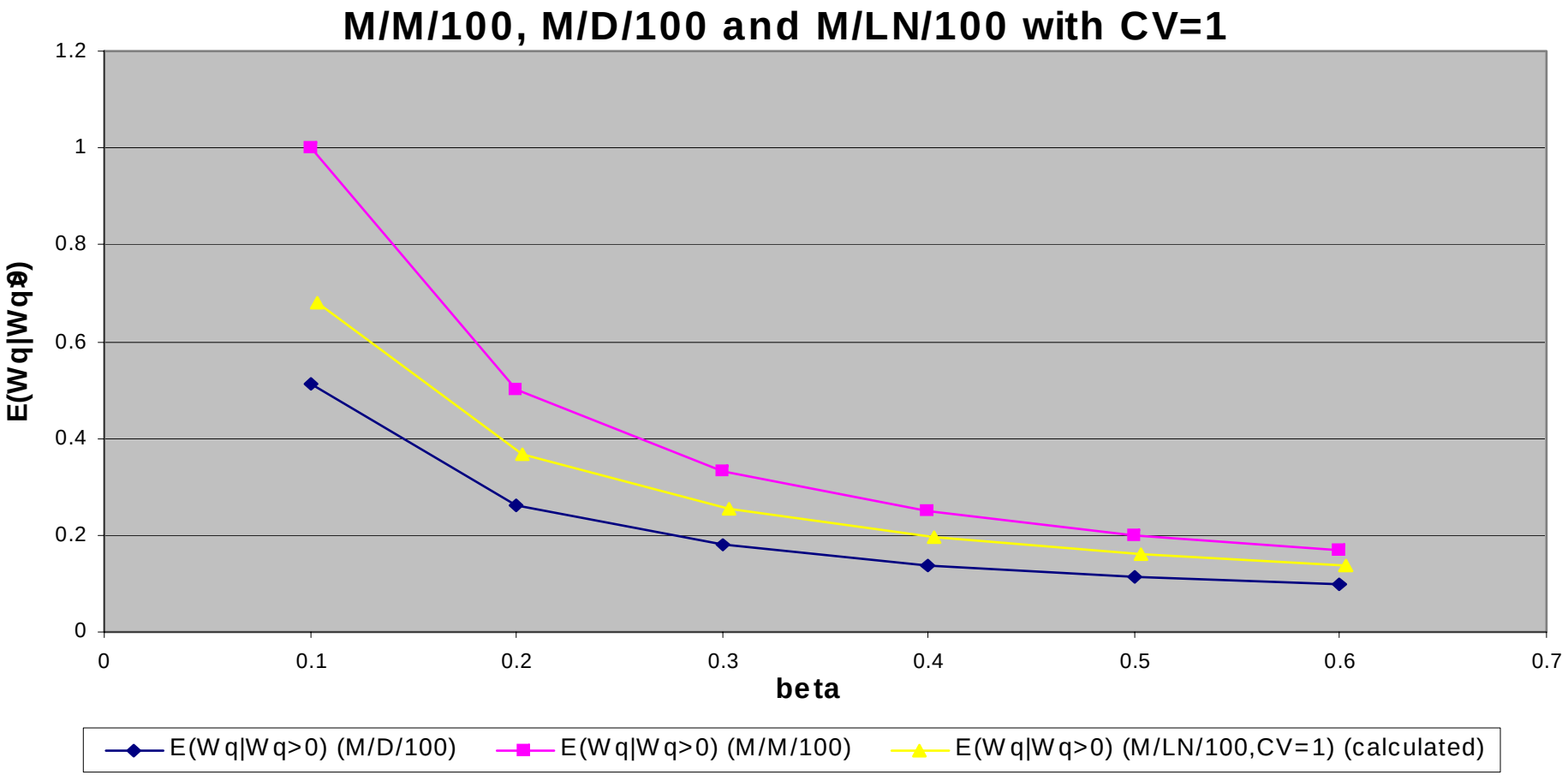
$\Rightarrow \frac{d}{c} = (y^*)^{-1}(2.38) = 75$: very high service index

Valuation of customers' time as being worth 75-fold of agents' time seems reasonable only in extreme circumstances:

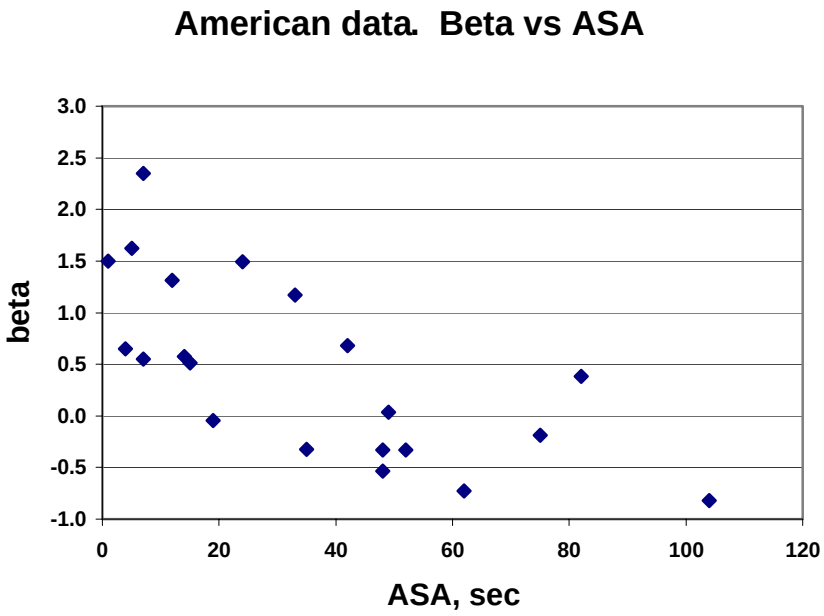
- Cheap servers (IVR)
- Costly delays (Emergency)

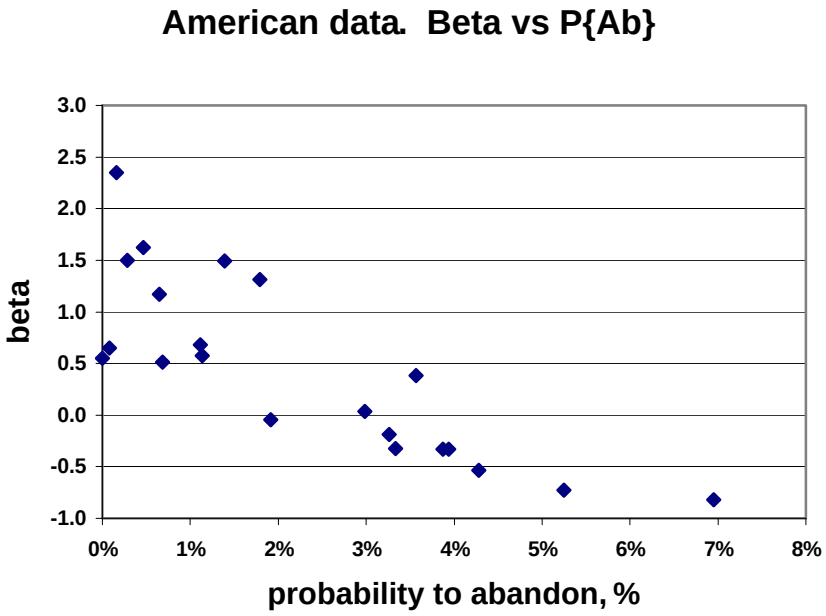
Note: **Satisfization easier to model but Costs easier to grasp**

E(Wq|Wq>0) vs. β


M/M/100, M/D/100 and M/LN/100 with CV=1
E(Wq|Wq>0)
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
beta
E(Wq|Wq>0) (M/D/100)
E(Wq|Wq>0) (M/M/100)
E(Wq|Wq>0) (M/LN/100,CV=1) (calculated)

Empirical Service Grade (Beta)





QED Relevance: One of 12 Call Centers

6/13/00 - Tue

Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
Total	**20,577**	**19,860**	**~3.0%**	**30**	**307**	**95.1%**	**85.4%**	**222.7**	**234.6**	**95.0%**
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	866	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	21	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1.286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	**1,364**	**1,338**	**1.9%**	**33**	**296**	**99.0%**	84.1%	**222.5**	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	**1,174**	**1,160**	**1.2%**	**10**	**302**	**95.5%**	93.6%	**203.8**	209.8	97.1%
13:00	**1,018**	**999**	**1.9%**	**9**	**314**	**95.4%**	91.2%	**182.9**	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	86.0%	**206.1**	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	89.1%	5.8	1.4	416.2%

Call Centers with Impatient Customers

2. Supporting Material (Downloadable)

Gans, Koole, and M.: “Telephone Call Centers: Tutorial, Review and Research Prospects.” *MSOM*.

Garnett, M. and Reiman: "Designing a Call Center with Impatient Customers." *MSOM*.

M. and Zeltyn: "The Impact of Customer Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/N+G Queue." Submitted.

Zeltyn: Ph.D. thesis, in progress.

Palm: "Methods of judging the annoyance caused by congestion.
Written earlier (1943?, in German),
and Translated in 1953 (English).
Still excellent reading.

Abandonment Important Practically

One of 2 Customer-Centric performance measures (2^{nd} = Redials)

- Lost business (now)
- Poor service level (future losses)
- 1-800 costs decrease (out-of-pocket vs. alternative)
- Self-selection: the "fittest" survive and wait less
- Must account for (carefully) in models and measures
  - Otherwise wrong picture of reality
  - Misleading performance measures, hence staffing
  - Unstable models (vs. Robustness)

But Abandonment also Interesting Theoretically

- Queueing Science
  (Paradigm: experiment, measure, model, validate)
- Research: OR + Psychology + Marketing
  (Modelling: steady-state, transient, equilibrium)
- Wide Scope of Applications: in addition to Phone,
  - VRU/IVR: opt-out-rates
  - Internet: business-drivers (60% and more)

Operational Aspects of Impatience

With less agents, the "fittest" survive and wait less – much less!

Recall earlier Q, E and QED Scenarios (E(S) = 3:45):

λ/hr	N	OCC	ASA	% Wait ≤ 2 sec
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
		BUT	with	Impatience
				%Abandonment
1600	100	97.3%	0:23	2.7 %
1600	95	98.4%	0:23	6.5%
1800	105	97.7%	0:23	3.4%

QED with Impatient Customers: Theory in the Making, but

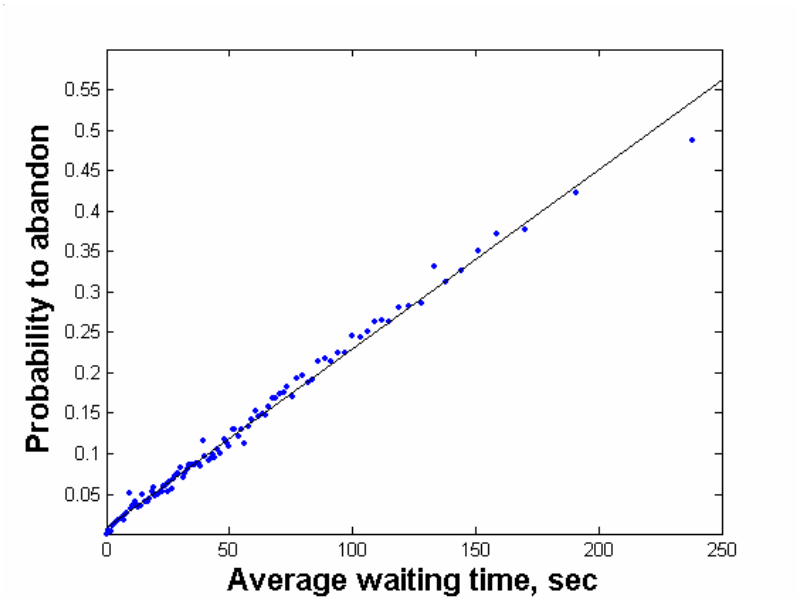
Erlang-A: in **4CallCenters**

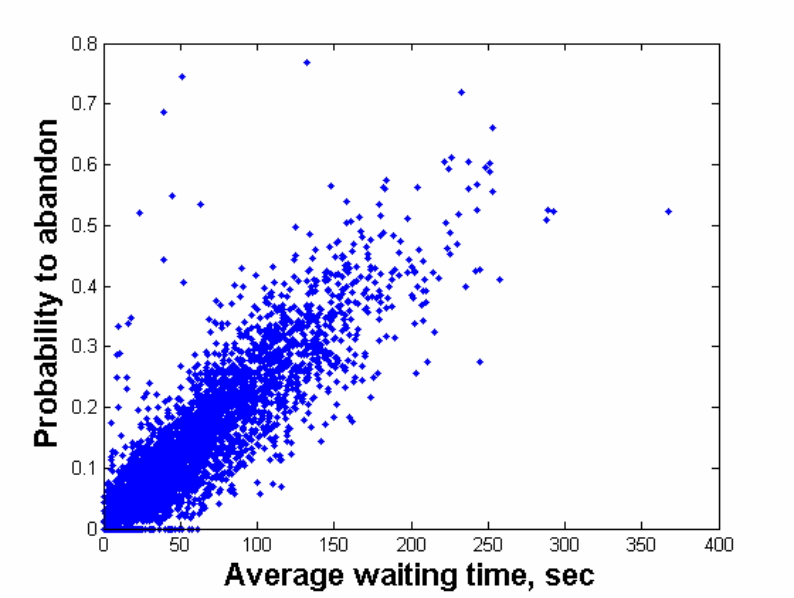
Downloadable from website and Easily Usable

Empirically-Based Theory

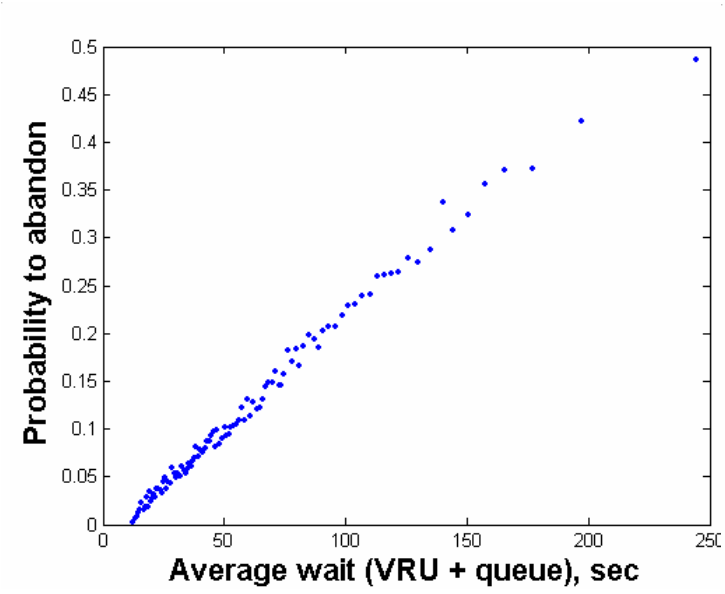
Linear pattern observed: $P\{\text{Abandon}\} = C \bullet E[\text{Wait}]$

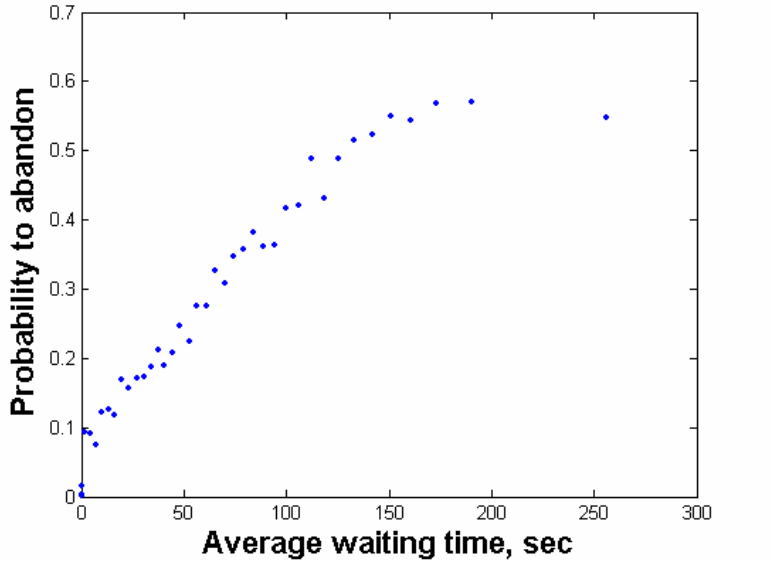
Theory: Average Patience = 1/C in Erlang-A, else?





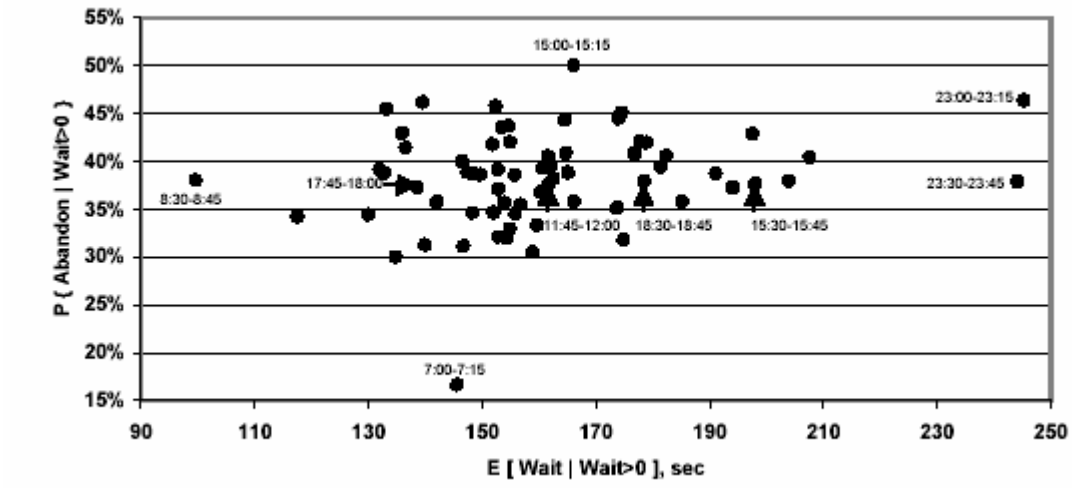
Human behavior





Delayed Abandons (IVR)

Balking (New Customers)



Learning (Internet Customers)

Understanding (Im)Patience

- **Observing** (Im)Patience: Heterogeneity.
  Under a single roof, the fraction
  abandoning varies from 6% to 40%,
  depending on the type of service/customer .

- **Describing** (Im)Patience Dynamically:
  Irritation proportional to Hazard Rate (Palm);

- **Managing** (Im)Patience:
  VIP vs. Regulars: who is more "Patient"?
  What are we actually measuring?
  (Im)Patience-Index = "How long **expect** to wait"
  relative to
  "How long **willing** to wait"

- Estimating (Im)Patience: Censored Sampling.
- Modelling (Im)Patience: Equilibrium Models

Common Performance

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN
Skill: 37
Skill Name: !BA AUTH1

Date: 7:00 pm WED MAR 10, 1999

Acceptable Service Level: 30

DAY	ACD CALLS	AVG SPEED ANS	ABAND CALLS	AVG ABAND TIME	AVG TALK TIME	TOTAL AFTER CALL	FLOW IN	FLOW OUT	TOTAL AUX/ OTHER	AVG STAFF	% IN SERV LEVL
3/04/99	637	0:19	219	0:26	1:57	92:05	0	0	4310:06	8.7	66
3/05/99	849	0:06	135	0:06	1:35	179:58	0	0	4299:43	11.3	85
3/06/99	1330	0:11	363	0:13	1:42	280:22	0	0	5592:29	13.2	73
3/07/99	1213	0:12	358	0:18	1:46	226:20	0	0	4830:15	11.5	72
3/08/99	631	0:26	382	0:33	1:57	150:50	0	0	3743:04	7.9	49
3/09/99	570	0:40	487	0:43	1:52	148:41	0	0	3979:04	6.7	38
3/10/99	512	0:29	292	0:28	1:41	243:06	0	0	3046:00	7.9	50
SUMMARY	5742	0:18	2236	0:26	1:46	1321:22	0	0	****:**	9.6	63

Arrivals Abandons 40 %

Switch Name: FDC/HAMPDEN
Skill: 46
Skill Name: !BA AUTHORIZATION

Date: 7:00 pm WED MAR 10, 1999

Acceptable Service Level: 30

DAY	ACD CALLS	AVG SPEED ANS	ABAND CALLS	AVG ABAND TIME	AVG TALK TIME	TOTAL AFTER CALL	FLOW IN	FLOW OUT	TOTAL AUX/ OTHER	AVG STAFF	% IN SERV LEVL
3/04/99	1185	0:22	479	0:31	2:08	190:16	0	0	4213:22	8.4	61
3/05/99	1805	0:05	.308	0:04	1:38	337:20	0	0	4299:43	11.3	84
3/06/99	2437	0:12	642	0:12	1:51	444:03	0	0	5592:29	13.2	73
3/07/99	2260	0:13	558	0:14	1:46	326:33	0	0	4830:14	11.5	74
3/08/99	1260	0:35	676	0:28	2:06	308:19	0	0	3743:04	7.9	48
3/09/99	1126	0:40	653	0:34	2:10	250:40	0	0	3979:04	6.7	44
3/10/99	890	0:30	472	0:32	2:16	162:13	0	0	3046:00	7.9	51
SUMMARY	10963	0:19	3788	0:22	1:55	2019:24	0	0	****:**	9.6	65

30 %

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN
Skill: 33
Skill Name: GA Authorization

Date: 7:01 pm WED MAR 10, 1999

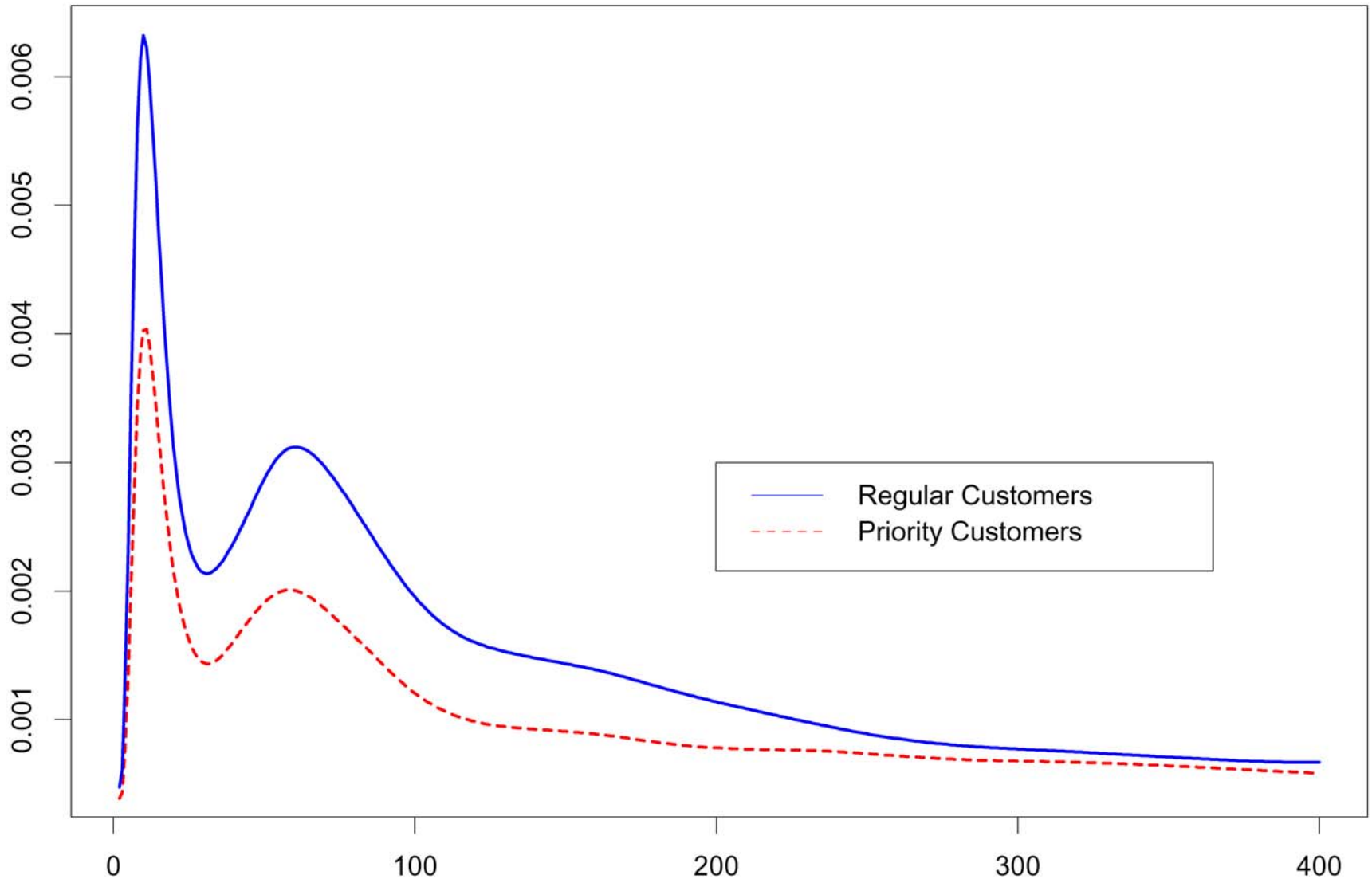
Acceptable Service Level: 30

DAY	ACD CALLS	AVG SPEED ANS	ABAND CALLS	AVG ABAND TIME	AVG TALK TIME	TOTAL AFTER CALL	FLOW IN	FLOW OUT	TOTAL AUX/ OTHER	AVG STAFF	% IN SERV LEVL
3/04/99	1248	0:27	61	0:42	1:57	330:04	0	0	4390:04	9.5	72
3/05/99	1521	0:14	37	0:20	1:58	353:48	0	0	6035:35	13.0	85
3/06/99	2388	0:20	130	0:34	2:10	550:16	0	0	6369:58	14.4	76
3/07/99	1748	0:14	66	0:30	2:08	432:16	0	0	4616:11	11.7	82
3/08/99	925	0:18	50	1:00	1:53	191:06	0	0	3835:19	8.4	81
3/09/99	856	0:26	57	0:53	1:54	125:16	0	0	4388:02	8.1	73
3/10/99	959	1:15	125	1:55	1:48	186:44	0	0	4198:39	8.9	53
SUMMARY	9645	0:25	526	0:57	2:02	2169:30	0	0	****:**	10.6	76

6%

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN

Date: 7:02 pm WED MAR 10, 1999


Hazard Rate: Empirical (Im)Patience
0.001
0.002
0.003
0.004
0.005
0.006
0
100
200
300
400
Regular Customers
Priority Customers

PATIENCE INDEX

- How to Define? Measure? Manage?

Statistics	Time Till	Interpretation
360K served (80%)	2 min.	**?** must = **expect**
90K abandon (20%)	1 min.	**? willing** to wait

"Time willing to wait" of served is **censored** by their "wait".

"Uncensoring" (simplified)

Willing to wait $1 + 2 \times \frac{360K}{90K} = 1 + 2 \times 4 = \mathbf{9}$ min.

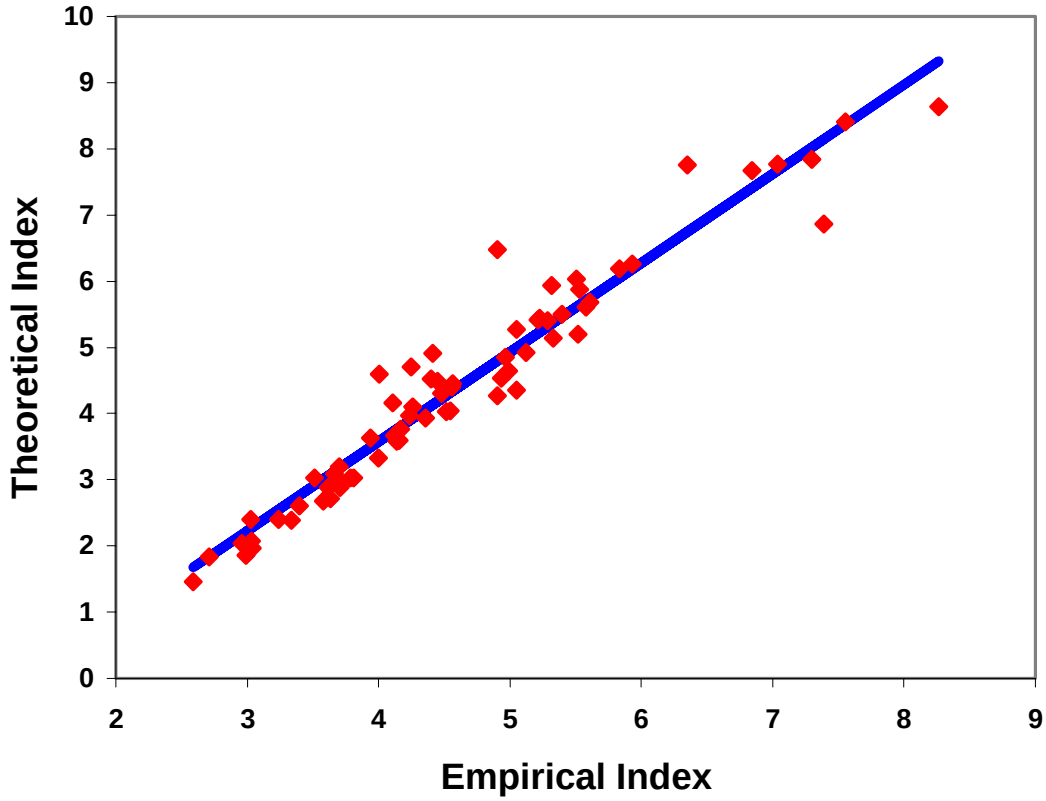
Expect to wait $2 + 1 \times \frac{90K}{360K} = 2 + 1 \times \frac{1}{4} = \mathbf{2.25}$ min.

Patience Index $= \frac{\text{time willing}}{\text{time expect}} = 4 = \frac{\#\,\text{served/wait} > 0}{\#\,\text{abandon/wait} > 0}$

↑ definition ↑ measure

- Supported by ongoing research (Wharton).

Patience Index: Excellent Fit



Erlang-A: Input, Inference

Input parameters:

Number of Agents (**N**) : in ACD data

Arrival rate (**λ**): ACD

Average Service time (**M**): ACD

Average Patience (**T**) estimated from ACD data via:

$$T = \frac{(\#\,\text{served}) \times \begin{pmatrix}\text{average wait}\\ \text{of served}\end{pmatrix} + (\#\,\text{abandon}) \times \begin{pmatrix}\text{average wait}\\ \text{of abandon}\end{pmatrix}}{\#\,\text{abandon}}$$

$$= \frac{\text{Average wait (overall)}}{\%\ \text{abandon}}$$

[Can be estimated via linear regression of (Avg Wait, % abandon)]

For square-root safety staffing:

$$\mathbf{y} = \frac{N - R}{\sqrt{R}}$$, possibly negative (via $N = R + y\sqrt{R}$)

where $\mathbf{R} = \lambda \cdot M$ is the Offered Workload

Abandonment's Impact (+ Busy Signals)

4CallCenters.com (New Improved)

Call Center iProfiler™

Performance Profiler | Staffing Profiler | Settings | Edit Account | Send Feedback

Performance Profiler Tool - Find out the Performance Level of your Call Center.

Performance Profiler Tool allows you to determine and optimize the Performance Level of your Call Center.
Please enter Your Call Center's parameters below.

Number of **Agents** in your call center	164 Agents.	Features:	Trunks,Abandons.
Average **Time** to **Handle** one call (mm:ss)	5 : 6	Basic Interval:	30 Minutes.
Number of **Calls** per 30 minutes	1061 Calls.	Target Time:	Not Defined.
Average callers' **Patience** (mm:ss)	12 : 0		
Number of **Trunks** in your call center	264 Trunks.		

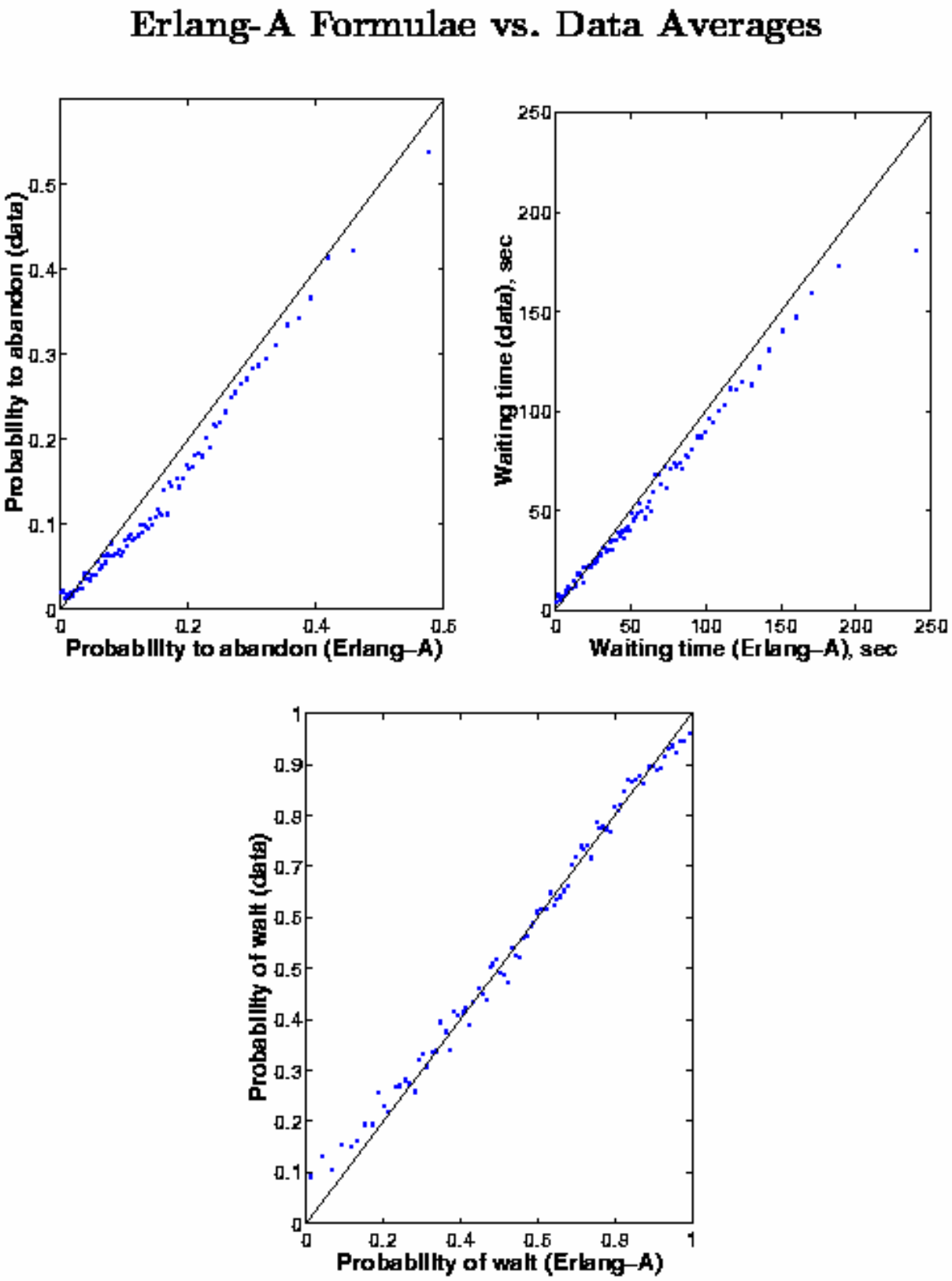
Add To Table | **Compute**

	Basic Interval	Number of Agents	Average Handling Time	Calls per Interval	Average Patience	Number of Trunks	Agent's Occupancy	% Answered	% Blocked	% Abandoned	Average Trunks Utilized	Average Speed of Answer	Average Time in Queue	Average Queue Length
	30	**164**	**05:06**	**1061**	**12:00**	**264**	**99.9%**	**90.9%**	**0.0%**	**9.1%**	**202.6**	**01:08.2**	**01:05.8**	**38.8**
☐	30	164	05:06	1061	12:00	264	99.9%	90.9%	0.0%	9.1%	202.6	01:08.2	01:05.8	38.8
☐	30	164	05:06	1061	12:00	253	99.9%	90.8%	0.1%	9.1%	202.3	01:07.6	01:05.3	38.5
☐	30	164	05:06	1061	12:00	220	99.9%	90.8%	1.7%	7.5%	195.8	00:56.8	00:55.1	32.0
☐	30	164	05:06	1061	12:00	209	99.8%	90.8%	2.9%	6.3%	190.6	00:48.0	00:46.8	26.8
☐	30	164	05:06	1061	12:00	165	96.9%	88.1%	11.9%	0.0%	159.0	00:00.2	00:00.2	0.1
☐	30	164	05:06	1061	12:00	-	99.9%	90.8%	-	9.2%	-	01:08.3	01:06.0	38.9
☐	30	164	05:06	1061	10:00	-	99.9%	90.8%	-	9.2%	-	00:57.1	00:55.2	32.6
☐	30	164	05:06	1061	05:00	-	99.6%	90.5%	-	9.5%	-	00:29.1	00:28.4	16.7
☐	30	164	05:06	1061	01:00	-	98.5%	89.6%	-	10.4%	-	00:06.1	00:06.3	3.7
☐	30	219	05:06	1179	-	-	91.5%	-	-	-	-	00:02.2	-	1.4
☐	30	218	05:06	1179	-	-	91.9%	-	-	-	-	00:02.6	-	1.7
☐	30	135	05:28	615	-	-	83.0%	-	-	-	-	00:00.3	-	0.1

Delete All | Delete Selected

- Current Result - Settings - Call Center Parameters - Performance Indicators

Fitting a Simple Model to a Complex Reality



BONUS SUPPLEMENT: E-TAILING'S FUTURE

e.b GEN

www.businessweek.com

BusinessWeek

OCTOBER 23, 2000 A PUBLICATION OF THE McGRAW-HILL COMPANIES

Mutual Funds
How to avoid a big tax bill

Wall Street
Will tech's slide keep spreading?

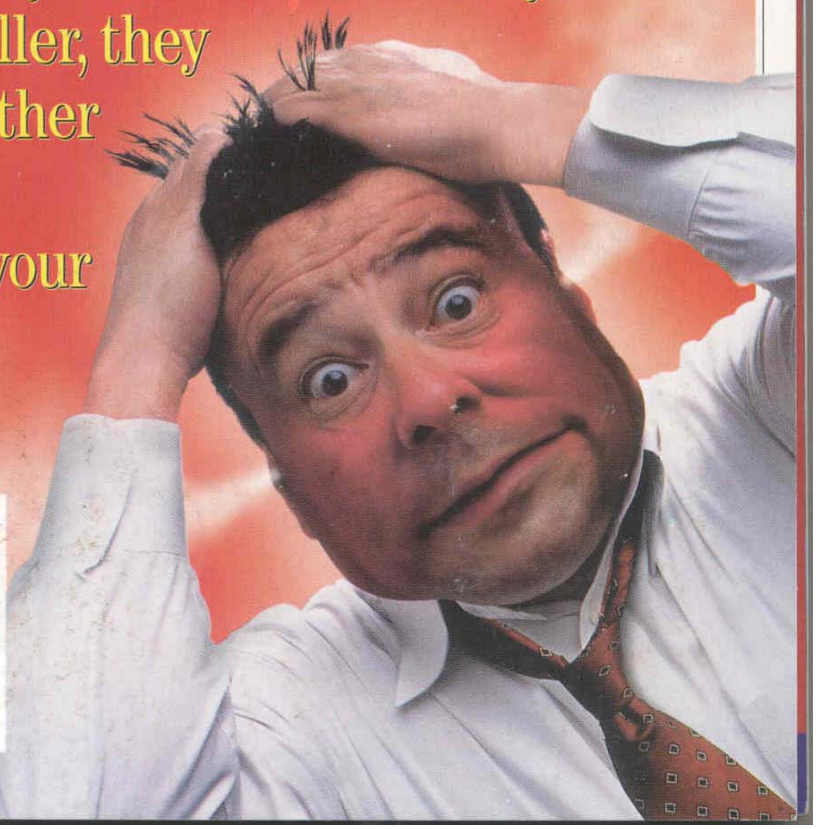
Dot-coms
The search for new business models

Managed Care
Employers seek a new solution

WHY SERVICE STINKS

Companies know just how good a customer you are—and unless you're a high roller, they would rather lose you than fix your problem

PAGE 118



AOL Keyword: BW

Skills-Based Routing and its Operational Complexities

3. Supporting Material (Downloadable)

Gans, Koole, and M.: "Telephone Call Centers: Tutorial, Review and Research Prospects." MSOM.

Garnett and M.: "An Introduction to Skills-Based Routing and its Operational Complexities", Teaching Note, 2000; under revision.

M. and Stolyar: "Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized cμ-Rule." Accepted to OR, 2003. (Efficiency-Driven SBR – **General Architecture**)

Armony and M.: " Design, Staffing and Control of Large Service Systems: The Case of a Single Customer Class and Multiple Server Types," in preparation. (**Reversed-V**)

Gurvich: "Staffing and Control of the M/M/N Queue with Multi-Type Customers and Many Servers", M.Sc. Thesis, ongoing. (**V-Design**).

Contents

1. Introduction to Skills-Based-Routing (SBR):

   Examples: CRM, Distributed Call Centers

   Truly a Multi-Disciplinary Challenge

2. Focus: Agent Scheduling, Customer Routing and

   Workforce Staffing.

3. E-Driven SBR: Index strategies in the General Case

4. QED SBR: Special Cases (V, Upside-Down V, N)

5. Dimensioning V and reversed-V: Square-Root Staffing

NationsBank CRM:
What are the relationship groups?

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

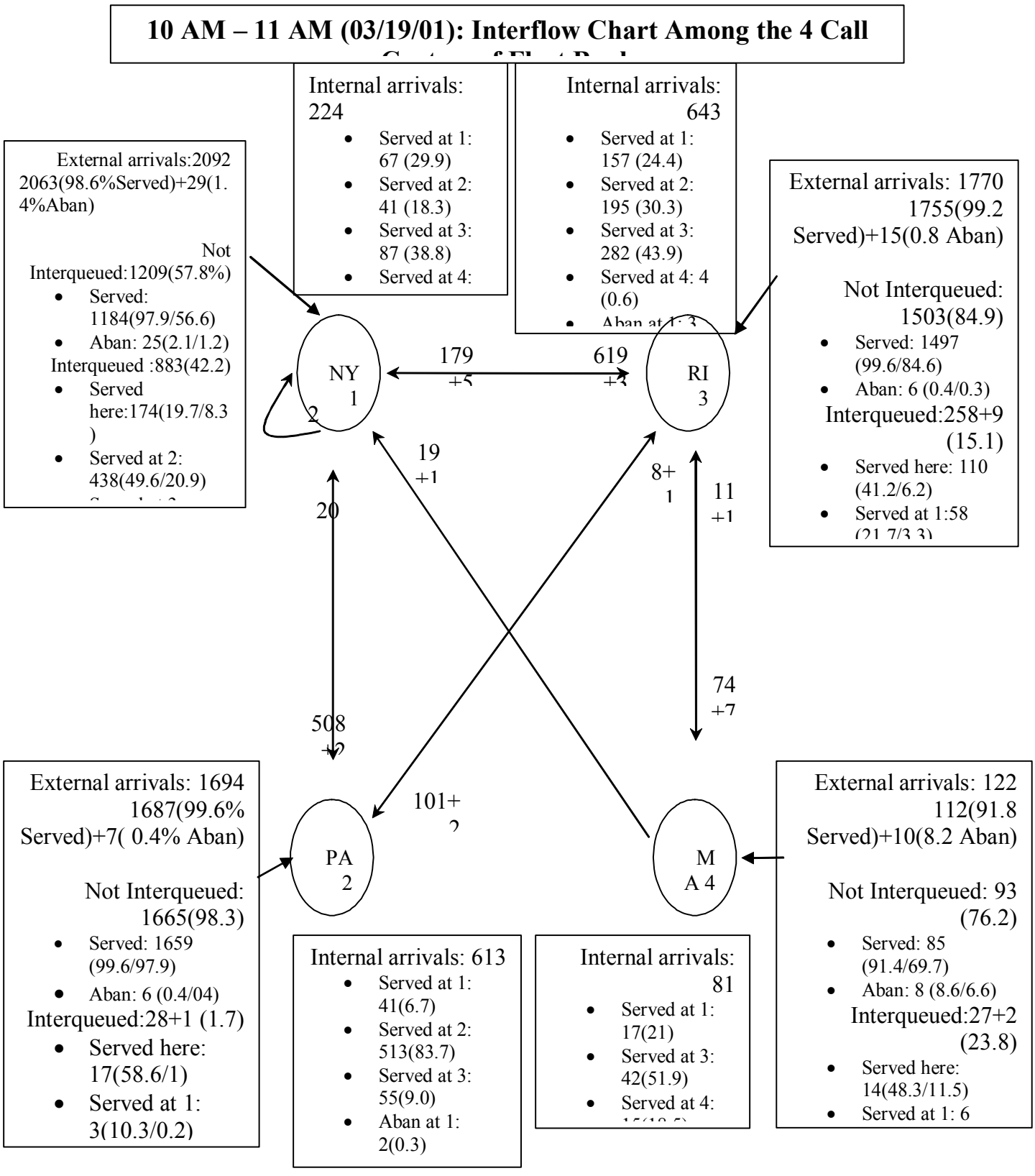
- What does it mean for a customer in each group to be profitable? Customer Revenue Management

NationsBank's Design of the Service Encounter

Examples of Specifications: Assignable Grade Of Service (AGOS)

	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

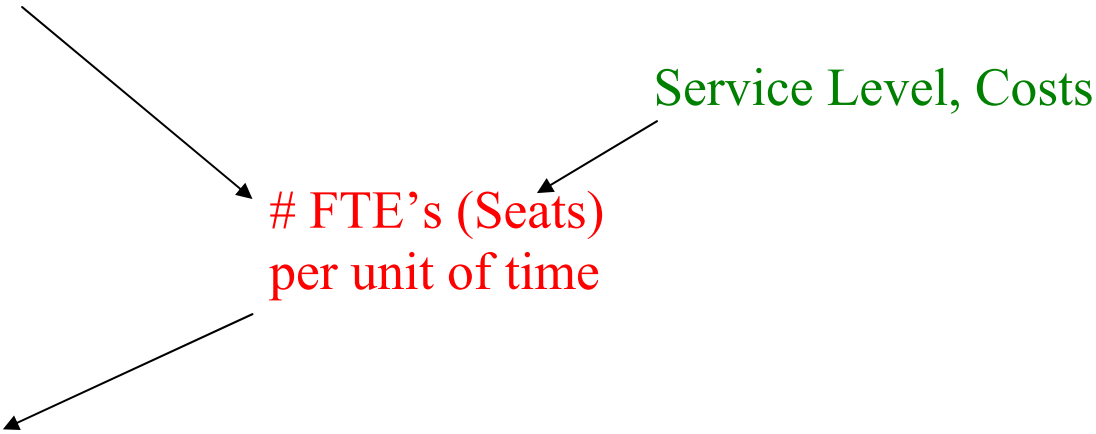
Distributed Call Center (U.S. Bank)



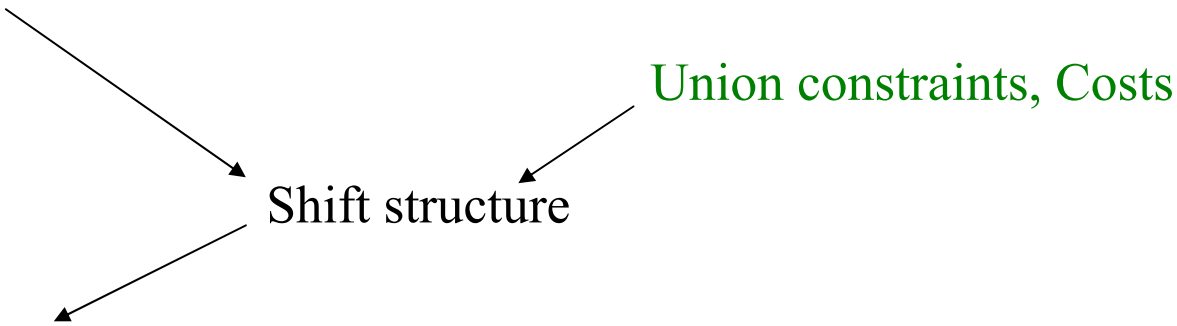
Workforce Management: Hierarchical Operational View

Forecasting Customers: Statistics, Time-Series
Agents : HRM (Hire, Train; Incentives, Careers)

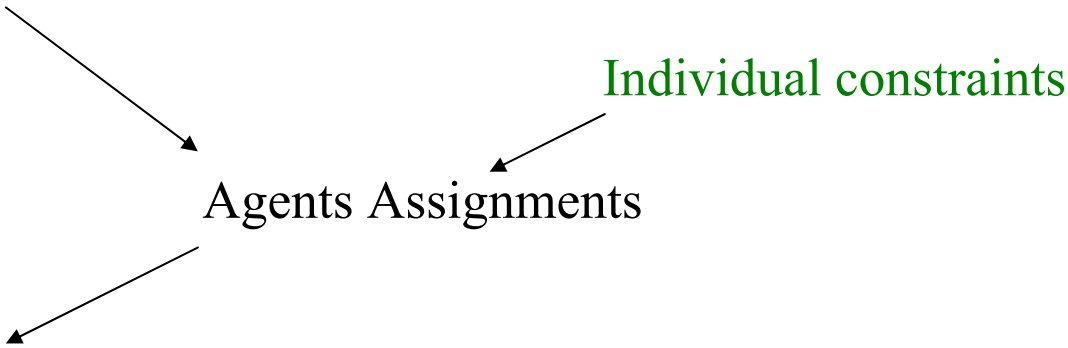
Staffing: Queueing Theory



Shifts: IP, Combinatorial Optimization; LP



Rostering: Heuristics, AI (Complex)



Skills-based Routing: Stochastic Control

Service Engineering
May 2000; Under Revision

An Introduction to Skills-Based Routing and its Operational Complexities

By Ofer Garnett and Avishai Mandelbaum

Technion, ISRAEL

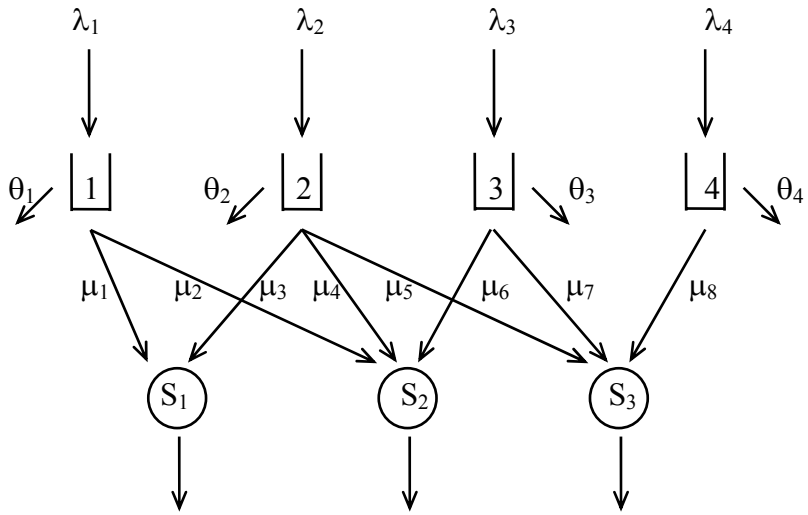
(**Full** Version)

Contents:

Acknowledgement: This teaching-note was written with the financial support of the Fraunhofer IAO Institute in Stuttgart, Germany. The authors are grateful to Dr. Thomas Meiren and Prof. Klaus-Peter Fähnrich of the IAO for their assistance and encouragement.

Introduction

Multi-queue parallel-server system = schematic depiction of a **telephone call-center**:



Here the λ's designate arrival rates, the μ's service rates, the θ's abandonment rates, and the S's are the number of servers in each server-pool.

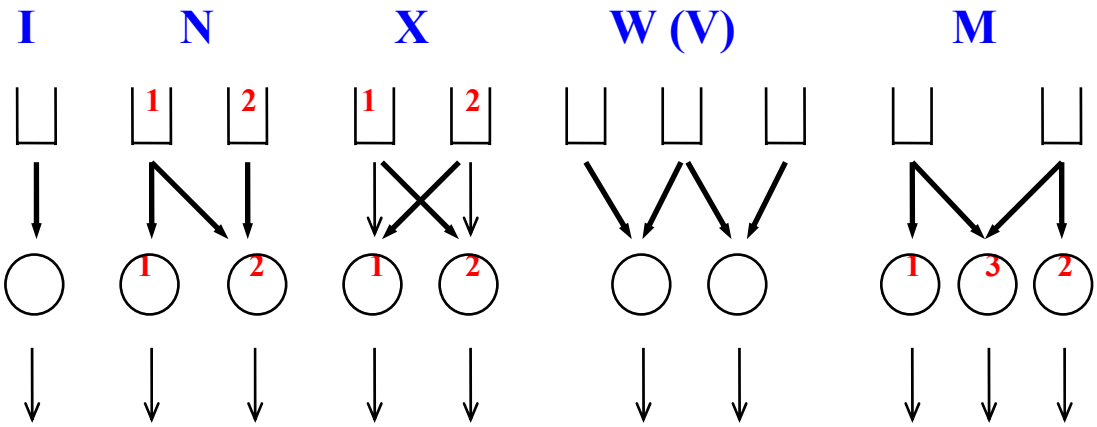
Skills-Based Design:

- **Queue**: "customer-type" requiring a specific type of service;
- **Server-Pool**: "skills" defining the service-types it can perform;
- **Arrow**: leading into a server-pool define its skills / constituency.

For example, a server with skill 2 (**S2**) can serve customers of type 3 (**C3**) at rate μ_6 customers/hour.

Customers of type 3 arrive randomly at rate λ_3 customers/hour, equipped with an impatience rate of θ_3.

Some Canonical Designs - Animation



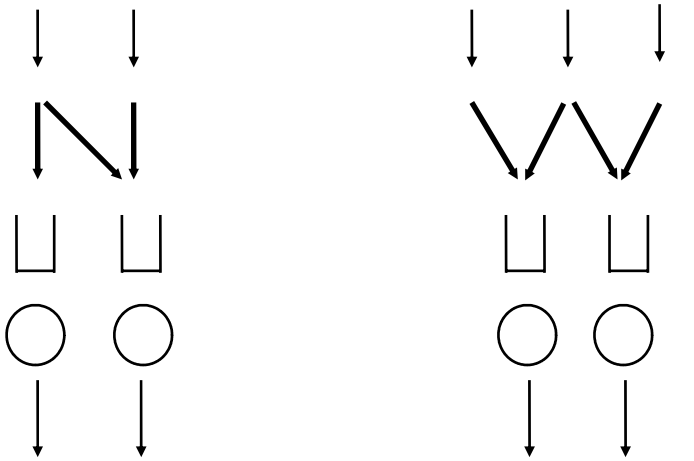
I – dedicated (specialized) agents

N: for example,

- C1 = VIP, then S2 are serving C1 to improve service level.
- C2 = VIP, then S2 serve C1 to improve efficiency.
- S2 = Bilingual.

X: for example, S1 has C1 as Primary and C2 as Secondary Types.

V: Pure Scheduling; **Upside-down V**: Pure Routing.



Major Design / Engineering Decisions

1. Classifying customers into **types** (**Marketing**):
   Tech. support vs. Billing, VIP vs. Members vs. New
2. Determining server **skills, incentives, numbers** (**HRM, OM, OR**)
   Universal vs. Specialist, Experienced / Novice, Uni- / Multi-lingual;
   Staffing: how many servers?
3. Prerequisite Infrastructure - MIS / IT / Data-Bases (**CS, Statistics**)
   CTI, ERP, Data-Mining

Major Control Decisions

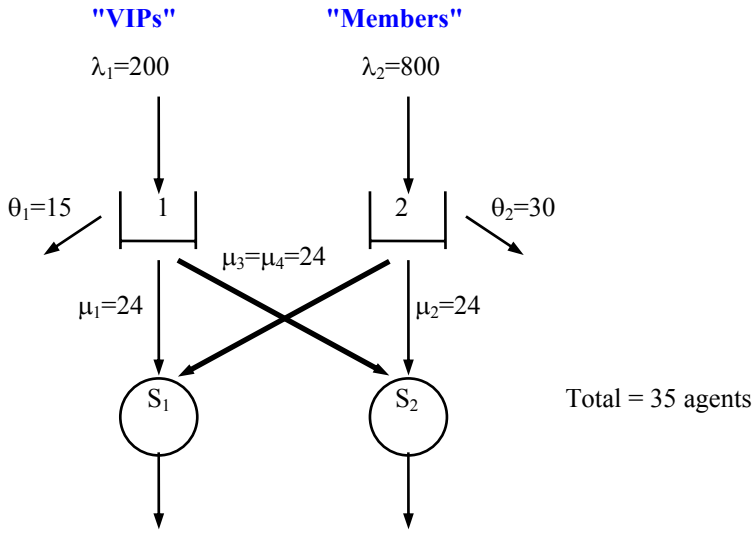
4. Matching customers and agents (**OR**)
   - Customer **Routing**: Whenever an agent turns idle and there are queued customers, which customer (if any) should be routed to this agent.
   - Agent **Scheduling**: Whenever a customer arrives and there are idle agents, which agent (if any) should serve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

Multidisciplinary Challenge

Skills-Based Routing: protocol for online matching of S's and C's.

- **Prevalent**: Static Priorities of customer types and agent skills
- **Index**-based: Dynamic Priorities via continuous review
- **Threshold**-based: Dynamic Management by Exception
- **Others**: discrete review, credit schemes (SLA), scripts; call backs

Example**: Scripts** for Staffing, Scheduling, Routing



Setup A : (X-design)

"VIP" servers : $S_1 = 20$

- If "VIP" queue not empty serve the "VIP" queue + all "Members" waiting more than **40** seconds, as a single FIFO queue.
- If "VIP" queue is empty, serve the first in the "Member" queue.

"Member" servers : $S_2 = 15$

- If "Member" queue not empty serve the "Member" queue + all "VIPs" waiting more than **6** seconds, as a single FIFO queue.
- If "Member" queue is empty, serve the first in the "VIP" queue.

Setup C : (V-design; feasible since servers are assumed equally skilled.)

Total servers: 35

- Serve as a FIFO queue, but "VIPs" enter the queue with a virtual **15** second wait (i.e. as if they had joined the queue 15 seconds earlier).

Chart 2 : 1000 Calls/hour - ASA

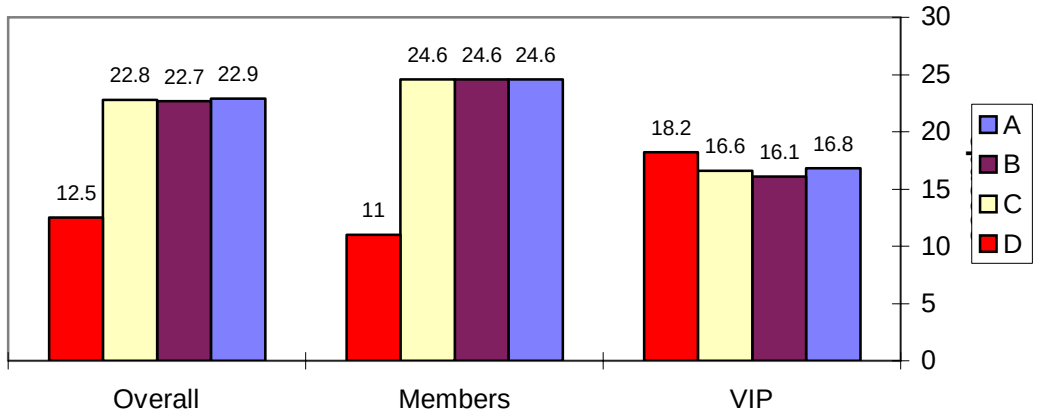


Chart 3 : 1000 Calls - Abandonment

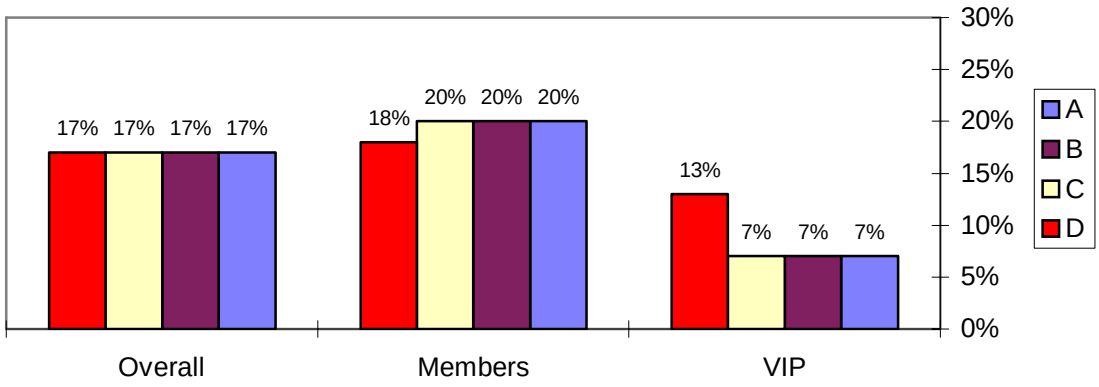
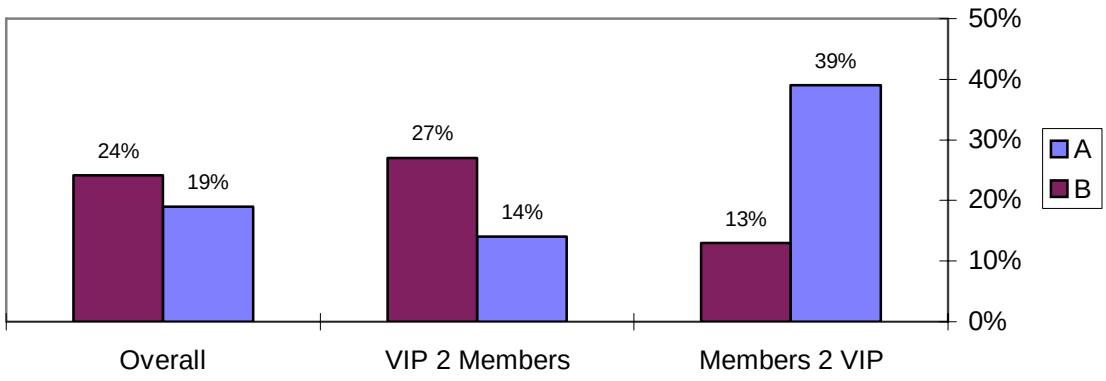


Chart 4 : 1000 Calls - Overflows



WHAT IF : 1500 Calls/hour - ASA

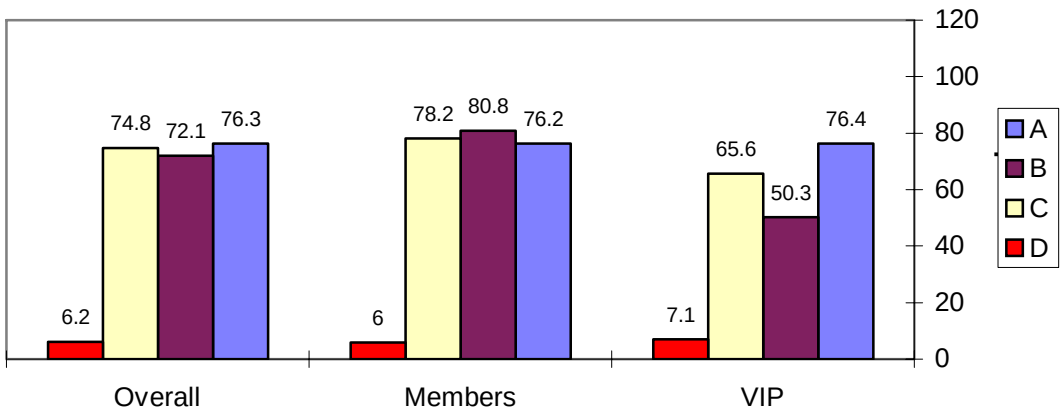


Chart 7 : 1500 Calls - Abandonment

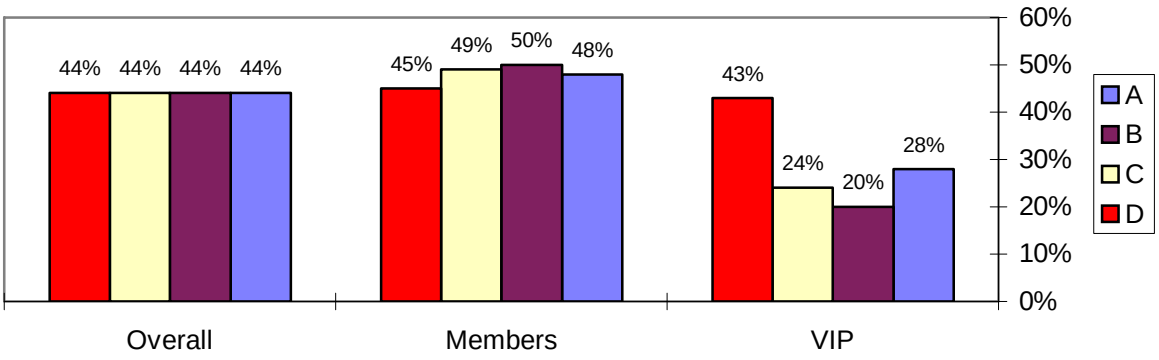
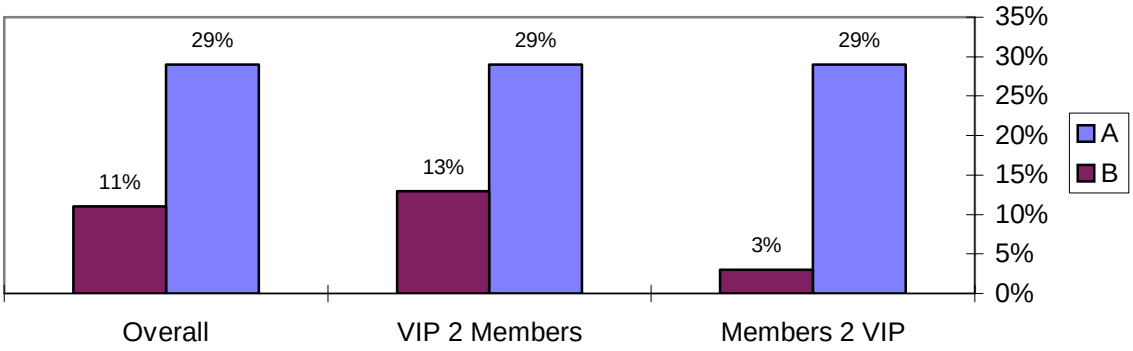


Chart 8 : 1500 Calls - Overflows



Efficiency-Driven SBR - the "EASY" Case (Stolyar)

Examples: Scarce agents, hence must be well utilized.

Email-dominance, hence can delay response.

Classical special case: **V**-design

- Agent Scheduling: upon service completion, if
  1. Same mean service times: serve the costliest queue (largest c)
  2. Same delay costs: serve the shortest service (smallest m)
  3. Generally: serve the largest **c/m** (= index).

General (N, X, W, M, …) solution: Index Control is optimal, under sufficient skills-overlap (complete resource pooling; Harrison, Lopez).

- Customer Routing: irrelevant, since essentially all customers wait.
- Agent Scheduling: upon service completion, the server chooses the queue with the largest index and serves its "oldest" customer.
- Index: marginal waiting-cost per unit of average service-time

  (Example: "waiting-time" of "oldest" customer in queue)

However: well-managed telephone services are not

Reality

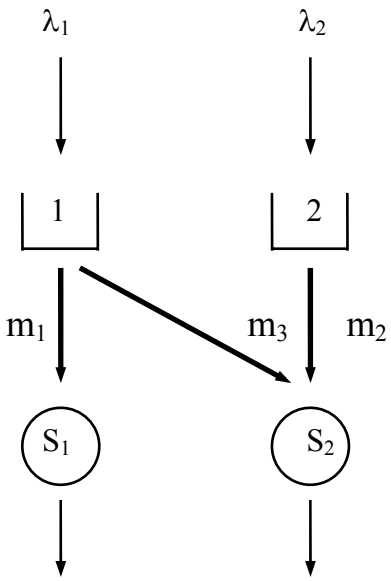
- Technology enables smart systems
- Reality becomes increasingly complex
- Solutions are urgently needed
- Theory lags significantly behind needs

- Ad-hoc methods: heuristics, simulation-based

Research Status

- Efficiency-driven SBR well understood and solved
- QED SBR is challenging and advancing
- **Small yet significant models for theoretical insight**
- Principles/Guidelines for design, staffing, control
- Implementation: fine-tuning of parameters, scale-up

Static Priorities (Cross-Training): Some Subtleties

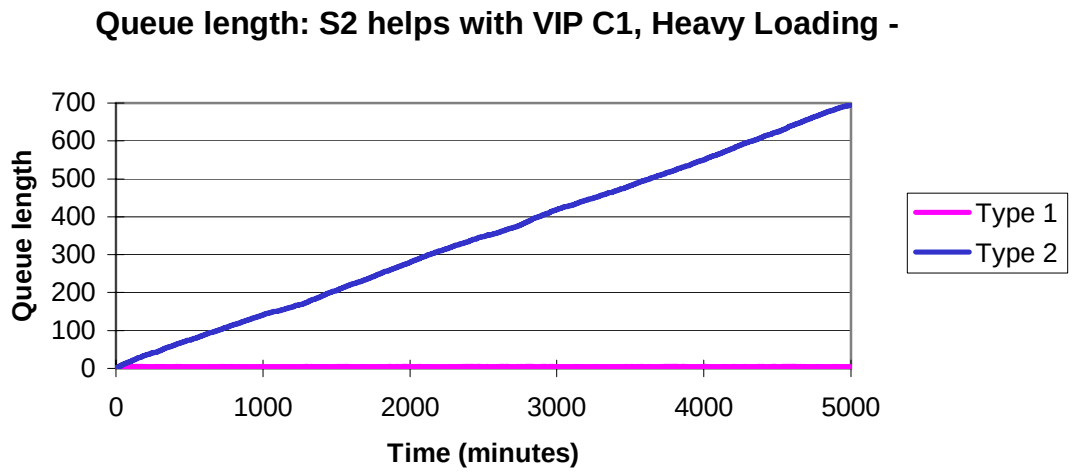


$\lambda_1 < 1.3 \quad \lambda_2 <= 0.4$

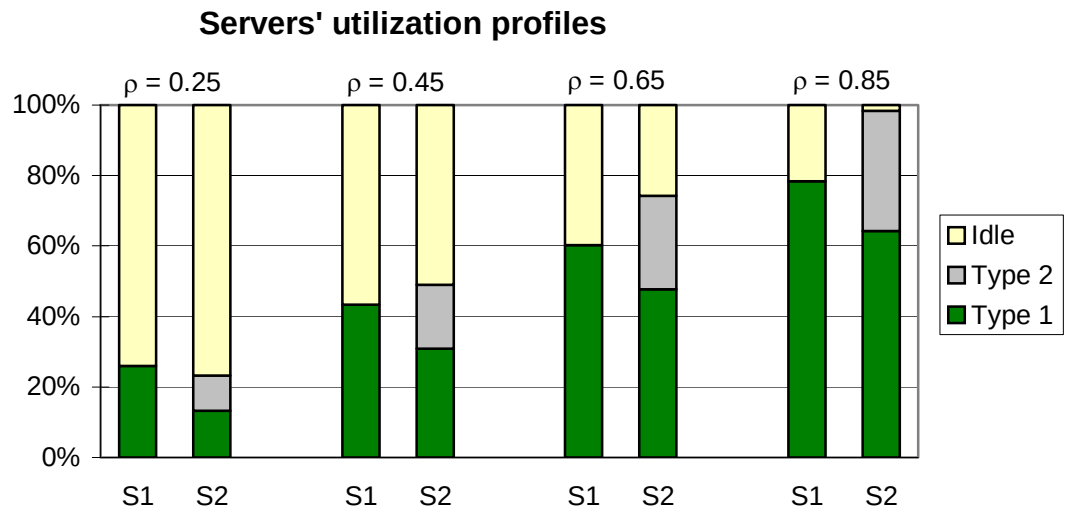
$m_1 = m_2 = 1, \quad \mathbf{m_3 = 2}$

$S_1 = S_2 = 1$

- C1 are **VIP**, hence S2 helps S1 by giving priority to C1 over C2.
- If both servers are idle - **Ci** customers are routed to server **Si**



Q2 "explodes, while Q2 is negligibly small – why ?

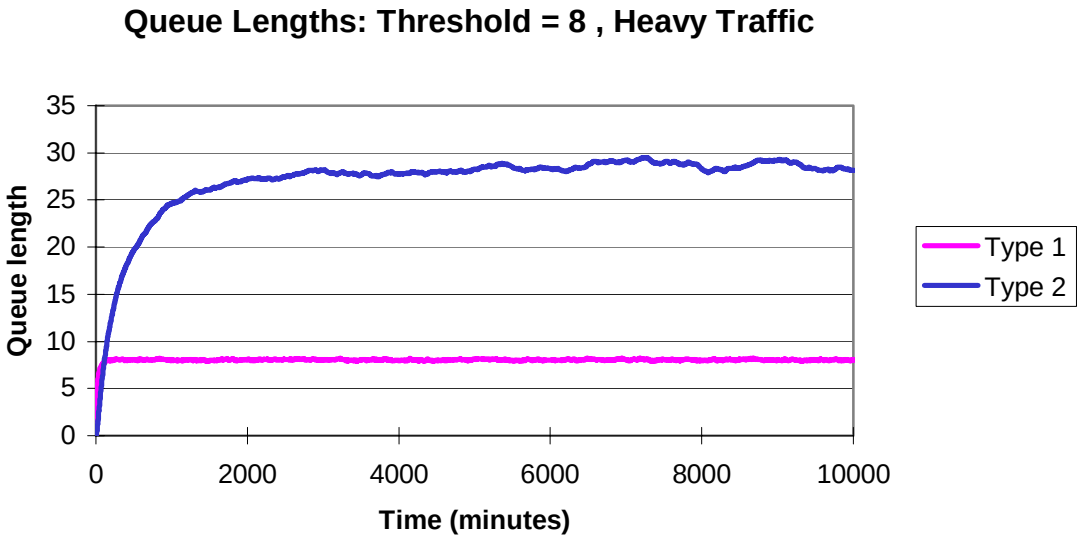


Instability: S2 **overworked** serving C1 and neglecting C2,

while S1 is **20%** idle.

To avoid "overzealous help", apply **Threshold Control**:

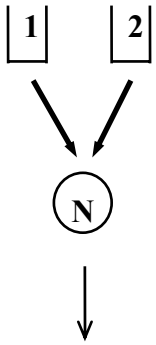
S2 assists **S1** only when **Q1 is at or above a certain threshold**



Both Q1 and Q2 are stable.

Now fine-tuning of the threshold value

V-Design: Pure Scheduling



N agents, fully flexible

C1 = VIP

Optimal Scheduling: **Agent Reservation** (Yahalom)

- C1(=VIP) always served, when possible;

- C2 served only if # of idle agents exceeds a threshold.

QED regime: $\sqrt{\cdot}$ Safety-Staffing, as before (Gurvich)

Threshold Size (relative to N) determines Service Levels:

- Large: C1 is Q-served, C2 is E-served

- Small: C1 and C2 indistinguishable QED

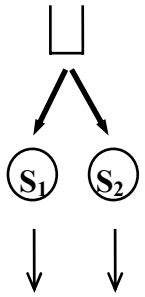
- Moderate: C1 is Q-served, C2 is QED

$\sqrt{\cdot}$ Safety-Staffing is asymptotically optimal.

Reversed-V Design: Pure Routing

Homogeneous Customers

Heterogeneous Agents: **S2 = Faster**



Optimal Routing: **"Slow-Server" phenomenon** (Rykov)

- S2(=Fast) always employed, if possible;

- S1(= Slow) employed if # in queue exceeds a threshold.

QED regime: $\sqrt{\cdot}$ Safety-Staffing – see below (Armony)

- No threshold needed: just have all servers work

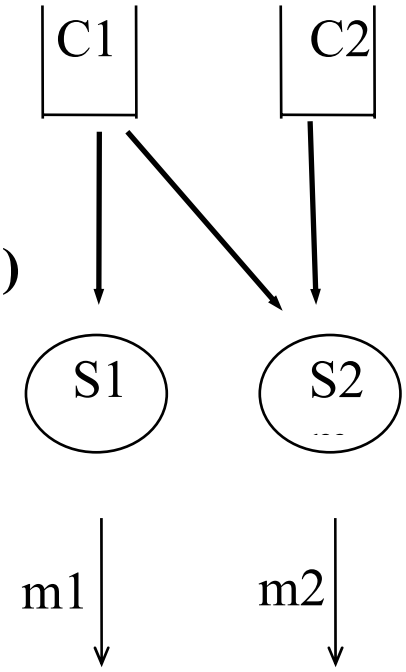
  when possible, ensuring that the "fast" get the priority.

Asymptotically optimal staffing:

1. Given a delay probability, determine S1 + S2 via $\sqrt{\cdot}$ Safety.
2. Given staffing costs, determine S1 / S2.

Distributed call centers: in progress.

N-Design: Routing and Scheduling



Heterogeneous Customers: **C2=VIP**

Heterogeneous Agents: **S2 = Faster (m1>m2)**

Costs: $\mathbf{D_1, D_2}$ Delay costs
$\mathbf{H_1, H_2}$ Staffing costs

Assume: $\mathbf{D_2 >> D_1}$ **(Truly VIP)**

Assume: $\mathbf{H_1 m_1 < H_2 m_2}$ (Otherwise V)

QED regime: $\sqrt{\cdot}$ Safety Staffing – see below (Gurvich).

- $(C_1, C_2; S_2)$ operate as V-model, with "idle-thresholds"

- $(C_1; S_1, S_2)$ operate as Λ, but without "queue-thresholds"

Asymptotically optimal staffing:

1. Given a delay probability (service level), determine
   $\mu_1 S1 + \mu_2 S2$ via $\sqrt{\cdot}$ Safety;
2. Given staffing costs, determine S1 / S2 via Math. Prog.

Ultimately: $\sqrt{\cdot}$ Safety-Staffing is asymptotically optimal.