Telephone Services:

Science, Engineering, Management and Teaching

1st International Symposium

Service Engineering & Management

Fraunhofer, Nov. 26, 2002

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

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

Tool : http://4CallCenters.com (register & use)

Supporting Material (in Web-Site)

Gans, Koole, and M.: "Telephone Call Centers: A Tutorial and Literature Review." Invited review to *MSOM,* 2002.

M., Sakov, and Zeltyn: "Empirical Analysis of a Call Center." Technical report, Technion, 2000.

Brown, Gans, Haipeng, M., Sakov, Zeltyn, Zhao: "Statistical Analysis of a Call Center: a **Queueing Science** Perspective." Submitted to *JASA*, 2002.

Borst, M. and Reiman:. "Dimensioning Large Telephone Call Centers." Accepted to *OR*, 2002.

Garnett, M. and Reiman: "Designing a Telephone Call-Center with Impatient Customers." *MSOM,* **4**, 208-227, 2002.

Jennings, M., Massey and Whitt: "Server staffing to meet time-varying demand." *Mgt. Sc.* 42, 1383–1394, 1966.

Atar, M. and Reiman: "Scheduling a Multi-Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy-Traffic." Submitted to *Annals Appl Prob*, 2002.

M. and Stolyar: "Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized cμ-Rule." Under revision to *OR*, 2002.

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 and efficiency, from the (often conflicting) views of customers, servers and managers.

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

Example: Skills-Based Routing in multi-media centers

e.g. Support + Sales via Telephone + IVR + e.mail + Chat, for VIP and others.

Example: Staffing the modern Call Center

e.g. How many agents to balance service- and efficiency-levels.

Significant: scientific, economic, social, psychological.

Multi-Disciplinary: typically OR/OM, Marketing, HRM. MIS.

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 highway rush hour, for green light
- Checks, waiting to be processed, cashed

Queues

- Operational Causes: Scarce Resources, Synchronization Gaps
- Prevalent, Significant (Economically, Psychologically)
- Costly, yet Here to Stay (e.g. 1-800 costs)
- Visible levers: proxy to the invisible

  (Means, rather than goals)

Tele-Nets: Call/Contact Centers

Scope	Examples
Information	#411, Tele-pay, Help Desks
Business	Tele-Banks, #800-Retail
Emergency	Police #911
Mixed	
Info + Emerg.	Utility Companies, City Halls
Info + Bus.	Airlines

Scale

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

Trends: THE interface with customers

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

BONUS SUPPLEMENT: E-TAILING'S FUTURE

e.bGEN

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?

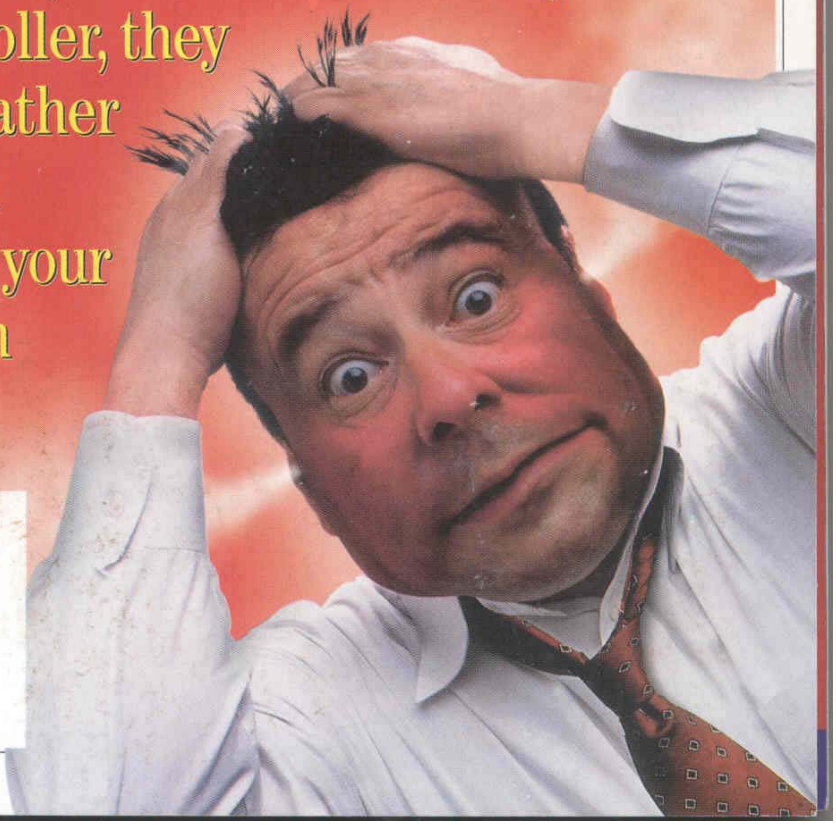
Dot-coms
The search for new business models

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

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

An Introduction to Skills-Based Routing and its Operational Complexities

By Ofer Garnett and Avishai Mandelbaum

Technion, ISRAEL

(Full Version)

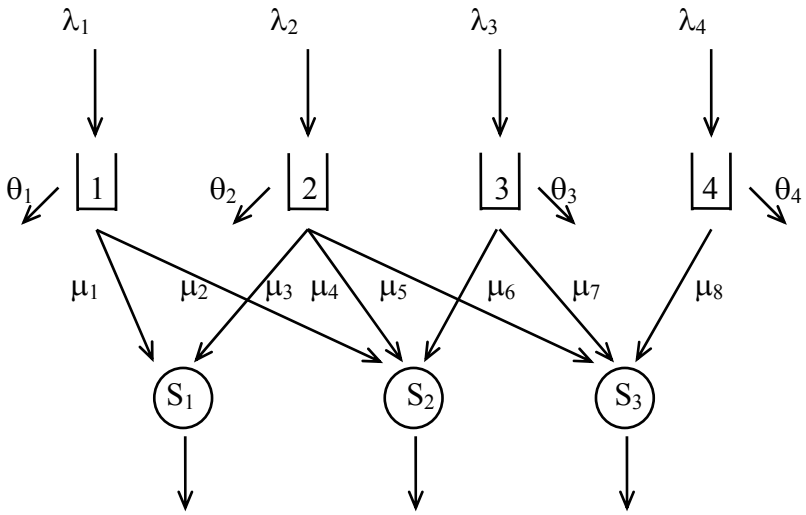
Contents:

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

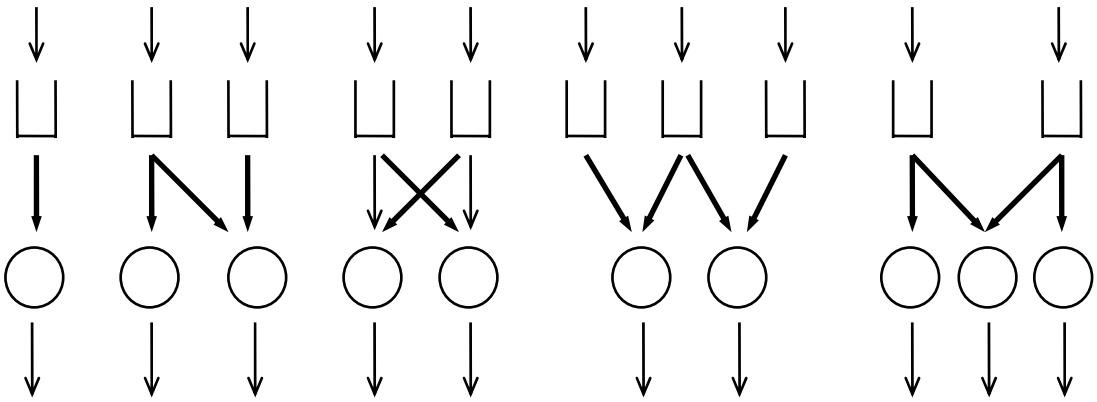
Consider the following multi-queue parallel-server system (animated, for example, by a **telephone call-center**):



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.

Such a design is frequently referred to as a Skills-Based Design since each queue represents "customers" requiring a specific type of "service", and each server-pool has certain "skills" defining the services it can perform.

In the diagram above, the arrows leading into a given server-pool define its skills. (For example, a server from pool 2 can serve customers of type 3 at the of rate μ_6 customers per unit of time) .
Some canonical designs are: I (I^k), N, X, W, M (V).



Major Decisions when implementing skills-based systems:

1. Who are the customers - defining customer **types** (*design*, **Marketing**)
2. Who are the servers - their **skills** and **numbers** (*design*, **HRM, Op Mgt**)
3. How are customers **routed** to servers (*control*, **Op Mgt, OR**)

A prerequisite for the above is a supporting **MIS / IT**

Truly a multi-disciplinary challenge.

System's design (engineering) consists of

classifying the customers and determining the servers' required skills.

A particular design can have alternative interpretations, for example:

- Different customer types can represent customers requiring different services (e.g. **technical support vs. billing**) or customer priorities (**VIP vs. Members**).
- Separate server-pools can be due to servers' level of capability / training / experience (e.g. Hebrew/Arabic speaker vs. **Hebrew/Russian/English** speaker, **generalist vs. specialist**, expert vs. novice).

Skills-Based Routing:protocol for online routing of customers.

SBR: State of the art (Stolyar; Atar, Reiman)

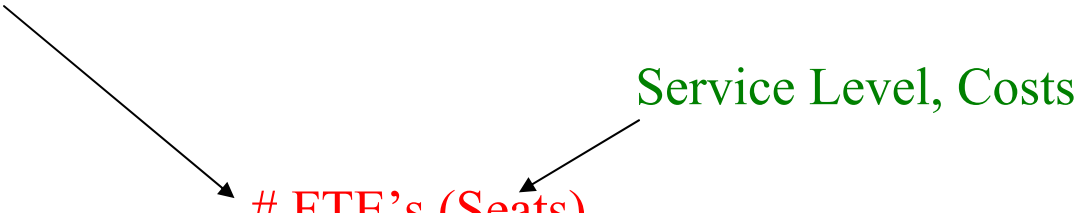
Complete (surprising) success for Efficiency-Driven services (e.g. email call centers): upon service completion, the server chooses the queue with the largest marginal waiting-cost per unit of service-time.

But well-managed telephone services are not (at least should not be) Efficiency-Driven: for these, SBR is truly challenging – an active research area.

Workforce Management = Staffing: Hierarchical Operational View

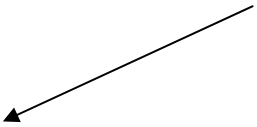
Forecasting Customers: Statistics, Time-Series
Agents : HRM (Hire, Train; Incentives, Careers)

Staffing: Queueing Theory (Erlang-**A** and beyond)

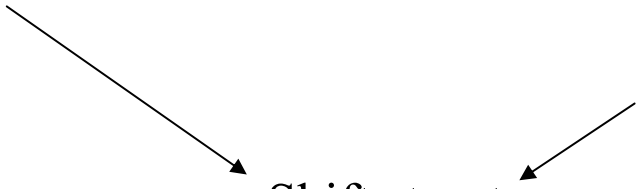


Service Level, Costs

FTE's (Seats) per unit of time

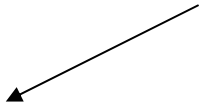


Shifts: IP, Combinatorial Optimization; LP

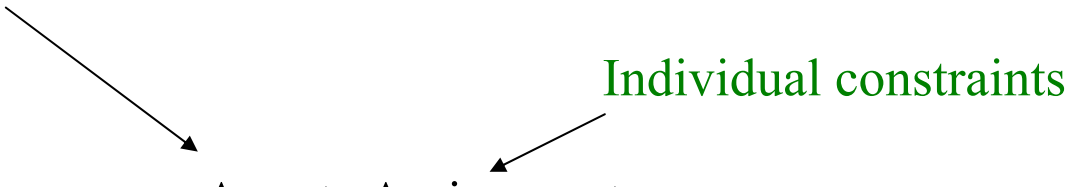


Union constraints, Costs

Shift structure

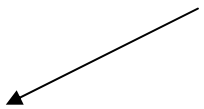


Scheduling: Heuristics, AI (Complex)



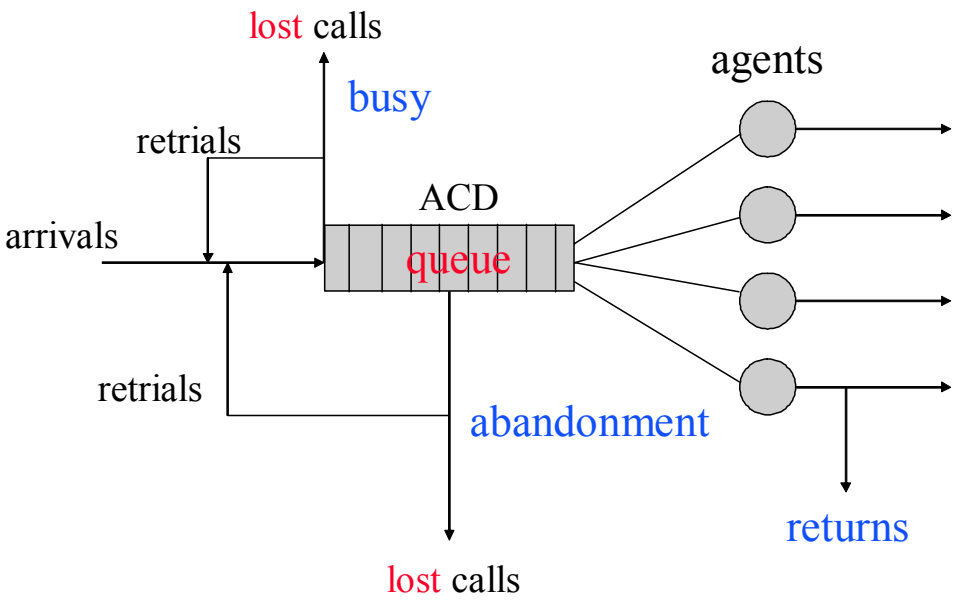
Individual constraints

Agents Assignments

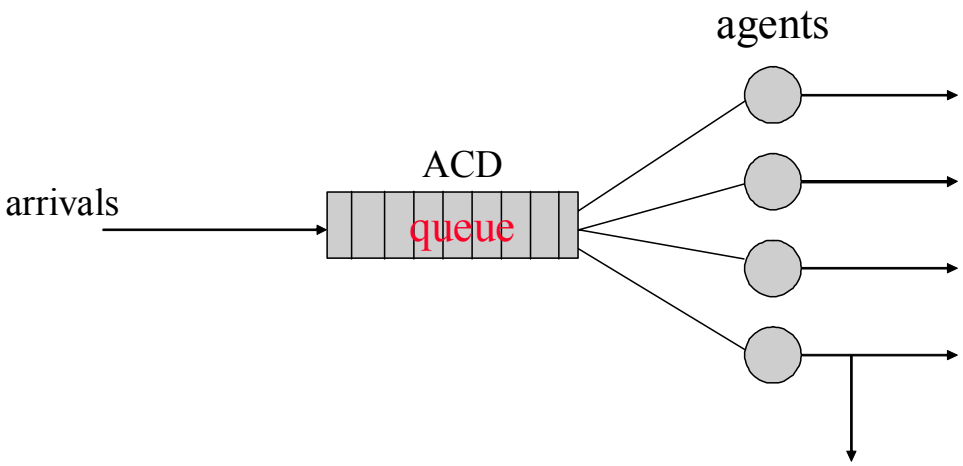


Online Skills-based Routing: Stochastic Control (ongoing)

The Basic Call Center



Erlang-C = M/M/N



Tool: ErlangA @ 4CallCenters.com

4CallCenters.com™

Workforce Management On The Web!

Call Center iProfiler

Workforce Management tools for inbound call centers on the Web!

New Member?

Sign-up Now for FREE!

Take our Tour

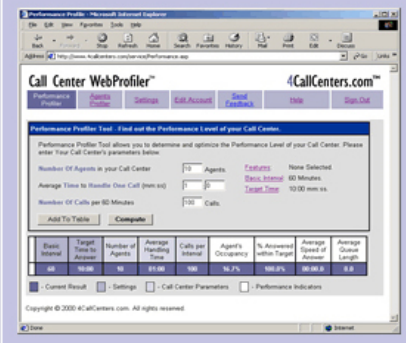
About 4CallCenters.com

Current Member?

Account No:

Password:

Login



Click above to Take our Tour.

Call Center Charisma™

Call Center Charisma™ is a Windows application offering powerful tools for profiling and optimizing your inbound call center.

News

October 31, 2000 - 4CallCenters.com today announced the launch of Call Center iProfiler, a unique online service offering powerful performance profiling and staffing optimization tools for inbound call centers of all sizes... (read more...)

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: 10 Agents.

Average Time to Handle one call (mm:ss): 1 : 0

Number of Calls per 30 minutes: 100 Calls.

Features: None Selected.
Basic Interval: 30 Minutes.
Target Time: Not Defined.

Add To Table | Compute

Basic Interval	Number of Agents	Average Handling Time	Calls per Interval	Agent's Occupancy	Average Speed of Answer	Average Queue Length
-	-	-	-	-	-	-

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

What can be achieved

At what cost

Copy of Summary Interval - Order PK

Date: 7/7/97
Split/Skill: Order PK

Time	Avg Speed Ans	Avg Aban Time	ACD Calls	Avg ACD Time	Avg ACW Time	Aban Calls	% ACD Time	% Ans Calls	Avg Pos Staf	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time
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	46	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	18
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	361	: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	:03:36	: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	6	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 = Average Speed of Answer

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)

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)

Intuition: at 100% utilization, N servers = 1 fast server.

$$W_q \approx W_q \mid W_q > 0 = \frac{1}{N} \cdot \frac{\rho_N}{1-\rho_N} \cdot E(S) \to E(S)/\gamma \;!$$

$$N(1-\rho_N) \sim \gamma \qquad (\rho_N \to 1)$$

Changing N (Staffing)

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** operational regime

Heavy traffic, in the sense that OCC > 95%;

Light traffic, 50% answered immediately.

Rationalized Operation: high service + efficiency levels

QED Regime = Quality- and Efficiency-Driven Regime

Enabler: Economies of Scale in a

Frictionless Environment (e.g. Call Center)

Rules of Thumb: Operational Regimes

$R = \lambda \times E(S)$ units of work per unit of time (load)

Efficiency-driven $(\%\{\text{Wait} > 0\} \to 100\%)$

$N = \lceil R + \gamma \rceil$, $\gamma > 0$ service grade

Quality-driven $(\%\{\text{Wait} > 0\} \to 0)$

$N = \lceil R + \delta R \rceil$, $\delta > 0$

QED Regime $(\%\{\text{Wait} > 0\} \to \alpha,\ 0 < \alpha < 1)$

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

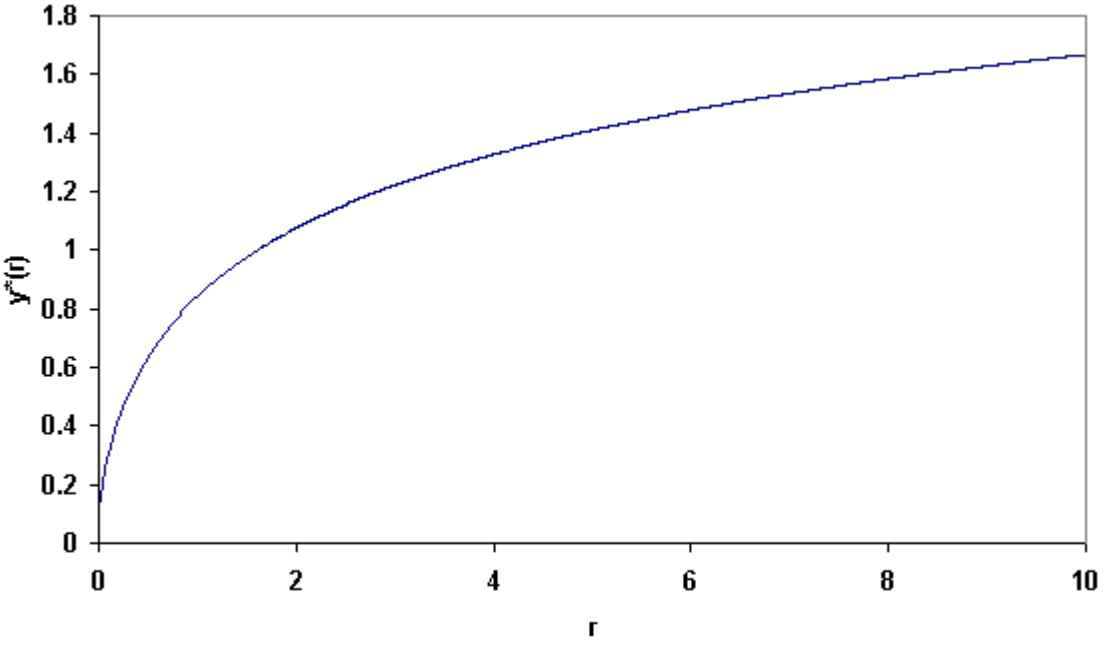
Determine Regimes (Strategy), Parameters (Economics)

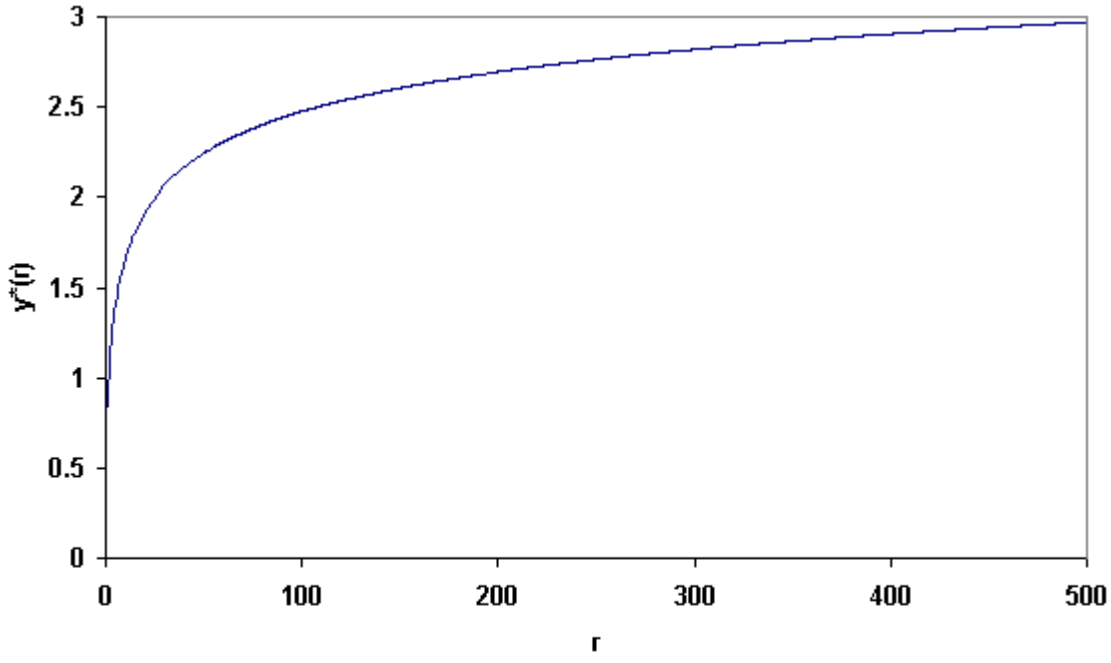
Strategy: Managers, Agents (Unions), Customers

Economics: Minimize agent salaries + waiting cost

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

r **= cost of delay (1-800) / cost of staffing (salary)**





√· Safety-Staffing (QED): Overview

Simple Rule-of-thumb: $\mathbf{N}^* \approx \mathbf{R} + y^*\left(\frac{d}{c}\right)\sqrt{\mathbf{R}}$

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

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

Relevant: Large call centers are QED

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

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 Min R !

Performance:	N^*	% wait > 20 sec.	Utilization
2.5%	400 + **11**	20%	97%
7.5%	400 + **29**	1%	93%

$\Rightarrow$ **Forecsting** R (workload) is the "real" problem

Relevance: QED Operation

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%	218	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%

Operational Aspects of Impatience

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:

Erlang-A: Theoretical performance analysis

Free Internet implementation (4CallCenters.com)

Theorem (Halfin-Whitt '81; Garnett, M. and Reiman '02):

Consider a queue attended by N servers, N "large".

Then the following **points of view** are equivalent:

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

QED performance (ASA, ...) is very easily computable, all in terms of β (the square-root safety staffing level).

Covers also the Extremes:

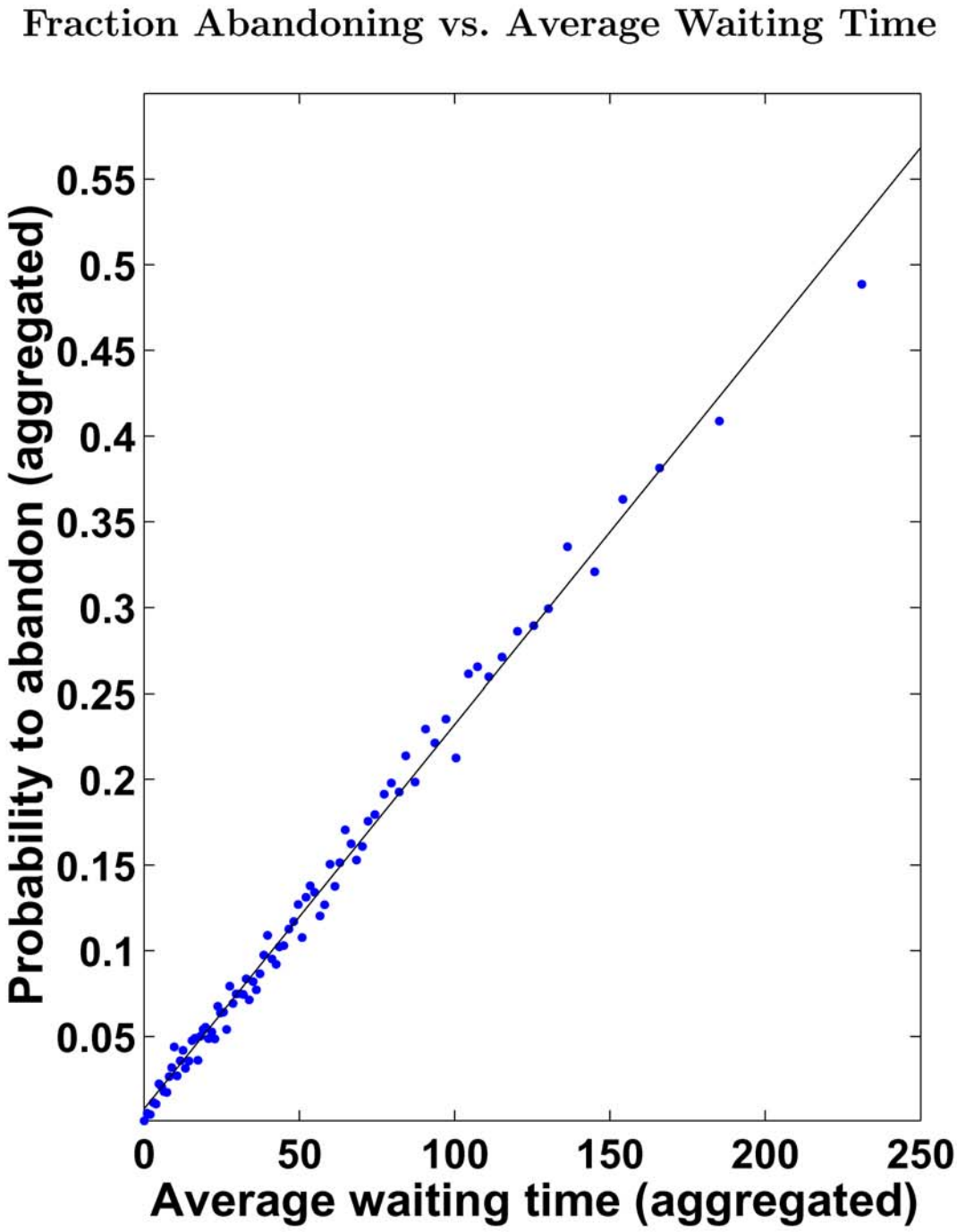
$\alpha = 1$: N = R - γ R **Efficiency-driven**

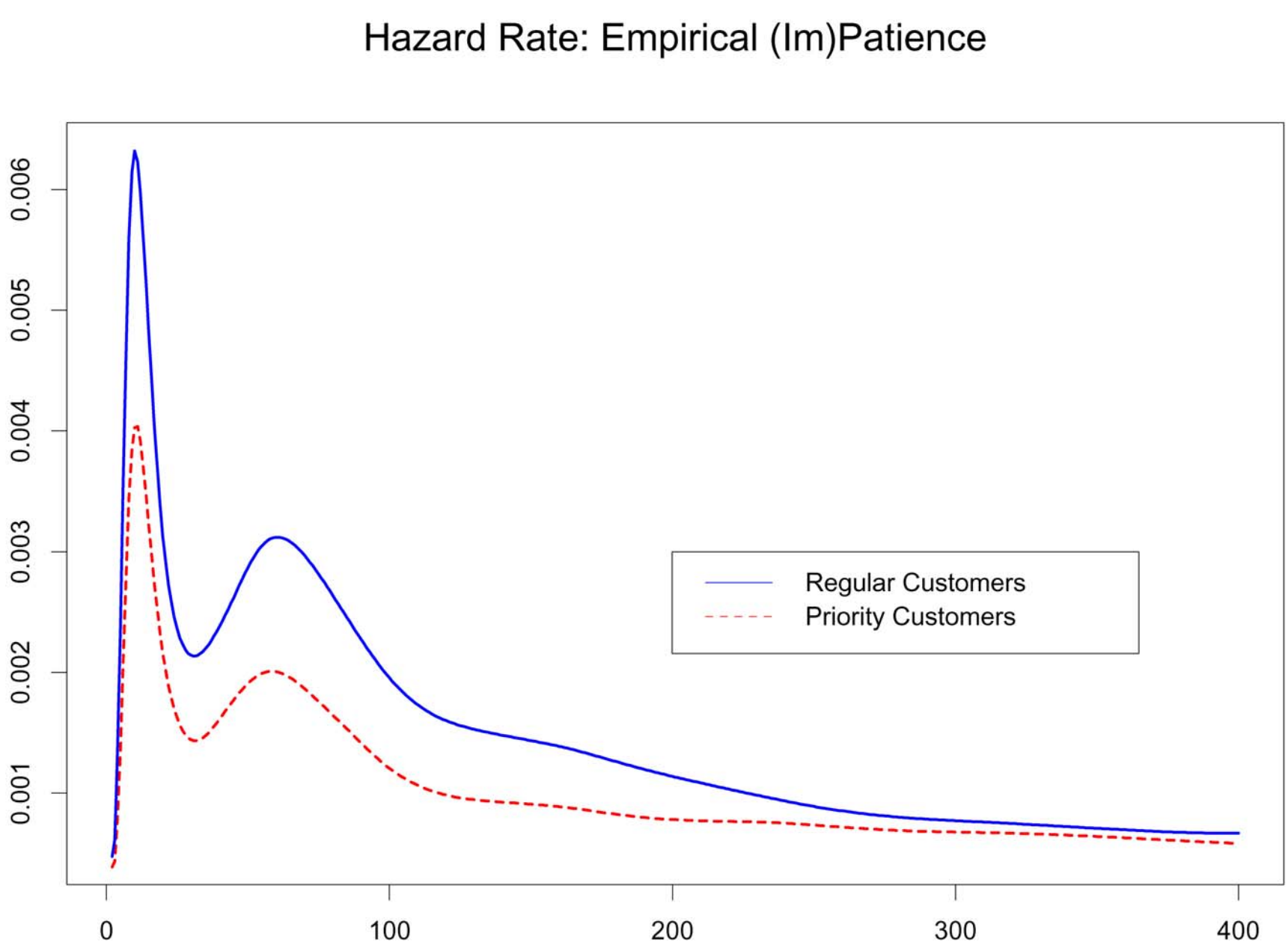
$\alpha = 0$: N = R + γ R **Quality-driven**

Laws of Congestion: Patience

Censored Sampling, or equivalently (under exp)

%(Abandon) = E(Wait) / E(Patience)




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

Abandonment Important

- 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
  - Unstable models (vs. Robustness)

But Abandonment also Interesting & Challenging

- Queueing Science
  (Paradigm: experiment, measure, model, validate)
- Research: OR + Psychology + Marketing
  (Modelling: steady-state, transient, equilibrium)
- Scope of Applications
  - VRU/IVR: opt-out-rates
  - Internet: business-drivers (60% and more)
  - Call Centers: customer-centric performance measures

Staffing the "Modern" Basic Call Center

Erlang-C $N \approx R + y\sqrt{R}, \quad y > 0$

Erlang-A Abandonment, with $-\infty < y < \infty$

Time-Varying y(t) varies with time, over Finite Horizon

General Service Time

And Beyond (conventional Q-Theory):

Skills-Based Routing

Networked (Virtual) call centers

Human Factors (e.g. Information, Learning)

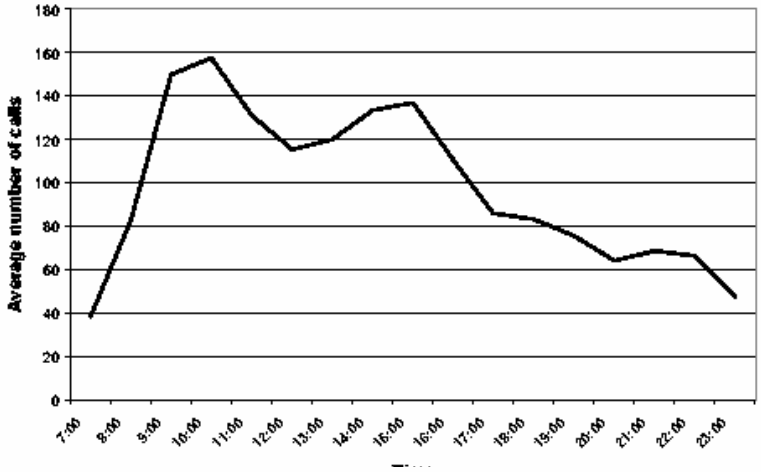
Forecasting

Economics

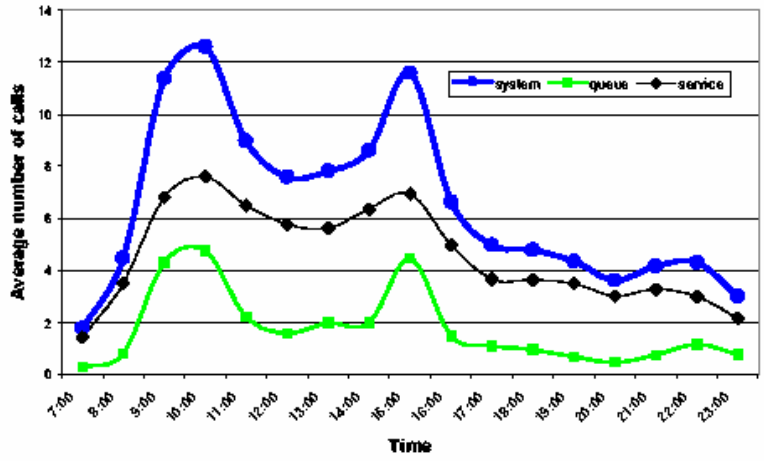
- Paradigm: Theory + Real Data + Experiments = Multi-Disciplinary Queueing Science, in Support of Service Engineering

Finite-Horizon Queues: Predictable Variability

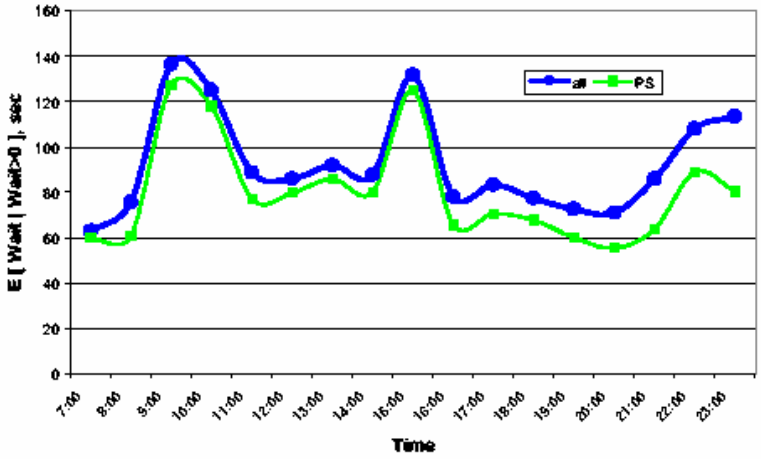
Arrivals



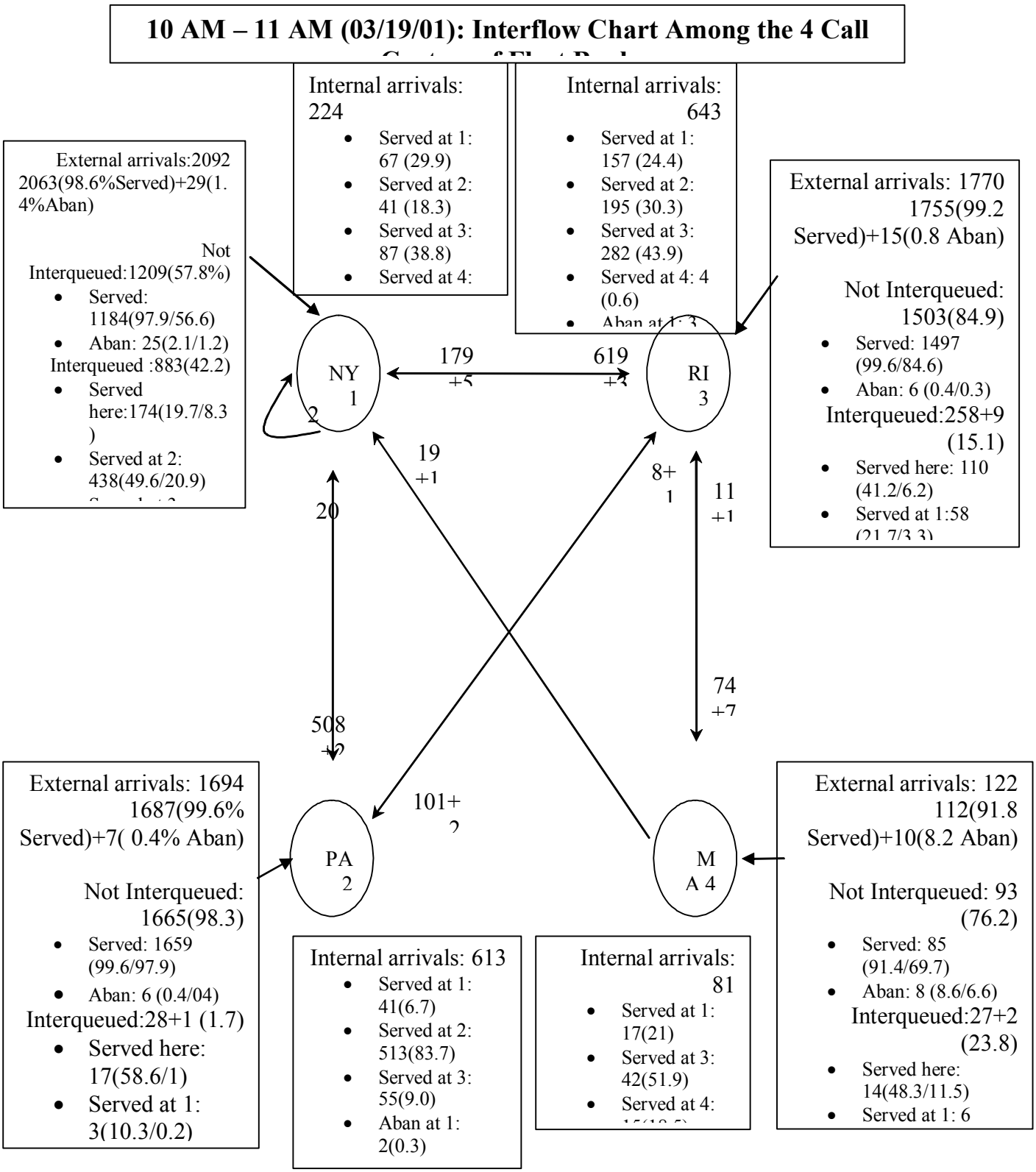
Queues



Waiting



Distributed Call Center: Simultaneous Queueing



Service Engineering (and Management) – A Tentative Syllabus

(November 2002; Previous versions: http://ie.technion.ac.il/serveng)

1. Introduction: Services (vs. Manufacturing), Service Engineering (vs. Science, Management)

- Service Nets = Queueing Nets.

HW: Read "Production Line Approach to Services", by T. Levitt, HBR, 1972.
HW: Read selected sections of the Review on Call Centers, by Gans, Koole and M, 2002.
HW: Read on the Service Economy, and on Services in General.
Rec: Rules of the game (e.g. Group work); Logistics (e.g. computer accounts).

2. The First Prerequisite: Measurements

- Measures of Performance;
- Transaction-based Data: Call Center, Bank Branch, Internet Site;
- Modeling a Service Station over a Finite Horizon: Empirical (Fluid) Models.

Rec: Bank KeyCorp Video (SEMS = Service Excellence Measurement Systems).
HW: Capacity Utilization, Inventory Buildups, Bottleneck Analysis, Little's Law
HW: Read, Prepare "Improving the N.Y. Arrest to Arraignment System", by R. Larson, 1993.
HW: Read, Prepare "Modeling Court Delays", by S. Flanders, Law and Policy, 1980.

3. The Second Prerequisite: Models

- The Processing Network Paradigm (BPR): Dynamic Stochastic DS/PERT Networks; Project Management; Product and Service Development;
- The Operational Sources of Congestion: Scarce Resources, Synchronization Gaps.

Rec: Forecasting demand (workloads).
HW: Forecasting workload for a call center.
HW: Read on New Service Design and Development

4. Deterministic (Fluid) Process Models of a Service Station, Part 2: Networks

- Lindley's and Skorohod 's representations;
- Transient Analysis: rush hour; Fluid Approximations (e.g. a call center);
- Bottlenecks in steady state – the Traffic Equations;

Rec: Descriptive models of a face-to-face service in a bank.
HW: Descriptive models of a telephone call center

5. Demand: Dynamic Randomness (The Poisson Process)

- Scaling and centering of data: the framework of Levy Processes; (the Brownian Motion).
- Poisson Processes: Predictable and stochastic variability; PASTA, Biased Sampling.

Rec: Statistical analysis of arrivals to a bank (face-to-face) service
HW: The cost of Stochastic Variability – Biased Sampling

6. Service Durations and Customer Patience

- Job Design (IE, HRM); The effect of service-durations on performance.
- Human Factors (Psychology): The effect of abandonment on performance.
- Review: Markov Jump-Processes (mainly Absorbing);
- Hazard Rates: a dynamic characterization of a distribution.

Rec: AT&T CAPS Video of support tools for call centers. Review and Preview of HW.
HW: Statistical analysis of a bank call center (Arrivals, Services, Abandonment).
HW: Read "The Psychology of Waiting Times", D. Maister, 1985.

7. A Service Station in Steady State: Markovian Models

- Modeling a Service Station in Steady-State: Stochastic (Birth & Death) Models
- Multi-server queues (Erlang-B, C, A; Ample Servers)
- Human Factors: Impatience, Slow Servers; Ancillary Activities

Rec: Review Markov Jump-Processes (mainly Ergodic)
HW: Theoretical Analysis of (Markov Jump-Process Models for) Service Operations

8. A Service Station in Steady State (Continued): G/G/. Queues
Laws of Congestion

- The Cost of Variability – continued (M/M/. vs. M/D/; Appointments/Reservations)
- Congestion/Accessibility Curves – trading-off (operational) service-quality vs. efficiency
- The first Laws of Congestion

Rec: Solution of Theoretical Analysis
HW: GazolCo' Call Center

9. Workforce Management: Staffing

- Regimes of Operations: Quality-Driven, Efficiency-Driven, QED
- Economies of Scale
- Ample-Server Heuristics: Stationary, Time-Varying
- Dimensioning Medium to Large Service Operations

Rec: The Queueing Inference Engine (QIE) – Larson, and Network Extensions
HW: Staffing a Call Center: Small (Israel), Medium (Italy), Large (U.S.A.)
HW: Read sections in Call Center Review Paper

10. Service Networks

- Design, Control (Jackson Nets; QNA = Non-Parametric Jackson Nets)
- Process and Organization Design: Pooling (Specialization vs. Flexibility)
- Priorities: exact and approximate analysis

Rec.: Compromises in modeling: Pronto (Anonymous) Pizza.
HW: Hand out FINAL PROJET = Simulation + Analysis of a Q-NET
HW: Introduction to Skills-based Routing and its Operational Complexities
HW: Prepare Nations-Bank Case
HW: Read sections in Call Center Review Paper

11. Customer Relations Management; Skills-Based Routing

- Skills-based routing: teaching note
- Skills-based routing: efficiency-driven and QED
- CRM: Marketing and/or MIS perspective, or

(See Chapter 6 in Hope and Muhlemann, for a chapter on the interface with Marketing, in the context of Service Design).
Rec.: QDF, and other tools from Marketing Engineering
WH: Nations Bank Analysis

---

12. Revenue Management; Pricing

- Life-time value of customers; guarantees
- Field support, Technical support

13. Design and Development of New Service (Processes)

14. Quality Management; HRM; Designing the User Interface

- Design of the Intangibles (Emotions, Culture,…) – The Psychological Cost of Delay
  Or Selection, Training, Incentives
- Introduction to Human Resource Management/Engineering

(See Chapter 3 of Voss, Armistead, … for a Chapter on HRM in the context of Operations)
HW: AT&T Universal Card (HBS Case, on Over-Measurements and Incentives)

---

15. Service Technology and Automation: at the User Interface and Within

- The "New-Age" Industrial Engineer, as manifested in Banking
- E-Commerce; Tele-Marketing;
- Multi-media (Contact) Centers

16. Supply-Chain Management; The Service Value-Chain

17. Specialized Service Sectors:

- Banking, Insurance, Healthcare, Public

18. Functional Interfaces:

- Human Resources Management (Incentives, Hiring, Training)
- Consumer Psychology (Psychology, Marketing)
- Marketing Interface: CRM, New Products and Processes, Techniques (QFD,…)
- MIS (Self-Service)
- Industrial Engineering: Ergonomics, Service Efficiency (DEA),…

The Future: Research, Teaching, Practice

Research

Multi-Disciplinary

IE/OR/OM, Marketing, HRM, MIS

Teaching

Advanced courses, leading ultimately to Core

Data-Based ("Data = The Language of Nature")

Syllabus for an "Introduction to Service Engineering"?

Research- and Case-Based

Practice

Increased professionalism (Outsourcing)

A profession: Service Engineering

Who, Where: Engineering vs. Business Schools

END OF LECTURE

QED: $\sqrt{\cdot}$ Safety-Staffing

$$N = R + \underbrace{\beta\sqrt{R}} \qquad \beta = \text{“service-grade”} > 0$$

$$= R + \Delta \qquad \sqrt{\cdot} \quad \text{safety-staffing}$$

$$R = \lambda \times E(S) \qquad \text{Offered load} \quad \text{(Erlangs)}$$

Expected Performance (trivial to calculate):

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

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

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

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

Economies of Scale

Base case: M/M/N with parameters λ, μ, N

Scenario: $\lambda \to m\lambda$ $(R \to mR)$

	Base Case	Efficiency-driven	Quality-driven	Rationalized
Offered load	$R = \frac{\lambda}{\mu}$	mR	mR	mR
Safety staffing	Δ	Δ	$m\Delta$	$\sqrt{m}\Delta$
Number of agents	$N = R + \Delta$	$mR + \Delta$	$mR + m\Delta$	$mR + \sqrt{m}\Delta$
Service grade	$\beta = \frac{\Delta}{\sqrt{R}}$	$\frac{\beta}{\sqrt{m}}$	$\beta\sqrt{m}$	$\boxed{\beta}$
Erlang-C = P{Wait>0}	$P(\beta)$	$P\left(\frac{\beta}{\sqrt{m}}\right) \uparrow 1$	$P(\beta\sqrt{m}) \downarrow 0$	$\boxed{P(\beta)}$
Occupancy	$\rho = \frac{R}{R+\Delta}$	$\frac{R}{R+\frac{\Delta}{m}} \uparrow 1$	$\boxed{\rho = \frac{R}{R+\Delta}}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \uparrow 1$
$\text{ASA} = \text{E}\left[\frac{\text{Wait}}{\text{E}(S)} \middle\vert \text{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \text{ASA}}$	$\frac{1}{m\Delta} = \frac{\text{ASA}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\text{ASA}}{\sqrt{m}}$
$\text{TSF} = \text{P}\left\{\frac{\text{Wait}}{\text{E}(S)} > T \middle\vert \text{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \text{TSF}}$	$e^{-mT\Delta} = (\text{TSF})^m$	$e^{-\sqrt{m}T\Delta} = (\text{TSF})^{\sqrt{m}}$

Strategy: Sustain Regime through Pooling

PATIENCE INDEX

- How to Define? Measure? Manage? (Israeli Data Base)

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

$\uparrow$ definition $\qquad$ $\uparrow$ measure

- Supported by ongoing research (with Brown, Haipeng, Zhao).

Designing Call/Contact Centers with Impatient Customers:

10 Years History, or A Modeling Panorama

1. Kella, Meilijson: Practice ⇒ Abandonment important
2. Shimkin, Zohar: No data ⇒ Rational patience in Equilibrium
6. Carmon, Zakay: Cost of waiting ⇒ Psychological models
7. Garnett, Reiman: Palm/Erlang-A to replace Erlang-C/B as the standard Steady-state model
8. Massey, Reiman, Rider, Stolyar: Predictable variability ⇒ Fluid models, Diffusion refinements
9. Ritov, Sakov, Zeltyn: Finally Data ⇒ Empirical models
10. Brown, Gans, Haipeng, Zhao: Statistics ⇒ Queueing Science
11. Garnett, Atar, Reiman: Skills-based routing ⇒ Control models
12. Nakibly, Meilijson, Pollatchek: Prediction of waiting ⇒ Online Models and Real Time Simulation
13. Garnett: Practice ⇒ 4CallCenters.com