Call Centers

Queueing

Theory, Science, Practice

Service Engineering

Madrid, July 3, 2002

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

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

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

Supporting Material

Koole, and M.: "Queueing Models of Call Centers: An Introduction." *AOR* (MCQT '02).

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

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

Jelenkovic, M. and Momcilovic: "The GI/D/N Queue in the QED (Quality and Efficiency Driven) Regime." Under preparation.

Borst, M. and Reiman:. "Dimensioning Large Telephone Call Centers." Under revision to *Operations Research*.

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

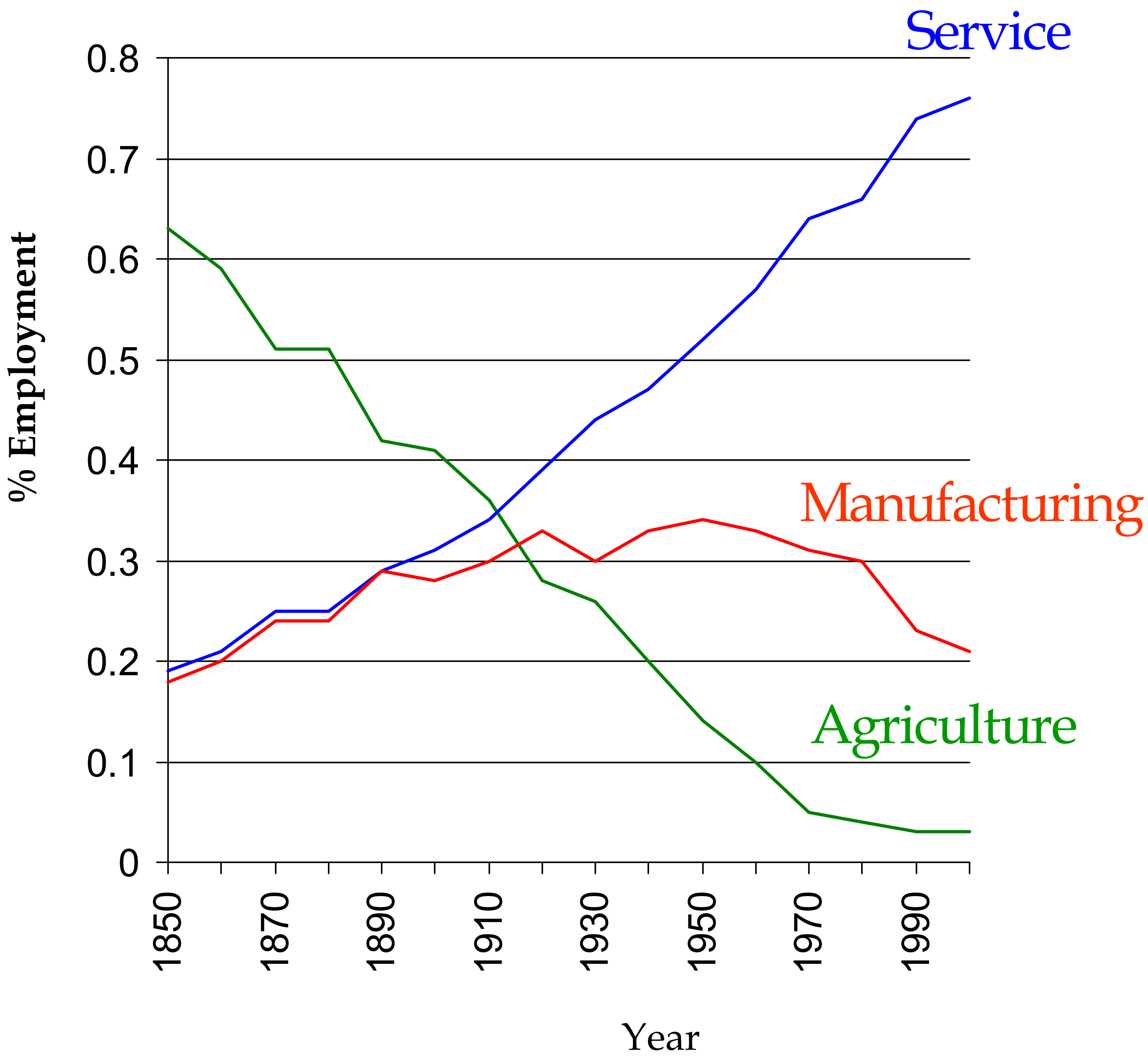
Jennings, M., Massey and Whitt: "Server staffing to meet time-varying demand." *Management Science* **42**, 1383–1394, 1966.

Atar, M. and Reiman: "Scheduling a Multi-Class Queue with Many Exponential Servers: Asymptotic Optimality in Heavy-Traffic." Technical Report, Technion, 2002.

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

Contents

Employment History: 1850-2000



Service Engineering

- Contrast with the traditional and prevalent

Service Management	(Business Schools)
Industrial Engineering	(Engineering Schools)

- 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: Designing Techology-Intensive User-Interfaces

- Support + Sales via Telephone + Chat + e.mail

Example: Staffing the Modern Call Center

- People = 60-80% costs of running a call center

(±1% of 1000 agents = 10 salaries; 2% U.S. workforce.)

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

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)

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

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

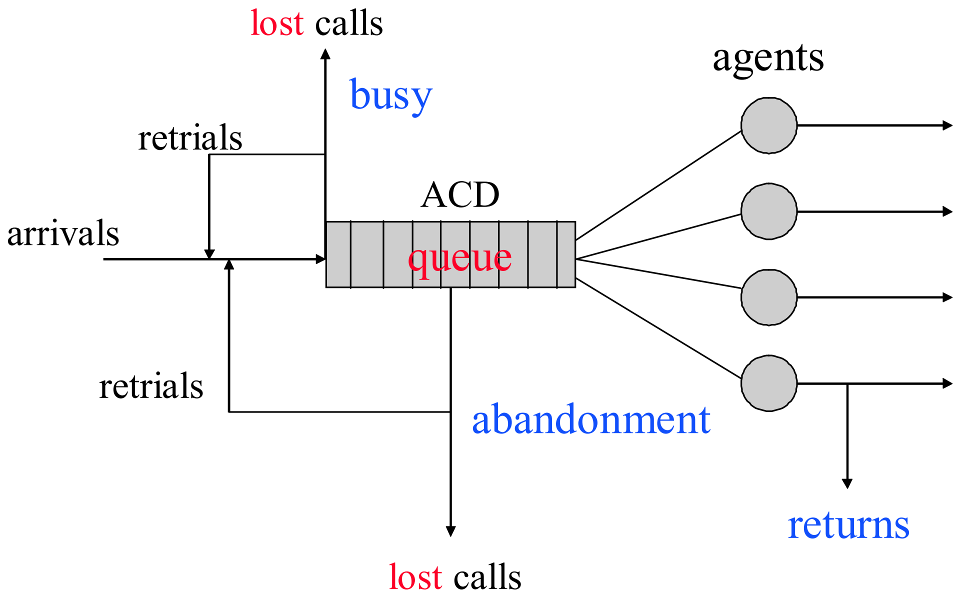
Scheduling: Heuristics, AI (Complex)

Individual constraints

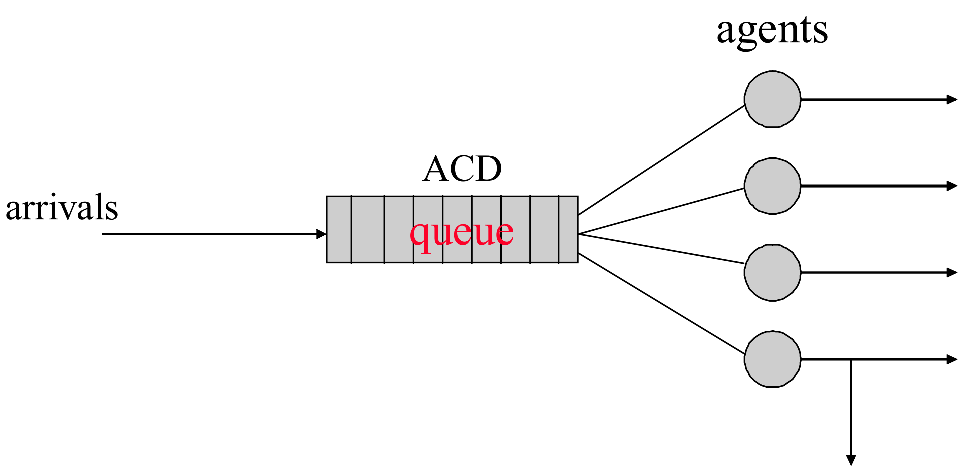
Agents Assignments

Online Skills-based Routing: Stochastic Control (ongoing)

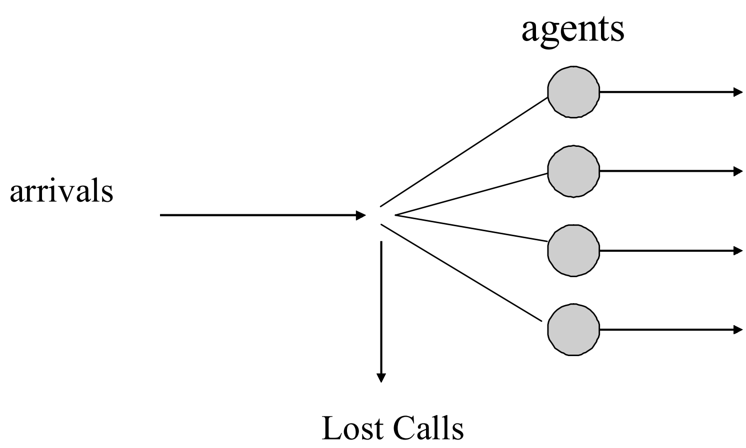
The Basic Call Center



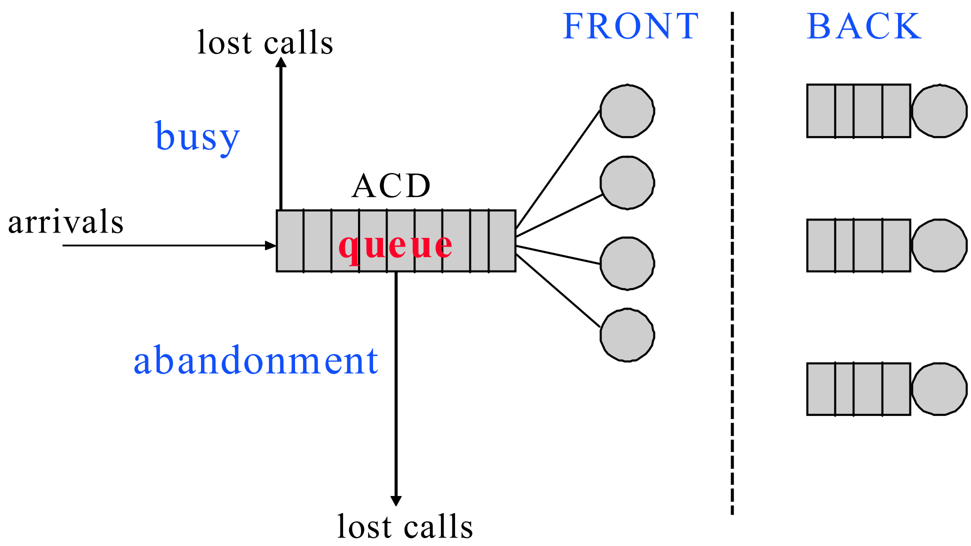
Erlang-C = M/M/N



Erlang-B



Erlang-A <4CallCenters.com>



iProfiler @ 4CallCenters.com

4CallCenters.com™
Workforce Management On The Web!

Call Center iProfiler

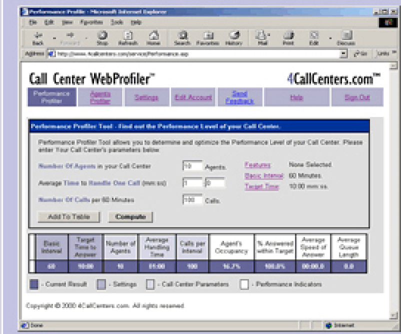
Workforce Management tools for inbound call centers on the Web!

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

Workforce Management (Staffing): Hierarchical Operational View

Forecasting Customers (Statistics), Agents (HRM)

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)

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	28	: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	:00:26	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	:03:53	: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		379	: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

Page 1 of 2

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** = $\lambda \times M$

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

= 50 Erlangs

Occupancy ρ = R/N

= 50/100 = 50%

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

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

Quality-driven: 100 agents, 50% utilization

$\Rightarrow$ **Can** increase offered load - but **by how much?**

M/M/N **N=100** **E(S) = 3:45 min.**

$\underline{\lambda}$/hr	$\underline{\rho}$	$E(W_q)$ = ASA	% Wait $\leq$ 2 sec
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 = 1fast 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 ≤ 2 sec
1585	100	99.1%	3:34	12%
1599	**100**	99.9%	**59:33**	1%
1599	**100+1**	98.9%	**3:06**	13%
1599	102	98.0%	1:24	24%
1599	105	**95.2%**	0:23	**51%**

=> **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-Driven + Efficiency-Driven

Enabler: Economies of Scale in a

Frictionless Environment (e.g. Call Center)

Theorem (**Halfin-Whitt**, 1981):

Consider a sequence of M/M/N models, **N=1,2,3,…**

Then the following **3 points of view** are equivalent:

- Customer $\lim_{N\to\infty} P_N\{\text{Wait} > 0\} = \alpha\,, \qquad 0 < \alpha < 1;$
- Server $\lim_{N\to\infty} \sqrt{N}(1-\rho_N) = \beta\,, \qquad 0 < \beta < \infty;$
- Manager $N \approx R + \beta\sqrt{R}\,, \qquad R = \lambda \times \mathrm{E(S)}$ large;

Here $$\alpha = \left[1 + \frac{\beta\phi(\beta)}{\varphi(\beta)}\right]^{-1},$$

where $\varphi(\cdot)/\phi(\cdot)$ is the standard normal density/distribution.

Extremes:

Everyone waits: $\alpha = 1 \quad \Leftrightarrow \quad \beta \le 0$ **Efficiency-driven**

No one waits: $\alpha = 0 \quad \Leftrightarrow \quad \beta = \infty$ **Quality-driven**

Theorem (**Halfin-Whitt**, 1981):

Consider an M/M/N (Erlang-C) model, **N large.**

Then the following **3 points of view** are equivalent:

- Customers $\%\{\text{Wait} > 0\} = \alpha$, $0 < \alpha < 1$;
- Agents $\rho \approx 1 - \frac{\beta}{\sqrt{N}}$, $0 < \beta < \infty$;
- Managers $N \approx R + \beta\sqrt{R}$, $R = \lambda \times \mathrm{E(S)}$ large;

Here $$\alpha = \left[1 + \frac{\beta\phi(\beta)}{\rho(\beta)}\right]^{-1},$$

where φ / ϕ are the standard normal density/distribution

Extremes:

Everyone waits: $\alpha = 1 \Leftrightarrow \beta \leq 0$ **Efficiency-driven**

No one waits: $\alpha = 0 \Leftrightarrow \beta = \infty$ **Quality-driven**

$\sqrt{\cdot}$ **Safety-Staffing**: Performance

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

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

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

Expected Performance:

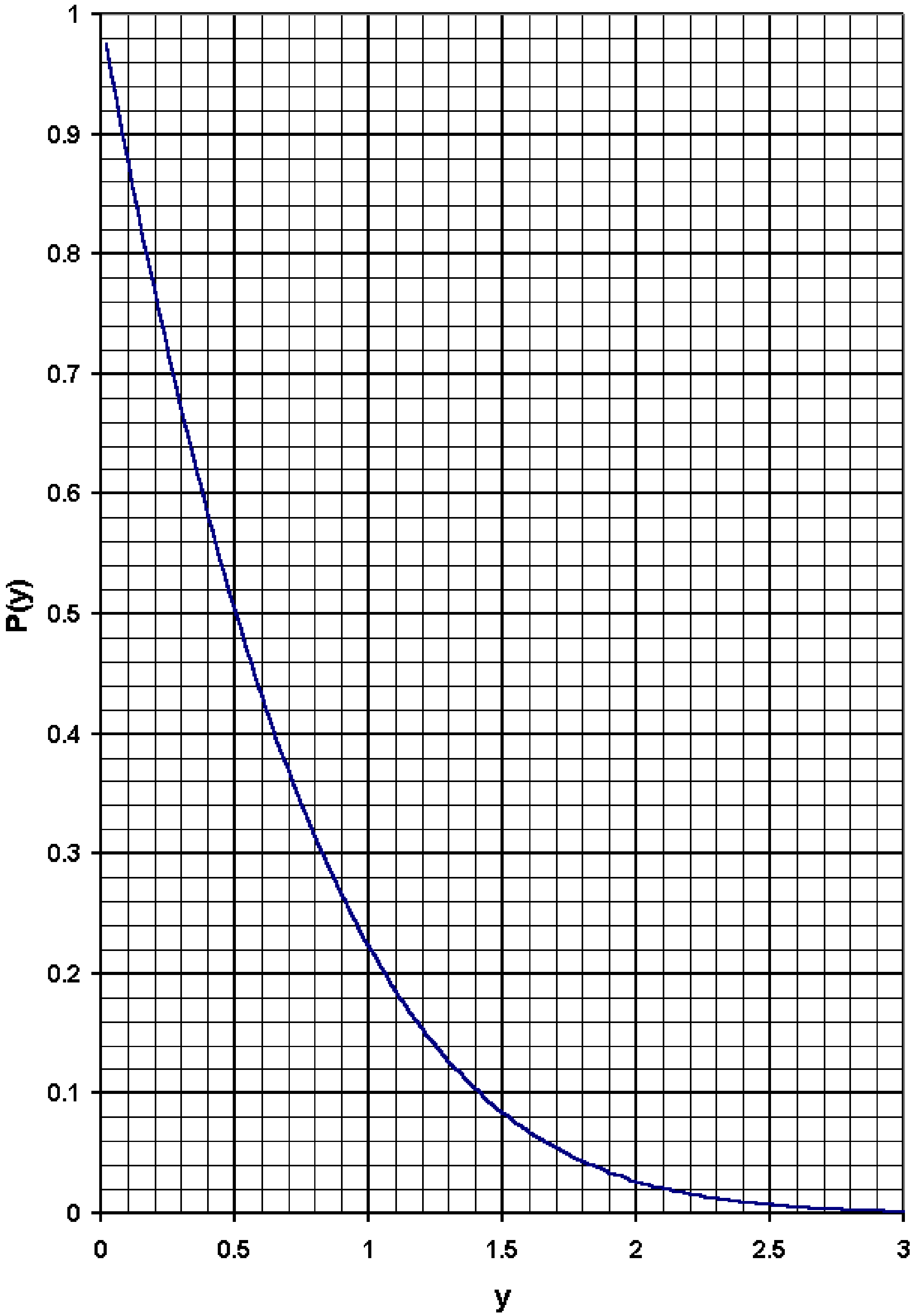
$$\%\ \text{Delayed} \approx P(\beta) = \left[1 + \frac{\beta\phi(\beta)}{\varphi(\beta)}\right]^{-1}, \quad \beta > 0$$ $\approx$ Erlang-C

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

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

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

The Halfin-Whitt Delay Function *P(y)*



QED : Intuition (Assume $\mu = 1$)

M/M/N: $W_N \mid W_N > 0 \overset{d}{=} \exp\left(\text{mean} = \frac{1}{N} \frac{1}{1-\rho_N} \right)$

$$\sqrt{N}\, W_N \mid W_N > 0 \overset{d}{=} \exp\left(\sqrt{N}\,(1-\rho_N)\right) \Rightarrow \exp(\beta)$$

But why $P(W_N > 0) \to \alpha, \quad 0 < \alpha < 1$? answer via

M/D/N: (with **P. Jelenkovic** and **P. Momcilovic**)

Observation: Cyclic assignment does not alter waiting times

$\Rightarrow$ Same waiting as in $E_N/D/1$!

QED $N = R + \beta\sqrt{R}$ and consider one of the $E_N/D/1$:

Interarrivals $A_N \approx 1 + \frac{\beta}{\sqrt{N}} + \frac{Z}{\sqrt{N}}, \qquad Z \overset{d}{=} N(0,1)$

Lindley $W_N = (W_N + 1 - A_N)^+ \qquad (\sqrt{N}\; W_N \Rightarrow W)$

$$P(W_N \le 0) = P(W_N + 1 - A_N \le 0) \approx$$

$$\approx P\left(W/\sqrt{N} + 1 - 1 - \beta/\sqrt{N} - Z/\sqrt{N} \le 0\right) = P(W - \beta \le Z)$$

$$P(W_N > 0) \to P(Z < W - \beta) = E\phi(W - \beta) \quad < \quad 1$$

(Efficiency: $N = R+\gamma$ (HT); Quality: $N = R+\delta R$ (D/D/1))

Rules of Thumb: Operational Regimes

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

Efficiency-driven $(P\{\text{Wait} > 0\} \to 1)$

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

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

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

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

$$N = \lceil R + \beta\sqrt{R} \rceil, \qquad \beta > 0 \quad \textbf{service grade}$$

How to determine parameters? **regimes** ?

via Strategy, Economics

Strategy: Sustain Regime through Pooling

Economies of Scale

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

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

	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 - Example

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

$$R = 300 \times \frac{5}{60} = 25, \quad OCC = 83.3\% \quad ASA = 15 \text{ sec}$$

$$y = (N - R)/\sqrt{R} = (30 - 25)/\sqrt{25} = 1, \quad P(1) = 22\%$$

4 CC: $\lambda = 1200$, AHT = 5, R = 100; N=?

Quality-Driven: maintain OCC at 83.3%.

$$N = 120, \quad ASA = .5 \text{ sec}, \quad y = (120 - 100)/10 = 4$$

Efficiency-Driven: maintain ASA at 15 sec.

$$N = 107, \quad OCC = 95\%, \quad y = 0.8$$

QED: maintain %{Wait>0}) at 22% (y at 1).

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

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

Q: N = 271

E: N = 233

QED: $N = 225 + 1 \cdot \sqrt{225} = 240$, OCC = 94%, ASA = 47 sec

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

Service-Quality vs. Operation Efficiency

With S. Borst, M. Reiman (1997-2002)

Quality	**D**(t)	delay cost	(t = delay time)
Efficiency	**C**(N)	staffing cost	(N = # agents)

Optimization: N* that minimizes total costs

(Satisfization: N* least that adheres to a cost constraint)

- **C >> D** : Efficiency-driven
- **C << D** : Quality-driven
- **C ≈ D** : Rationalized: QED

Framework: Asymptotic theory of M/M/N, $N \uparrow \infty$.

Economics: Linear Costs Model

Expected cost / unit of time =

$$E(N, \lambda) = C(N) + \lambda \cdot P\{W_q > 0\} \cdot E[D(W_q) \mid W_q > 0]$$

Change of variables $\boxed{N \to N_\lambda(x) = \frac{\lambda}{\mu} + x\sqrt{\frac{\lambda}{\mu}}, \quad x > 0}$

Erlang-C Formula $\quad P\{W_q > 0\} = \pi\left(N, \frac{\lambda}{\mu}\right) \to \pi_\lambda(x)$

Linear costs $\quad C(N) = c \cdot N, \qquad D(t) = d \cdot t$

Then $\quad E(N, \lambda) = c \cdot N + \lambda\pi\left(N, \frac{\lambda}{\mu}\right)\frac{d}{N\mu - \lambda}$

$$= c\frac{\lambda}{\mu} + cx\sqrt{\frac{\lambda}{\mu}} + \pi_\lambda(x)\frac{d}{x}\sqrt{\frac{\lambda}{\mu}} \quad .$$

Continuous Approximation of original **discrete** problem:

$$x_\lambda^* = \arg\min_{x>0}\left\{cx + \frac{d_\lambda}{x}\pi_\lambda(x)\right\} \quad \text{(c-fixed, d varies with } \lambda\text{)}.$$

Economics: Linear Costs Asymptotics

Efficiency-driven: $d_\lambda = d\lambda^{-1/2}$; then $x_\lambda^* \to 0$, $\pi_\lambda(x_\lambda^*) \sim 1$.

Let $$y_\lambda^* = \arg\min_{y>0} \left\{ cy + \frac{d}{y}\lambda^{-1/2} \right\}$$

Quality-driven: $d_\lambda = d\lambda^{1/2}$; then $x_\lambda^* \to \infty$, $\pi_\lambda(x_\lambda^*) \sim \dfrac{\varphi(x_\lambda^*)}{x_\lambda^*}$.

Let $$y_\lambda^* = \arg\min_{y>0} \left\{ cy + \frac{d}{y^2}\lambda^{1/2}\varphi(y) \right\}$$

QED: $d_\lambda \equiv d$;

then $x_\lambda^* \to x^*$ $(0 < x^* < \infty)$, $\pi_\lambda(x_\lambda^*) \sim P(x_\lambda^*)$.

Let $$\boxed{y^* = \arg\min_{y>0} \left\{ cy + \frac{d}{y} P(y) \right\}}$$

<u>Theorem</u>: Asymptotic Optimality of $N_\lambda(y_\lambda^*) = \dfrac{\lambda}{\mu} + y_\lambda^* \sqrt{\dfrac{\lambda}{\mu}}$

(Roughly) $$\frac{E(N_\lambda(y_\lambda^*), \lambda) - C\left(\dfrac{\lambda}{\mu}\right)}{E(N_\lambda^*, \lambda) - C\left(\dfrac{\lambda}{\mu}\right)} \to 1, \quad \text{as } \lambda \uparrow \infty .$$

Economics: Quality vs. Efficiency (Linear Costs)

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

where $\mathbf{d}$ = delay/waiting costs

$\mathbf{c}$ = service/staffing costs

Here $$y^*(\mathbf{r}) \approx \left(\frac{r}{1+r\left(\sqrt{\pi/2}-1\right)}\right)^{1/2}, \quad 0 < r < 10$$

$$\approx \left(2\ln\frac{r}{\sqrt{2\pi}}\right)^{1/2}, \quad r \text{ large.}$$

Performance measures**:** $\Delta = y^*\sqrt{\mathrm{R}}$ safety staffing

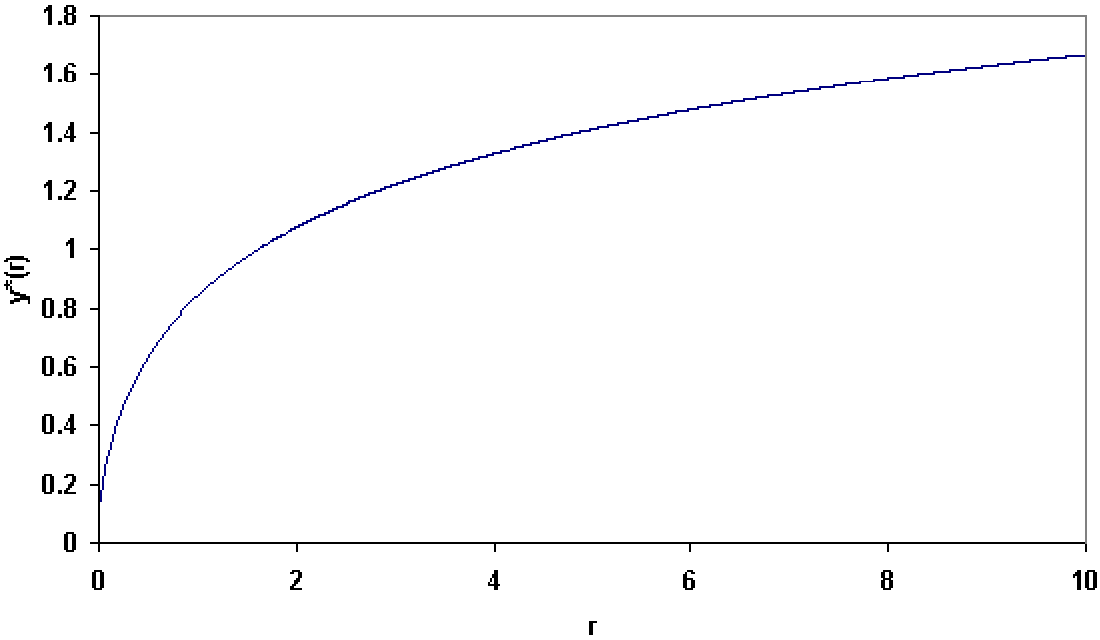
$$\%\{\text{Wait} > 0\} \approx \mathbf{P}(\boldsymbol{y}^*) = \left[1+\frac{y^*\phi(y^*)}{\varphi(y^*)}\right]^{-1}$$ **Erlang-C**

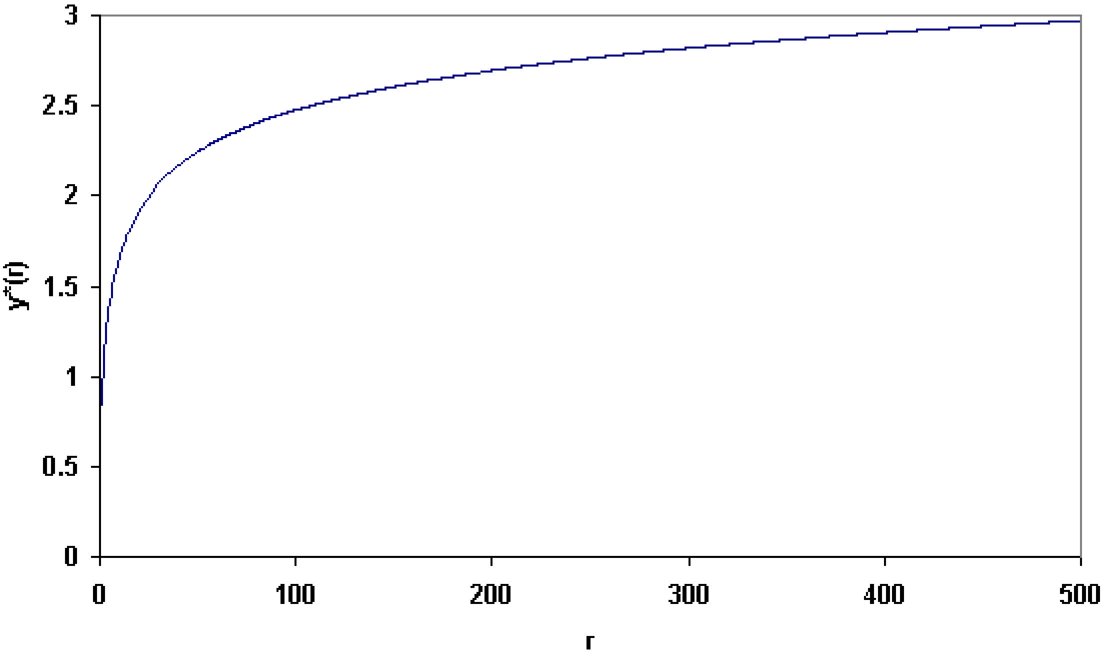
$$\text{TSF} = \%\left\{\frac{\text{Wait}}{\mathrm{E(S)}} > \mathrm{T} \,\middle|\, \text{Wait} > 0\right\} = e^{-\mathrm{T}\Delta}$$

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

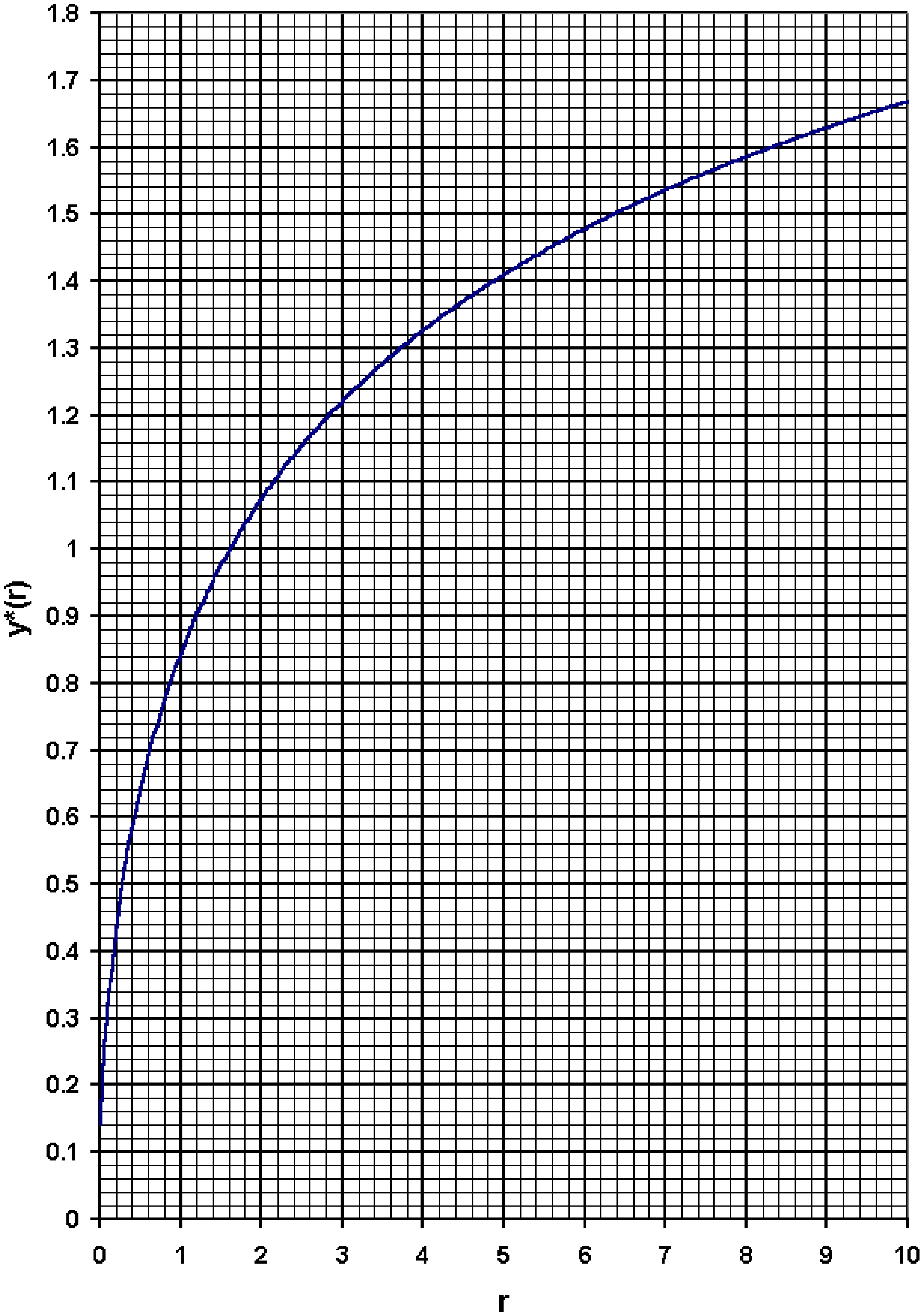
$$\text{Occupancy} = 1-\frac{\Delta}{N} \approx 1-\frac{y^*}{\sqrt{N}}$$

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





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



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

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

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

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

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

<u>Example</u>: 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" Problem ?

<u>Performance</u>:

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

Relevant: Large call centers do perform as above.

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 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)

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

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

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

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

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

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

Scenario Analysis: 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)

Figure 5
American data. Beta vs P{Ab}

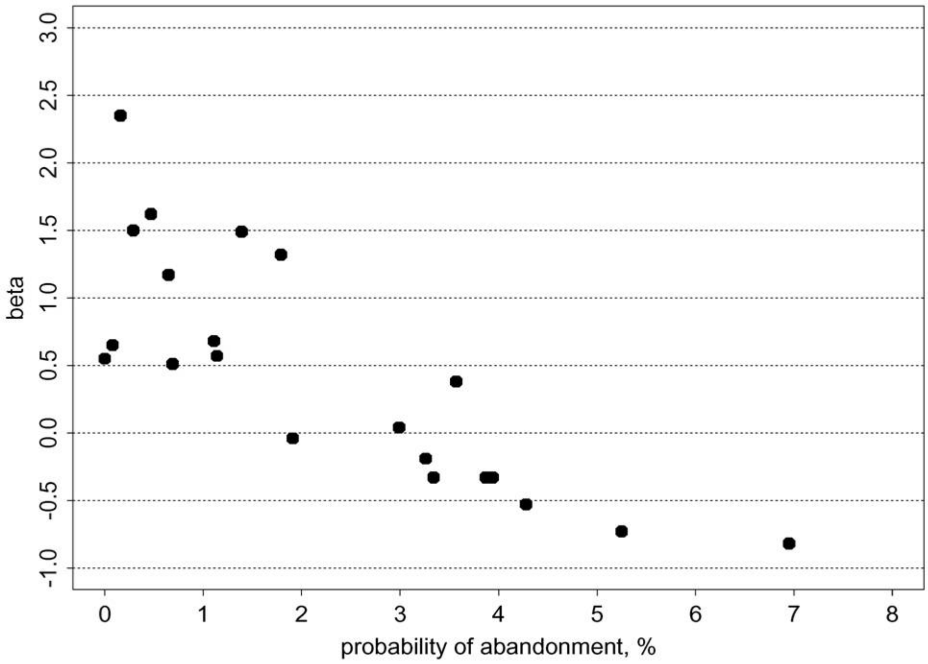
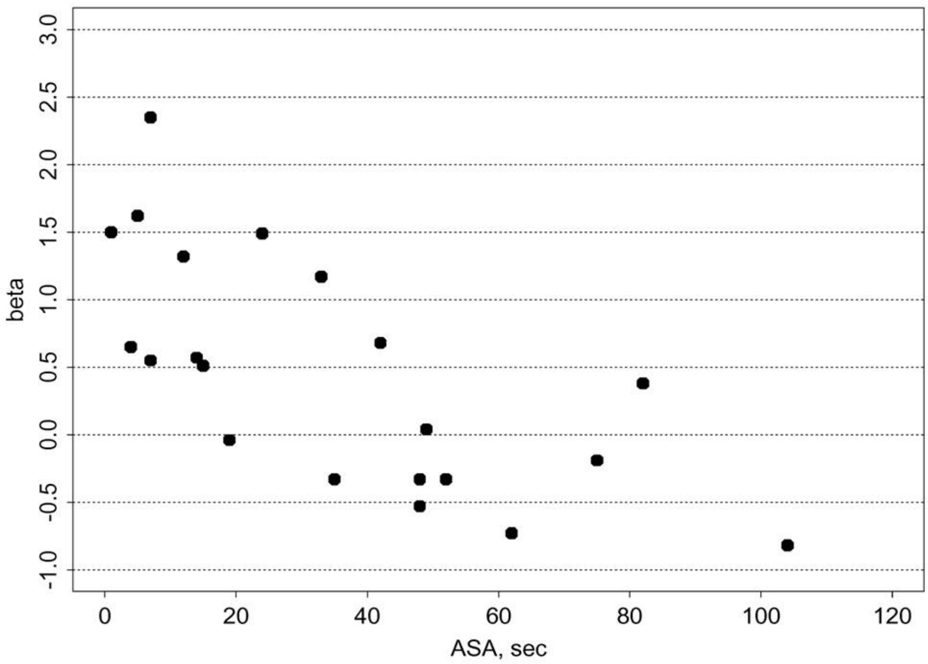


Figure 6
American data. Beta vs ASA



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 (**with Garnett & Reiman**):

Erlang-A: Theoretical performance analysis

Free Internet implementation (4CallCenters.com)

Prevalent in well-managed large call centers

Charlotte – Center

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%

Theorem (with **Garnett** and **Reiman**, 2001):

Consider a sequence of M/M/N+M (**Erlang-A**) models,

with parameters λ_N, μ,, θ, for N=1,2,3,…

Then the following **3 points of view** are equivalent:

- Customer $\lim_{N\to\infty} P_N\{\text{Wait} > 0\} = \alpha$, $0 < \alpha < 1$;
- Server $\lim_{N\to\infty} \sqrt{N}(1-\rho_N) = \beta$, $-\infty < \beta < \infty$;
- Manager $N \approx R + \beta\sqrt{R}$, $R = \lambda \times E(S)$ large;

$\Rightarrow$ **Serendipity** $\lim_{N\to\infty} \sqrt{N}\, P_N\{\text{Abandon}\} = \gamma$, $0 < \gamma < \infty$.

Here $\alpha(\beta;\, \mu,\theta)$, $\gamma(\beta;\, \mu,\theta)$ are easily computable.

Extremes:

$\alpha = 1$: $N = R - \gamma\ R$ **Eficiency-driven**

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

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, which does apply here,

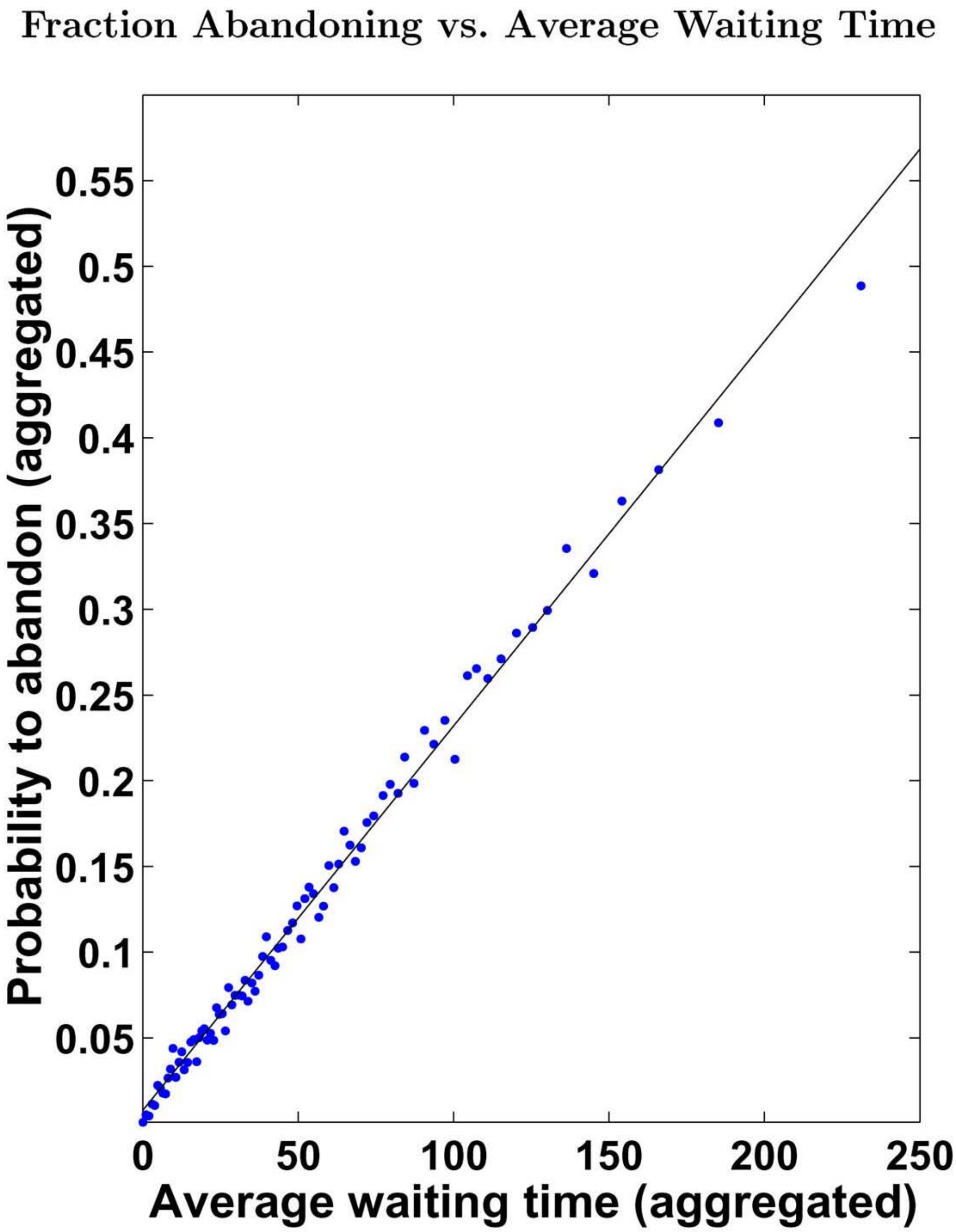
$\mathbf{y} = \frac{N - R}{\sqrt{R}}$, possibly negative (via $N = R + y\sqrt{R}$)

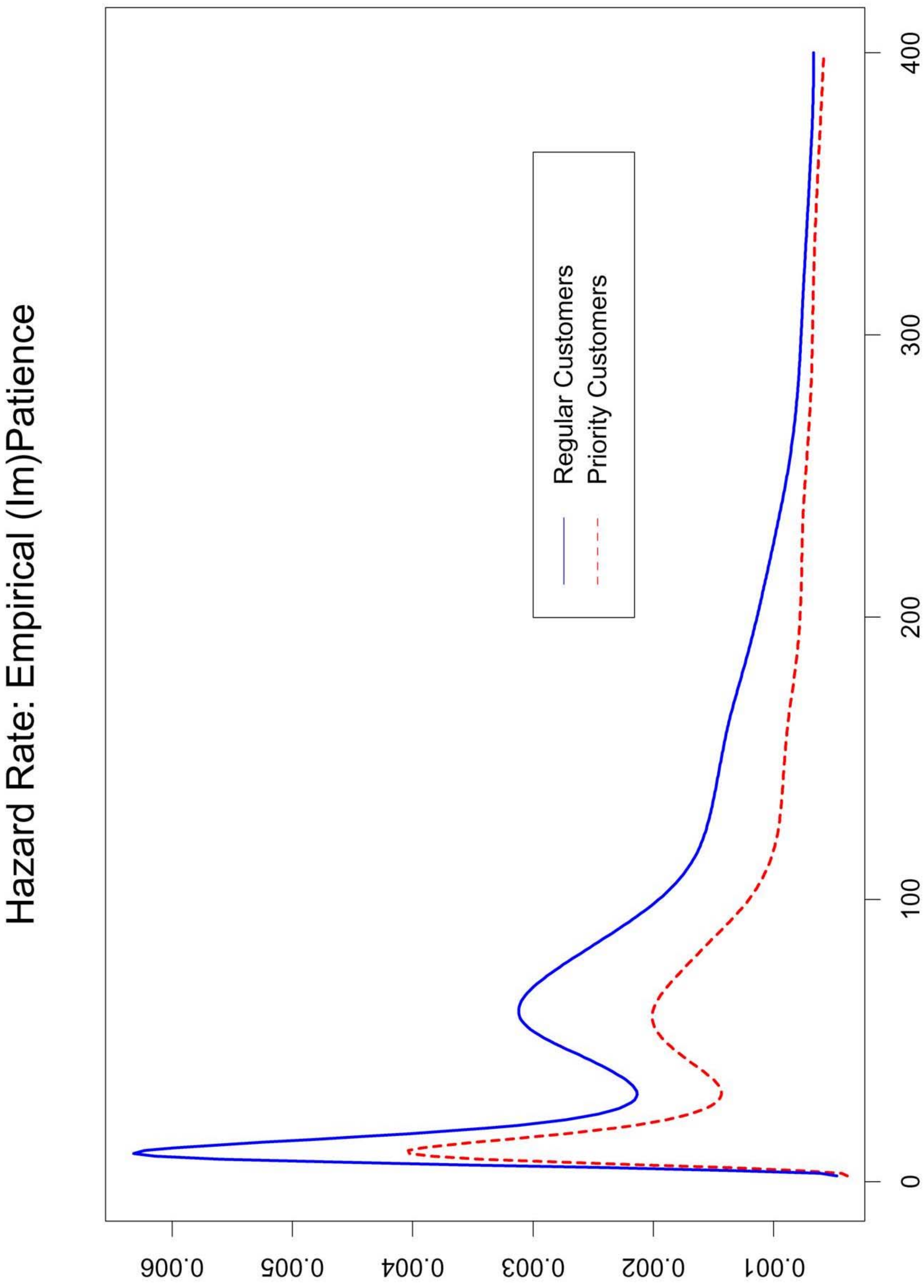
where $\mathbf{R} = \lambda \cdot M$ is the Offered Workload

Estimating Patience

Censored Sampling, or equivalently (under exp)

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




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

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
    (e.g. Censoring)
  - Misleading performance measures
    (e.g. LIFO in skills-based-routing)
  - 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)
- Applications
  - VRU/IVR: opt-out-rates
  - Internet: business-drivers (60% and more)
  - Call Centers: unique subjective performance measures

PATIENCE INDEX

- How to Define? Measure? Manage? (via 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.

$$\textbf{Patience Index} = \underset{\substack{\uparrow \\ \text{definition}}}{\frac{\text{time willing}}{\text{time expect}}} = 4 = \underset{\substack{\uparrow \\ \text{measure}}}{\frac{\#\,\text{served/wait} > 0}{\#\,\text{abandon/wait} > 0}}$$

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

Designing Call/Contact Centers

with Impatient Customers:

10 Years History, or A Modelling Panorama

1. Kella, Meilijson: Practice ⇒ Abandonment important

2. Shimkin, Zohar: No data ⇒ Rational patience in Equilibrium

4. Carmon, Zakay: Cost of waiting ⇒ Psychological models

5. Garnett, Reiman: Palm/Erlang-A to replace Erlang-C/B

   as the standard Steady-state model

6. Massey, Reiman, Rider, Stolyar: Predictable variability ⇒

   Fluid models, Diffusion refinements

7. Ritov, Sakov, Zeltyn: Finally Data ⇒ Empirical models

8. Brown, Gans, Haipeng, Zhao: Statistics ⇒ Queueing Science

9. Garnett, Atar, Reiman: Skills-based routing ⇒ Control models

10. Nakibly, Meilijson, Pollatchek: Prediction of waiting ⇒

    Online Models and Real Time Simulation

11. Garnett: Practice ⇒ 4CallCenters.com

Staffing the "Modern" Basic Call Center

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

   - Conceptual: Halfin & Whitt
   - Dimensioning: with Borst & Reiman $\Rightarrow y^*(d/c)$

2. Erlang-A (Abandonment, with $-\infty < y < \infty$)

   - Conceptual: with Garnett & Reiman
   - Dimensioning: with Borst & Reiman, in progress

3. Time-Varying (Non-homogeneous Poisson arrivals)

   - Ample-server heuristics: with Jennings & Massey & Whitt
   - Conceptual part: with Massey & Rider, in progress
   - Dimensioning: open (Stochastic Control ?)

4. General Service Time (for all the above)

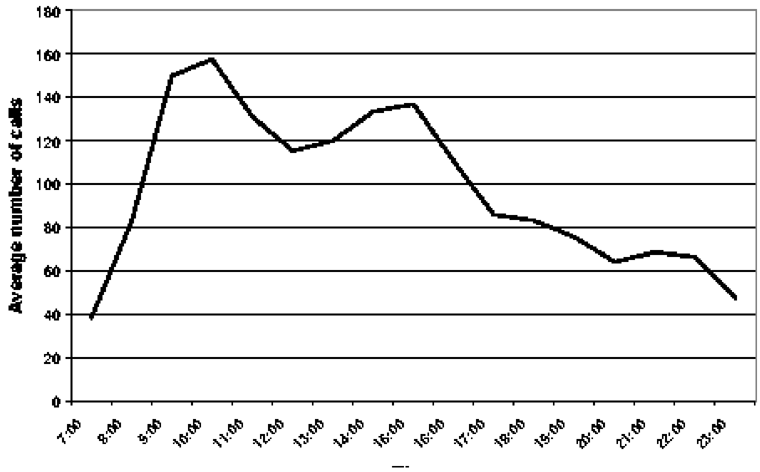
   - Conceptual supported by Puhalski & Reiman, M/PH/N
   - M/D/N: with Jelenkovic & Momcilovic, in progress
   - M/G/N open and challenging (measure-valued limit)

   (Beyond 2^{nd} moment theory in the QED regime!)

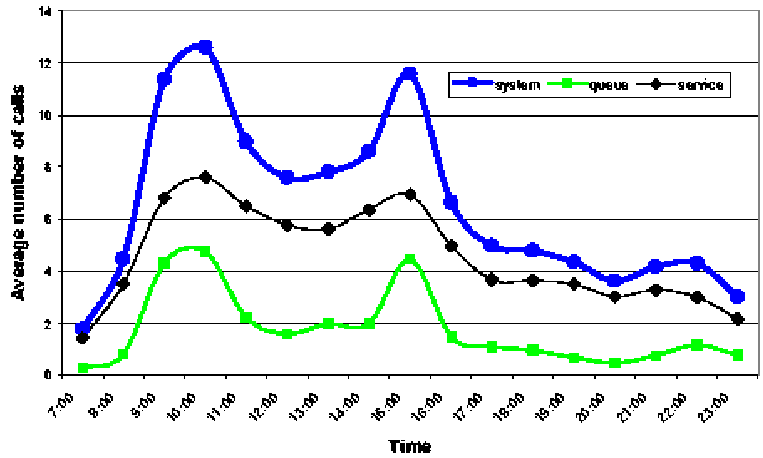
Time-Varying Queues: Predictable Variability

(with Jennings, Massey, Whitt)

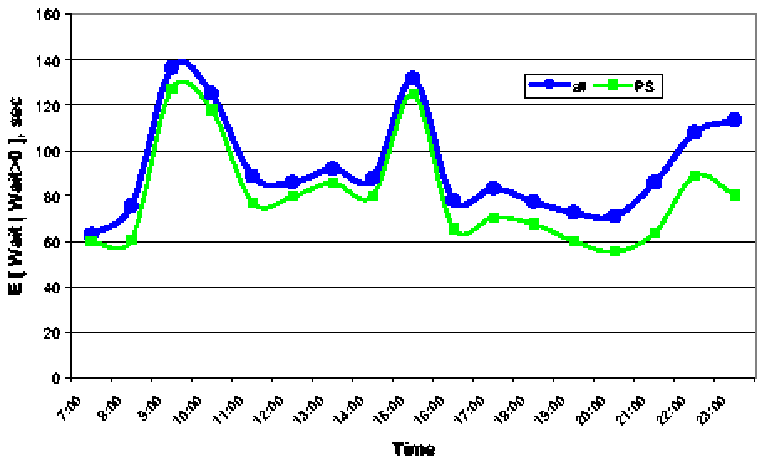
Arrivals



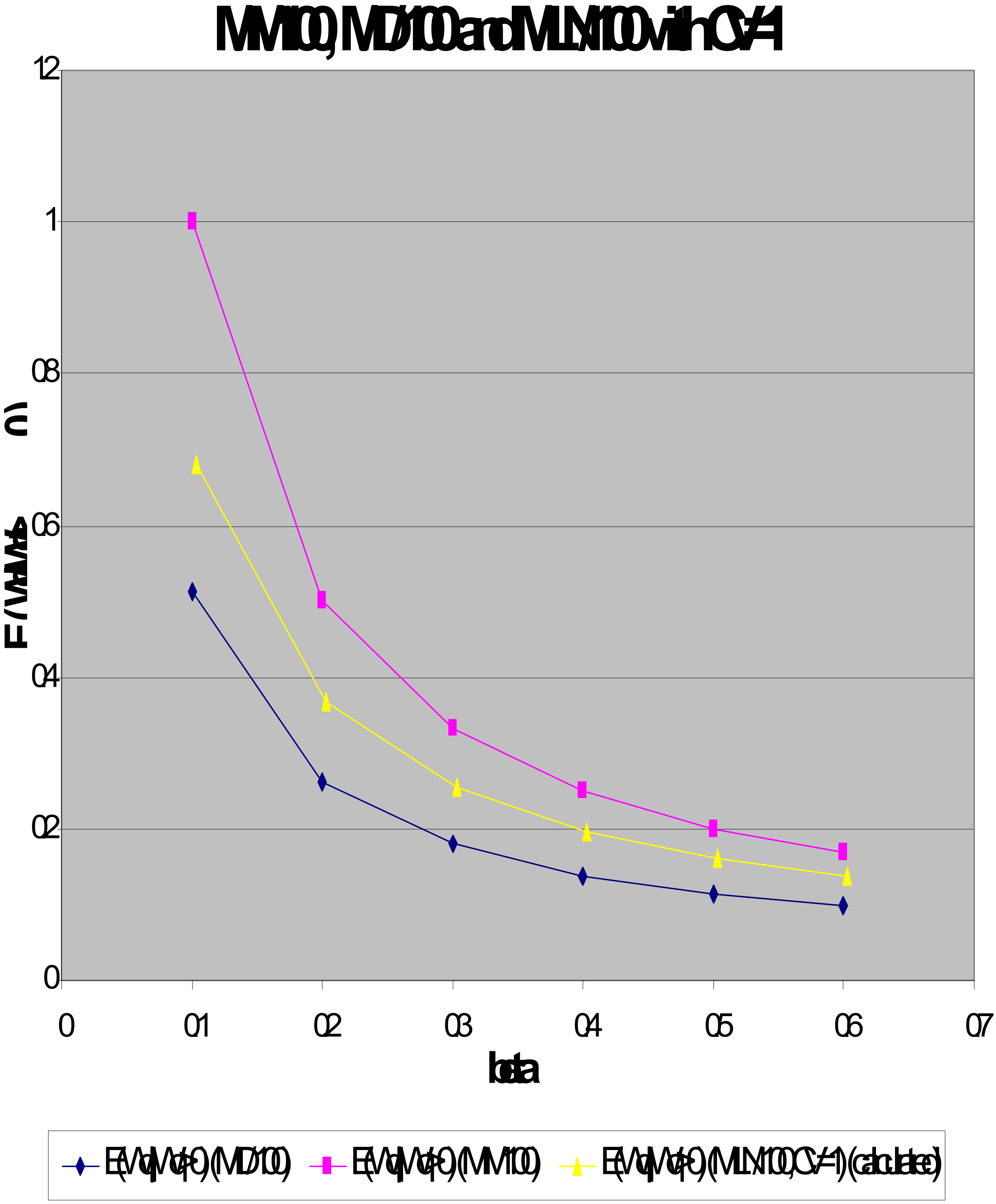
Queues



Waiting



E(Vq|M>0) vs β



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

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

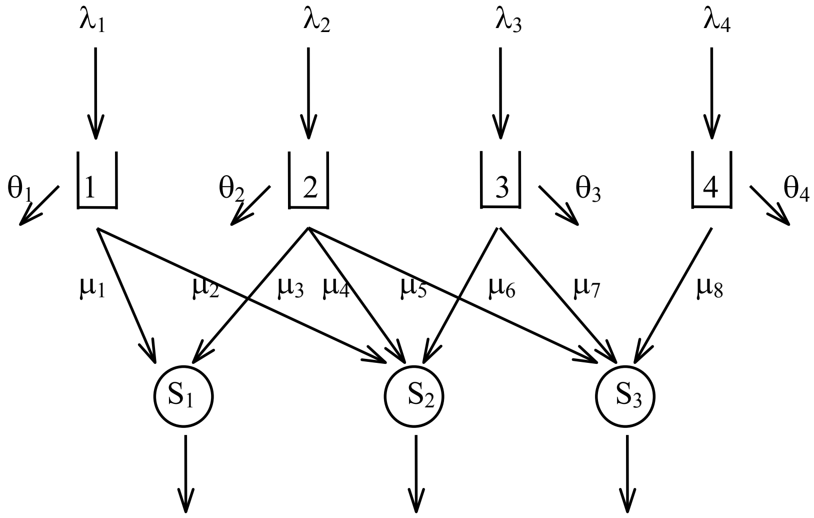
(Full Version)

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

Introduction

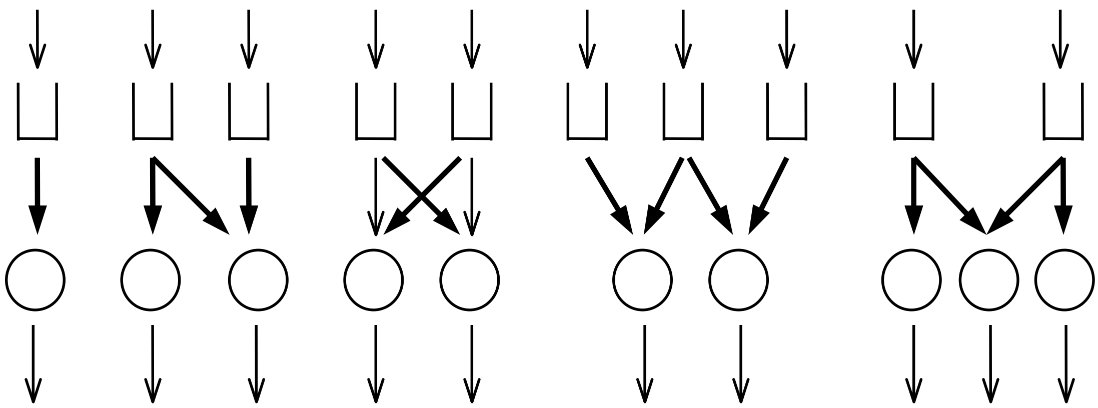
Consider the following multi-queue parallel-server system (animated, for example, by a telephone call-center):



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).



SBR in Efficiency-Driven Systems (with Stolyar)

Customer types i (renewal arrivals)

Server skills j (overlapping)

μ_{ij} =service rate of type i by server j (iid services)

$\mu_{ij} = 0$ if j cannot serve i ; ($1/\mu_{ij}$ = E[service time])

$C_i(w)$ = cost for type i waiting w units of time, Convex

($C_i(0) = C_i'(0+) = 0$; eg., quadratic, not linear)

Generalized cμ-rule : when becoming idle at time t,

server j chooses to serve type i for which

$$i \in \arg\max_i C_i'(W_i(t)) \ \mu_{ij}$$

$W_i(t)$ = head-of-line waiting time in queue i at time t.

Theorem In heavy traffic and with sufficient skills-overlap, Gcμ is asymptotically optimal: minimizes cumulative costs.

Special cases	single server:	Van Mieghem's Gcμ rule
	quadratic costs:	Kleinrock's aging factor

Idea: complete pooling into a single super-serve

SBR in the QED-Regime (with Atar, Reiman)

Agents' assignment to queues (upon service completion)
as well as
Customers' routing to idle servers (upon arrival)
are both significant.

Customer types i	(renewal arrival; exponential services)
Abandonment	(exponential patience)
N servers	(iid, in a V-Design)

Convex costs of queue-lengths (linear delay costs, abandons)

Theorem In the **QED Regime**, namely $N \approx R + \beta\sqrt{R}$,
Hamilton-Jacobi-Bellman policies are asymptotically optimal:
minimize cumulative discounted costs. $\left(R = \sum_i \lambda_i / \mu_i\right)$

Qualitative insights

- Preemption benefits are negligible.
- Queueing and waiting costs are "equivalent".
- Work-conservation is optimal for work-encouraging costs: optimal if optimal under preemption.
- No state collapse in general $\Rightarrow$ numerical insights (as in Harrison-Zeevi)

Beyond the “Basic” Call Center

- Skills-based Routing

  - Efficiency-Driven: with Stolyar, Gcμ optimal

  - QED: iid Servers, with Atar & Reiman, HJB-based

  - QED: Heterogeneous Servers, with Atar & Reiman

- Networks

  - IVR + ACD ; Retrials

  - Hierarchical Help Desk

  - Distributed Call Centers

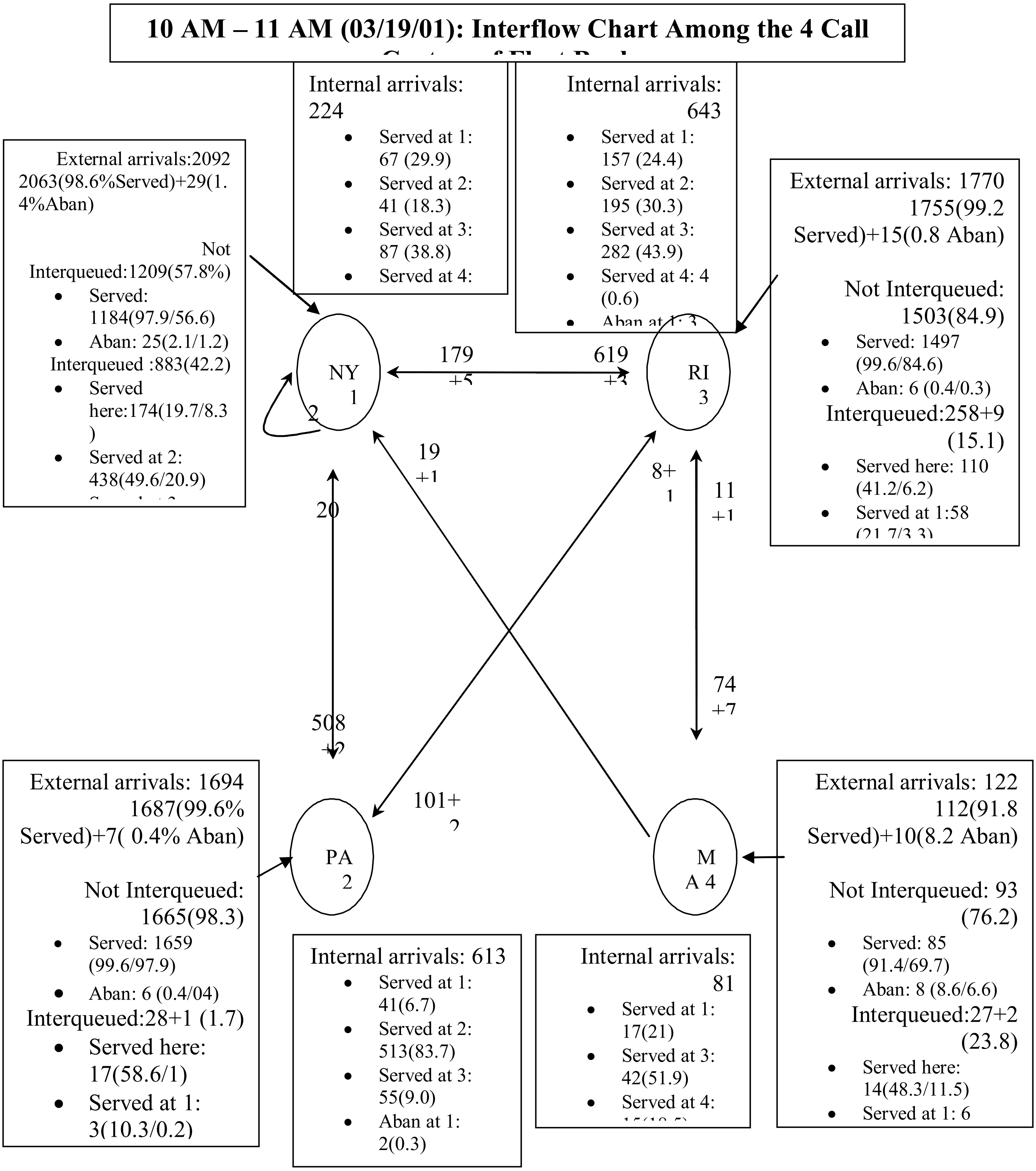
- **Staffing** SBR / Networks: Open

- Profit Contact Centers: $-driven multi-media interface

- Information to customers

- **Forecasting**: with Brown & Haipeng &Zhao: important

Distributed Call Center: Simultaneous Queueing



Beyond Traditional Queueing Theory

Some Characteristics of Services

- Time-varying conditions
  - Predictable variability dominant – Fluid View
  - Arrivals – typically given, Services – Staffing
- State-dependent responses
  - Skills-based routing
  - Finite Buffers
    - Physical = finite waiting room, busy-signal
    - Mental: customers balk, abandon
- Stability ? (9:00 – 17:00, Abandonment)
- Human factors
  - Equilibrium (decentralized) analysis
  - Fairness – FCFS often costly, unnecessary
  - **Tele-queues** – patience, information
- Approximations Fluid and Diffusion (Long- & Short-run)

- **Theory + Real Data + Experiments = Multi-Disciplinary Queueing Science**

Suggested "What Next"

- Register at www.4CallCenters.com, and play some

  (eg. Review lecture)

- Visit http://ie.technion.ac.il/serveng , then do

  - Homework 7: Gazolco
  - Homework 11: Staffing a Small, Medium, Large CC

- Feedback on Homework ($\Rightarrow$ I'll send solution)

- Download Charisma, and play/pay some