A Personal Tool for Workforce Management

Service Engineering

Garnett, O.	Technion
Zeltyn, S.	Technion
Borst, S.	CWI
Reiman, M.	Bell Labs

Wharton, Call Center Forum

May 16, 2002

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

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

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

Contents

1. Service Engineering – Teaching, Practice: Research-Based

2. Workforce Management: Hierarchical View

   The Role of **iProfiler** (or **Charisma**)

3. "Secrets of the Trade" – Recalling Previous Forums

   Operational Regimes

   QED: Beyond the Quality-Efficiency Tradeoff

   Rough Performance Analysis

4. Example: Pooling Call Centers, via Erlang-C

5. What is Lacking: Abandonment

   Patience: Understanding, Estimating, Managing

6. **Erlang-A** : Abandonment (Busy, Overflow)

7. Feedback: for Work? Classroom? Research?

8. Homework: HW7 in <**http://ie.technion.ac.il/serveng**>

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

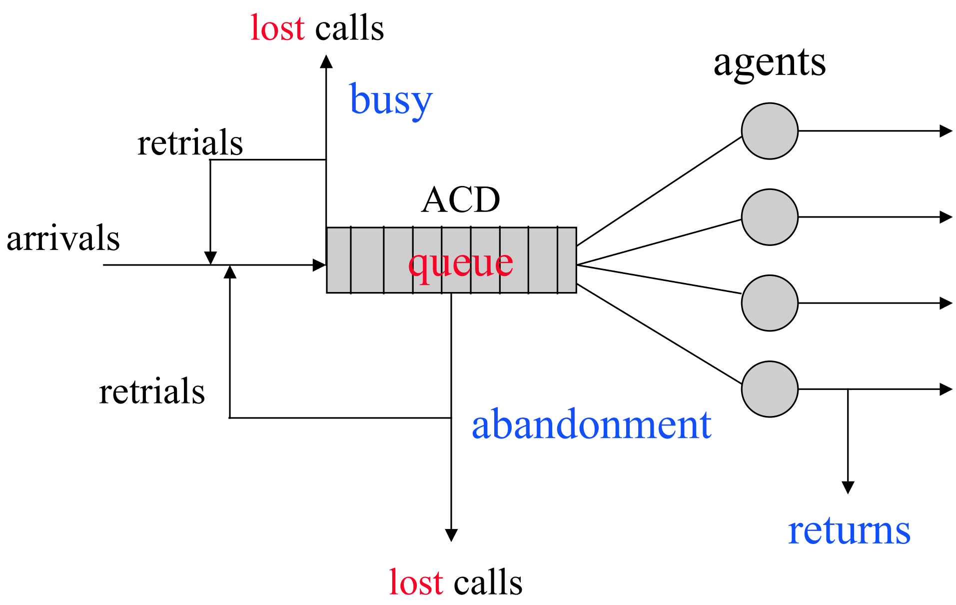
Example: Staffing the Modern Call Center

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

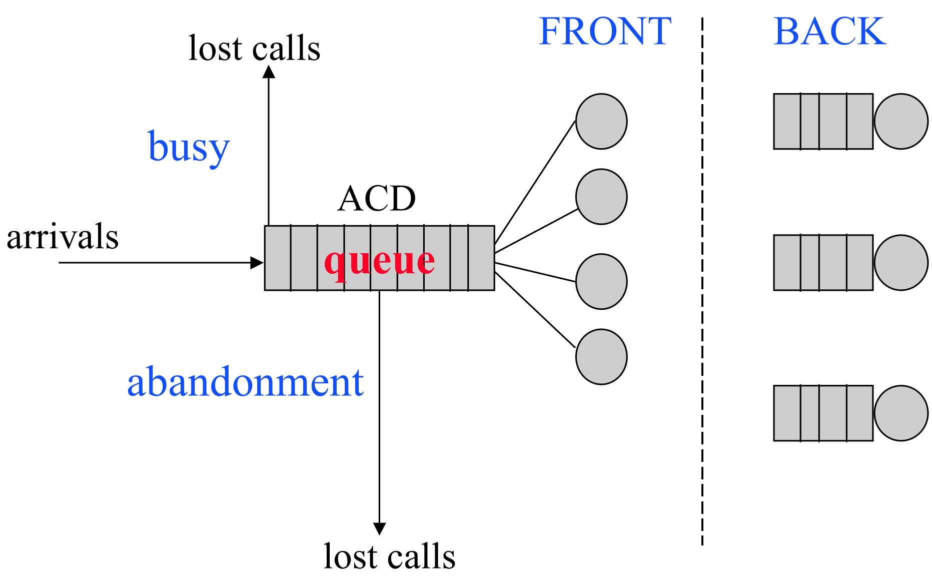
(±1% of 1000 agents = 10 salaries; 2% U.S. workforce.)

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

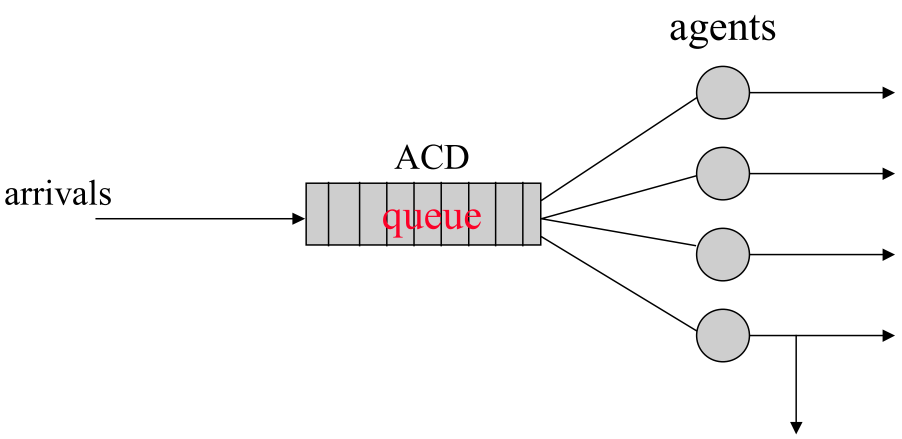
A Basic Call Center



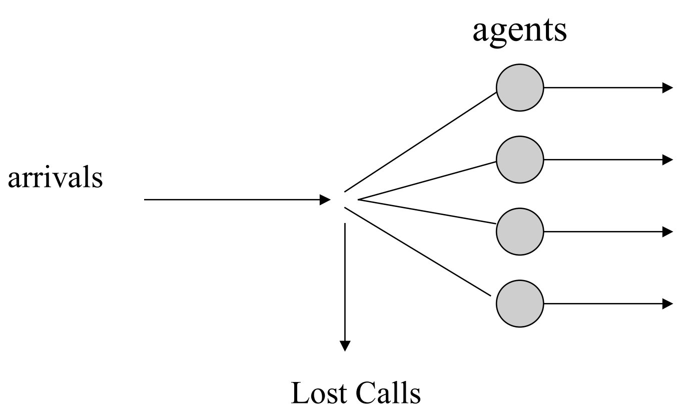
4CallCenters.com



Erlang-C



Erlang-B



Workforce Management: 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)

Potential Gain from "Perfect" Multi-Site Routing = Pooling

K = # sites

N = # agents/site

Calculate

$$R(K,N) = \frac{\text{ASA for } K \text{ isolated sites, each w/ } N \text{ agents}}{\text{ASA for FCFS with } K \times N \text{ agents}} \geq 1$$

Parameters: $K = 2, 16, 50$ # sites

$N = 10, 20, 50$ # agents/site

$= 0.9, 0.8$ occupancy

$= 0.9$

K \ N	10	20	50
2	2.43	2.67	3.35
16	83.6	206.5	X
50	2006	X	X

$= 0.8$

K \ N	10	20	50
2	3.20	4.21	8.87
16	1688	60,528	X
50	24,533,333	X	X

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.

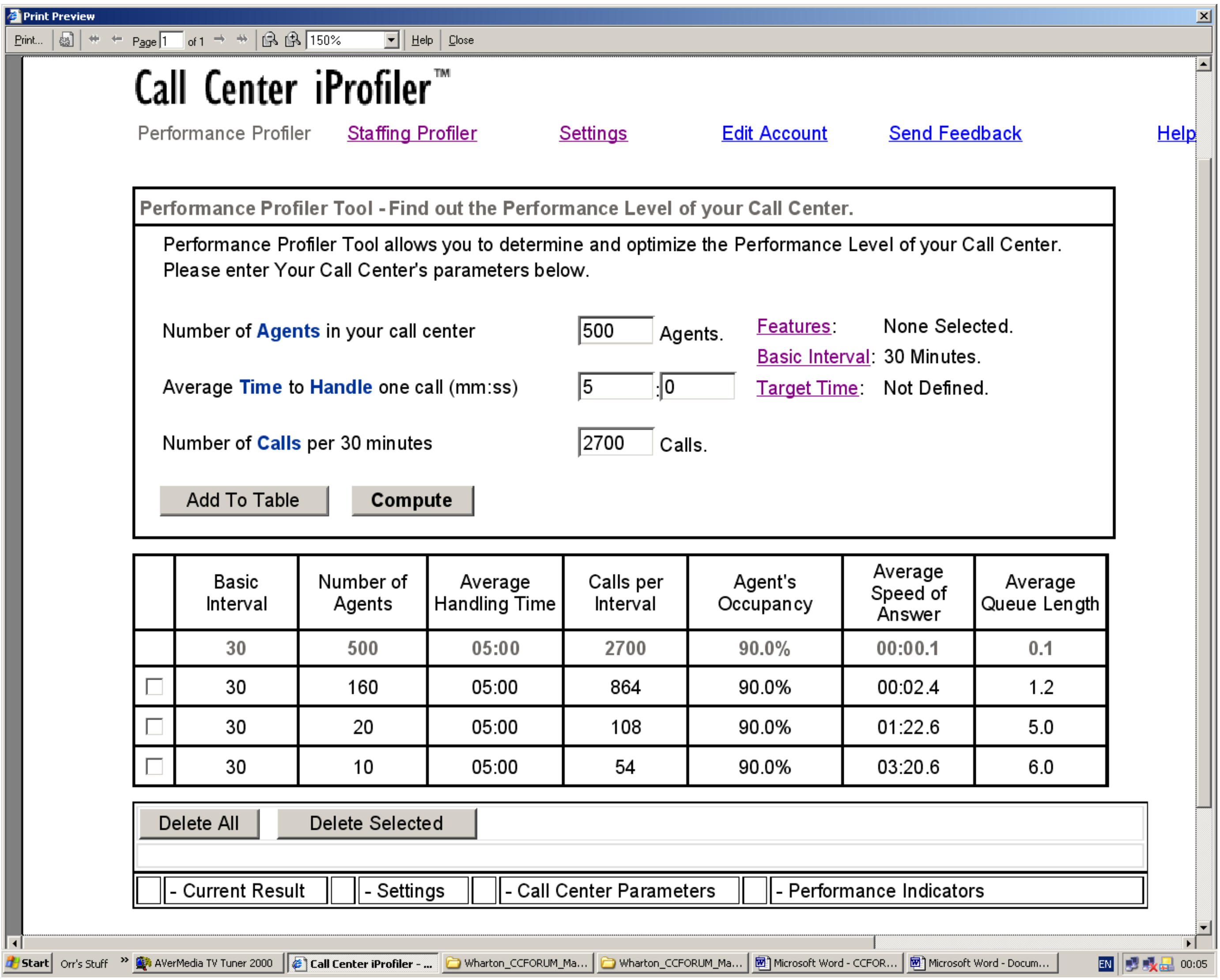
Features: None Selected.
Basic Interval: 30 Minutes.
Target Time: Not Defined.

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

Using iProfiler for calculating the first column of "Table = 0.9"


Print Preview

Print... Page 1 of 1 150% Help Close

Call Center iProfiler™

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

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: 500 Agents.

Average **Time** to **Handle** one call (mm:ss): 5 : 0

Number of **Calls** per 30 minutes: 2700 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
	30	500	05:00	2700	90.0%	00:00.1	0.1
☐	30	160	05:00	864	90.0%	00:02.4	1.2
☐	30	20	05:00	108	90.0%	01:22.6	5.0
☐	30	10	05:00	54	90.0%	03:20.6	6.0

Delete All | Delete Selected

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

$$3{:}20.6 / 1{:}22.6 = 200.6 / 82.6 = 2.43$$

$$3{:}20.6 / 0{:}02.4 = 83.6$$

$$3{:}20.6 / 0{:}00.1 = 2006$$

“First National City Bank Operating Group”

“By tradition, the method of meeting increased work load in banking is to increase staff. If an operation could be done at a rate of 80 transactions per day, and daily load increased by 80, then the manager in charge of that operation would hire another person; it was taken for granted…” (Harvard Case)

1:1 Staffing: Performance Analysis (Erlang-C = M/M/N)

8 transactions per hour $\Rightarrow$ E(S) = **7:30 minutes** (=M)

λ/hr	N Agents	ρ = OCC	Q-Length	ASA min
8	2	50%	0.3	2:30
16	3	67%	0.9	3:20
24	4	75%	1.5	3:49
32	5	80%	2.2	4:09

λ/hr	N	OCC	Q-Length	ASA
72	10	90%	60	5:01
120	16	93.8%	11	5:29
400	51	98%	42	6:18
640	81	98.8%	70	6:32
1,280	161	99.4%	145	6:48
2,560	321	99.7%	299	7:00
3,600	451	**99.8%**	423	**7:04**
↓	↓	↓	↓	↓
∞	∞	**1**	∞	**7:30 !**

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

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

Indeed $\overline{W}_q \approx \overline{W_q \mid W_q > 0} = \frac{1}{N} \cdot \frac{\rho_N}{1-\rho_N} \cdot E(S) \rightarrow E(S) = 7{:}30$ **!**

since $\rho_N = \frac{\lambda_N \times E(S)}{N} = \frac{8(N-1) \times 7.5/60}{N} = \frac{N-1}{N} = 1 - \frac{1}{N}$

$N(1-\rho_N) = 1 \quad , \quad \rho_N \rightarrow 1 \; .$

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	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	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	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	:03:38	: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		366	: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	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	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

1 abandonment (after 1 second)

Offered load $\mathbf{R} = \lambda \times M$

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

$= 50$ Erlangs

Occupancy $\rho = R/N$

$= 50/100 = 50\%$

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

$\Rightarrow$ Classical Queueing Theory (Erlang-C = M/M/N)

Quality-driven 100 agents, 50% utilization

⇒ **Can** increase offered load (more calls) - **by how much?**

Erlang-C

N=100 **M = 3:45 min.**

λ/hr	**OCC**	**ASA**	**% Wait ≤ 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 agents = 1fast agent.

Changing N (Staffing Level) in Erlang-C

			M = 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)

Rules of Thumb: Operational Regimes

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

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

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

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

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

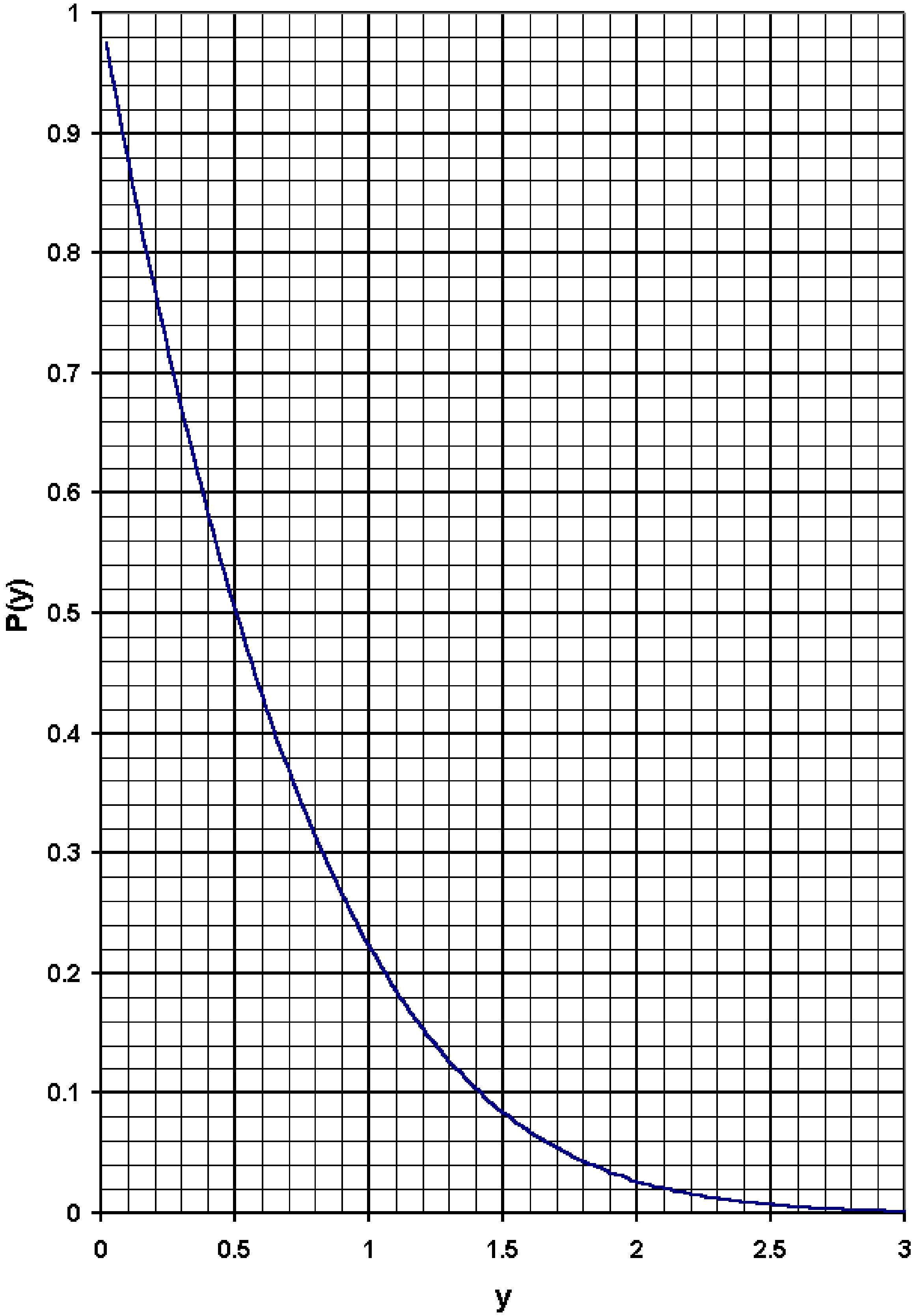
QED regime $(\%\{\text{Wait} > 0\} \to \alpha,\ 0 < \alpha < 1)$

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

%{Wait > 0} = P(y): Halfin-Whitt approx.for Erlang-C

How to determine **regimes** ? Strategy, Economics

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



Strategy: Sustain Regime under Pooling

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

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

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

4 CC: $= 1200$, $\quad \text{AHT} = 5$, $\quad R = 100$; $\quad N=?$

Quality-Driven: maintain OCC at 83.3%.

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

Efficiency-Driven: maintain ASA at 15 sec.

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

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

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

9 CC: $= 2700$, $\quad \text{AHT} = 5$, $\quad R = 225$

Q: $N = 271$

E: $N = 233$

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

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	$\rho = \frac{R}{R+\Delta}$	$\frac{R}{R+\frac{\Delta}{m}} \uparrow 1$	$\boxed{\rho = \frac{R}{R+\Delta}}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \uparrow 1$
$\text{ASA} = \text{E}\left[\frac{\text{Wait}}{\text{E}(S)} \,\middle\vert\, \text{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \text{ASA}}$	$\frac{1}{m\Delta} = \frac{\text{ASA}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\text{ASA}}{\sqrt{m}}$
$\text{TSF} = \text{P}\left\{\frac{\text{Wait}}{\text{E}(S)} > T \,\middle\vert\, \text{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \text{TSF}}$	$e^{-mT\Delta} = (\text{TSF})^m$	$e^{-\sqrt{m}T\Delta} = (\text{TSF})^{\sqrt{m}}$

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

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

where $\qquad$ **d** = delay/waiting costs

c = service/staffing costs

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

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

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

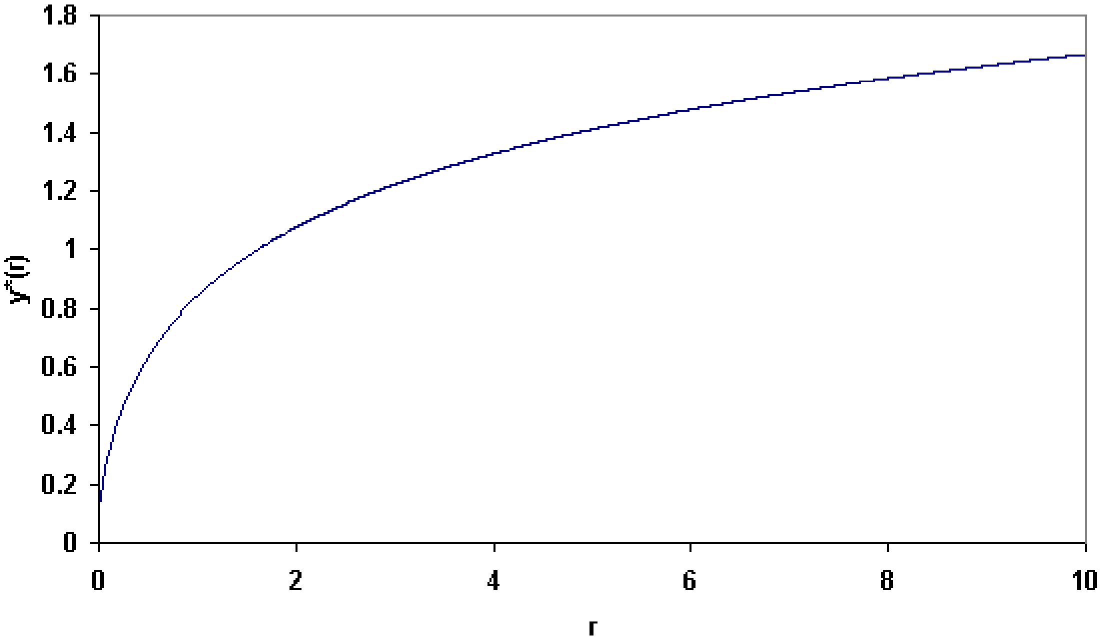
$\%\{\text{Wait} > 0\} \approx \mathbf{P}(\boldsymbol{y}^*) = \left[1+\frac{y^*\phi(y^*)}{\varphi(y^*)}\right]^{-1}$ $\qquad$ **Erlang-C**

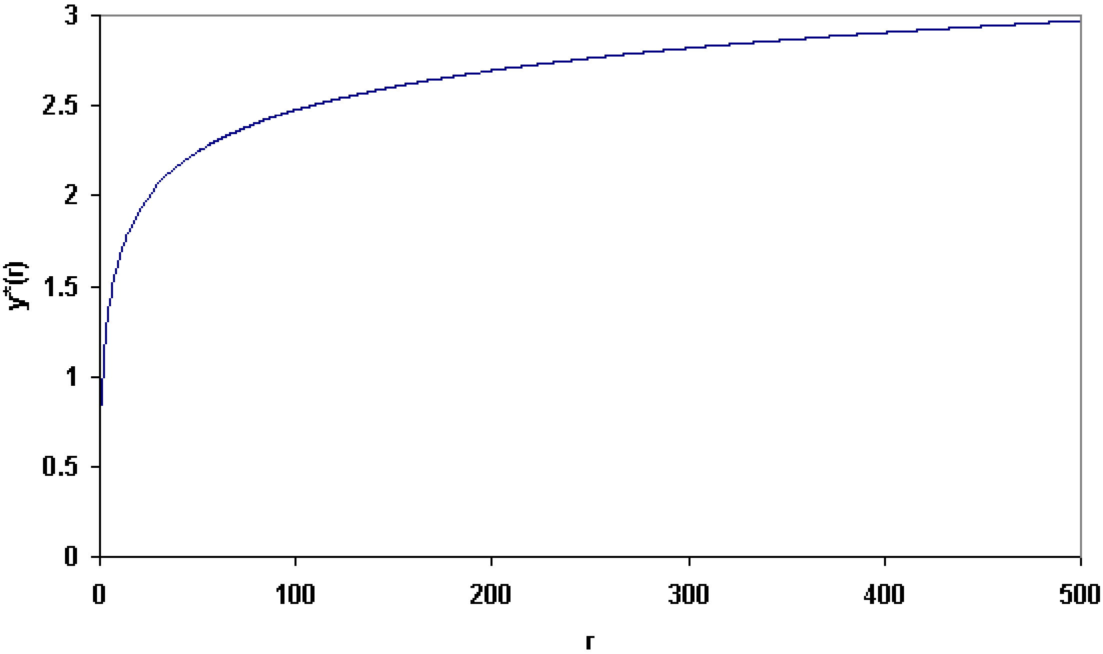
$\text{TSF} = \%\left\{\frac{\text{Wait}}{\text{E(S)}} > \text{T} \,\middle|\, \text{Wait} > 0\right\} = e^{-\text{T}\Delta}$

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

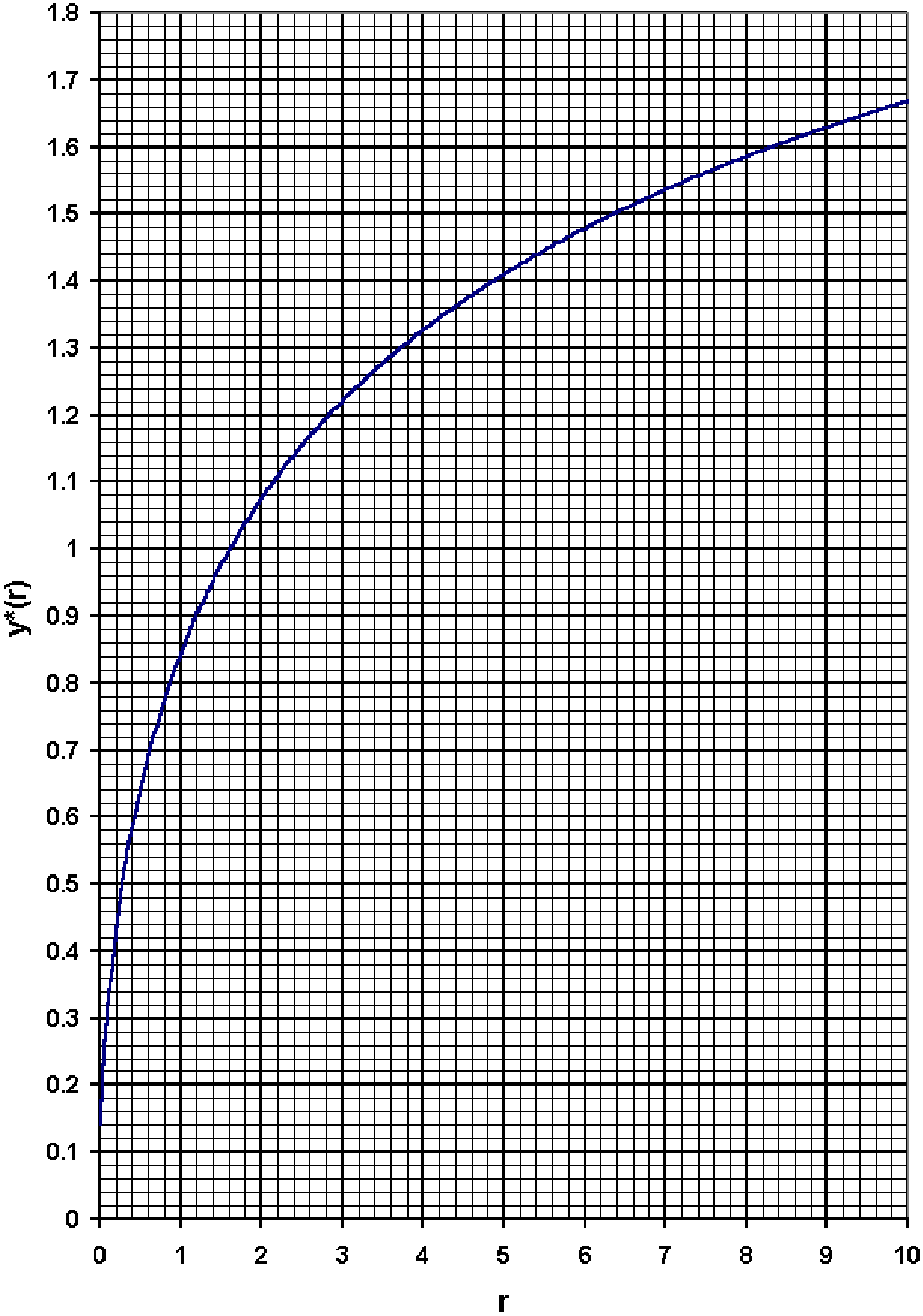
Occupancy $\qquad = 1 - \frac{\Delta}{N} \approx 1 - \frac{y^*}{\sqrt{N}}$

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**



$\surd$. Safety-Staffing: Overview

- Parameters

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

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

Δ : **safety-staffing**

- Simple, robust, accurate, relevant, instructive

$\surd$. Safety-Staffing: Overview (cont'd)

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

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

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

Relevant: Medium to Large CC do perform as above.

Instructive: In large call centers, 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 R=Min ! $\Rightarrow$ **"Real" Problem?**

Performance:

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

Scenario Analysis: A "Best-Practice" Call Center

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

Peak hour analysis:

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

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

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

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

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

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

Performance(via Erlang -C):

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

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

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

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

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

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

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

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

Low valuation of customers' time, at $\frac{1}{3}$ of servers' time, yet

reasonable 80:20 performance? enabled by scale!

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

N* = 429 agents (vs. 411 before)

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

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

Scenario Analysis: on Economies of Scale

"Best Practice" call center

$\lambda = 1{,}800$ calls per peak hour; **M = 4 min.**

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

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

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

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

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

- What if M = **30 sec.** (as in 411 services), $N^* = 126$

  suffices for the above performance, which implies

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

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

Scenario Analysis: "Satisfization" vs. Optimization

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

(Folklore: $N^* = R + \bar{\phi}^{-1}(\varepsilon)\sqrt{R}$, $\bar{\phi} = 1 - \phi$,

based on 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.

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

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

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

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

Scenario Analysis: Reasonable Service Level ?

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

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

$\mathrm{M} = 4$ min. service time (avg)

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

Service level constraint: 1 out of 100 delayed (avg), namely

99% answered immediately.

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

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

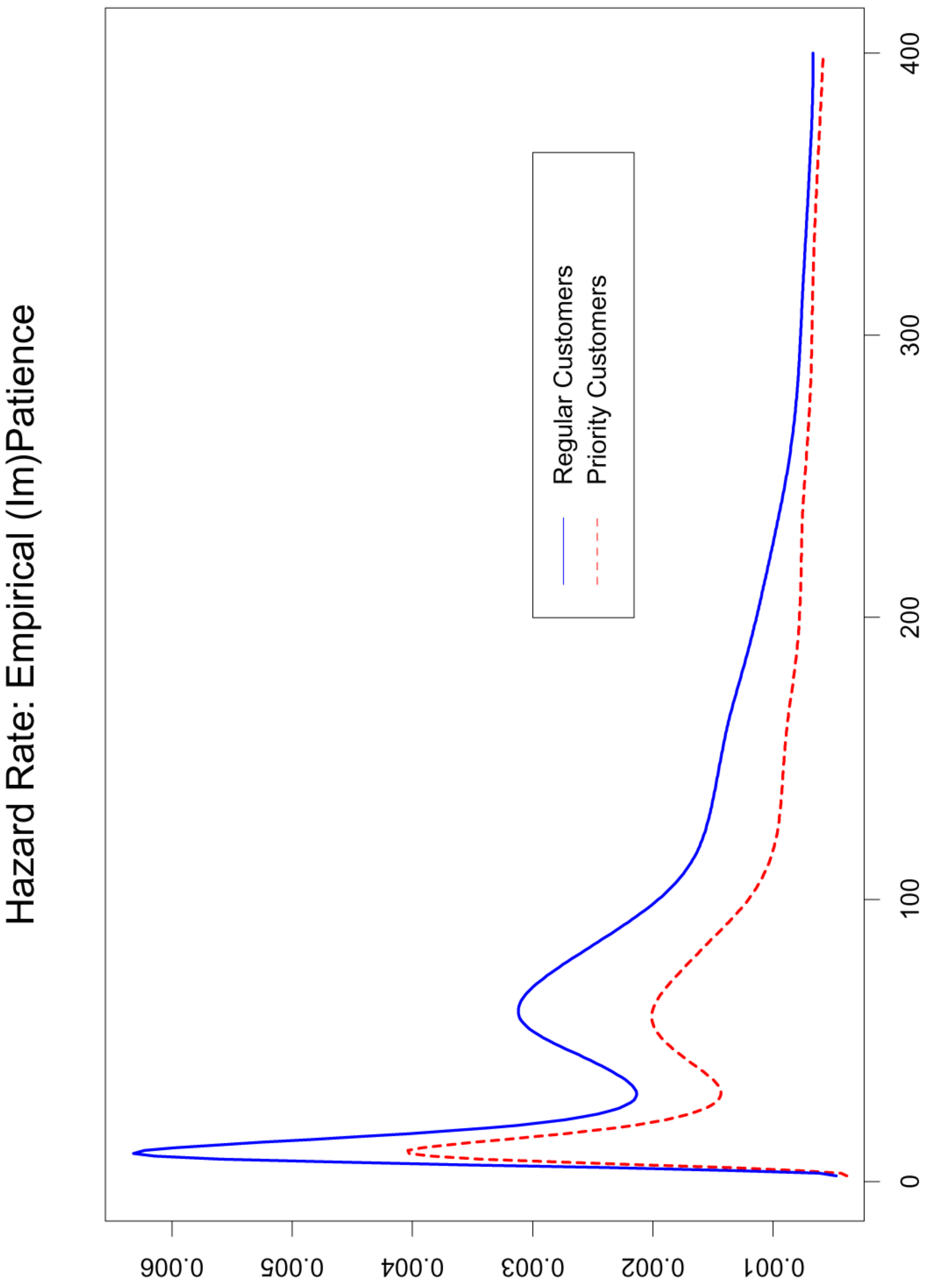
Valuation of customers' time as being worth 75-fold of agents' time seems reasonable only in extreme circumstances:

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

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%



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's Impact (+ Busy Signals)

Call Center iProfiler™

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

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

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

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

Add To Table | **Compute**

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

Delete All | Delete Selected

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

Erlang-A: Input, Inference

Input parameters:

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

Arrival rate ( ): 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,

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

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

Erlang-A: Pooling in the QED-Regime

Base Case: Call Center in the QED-Regime

with

%{Wait > 0} ~ satisfactory

%{Abandon} ~ "

ASA ~ "

Forecast: Pooling **m** call centers into a single one (m = 4)

(Load increases by a factor of m.)

Sustain present service level (remain QED) via:

- Observe N_{old} (110)
- Calculate offered load $R_{old} = \lambda \times M$ (100)
- Calculate service grade $y = (N - R)\sqrt{R}$ (1)
- New staffing level $N = mR + \sqrt{mR}$ (400+20)

Expect new performance as follows:

%{Wait > 0} unchanged

%{Abandon} & ASA improved by factor $\sqrt{m}$: EOS (2)

Figure 5
American data. Beta vs P{Ab}

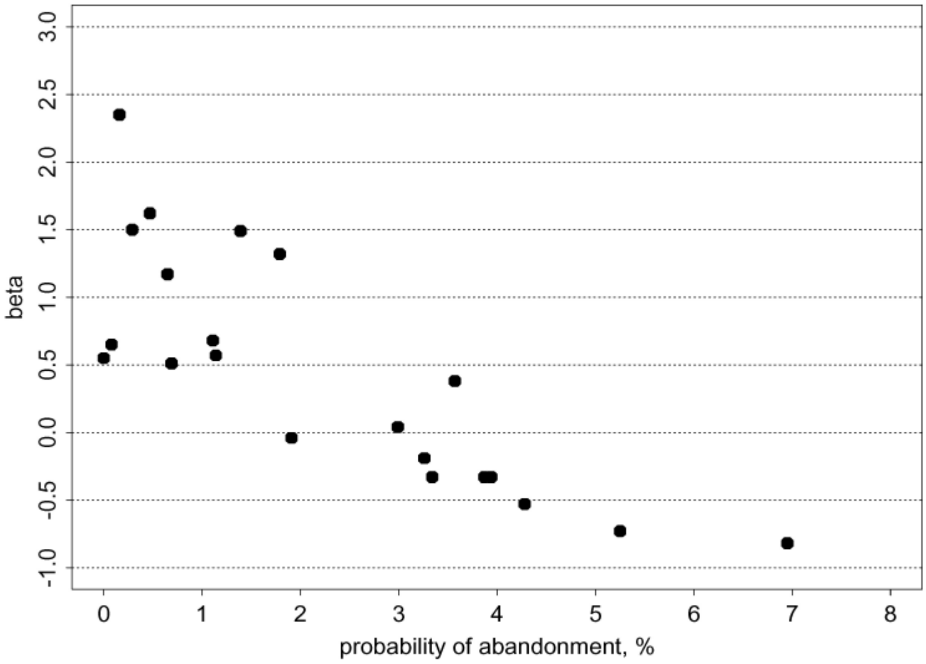
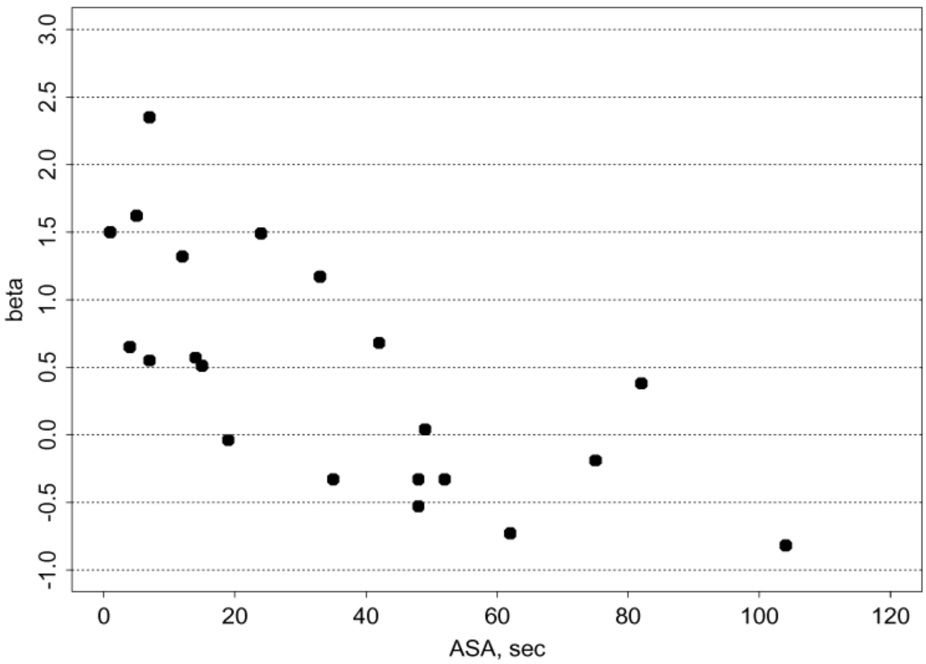
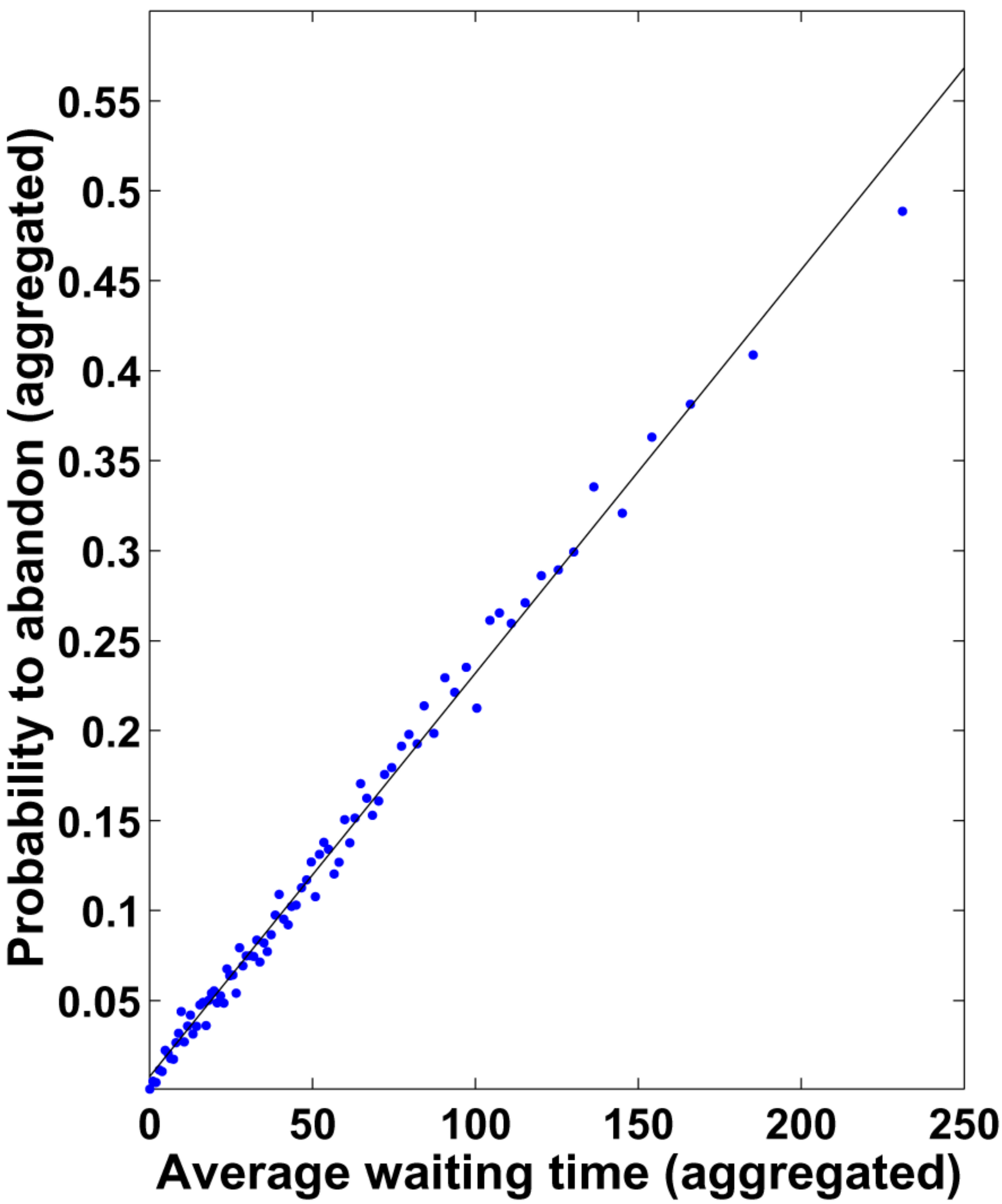


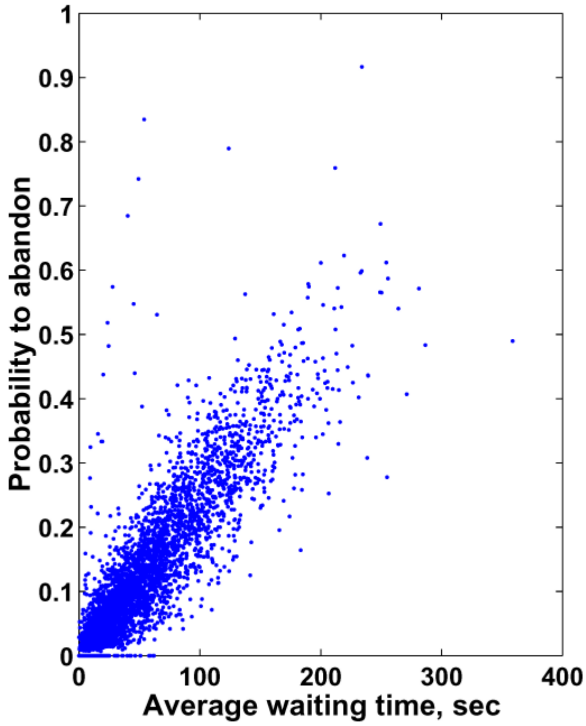
Figure 6
American data. Beta vs ASA

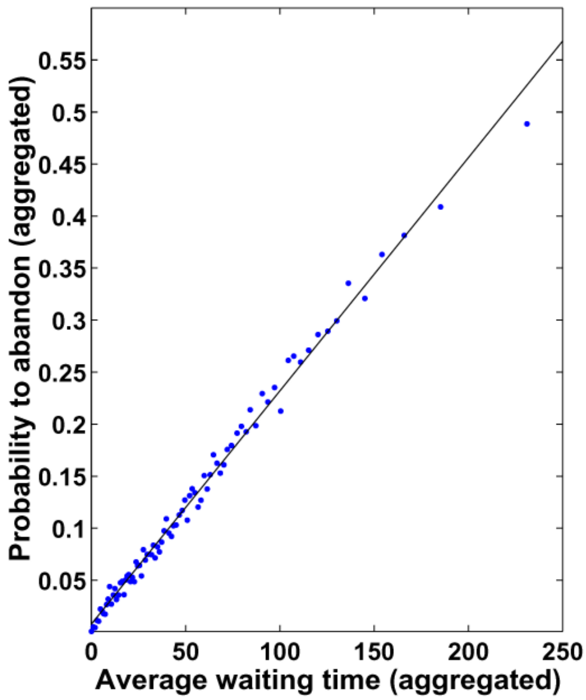


Fraction Abandoning vs. Average Waiting Time

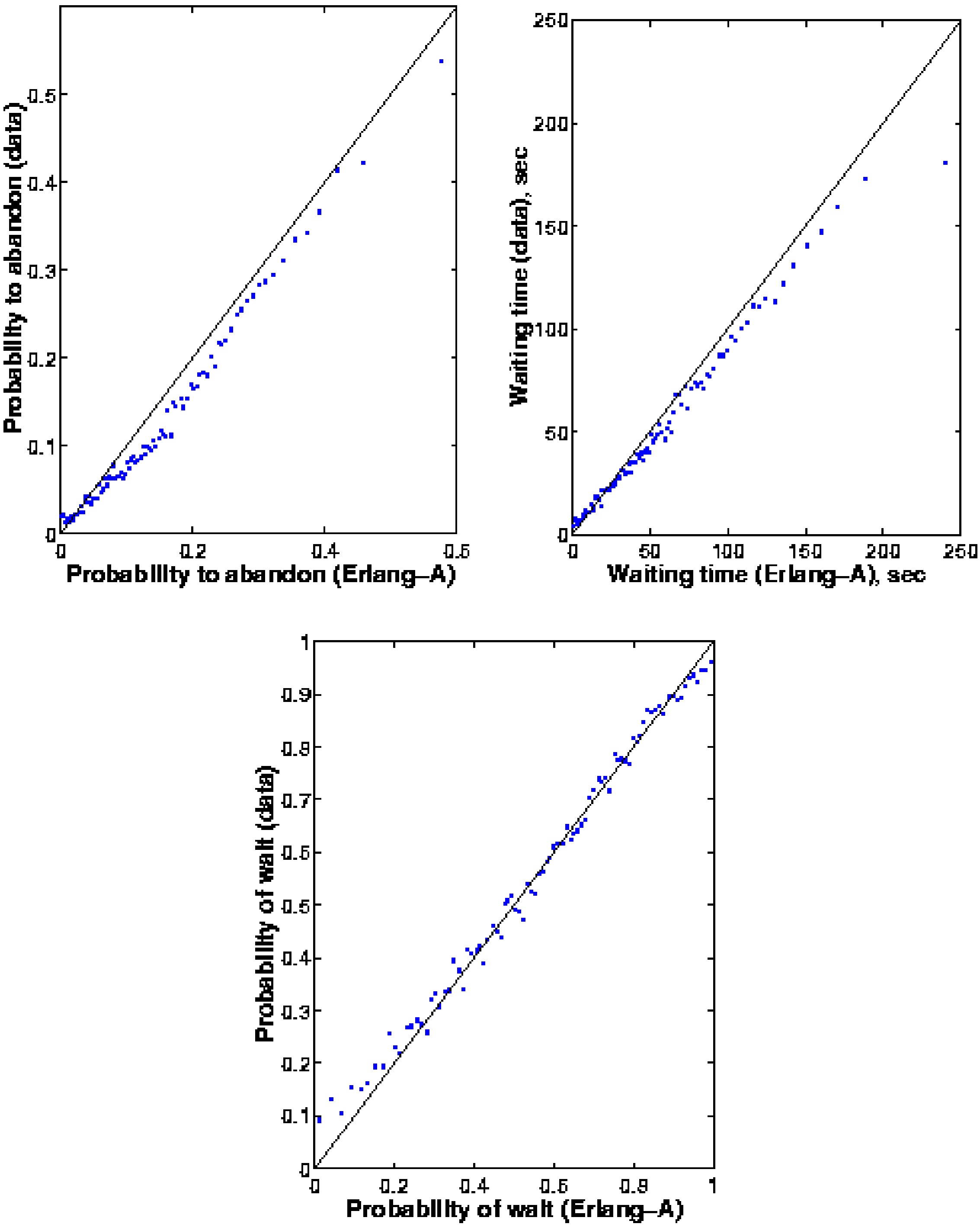


Fraction Abandoning vs. Average Waiting Time





Erlang-A Formulae vs. Data Averages



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.

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

- Supported by ongoing research (Wharton).

Designing Call/Contact Centers

with Impatient Customers:

10 Years History, or A Modelling Spectra

1. Kella, Meilijson: Practice $\Rightarrow$ Abandonment important
2. Shimkin, Zohar: No data $\Rightarrow$ Rational patience in Equilibrium
3. Carmon, Zakay: Cost of waiting $\Rightarrow$ Psychological models
4. Garnett, Reiman: Palm/Erlang-A to replace Erlang-C/B

   as the standard Steady-state model
5. Massey, Reiman, Rider, Stolyar: Predictable variability $\Rightarrow$

   Fluid models, Diffusion refinements
6. Ritov, Sakov, Zeltyn: Finally Data $\Rightarrow$ Empirical models
7. Brown, Gans, Haipeng, Zhao: Statistics $\Rightarrow$ Queueing Science
8. Garnett, Atar, Reiman: Skills-based routing $\Rightarrow$ Control models
9. Nakibly, Meilijson, Pollatchek: Prediction of waiting $\Rightarrow$

   Online Models and Real Time Simulation
10. Garnett: Practice $\Rightarrow$ 4CallCenters.com

Staffing the "Modern" Basic Call Center

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

   - Conceptual: Halfin & Whitt

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

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

   - Conceptual: Garnett & Reiman

   - Dimensioning: (Borst & Reiman) in progress

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

   - Ample-server heuristics: Jennings & Massey & Whitt

   - Conceptual part (Massey & Rider) in progress

   - Dimensioning: open

4. General Service Time (for all the above)

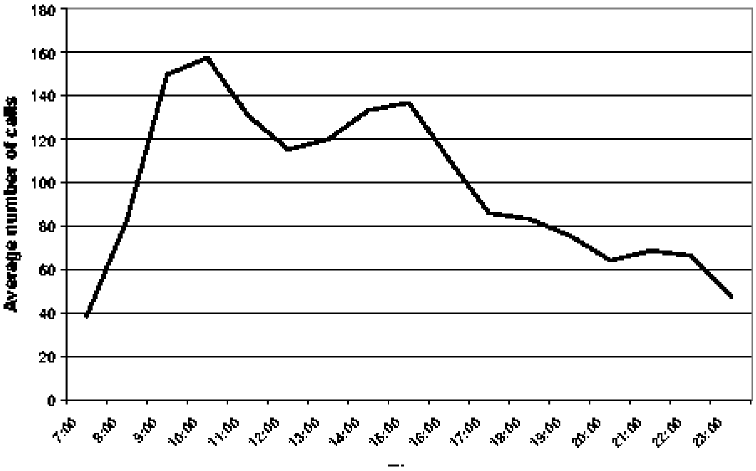
   - Conceptual supported by Puhalski & Reiman, M/PH/N

   - M/D/N (Jelenkovic & Momcilovic) in progress

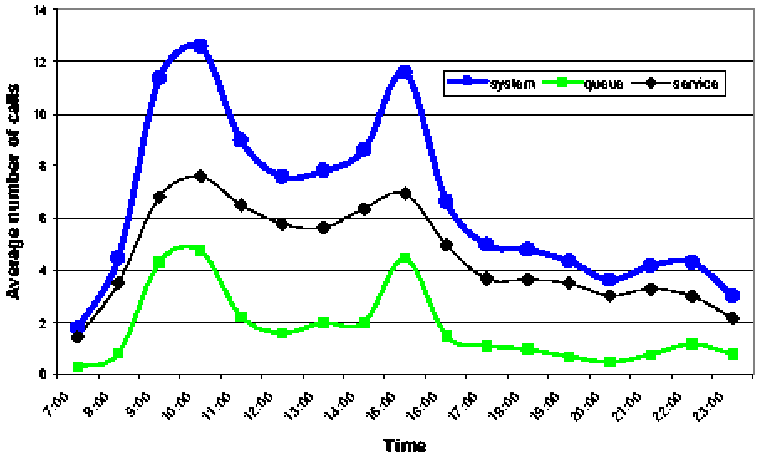
   - M/G/N open and challenging (measure-valued limit)

Time-Varying Queues: Predictable Variability

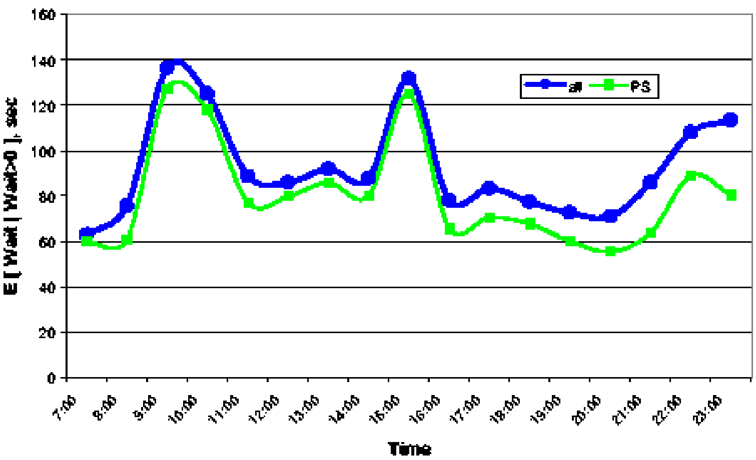
Arrivals



Queues



Waiting



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 | 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

Beyond the "Basic" Call Center

- Skills-based Routing:
  - Efficiency-drive (Stolyar): index control; ~ easy
    e.g., e-mail, chat(?)
  - QED (Atar, Reiman): $\sqrt{\cdot}$ - staffing; difficult
- Networks
  - IVR + ACD ; Retrials
  - Hierarchical Help Desk
  - Distributed Call Centers
- Information
- Profit Contact Centers: $-driven multi-media interface
- **Forecasting** (Brown, Haiping, Zhao): very important

An Introduction to Skills-Based Routing and its Operational Complexities

By Ofer Garnett and Avishai Mandelbaum
Technion, ISRAEL

(Full Version)

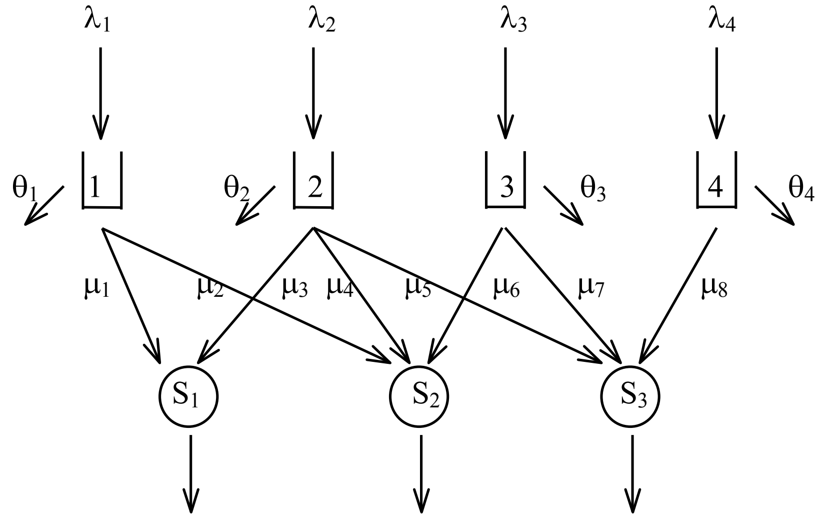
Contents:

1. **Introduction**
2. **N-design with single servers**
3. **X-design with multi-server pools and impatient customers**
4. **Technical Appendix: Simulations – the comutational effort**

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

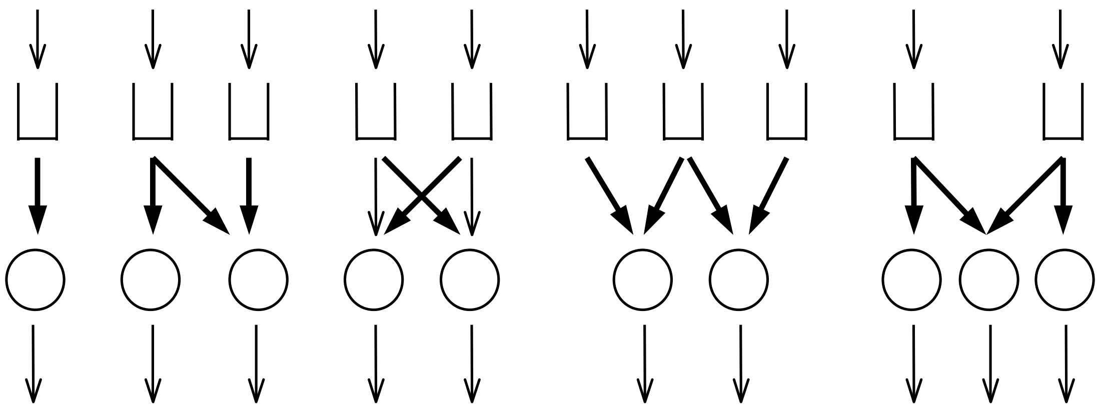
Consider the following multi-queue parallel-server system (animated, for example, by a telephone call-center):



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).



What Next

- Feedback: for work, for whom ?

  for classroom, for whom ?

  for Research ?

⇒ face-to face, note, e-mail = avim@tx.technion.ac.il

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

  (eg. Review lecture, do Scenarios)

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

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

- Feedback on Homework (⇒ I'll send solution)

- Download Charisma, and play some (eg. Redo HW).

Homework 7: *GazolCo's Call Center*[*]

Ten agents are busy answering calls at GazolCo's call center. Most calls are by customers calling to pay or inquire about their gas bills. Looking through recent ACD reports you see that the average handling time of each call is approximately 3.5 minutes. Methaney, the call center's manager, is sitting behind her desk playing with the screen saver's settings while awaiting the opening remarks of your analysis. As for you - your head is all clouded and you feel a bit queasy, but gradually you begin to recall a long forgotten assignment you once did for your Service Engineering course…

… calls are answered by 10 agents, the average handling time being 3.5 minutes. Normally the call volume is 150 calls per hour.

Start out with the iProfiler's "Performance Profiler" or Charisma's "Performance Profile".

1. Use the Erlang-C (M/M/N) model (no features) to answer the following questions:

   Record the change in the average speed of answer and agent's occupancy as the call volume gradually increases from 150 to 180 calls per hour (test at least 4 values). Can you explain the phenomenon you encounter in terms of system stability?

2. Continue your analysis using the Erlang-A (M/M/N+M) model (i.e. the Erlang-C model with the addition of exponential abandonment).
   (Select the "abandons" feature).

   a. Set the average patience parameter to a value that seems reasonable (keep in mind that the average handling time is 3.5 minutes). What value have you selected ?

   b. Repeat 1 and compare the results. What are the "positive" side-effects of abandonment?

   c. How do you expect the following performance indicators to change (increase/decrease) as the average patience parameter increases?
      I. % Abandoned
      II. Average speed of answer
      III. Average queue length
      IV. Agent's occupancy
      Test this with values of average patience ranging over 30, 90, 300, 450, 600 seconds (with the "normal" call volume).

[*] Prepared by Ofer Garnet; modified by Sergey Zeltyn.

d. How about the "fraction answered within 2 minutes"? Try and give a qualitative explanation to the phenomenon that you observe.

e. The average speed of answer (ASA) is a common "service measure", meaning that it is frequently regarded as a "score" given to the call center. It is thus constantly monitored and staffing levels are planned so as to meet given "service goals". Use 2c to argue against the use of ASA as an exclusive "service goal". (In light of 2c, how could you improve your call center's ASA?).

From here on assume that average patience is 2 minutes.

f. Repeat 2c but now vary the average handling time (use the same range 30-600 seconds as with patience). Variations of which parameter (patience or handling time) has a greater impact on performance?

g. Plot the fraction of calls <u>abandoned</u> within T seconds, T = 0, 10, 20, 30, 40, 60, 120, 180. Use this data together with the total fraction of calls abandoned to plot an approximate density function of the abandoning calls' waiting time.

h. Check what happens at the call center when there is a surge of calls which is double or triple the normal call volume (i.e. 300 or 450 calls per hour). Give a description of how "bad" things get, based on your results.

i. To maintain the original (i.e. with the "normal" call volume) fraction of abandoned calls when these surges occur, do you need more or less than double/triple the original number of agents? What is the reason for this? (use the iProfiler's "Staffing Profiler" or Charisma's "How Many Agents").

j. Garnet, Mandelbaum & Reiman (GMR), in their paper "*Designing a Call Center with Impatient Customers,*" suggest a staffing rule (*rationalized staffing*) that ensures *both* high quality and efficiency of service (given arrival rate to the call center is sufficiently large). GMR follow earlier work by Whitt, and both will be described later in our course. The present question, which continues Part i. above in some sense, demonstrates GMR's staffing rule.
Assume that performance measures of a given call center are considered reasonable. Call this the "Base Case", and assume for concreteness that this is the call center described above (normal circumstances, 2 minutes average patience). Suppose that the arrival rate increase by a factor *m*. (For example, by pooling *m* call centers into a single large call center.) It turns out now possible to both increase servers' utilization (*efficiency*) and improve service level (*quality*). (One typically expects to achieve only one of these two.)
Let ρ denote the offered load per server, where
offered load per server = (arrival rate * average service time) / (number of agents).

GMR rule: Choose the number of agents so that $(1-\rho)$ decreases by factor $\sqrt{m}$.

For example, consider our base case: 150 calls per hour, average handling time 3.5 minutes and 10 agents. Then the offered load per server is equal to 87.5%. If the arrival rate increases to 600 calls per hour (by factor 4), we should decrease $(1-\rho)$ by 2, namely $\rho = 93.75\%$. The closest approximations to this value of ρ are achieved with 37 agents ($\rho = 94.59\%$) and with 38 agents ($\rho = 92.11\%$).

Then theory predicts that the following changes in performance measures are expected (approximately):

- Probability to get service immediately P{Wait=0} is sustained on the same level as in the base case.
- ASA decreases by factor $\sqrt{m}$.
- Average queue length increases by factor $\sqrt{m}$.
- Probability of abandonment decreases by factor $\sqrt{m}$.

How can you explain the fact that ASA and the average queue change in the opposite directions? Which performance measure of the two is more important from a customer's point of view? Why could queue length be a significant performance measure in a call center?
The following table was partially filled in order to check the theoretical statements above:

Number of Agents	Service Time	Arrival Rate	Average Patience	Occupancy	P{Abandon}	ASA	P{Wait=0}	Average Queue
10	03:30.0	150	02:00.0					
37	03:30.0	600	02:00.0					
38	03:30.0	600	02:00.0					
82	03:30.0	1350	02:00.0					
83	03:30.0	1350	02:00.0					
144	03:30.0	2400	02:00.0					
145	03:30.0	2400	02:00.0					

Explain how arrival rates and number of agents in the four bottom lines were chosen. Fill in the table using Charisma or iProfiler and comment on the degree of compliance between the table and the theoretical statements above. What can you say about changes in occupancy?
Technical Remark. iProfiler does not allow to calculate P{Wait=0} (Charisma does.) If you use iProfiler, compute P{Wait > 2 sec} as a proxy instead.

3. One easy-to-implement mechanism for preventing extreme overloading as in 2h is to reduce the number of trunks available. (An arriving call with all trunks occupied encounters a "busy" tone.) So far, using the M/M/N and M/M/N+M models, we have assumed that the system has an unlimited waiting capacity. In reality the capacity is always finite, but is frequently large enough to practically eliminate "blocking". Use the M/M/N/B+M model (select the trunks and abandons features) for the following section which tests the behavior of a system with busy tones.

a. Test the performance with the various call volumes ("normal", "double" and "triple") and the following trunking levels: 10, 15, 20. Based on the results, which trunking level seems best to you? (remember that the objective is to achieve a "safety valve" effect).

b. One of this model's performance measures is the "Average Trunks Utilized". Construct a formula for this measure from the number of agents, agent's occupancy and average queue length.

c. The benefits of limiting a system's capacity are even more significant in the case of call centers accessed via toll-free numbers.
   I. Why is this? (who's paying for the call? …).
   II. Which performance measure can one use to estimate this expense?
   III. What fraction of this expense is saved with 10, 15 and 20 trunks, at 300 calls per hour? (compare to a system with unlimited capacity using II above and the formula from 3b).

d. Assuming the number of trunks is 15 and the call volume is 300 calls per hour, anticipate the change in the number of agents needed to reduce the fraction of blocked calls by 5%. Now check your answer.
(use "Staffing Profiler" / "How Many Agents").

e. Repeat 3d with an average patience parameter of 5 minutes (start out by finding the fraction of blocked calls in the system with this new patience parameter). Draw transition diagrams of the corresponding Markov processes and explain the inconsistent behavior you've just encountered.

4. Another mechanism for controlling the workload is to "overflow" calls out of the queue when their waiting time reaches a certain time limit. Overflowed calls might be transferred to a different group of agents (or call center) or, as sometimes done, to a voice box. (The latter clearly being not very desirable from a service-level point of view!)
Select the "Overflows" feature (note that the "Trunks" feature deactivates).

   a. What are the drawbacks (service-wise) of such a mechanism? Can you suggest similar more sophisticated/sensitive mechanisms?

   b. Test the performance with the various call volumes ("normal", "double" and "triple") and at least 4 time limits in the range of 15-200 seconds. What time limit would you select for this call center?

Technicalities:

Here are some technical instructions and information concerning the assignment.

1. Software:

To perform the analysis and various calculations required in this assignment you can use either **Call Center iProfiler™** or **Call Center Charisma™**. Both of these can be found at www.4callcenters.com . You will be using tools that determine a call centers performance ("Performance Profiler") and help set the staffing levels needed to meet performance goals ("Staffing Profiler" / "How Many Agents"). These tools support various queueing models from the basic Erlang-C to state-of-the-art models including abandons, blocking and overflows.

Here's how you get started:

a. To use Call Center iProfiler you need to "Login" to the service, and then register. For a general overview of this service, take the "Tour" offered. For more details try accessing the "Help" after you login.
Advantages: Does not require installation; Accessible from any computer with internet; Multiplatform (PC, UNIX, …).

b. Call Center Charisma is a Window's application which can be downloaded from this site (you get a 30 day trial version). Installing this software on a PC is easy - just follow the instructions. Call Center Charisma has basic instructions appearing in the header of each tool and additional more detailed "Help".
Advantages: Offers two more advanced tool that are not available in Call Center iProfiler; Can export results to files easily read (and then plotted) by any spreadsheet.
Note that Charisma has an "Indicators" setting determining which performance indicators are visible - you will need to change these settings for the assignment.

2. General:

a. Keep your answers short and clear.
b. Unless stated otherwise, answers should present your analysis results in either tables or graphs - try selecting only the more important/interesting performance indicators . In most cases you have the freedom to choose the format that seems clearest to you. Using a spreadsheet is recommended.
c. Within the assignment, instructions concerning the software have this special "Century Gothic" font.
d. You are asked to fill out and hand in the attached "Feedback Questionnaire."
e. Any questions or problems should be addressed to Avi Mandelbaum.at

avim@tx.technion.ac.il