Skills-Based Routing and its Operational Complexities

Service Engineering

Eurandom

September 8, 2003

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

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

4. 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**)

Atar, M. and Reiman: “Scheduling a Multi-Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy-Traffic.” Submitted to Annals Appl Prob, 2002. (**V-Design**, with customer-driven services); see also Harrison & Zeevi.

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**, with iid Servers).

Yahalom and M.: "Optimal Scheduling of a Queueing System with Heterogeneous Customers, Multiple Homogenous Servers and Non-preemptive Service", in preparation. (**V**, iid servers)

Contents

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

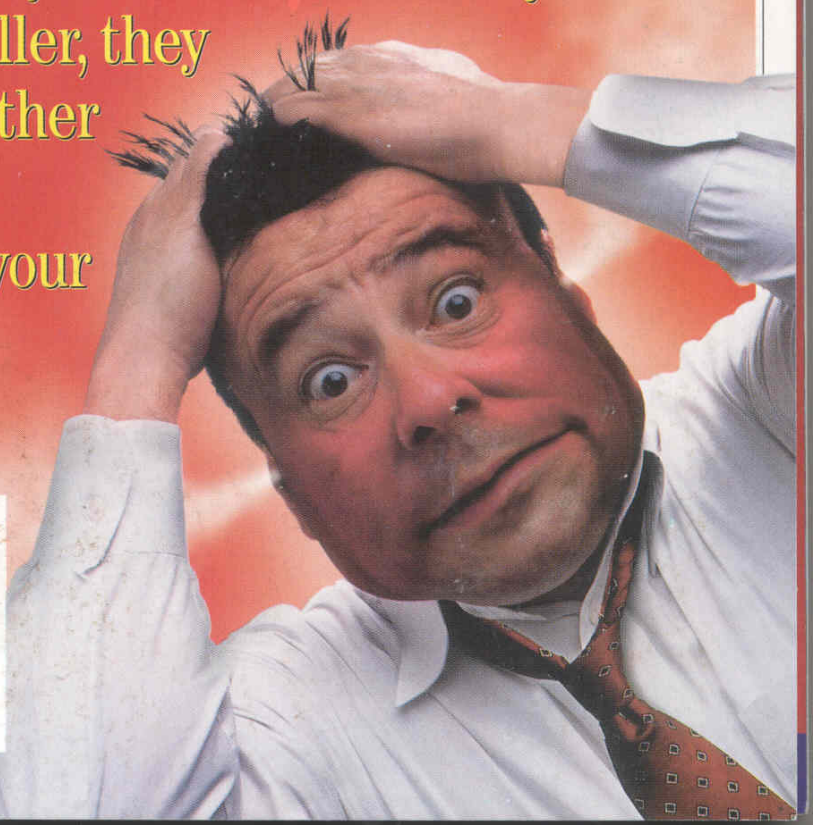
Managed Care

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

NationsBank CRM:
What are the relationship groups?

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

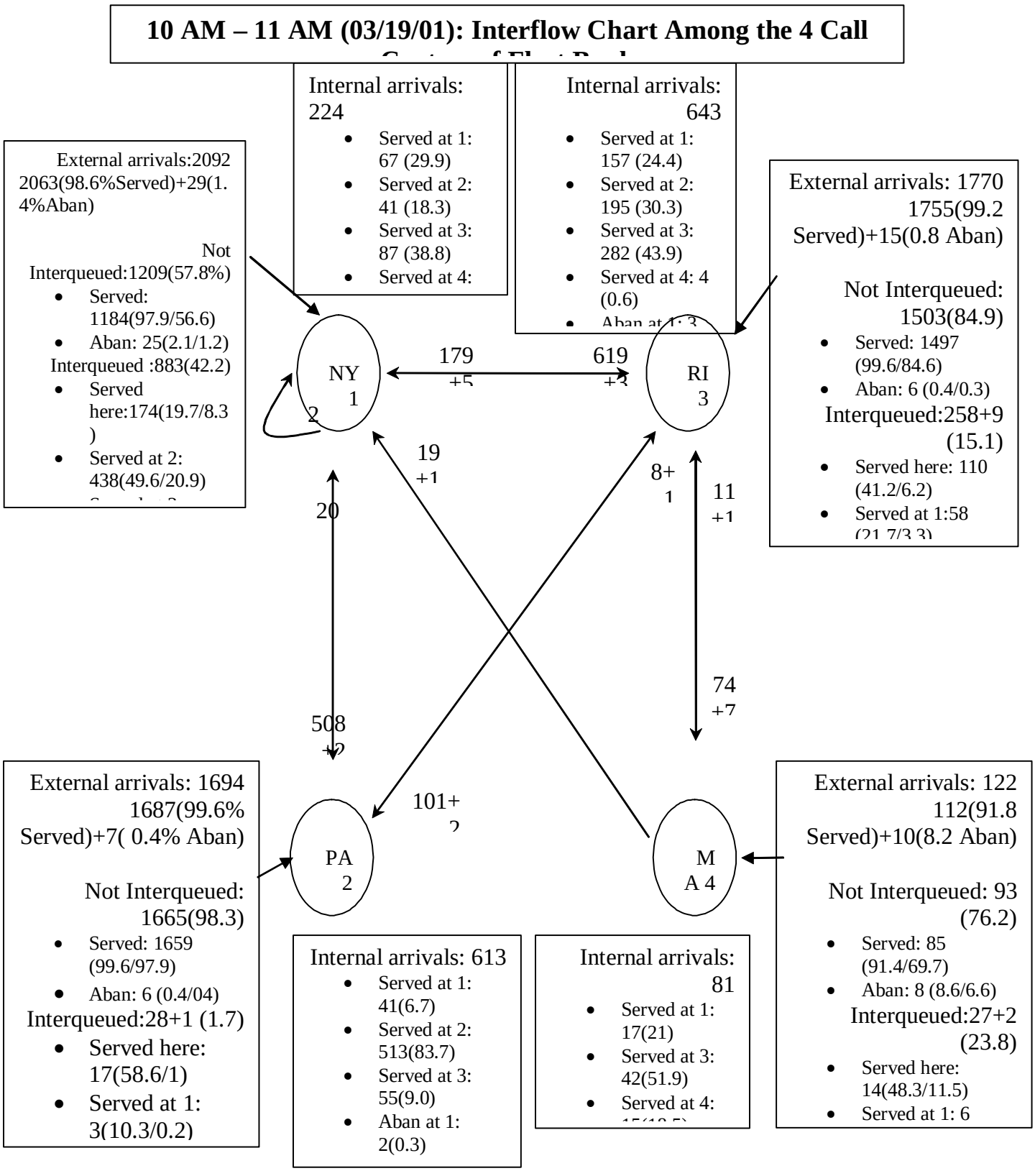
- What does it mean for a customer in each group to be profitable? Customer Revenue Management

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

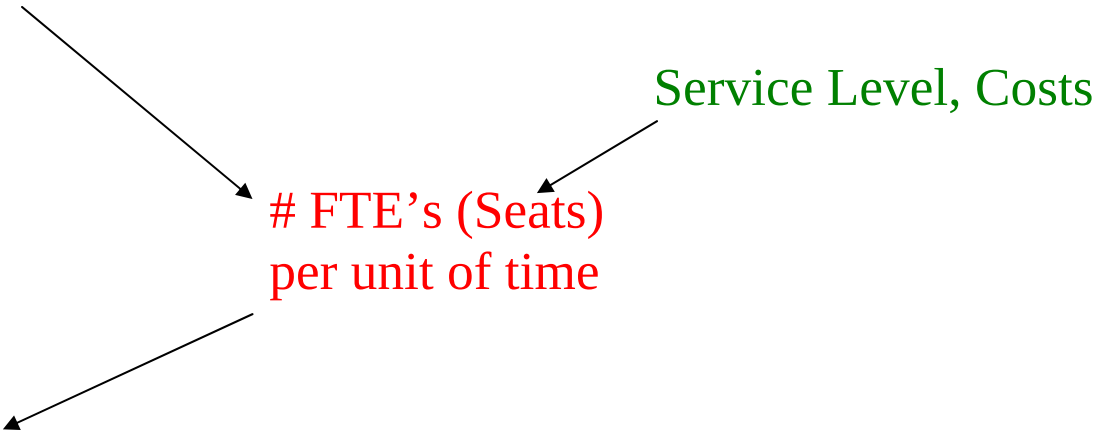
Distributed Call Center: Member1



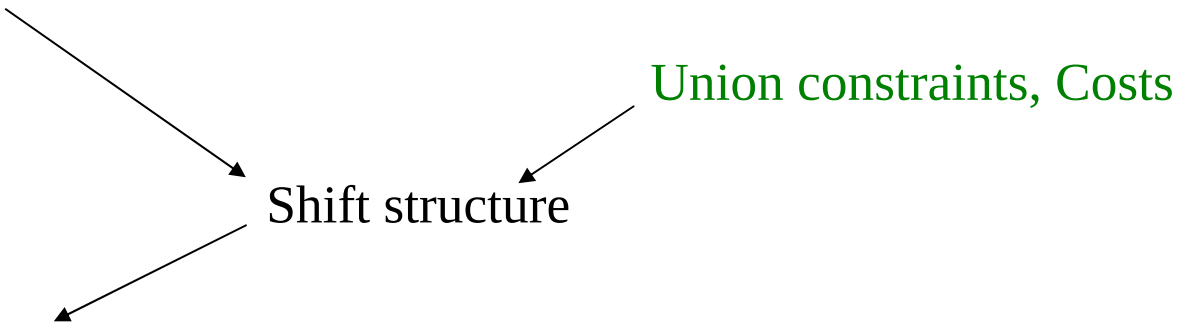
Workforce Management: Hierarchical Operational View

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

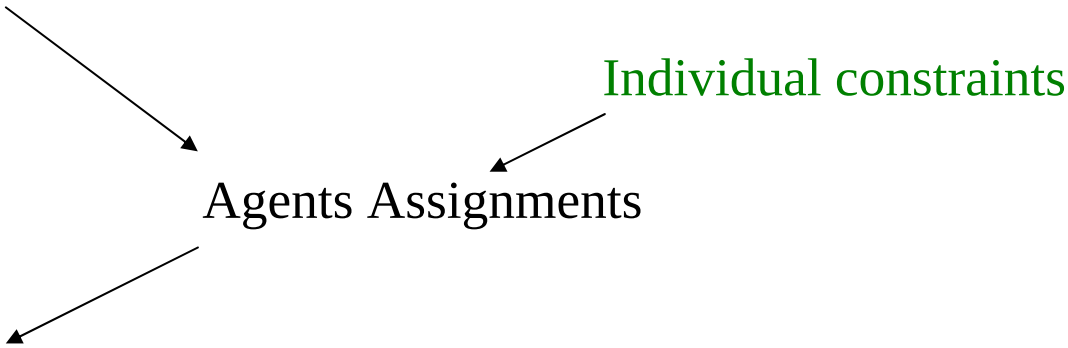
Staffing: Queueing Theory



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

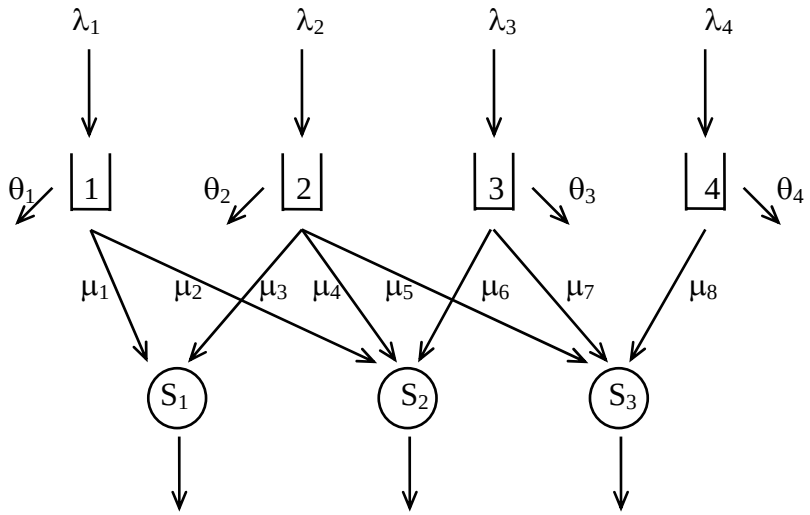
(**Full** Version)

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.

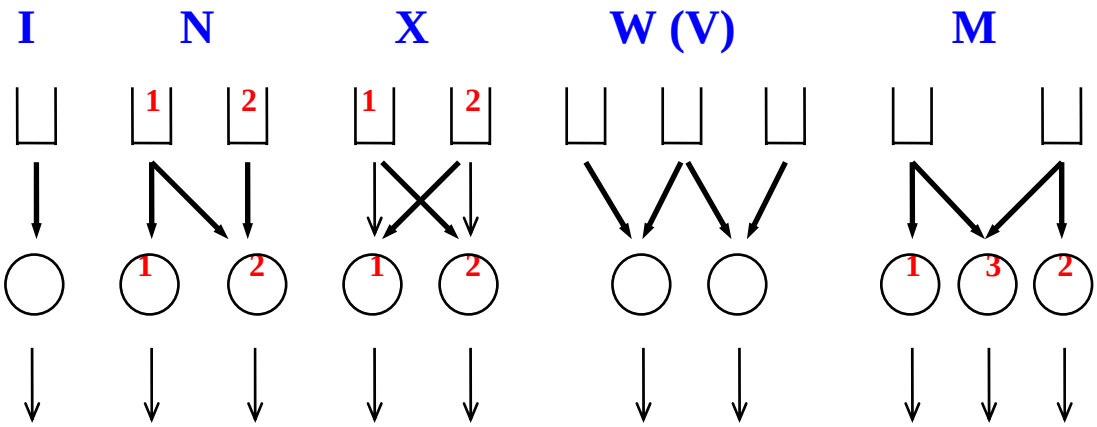
Skills-Based Design:

- **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



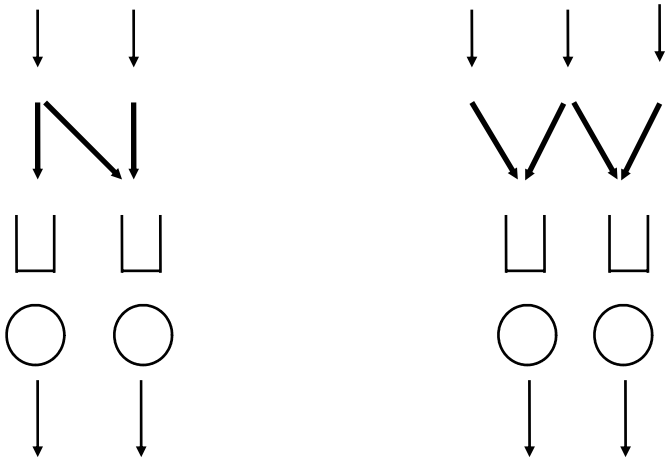
I – dedicated (specialized) agents

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

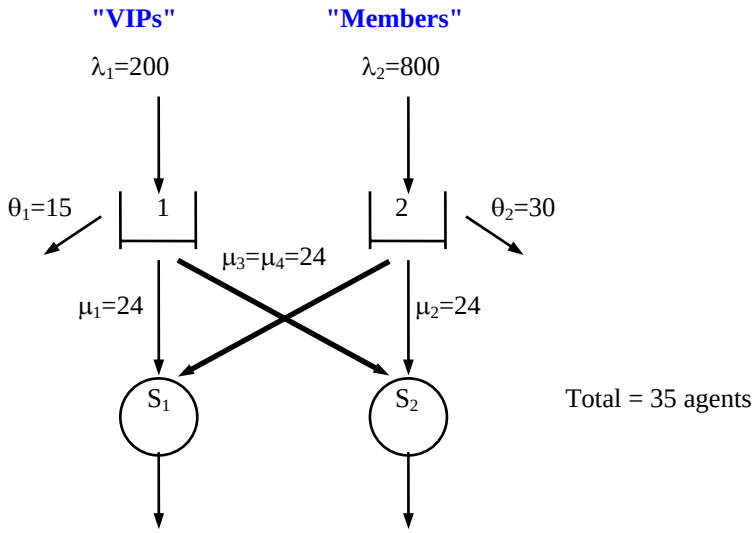
4. Matching customers and agents (**OR**)
   - Customer **Routing**: Whenever an agent turns idle and there are queued customers, which customer (if any) should be routed to this agent.
   - Agent **Scheduling**: Whenever a customer arrives and there are idle agents, which agent (if any) should serve 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

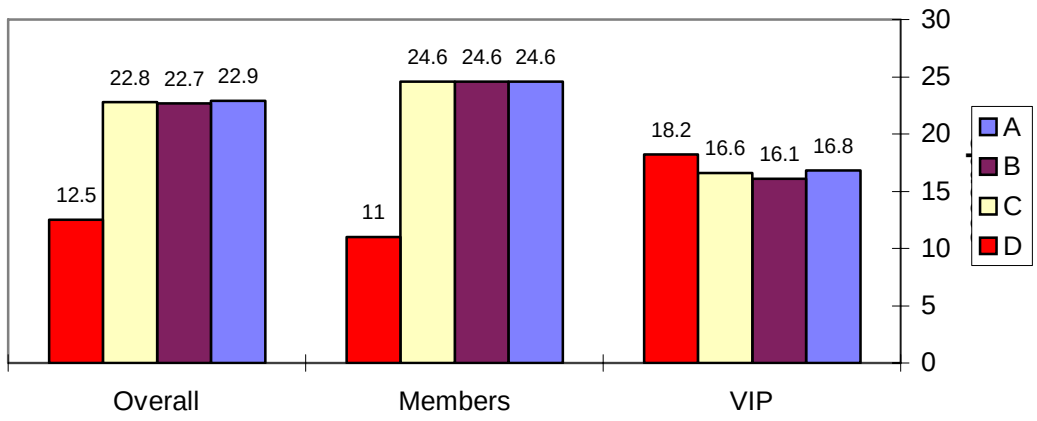


Chart 3 : 1000 Calls - Abandonment

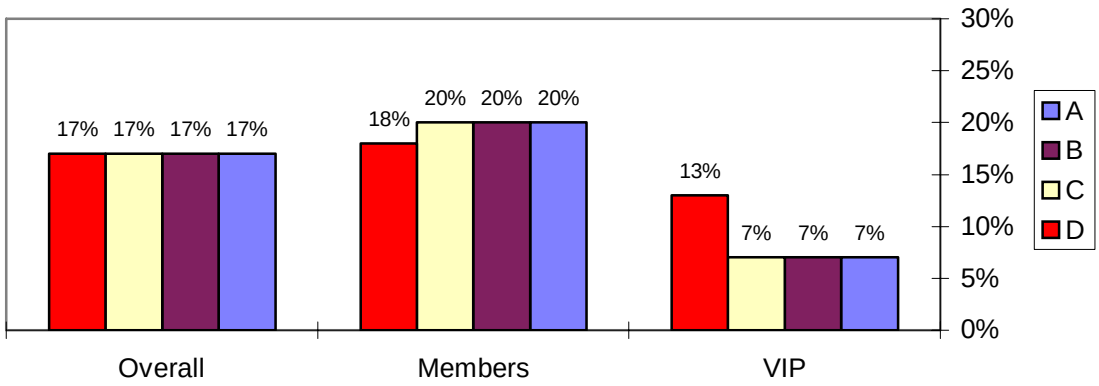
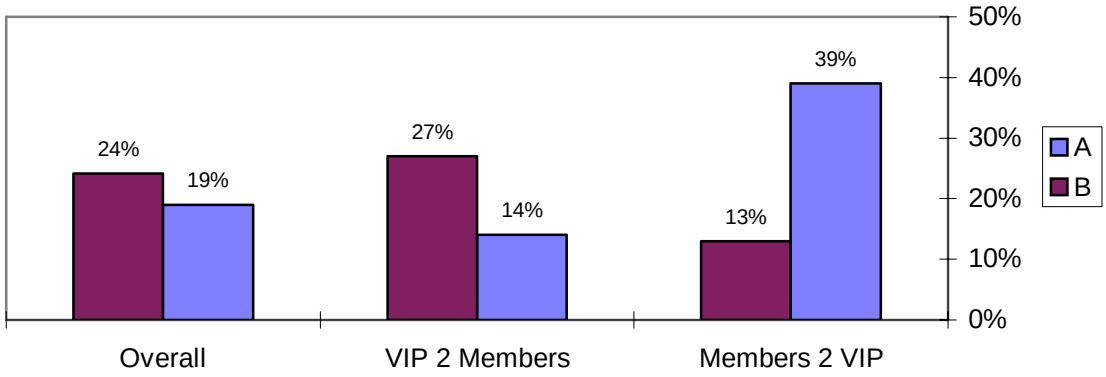


Chart 4 : 1000 Calls - Overflows



WHAT IF : 1500 Calls/hour - ASA

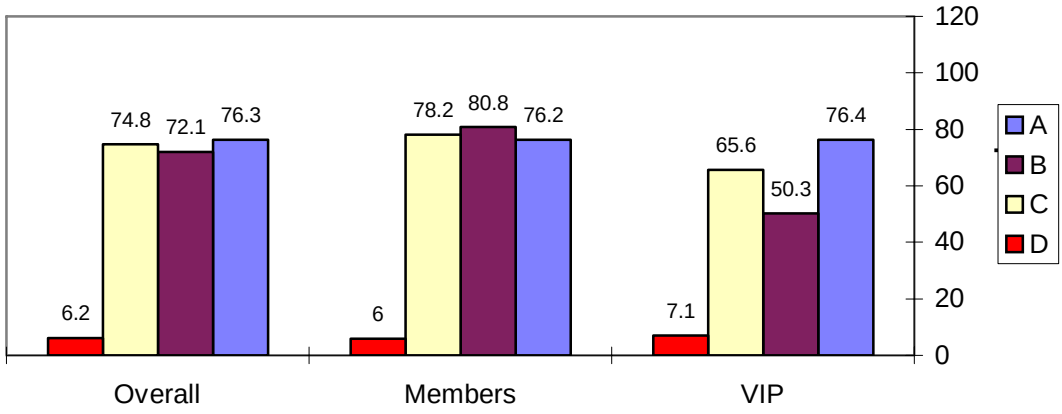


Chart 7 : 1500 Calls - Abandonment

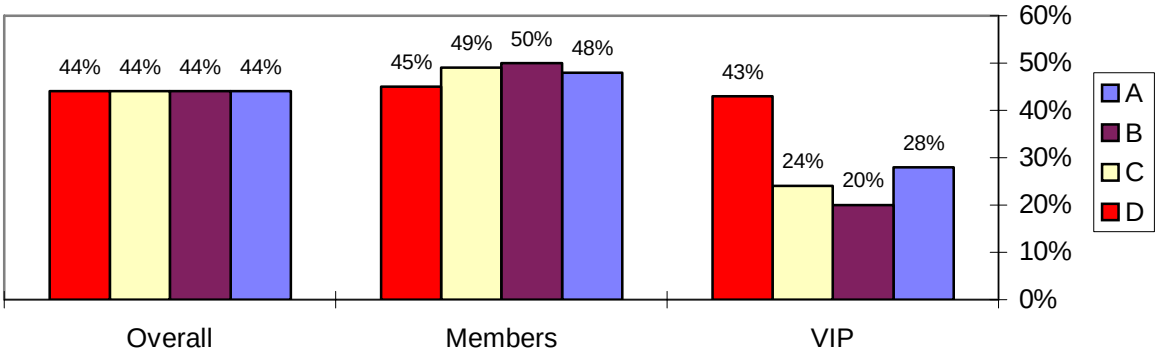
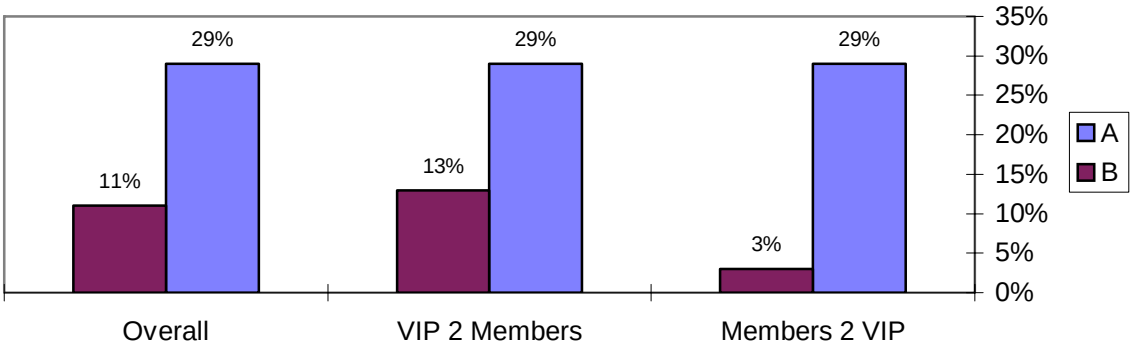


Chart 8 : 1500 Calls - Overflows



Reality

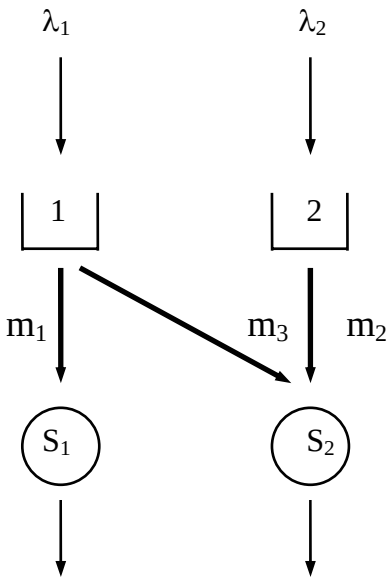
- Technology enables smart systems
- Reality becomes increasingly complex
- Solutions are urgently needed
- Theory lags significantly behind needs

- 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

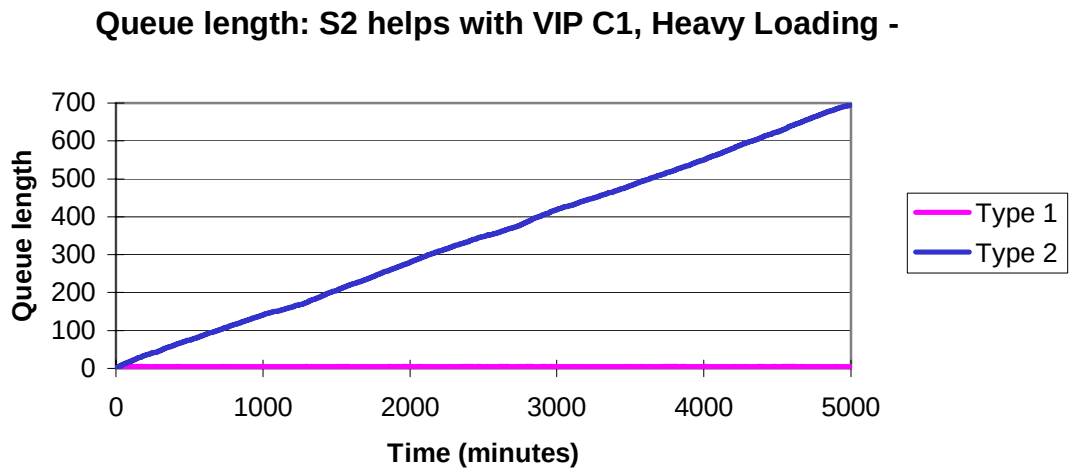


$\lambda_1 < 1.3 \quad \lambda_2 <= 0.4$

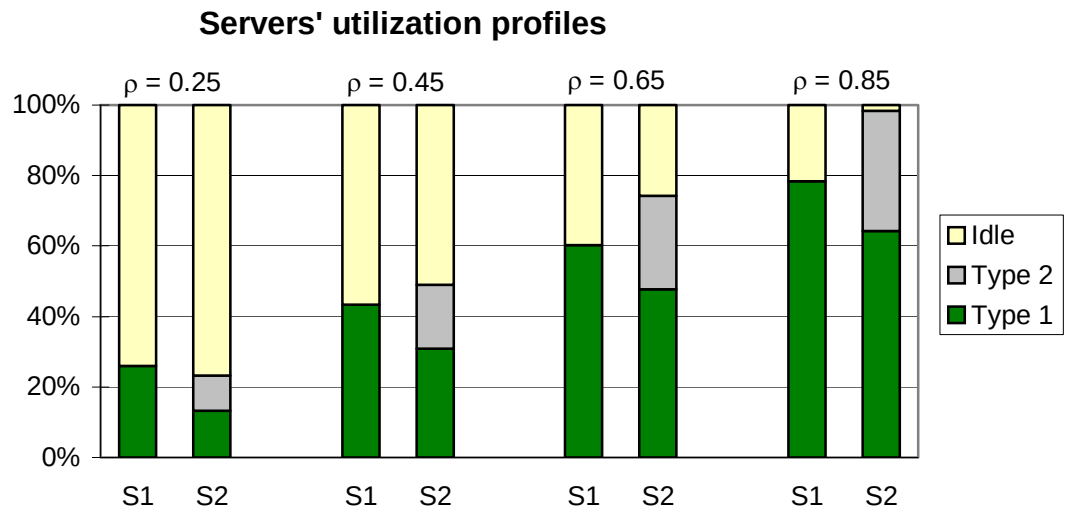
$m_1 = m_2 = 1, \quad \mathbf{m_3 = 2}$

$S_1 = S_2 = 1$

- 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 ?

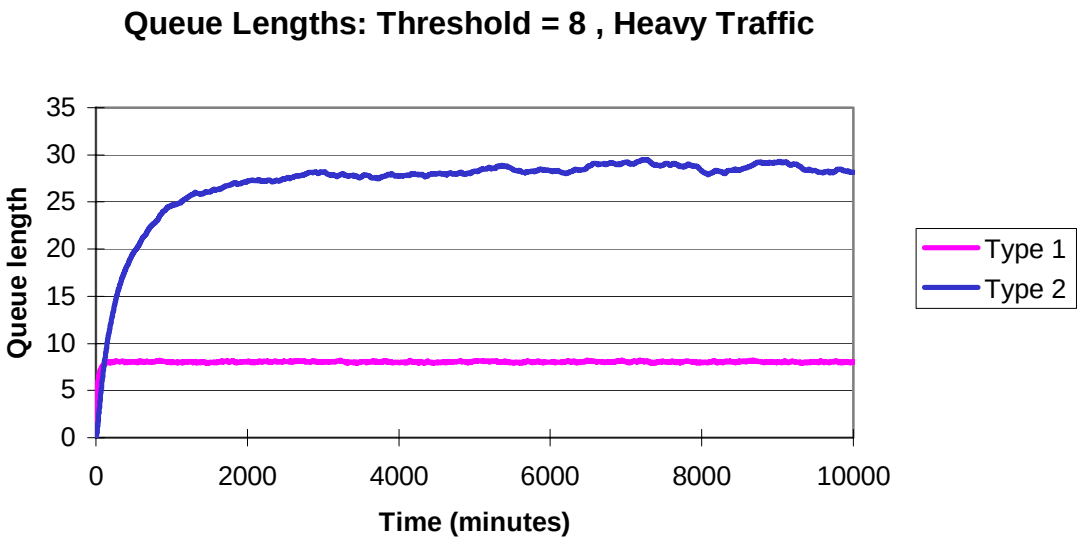


Instability: S2 **overworked** serving C1 and neglecting C2,

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

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).

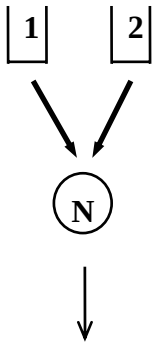
General (N, X, W, M, …) solution: Index Control is optimal, under sufficient skills-overlap (complete resource pooling; Harrison, Lopez).

- 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 (or, typically, should not be) E-Driven !?

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

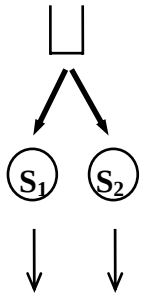
- Moderate: C1 is Q-served, C2 is QED

$\sqrt{\cdot}$ 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.