Skills-Based Routing and its Operational Complexities

Service Engineering

Wharton's Call Center Forum

May 9, 2003

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

Website: http://ie.technion.ac.il/serveng(2003)

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

Supporting Material (in Web-Site)

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

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

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

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.

Yahalom and M.: "Optimal Scheduling for a Multi-server Multi-class Non-preemptive Queueing System", in preparation.

Armony and M.: "Optimal Routing, Staffing and Networking in the QED (Halfin-Whitt) Regime: Homogenous Customers and Heterogeneous Servers", in preparation.

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 some

   Workforce Management (Staffing)

3. Operational Regimes: Quality-Driven, Efficiency-Driven

   QED (Quality and Efficiency Driven; Haflin-Whitt)

4. Square-Root Staffing Laws: Scale Economies (Erlang)

5. Efficiency-Driven SBR: Index strategies in the General Case

6. QED SBR: Only special Cases (I, N, V, Upside-Down V)

7. Beyond Conventional Queueing Theory:

   Abandonment, Retrials; Time-Varying Queues

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.

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

Common Performance

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN Date: 7:00 pm WED MAR 10, 1999
Skill: 37
Skill Name: !BA AUTH1 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 Date: 7:00 pm WED MAR 10, 1999
Skill: 46
Skill Name: !BA AUTHORIZATION 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 Date: 7:01 pm WED MAR 10, 1999
Skill: 33
Skill Name: GA Authorization 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

NationsBank CRM:
What are the relationship groups?

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

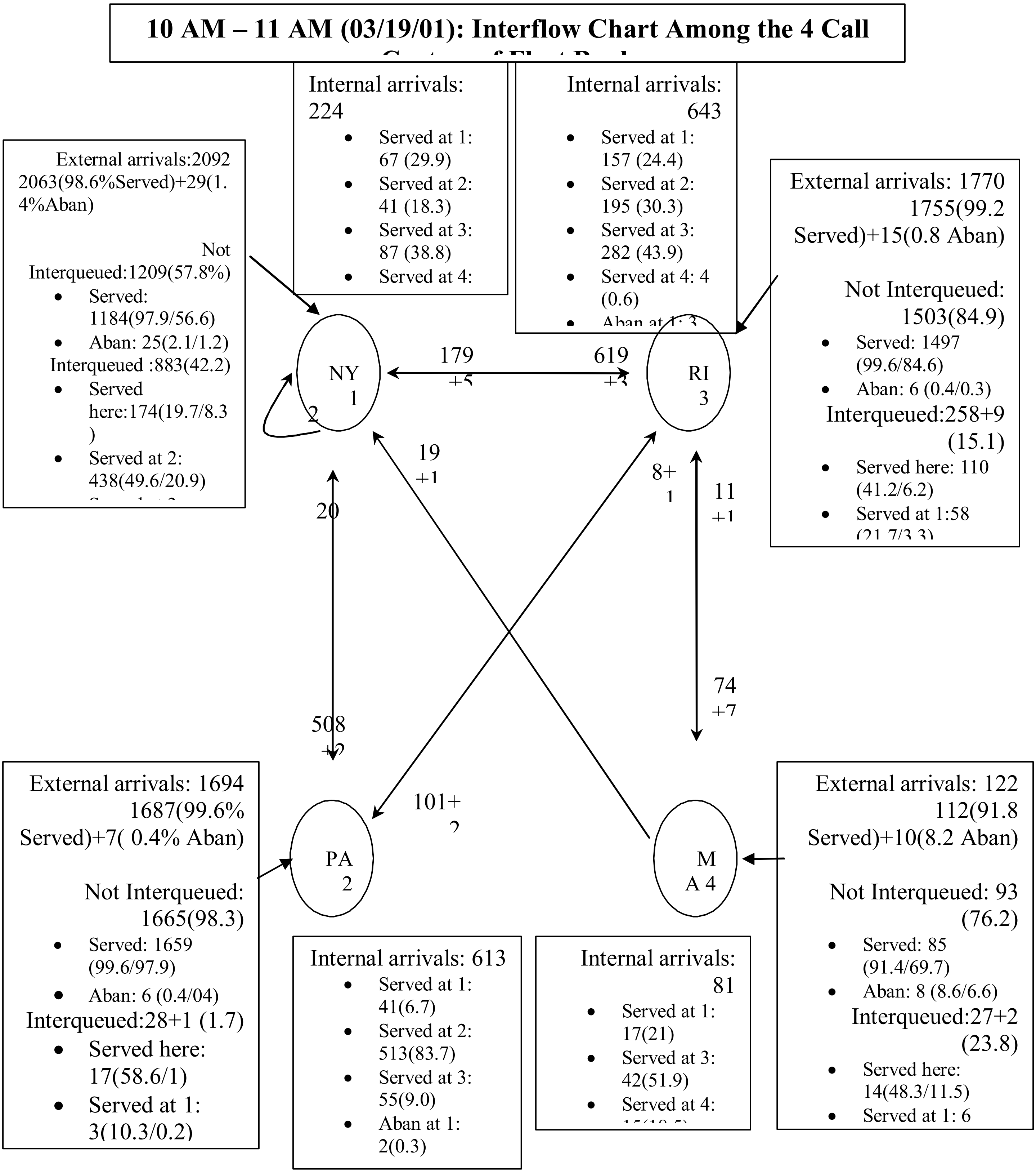
- What does it mean for a customer in each group to be profitable? Customer Revenue Management

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: Member1



Workforce Management: Hierarchical Operational View

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

Staffing: Queueing Theory

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

Service Engineering
May 2000; Under Revision

An Introduction to Skills-Based Routing and its Operational Complexities

By Ofer Garnett and Avishai Mandelbaum

Technion, ISRAEL

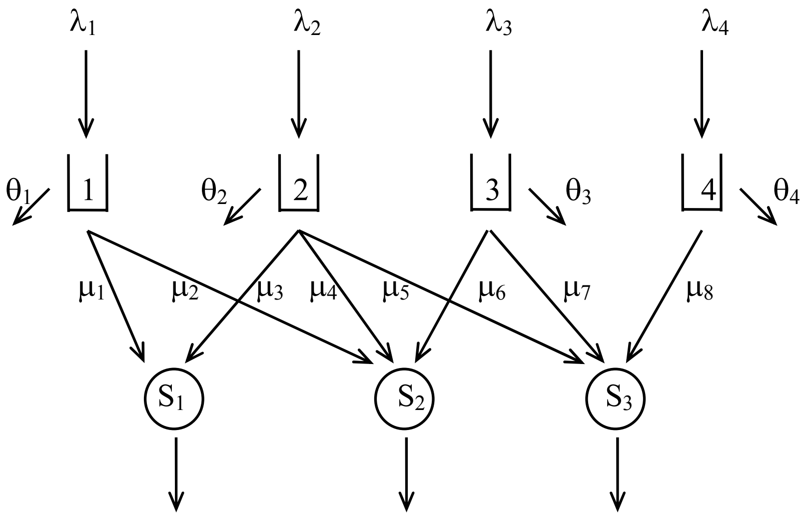
(**Full** Version)

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.

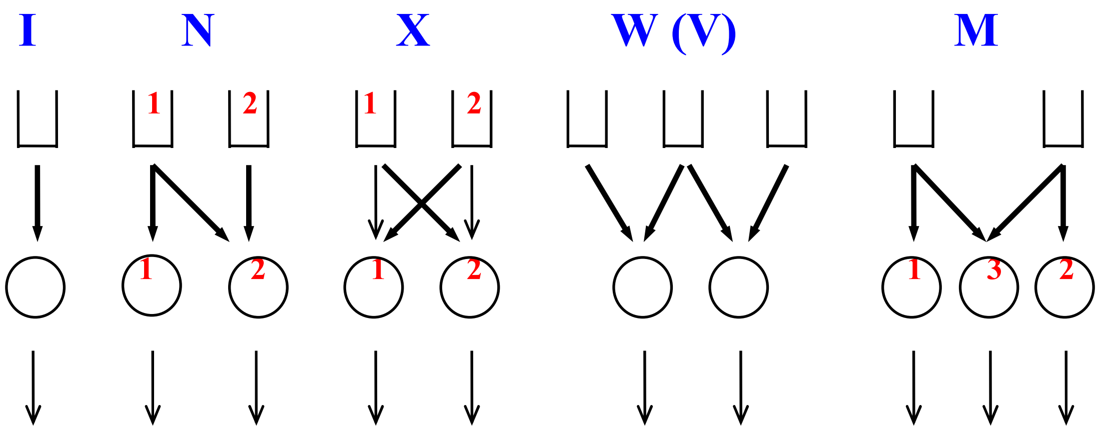
Skills-Based Design:

- **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



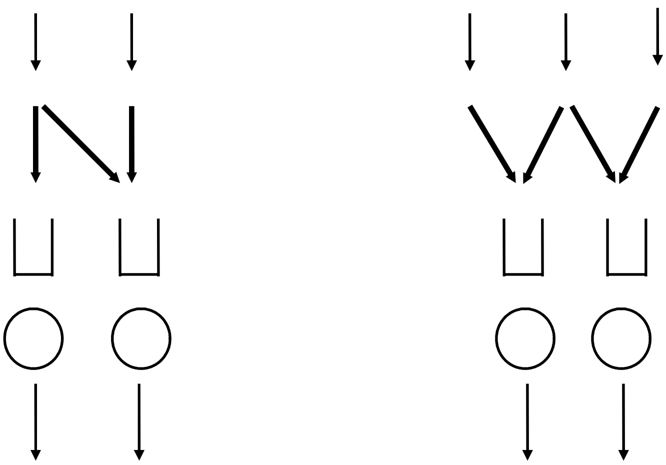
I – dedicated (specialized) agents

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
3. Prerequisite Infrastructure - MIS / IT / Data-Bases (**CS, Statistics**)
   CTI, ERP, Data-Mining

Major Control Decisions

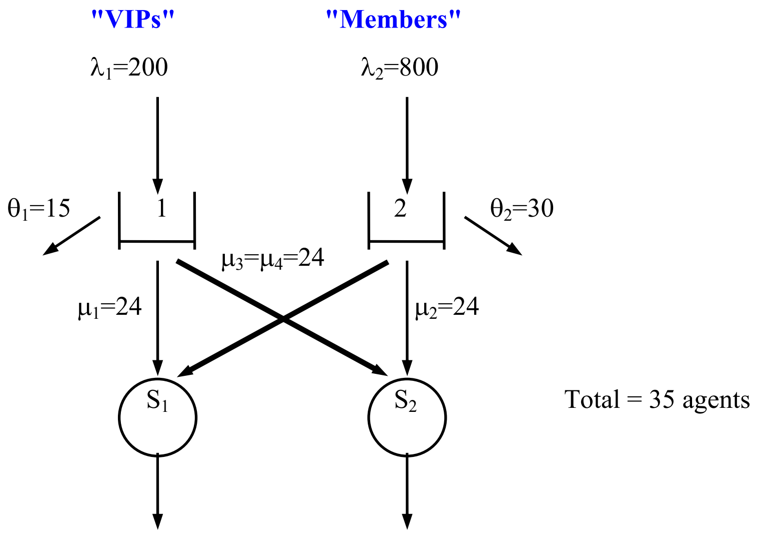
4. Matching customers and agents (**OR**)
   - Agent **Scheduling**: Whenever an agent turns idle and there are queued customers, which customer (if any) should be routed to this agent.
   - Customer **Routing**: Whenever a customer arrives and there are idle agents, which agent (if any) should serve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

Multidisciplinary Challenging Research

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

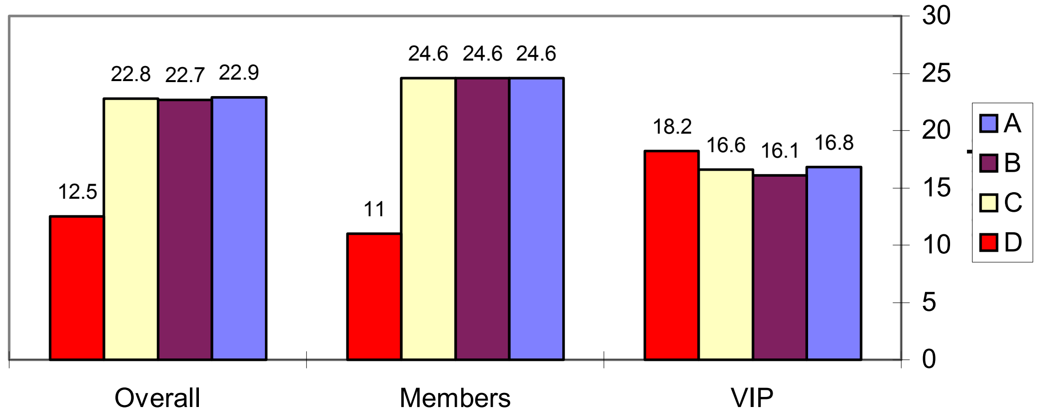


Chart 3 : 1000 Calls - Abandonment

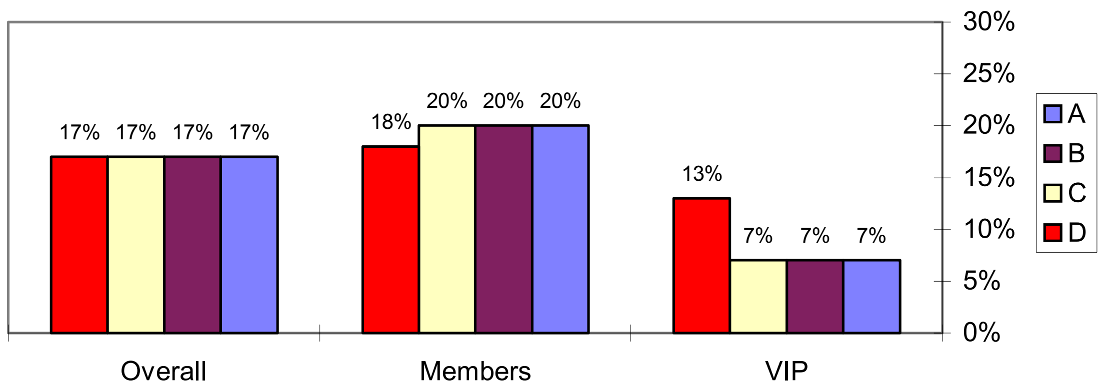
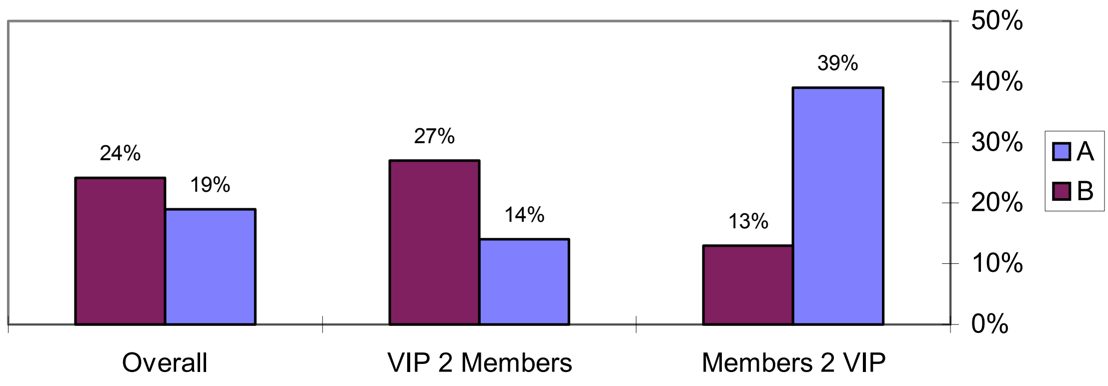


Chart 4 : 1000 Calls - Overflows



WHAT IF : 1500 Calls/hour - ASA

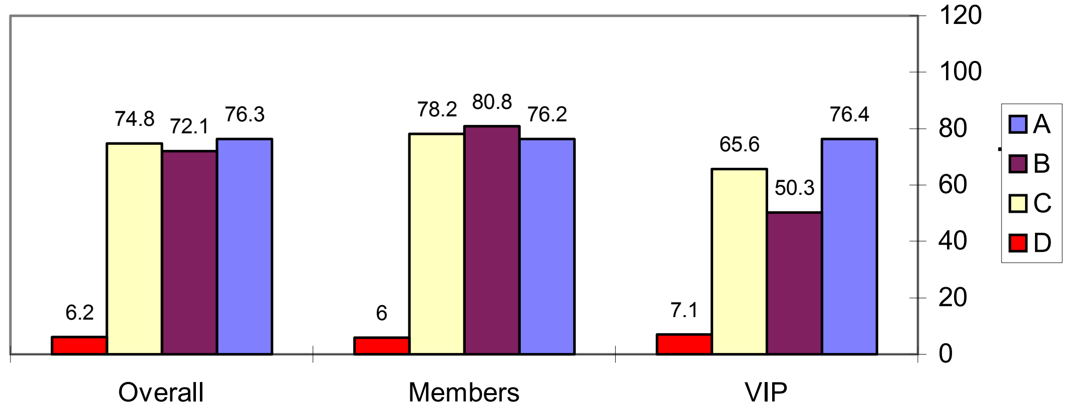


Chart 7 : 1500 Calls - Abandonment

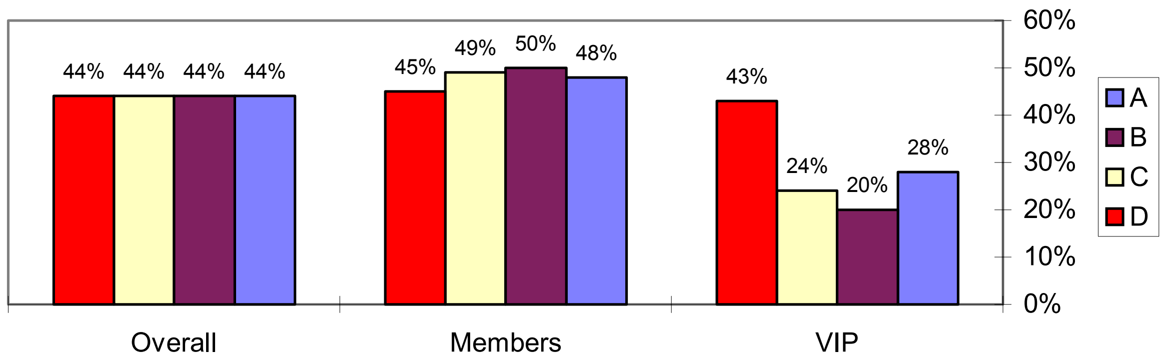
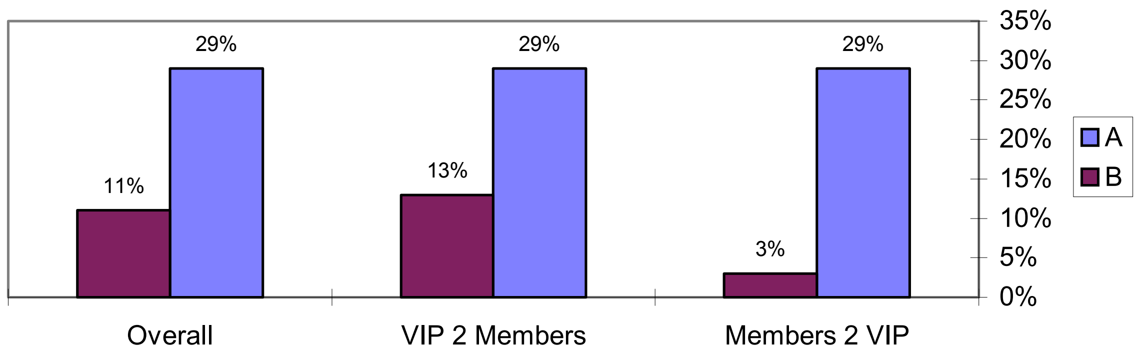


Chart 8 : 1500 Calls - Overflows



Reality

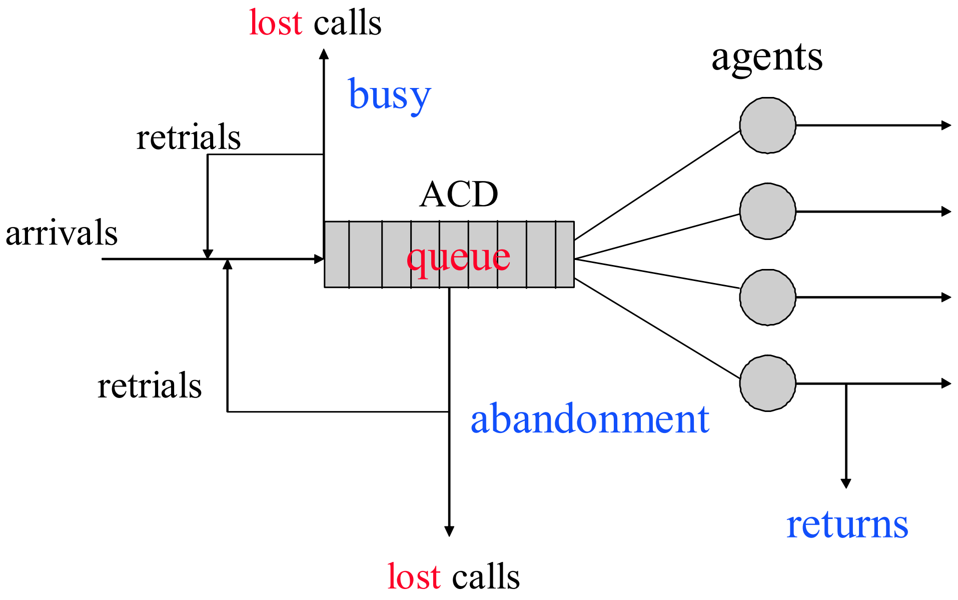
- Technology enables smart systems
- Reality becomes increasingly complex
- Solutions are urgently needed
- Theory lags significantly behind needs

- Ad-hoc methods: heuristics, simulation-based

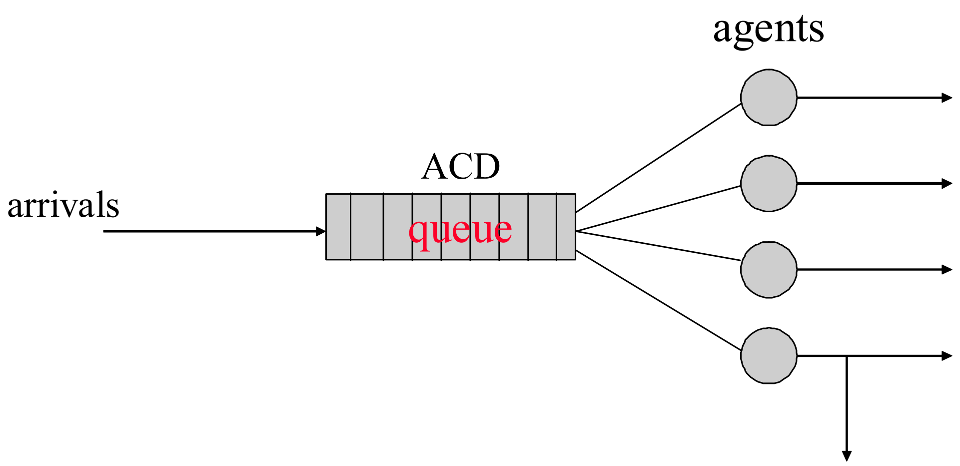
Research Status

- 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

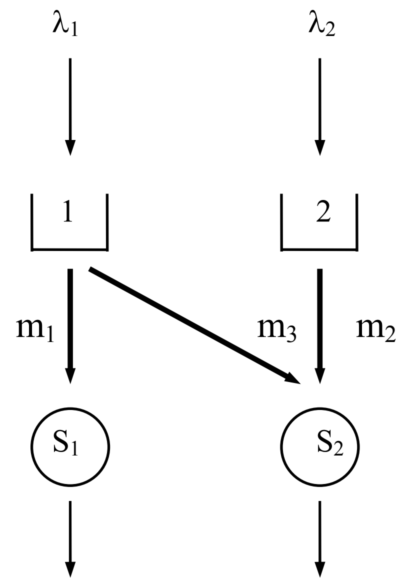
The Basic Call Center



Erlang-C = M/M/N



Static Priorities (Cross-Training): Some Subtleties

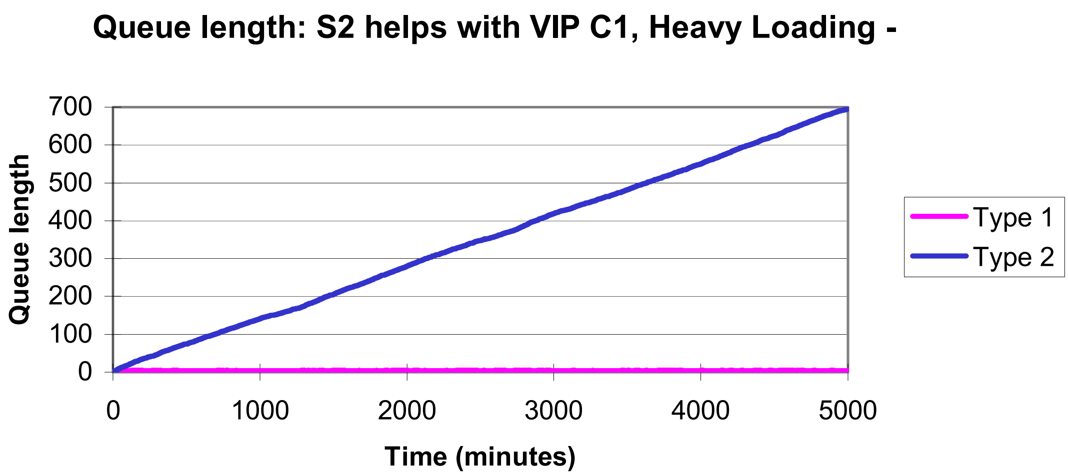


$\lambda_1 <$ 1.3 $\quad \lambda_2 <=$ 0.4

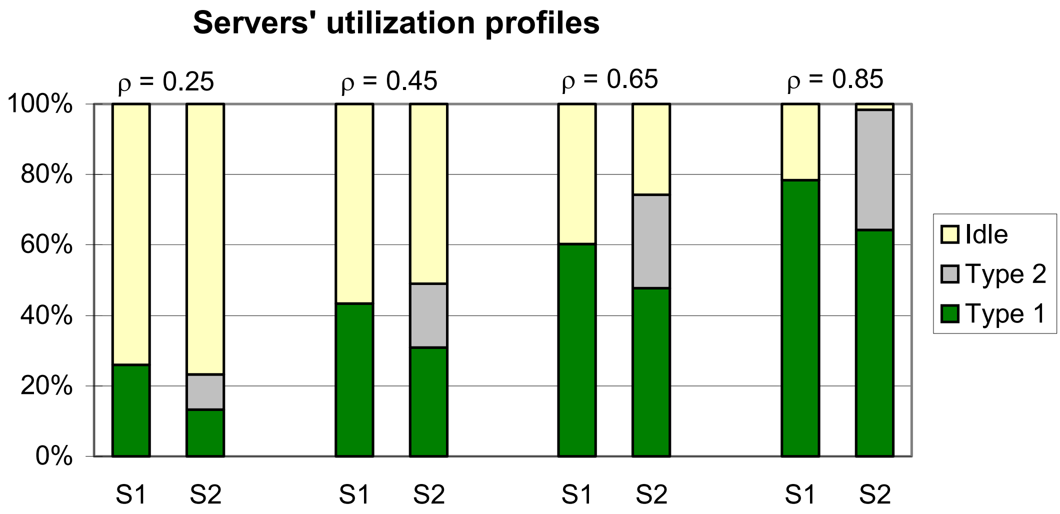
$m_1 = m_2 = 1$, $\quad$ **$m_3 = 2$**

$S_1 = S_2 = 1$

- 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 ?

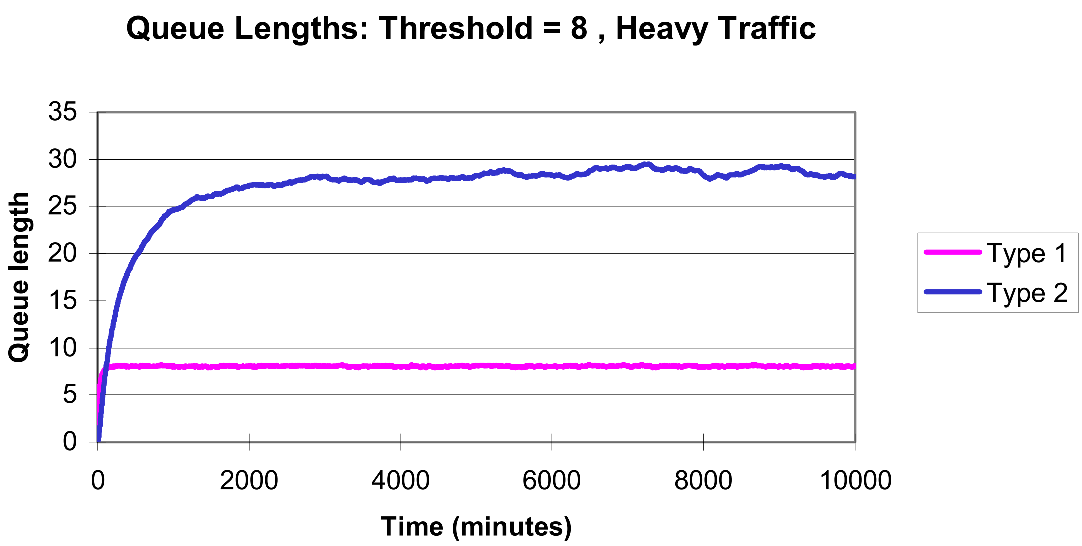


Instability: S2 **overworked** serving C1 and neglecting C2,

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

Efficiency-Driven SBR - the "EASY" Case

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

- 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

(at least should not be) Efficiency-Driven !?

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 W	Avg Aban Time	ACD Calls λ	Avg ACD Time 1/μ	Avg ACW Time	Aban Calls #Aban	% ACD Time	% Ans Calls	Avg Pos Staff 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	418	:03:49	:00:28	1	63	97	94	4	98	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	:03:59	:00:32	0	51	99	98	4	99	11	6	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	60	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:48	:01:37	0	54	100	98	4	100	6	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	46
5:30 PM*	:00:01		387	:03:58	:00:19	0	58	99	96	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

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

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

$= 50$ Erlangs

Occupancy $\rho = R/N$

$= 50/100 = 50\%$

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

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

$\Rightarrow$ **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%

$\Rightarrow$ **Efficiency-driven Operation** **(Heavy Traffic)**

Above: $R = 99$, $N = R + 1$, $\approx$ all delayed.

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

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%**

$\Rightarrow$ **New Rationalized Operation**

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

Light traffic, 50% answered immediately

QED Regime = Quality- and Efficiency-Driven Regime

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

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

Rules of Thumb: Operational Regimes

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

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

$$N = \lceil R + \gamma \rceil, \qquad \gamma > 0 \quad \text{service grade}$$

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

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

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

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

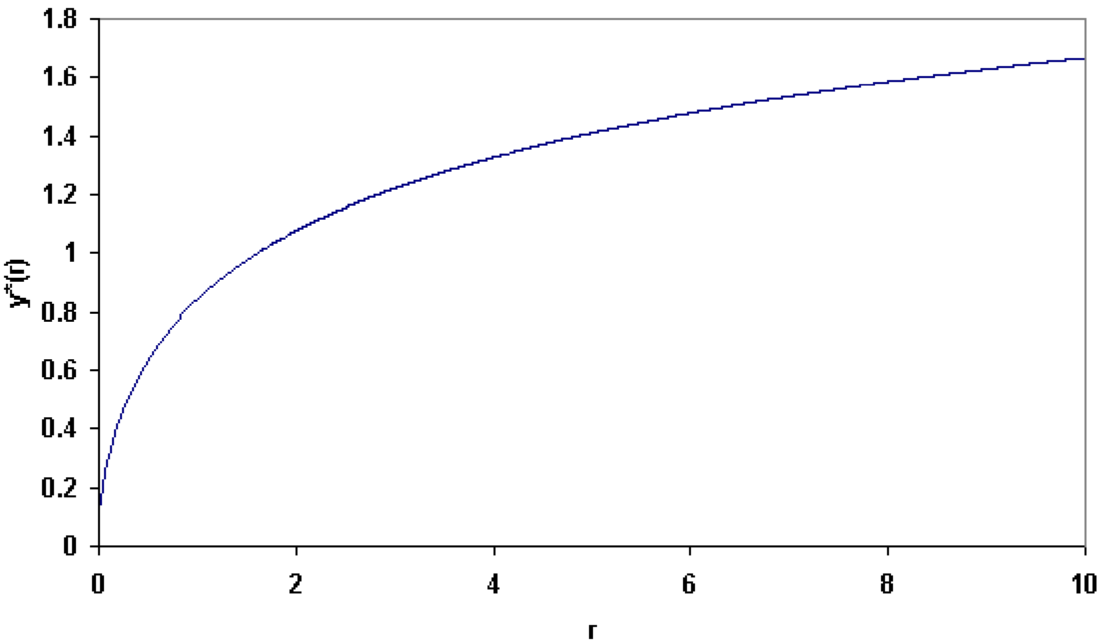
Determine Regimes (Strategy), Parameters (Economics)

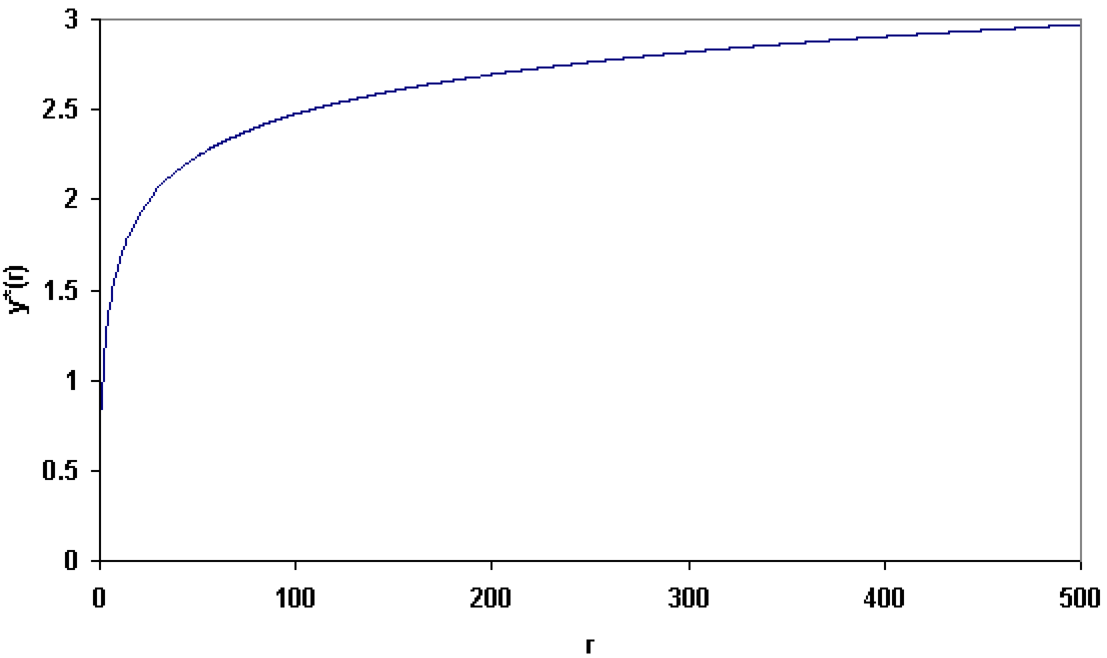
Strategy: Managers, Agents (Unions), Customers

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)



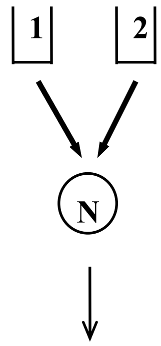


QED Relevance: Member2

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%

V-Design: Pure Scheduling



N agents, fully flexible

C1 = VIP

Optimal Scheduling: **Agent Reservation**

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

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

QED regime: $\sqrt{\cdot}$ Safety-Staffing, as before.

Threshold Size (relative to N) determines Service Levels:

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

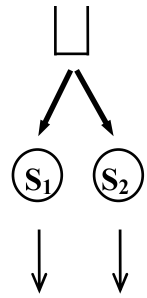
- Small: C1 and C2 indistinguishable QED

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

Upside-Down-V Design: Pure Routing

Homogeneous Customers

Heterogeneous Agents: **S1 = Faster**



Optimal Routing: **"Slow-Server" phenomenon**

- S1(=Fast) always employed, if possible;

- S2(= Slow) employed if # in queue exceeds a threshold.

QED regime: $\sqrt{\cdot}$ Safety-Staffing for S1+S2.

- No threshold needed: just have all servers work

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

- Can do also detailed staffing: how many S1 and S2.

- Distributed call centers: similar

But N-Design active challenging research

Beyond Erlang-C: (Im)Patience (Erlang_A)

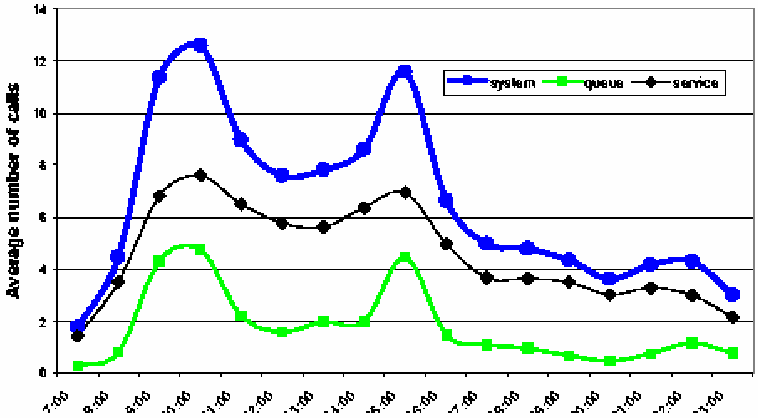
6/13/00 - Tue

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%

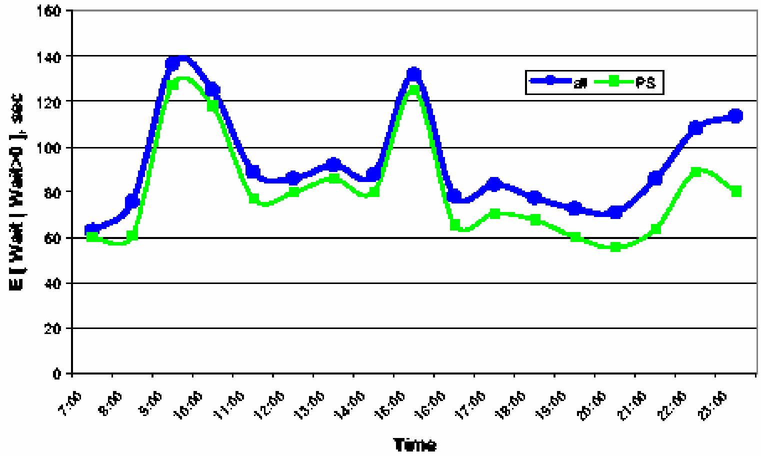
Beyond Erlang-C: Predictable Variability

Arrivals

Queues



Waiting



END OF LECTURE

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

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

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

- **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**

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