Dimensioning Large Call Centers

Avi Mandelbaum

Technion, IE&M, ISRAEL

(avim@tx.technion.ac.il ; 972 4 829 4504)

Atar, R.	Technion
Borst, S.	CWI
Garnett, O.	Technion
Jennings, O.	Stanford
Massey, W.	Princeton
Reiman, M.	Bell Labs
Rider, B.	Duke
Whitt, W.	AT&T
Zeltyn, S.	Technion

Service Engineering web site:
http://ie.technion.ac.il/serveng (+ instructors material)

Contents

Staffing the Modern Call Center

- Fundamental problem in service operations

- People = 60-70% costs of running a call center
- 3% U.S. workforce; 1000's agents in a Call Center

- Hierarchical (Classical: think M/M/N = Erlang-C)

- Forecasting: How many customers (Statistics, Time-Series)

* Staffing: How many agents (Queueing Theory)

- Shifts, Union constraints (LP, IP, Combinatorial Opt.)
- Individual Assignments (Heuristics, AI)

- The Modern Call Center (Q-Network)

- Hiring, Training (Aggregate Planning, Dynamic Prog.)
- Skills-based routing (Offline, Online)

- Research

- Scope: **many** servers, Erlang-**A**, time-varying, equilibrium
- Operational-regimes, Dimensioning, Control
- Service Engineering: Rules-of-thumb, Software

Workforce Management: Hierarchical Operational View

Forecasting (Customers, Servers)

Staffing: Queueing Theory (Erlang-A and beyond)

Service Level, Costs

FTE's
per unit of time

Shifts: IP, Combinatorial Optimization; LP

Union constraints, Costs

Shift structure

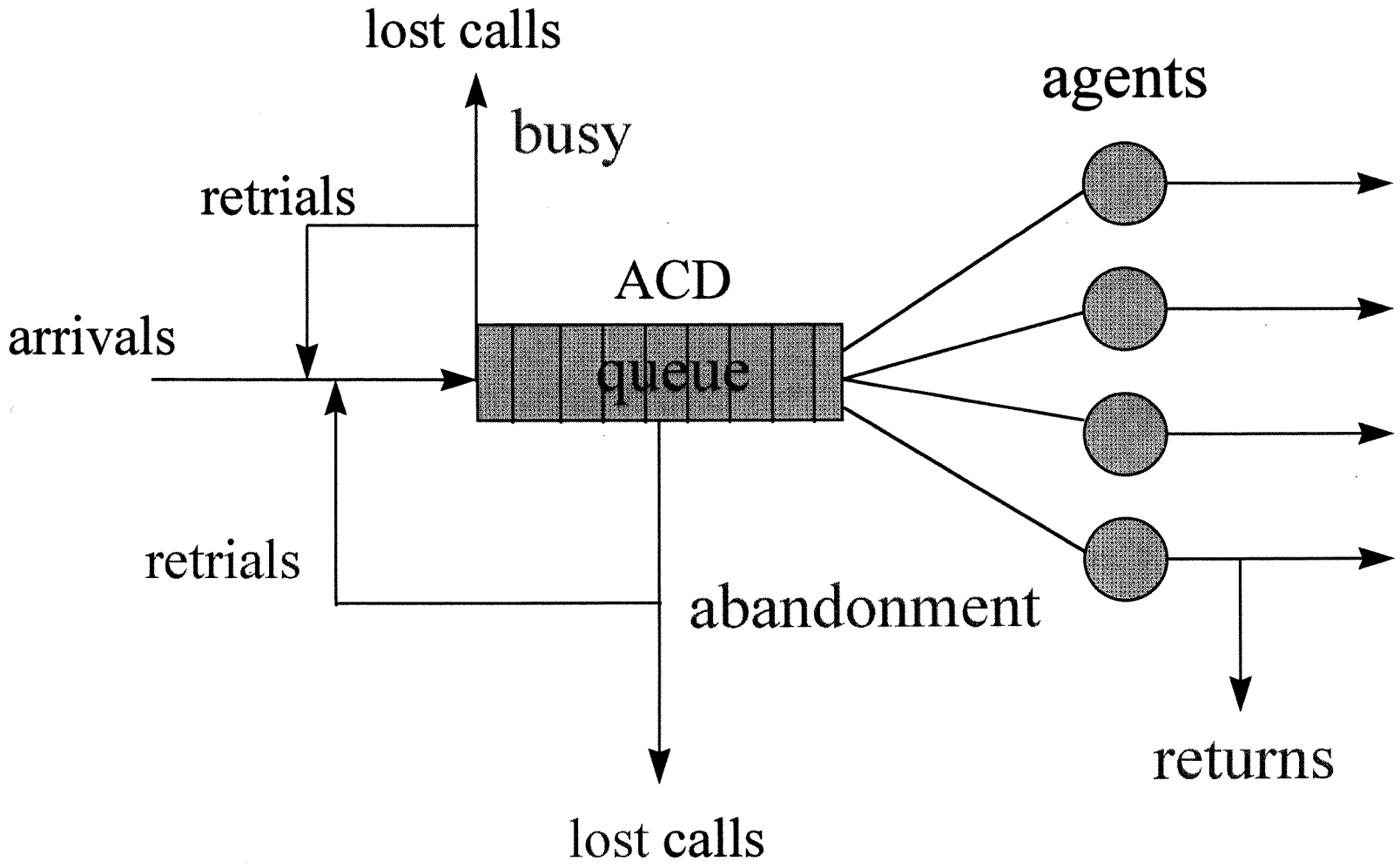
Scheduling: Heuristics, AI (Complex)

Individual constraints

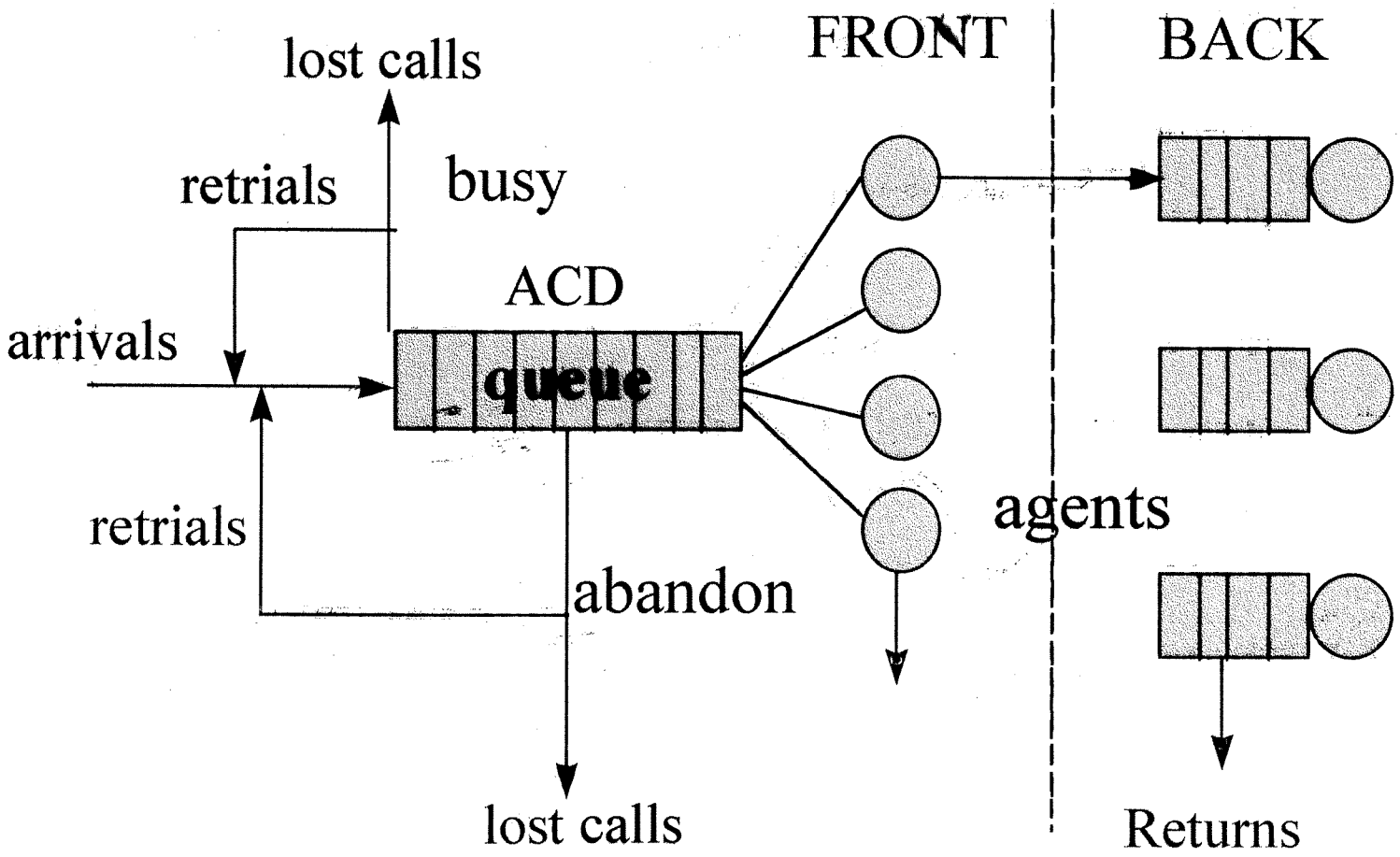
Agents Assignments

Online Skills-based Routing: Stochastic Control

A Simple Call Center



Telephone Network



What can be achieved

At what cost

Copy of Summary Interval - Order PK

Printed: 7/18/97 10:0

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

Time	Avg Speed Ans (W)	Avg Aban Time	ACD Calls (Σ)	Avg ACD Time (1/μ)	Avg ACW Time	Aban Calls (#Aban)	% ACD Time	% Ans Calls	Avg Pos Staff (N)	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time (P)
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	98	70	149		8		
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	51	7	4	51	2	16	61
12:30 AM*	:00:03	:04:10	14	:07:27	:00:33	1	89	52	5	3	48	1	26	63
1:00 AM*	:00:00		9	:04:54	:11:29	0	91	90	1	7	90	0	26	65
5:30 AM*			0			0	0		0	0		33	0	0
6:00 AM*	:00:00		12	:03:21	:00:19	0	21	100	7	2	100	9	2	19
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	29
7:00 AM*	:00:00		62	:03:34	:00:15	0	38	100	21	3	100	13	4	34
7:30 AM*	:00:00		93	:03:11	:00:34	0	36	100	30	3	100	7	4	32
8:00 AM*	:00:00		120	:03:37	:00:40	0	39	100	47	3	100	8	6	33
8:30 AM*	:00:00		193	:03:04	:00:14	0	44	100	61	3	100	10	7	37
9:00 AM*	:00:01		293	:03:25	:00:25	0	54	99	75	4	97	9	7	47
9:30 AM*	:00:02	:00:06	381	:03:45	:00:22	2	60	97	91	4	93	8	8	52
Peak → 10:00 AM*	:00:02	:00:01	416	:03:49	:00:28	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	6	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	6	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	8	44
1:30 PM*	:00:00		368	:03:52	:00:14	0	56	99	99	4	99	11	7	50
2:00 PM*	:00:01		393	:03:55	:00:17	0	51	100	106	4	100	10	5	46
2:30 PM*	:00:00		403	:03:58	:00:13	0	54	100	112	4	100	10	4	50
3:00 PM*	:00:00	:00:04	410	:04:02	:00:16	1	57	98	110	4	98	8	5	51
3:30 PM*	:00:00		347	:03:59	:00:14	0	50	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:48	:01:37	0	54	100	98	4	100	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	96	4	99	10	6	51
6:00 PM*	:00:01	:00:21	371	:03:28	:00:25	1	53	98	91	4	98	9	6	47
6:30 PM*	:00:00		260	:03:26	:00:13	0	41	100	90	3	100	8	4	37
7:00 PM*	:00:00		269	:03:24	:00:17	0	42	100	78	3	100	9	5	38

ACD

Rough Performance Analysis

Peak 10:00 – 10:30 a.m.

400 calls

3:45 minutes average service time

2 seconds ASA = Average Speed of Answer

1 abandonment (after 1 second)

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

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

$= 50$ Erlangs

Utilization $\rho = R/N$

$= 50/100 = 50\%$

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

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

Quality-driven: 100 agents, 50% utilization

⇒ **Can** increase offered load - but **by how much?**

M/M/N (Erlang-C)

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

λ/hr	ρ	$E(W_q)$ = 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 servers = 1fast server.

Changing N (Staffing Level) in M/M/N

E(S) = 3:45

λ/hr	N	ρ	$E(W_q)$	% 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

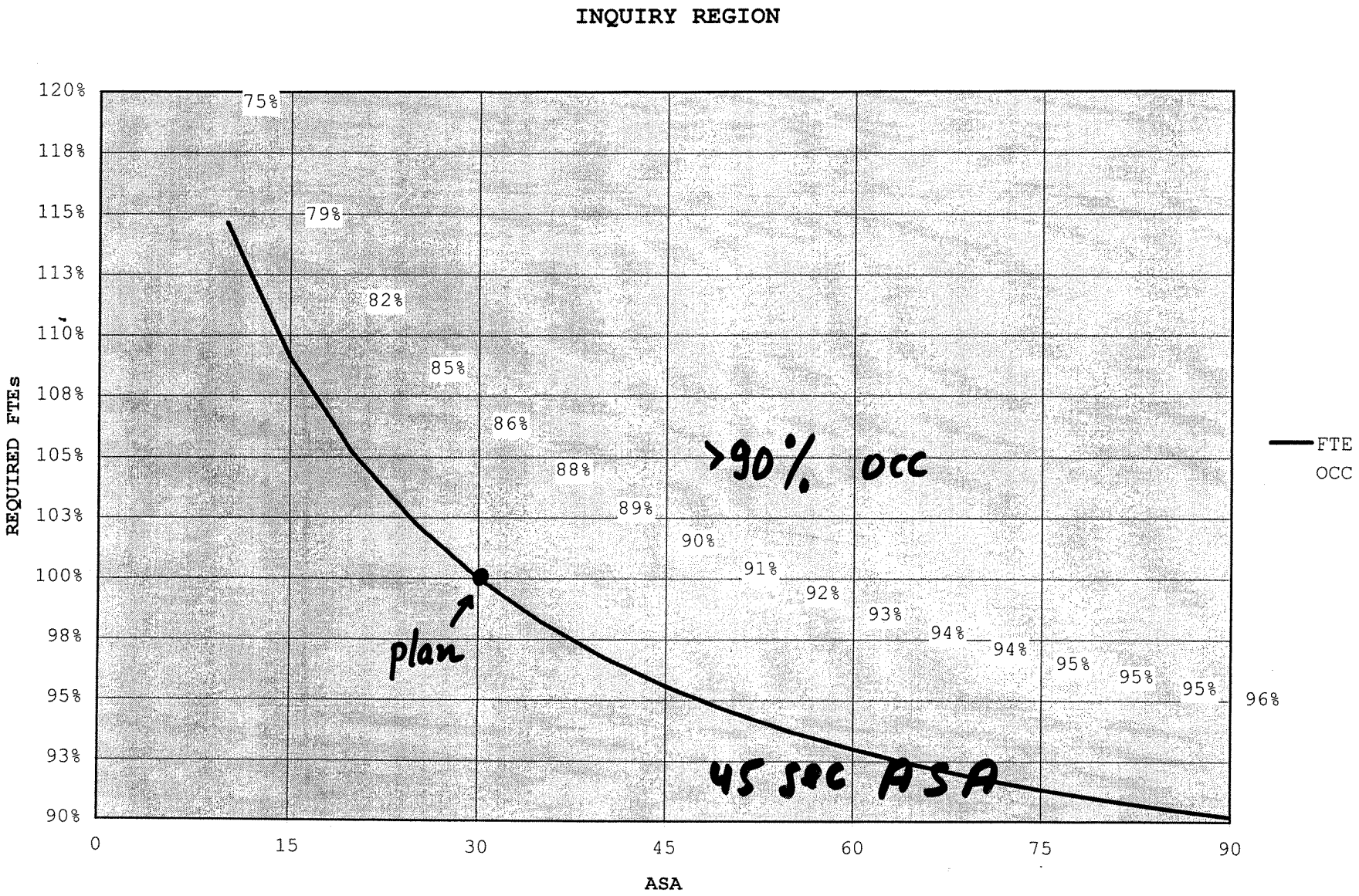
Heavy traffic, in the sense that $\rho > 95\%$;

Light traffic, 50% answered immediately.

⇒ **Rationalized Operation**: efficiency + high service-quality

Enabler: Economies of Scale in a

Frictionless Environment (e.g. Call Center)


INQUIRY REGION
REQUIRED FTEs
120%
118%
115%
113%
110%
108%
105%
103%
100%
98%
95%
93%
90%
75%
79%
82%
85%
86%
88%
89%
90%
91%
92%
93%
94%
94%
95%
95%
95%
96%
>90% occ
plan
45 sec ASA
FTE
OCC
0
15
30
45
60
75
90
ASA

Command Center Intraday Report

Updated Through: All Day

Date 06/13 - Tue		Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
Total:		129,960	126,321	2.8%	31	318	90.9%	88.4%	1531.7	1585.0	96.6%
INQ	Charlotte	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
INQ	Columbus MCSC	7,973	7,773	2.5%	36	314	94.9%	89.8%	89.2	94.5	94.4%
INQ	Phoenix	17,102	16,757	2.0%	31	298	92.7%	91.8%	187.3	194.8	96.2%
INQ	Scranton	1,257	1,254	0.2%	6	515	78.6%	28.9%	28.5	35.1	81.2%
INQ	Tampa	9,174	8,859	3.4%	42	366	91.5%	93.6%	123.1	125.9	97.8%
CEN	Bourbonnais	6,070	5,937	2.2%	33	362	86.7%	90.2%	86.0	88.4	97.3%
CEN	Bristol	10,667	10,505	1.5%	25	355	95.1%	93.1%	136.3	139.6	97.6%
CEN	Columbus Claims	5,258	5,153	2.0%	27	293	86.7%	89.8%	60.5	62.2	97.3%
STH	Atlanta	7,514	7,338	2.3%	40	318	82.1%	89.5%	98.6	99.8	98.8%
STH	Sherman	19,669	18,833	4.3%	46	252	93.8%	90.6%	175.5	174.9	100.4%
STH	Wilmington	10,422	9,888	5.1%	21	285	89.9%	92.1%	108.7	114.6	94.8%
WST	Visalia	14,277	14,164	0.8%	10	382	87.2%	85.0%	215.2	220.6	97.6%

12 CC's

6/13/00 - Tue

[redacted] - Center

Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
0	20,577	19,860	3.5%	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%	28	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1,286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	1,364	1,338	1.9%	33	296	99.0%	84.1%	222.5	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	1,174	1,160	1.2%	10	302	95.5%	93.6%	203.8	209.8	97.1%
13:00	1,018	999	1.9%	9	314	95.4%	91.2%	182.9	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	1,212	1,179	2.7%	23	304	96.6%	86.0%	206.1	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	89.1%	5.8	1.4	416.2%

Beyond Classical Browsing Theory 12

12

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: $\alpha = 1 \iff \beta = 0$ **Efficiency-driven**

No one waits: $\alpha = 0 \iff \beta = \infty$ **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 $\quad \mathrm{P}\{\mathrm{Wait} > 0\} \approx \alpha, \qquad 0 < \alpha < 1\ ;$
- Agents $\quad \rho \approx 1 - \dfrac{\beta}{\sqrt{N}}, \qquad 0 < \beta < \infty\ ;$
- Managers $\quad N \approx R + \beta\sqrt{R}\ , \quad R = \lambda \times \mathrm{E(S)}$ large;

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

where φ / ϕ are the standard normal density/distribution

Extremes:

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

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

The Halfin-Whitt Function

P(β) (later P(y))

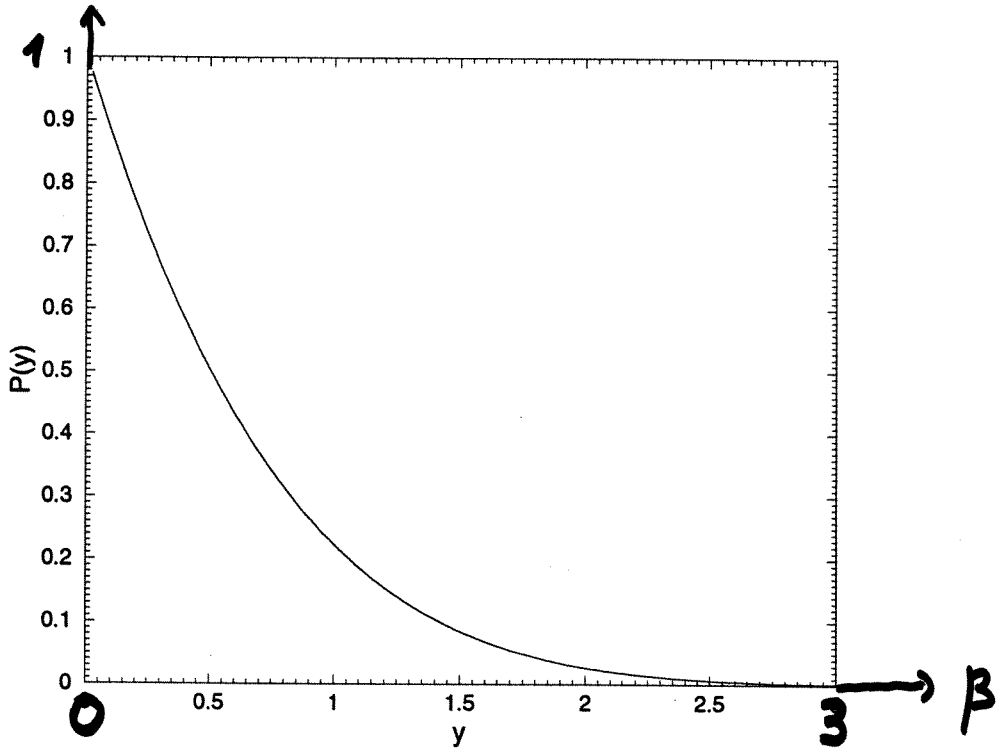


Figure 3: The Halfin-Whitt delay function $P(y)$.

hence the center operates with $y^* \approx 1.22$. Inverting $y^*(\cdot)$ in Figure 1 shows that, in this call center, an hour wait of customers is valued as 3 times the hourly wage of an agent. With this staffing level, it is expected that about 15% of the customers ($P(1.22) = 0.15$) are delayed; that 5% of the customers are delayed over 20 seconds (using (3) with $T = \frac{1}{12}$); and that, by (4), ASA equals 2.7 seconds (while those who were delayed actually averaged 18 seconds waiting).

□

But the staffing level in the example can be interpreted differently. To this end, recall that the prevalent alternative to the above optimization approach is constraint satisfaction. Specifically, in Example 8.5 it is shown that the least N that guarantees $\Pr\{\text{Wait} > 0\} < \epsilon$ is closely approximated by rounding up

$$N^* = R + P^{-1}(\epsilon)\sqrt{R}, \tag{5}$$

where $P(\cdot)$ is the Halfin-Whitt delay function introduced in (2). Returning to the above best-practice call center, $P^{-1}(\epsilon) = 1.22$ yields, as expected, $\epsilon = 0.15$.

Example 1.2

One should note that a constraint on the fraction of delayed customers is severe, hence it fits call centers that cater to say emergency calls. This can be nicely explained within our

15

$\sqrt{\cdot}$ Safety-Staffing: Conceptual Part

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

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

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

Expected Performance:

$$\%\ \text{Wait} \approx P(\beta) = \left[1 + \frac{\beta\phi(\beta)}{\varphi(\beta)}\right]^{-1}, \quad \beta > 0 \qquad (\alpha(\beta)\ \text{before})$$

$$\text{Congestion index} = \frac{E[\text{Wait}]}{E[\text{Service Time}]} \approx \frac{P(\beta)}{\Delta} \qquad \text{ASA}$$

$$\%\ \text{Wait} > T \times E(\text{Service Time}) \approx P(\beta) \times e^{-T\Delta} \qquad \text{TSF}$$

Conceptual insight: **Economies of Scale**

m-fold increase in R requires $\sqrt{m}$-fold increase in Δ to sustain % **Wait; But note improvement in ASA, TSF !**

√· Safety-Staffing: Performance

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

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

$= R + \Delta$ √· safety-staffing

Expected Performance:

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

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

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

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

Rules of Thumb: Operational Regimes

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

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

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

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

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

Rationalized $(P\{Wait > 0\} \to \alpha,\ 0 < \alpha < 1)$

$\mathbf{N = \lceil R + \beta\sqrt{R} \rceil}$, $\beta > 0$ **service grade**

How to determine β, or ε, or δ ?

More fundamentally, how to determine **regimes** ?

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

?

Service-Quality vs. Operations-Efficiency

With S. Borst, M. Reiman (1997 – 2001)

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

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

Example: 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 $P\{W_q > 0\} = \pi\left(N, \frac{\lambda}{\mu}\right) \to \pi_\lambda(x)$

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

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

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

Continuous Approximation of original **discrete** problem:

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

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

Rationalized: $d_\lambda \equiv d$;

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

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

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

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

√· Safety-Staffing: Performance

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

where **d** = delay/waiting costs

c = staffing costs

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

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

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

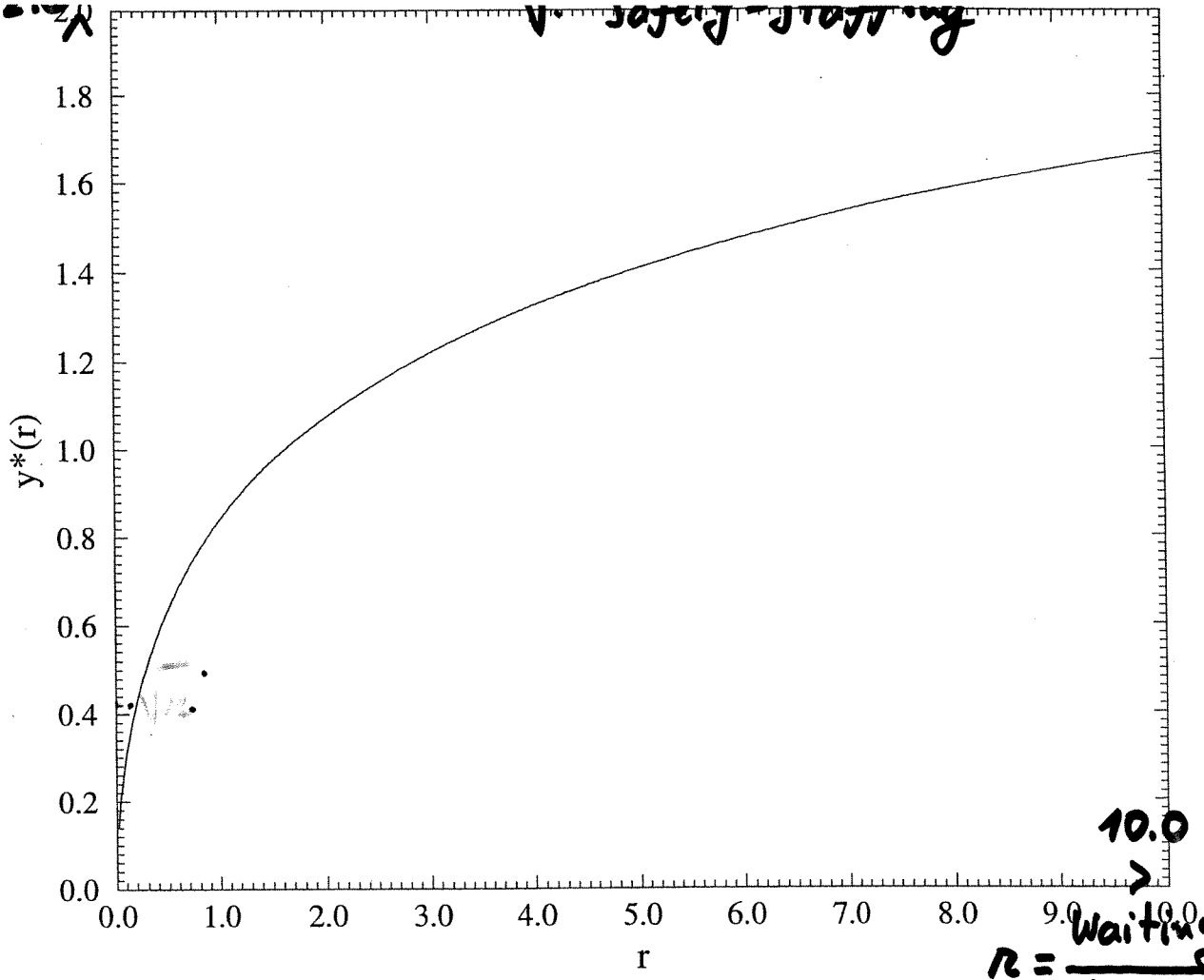
$$P\{\text{Wait} > 0\} \approx \mathbf{P}(y^*) = \left[1+\frac{y^*\phi(y^*)}{\varphi(y^*)}\right]^{-1} \quad \textbf{Erlang-C}$$

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

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

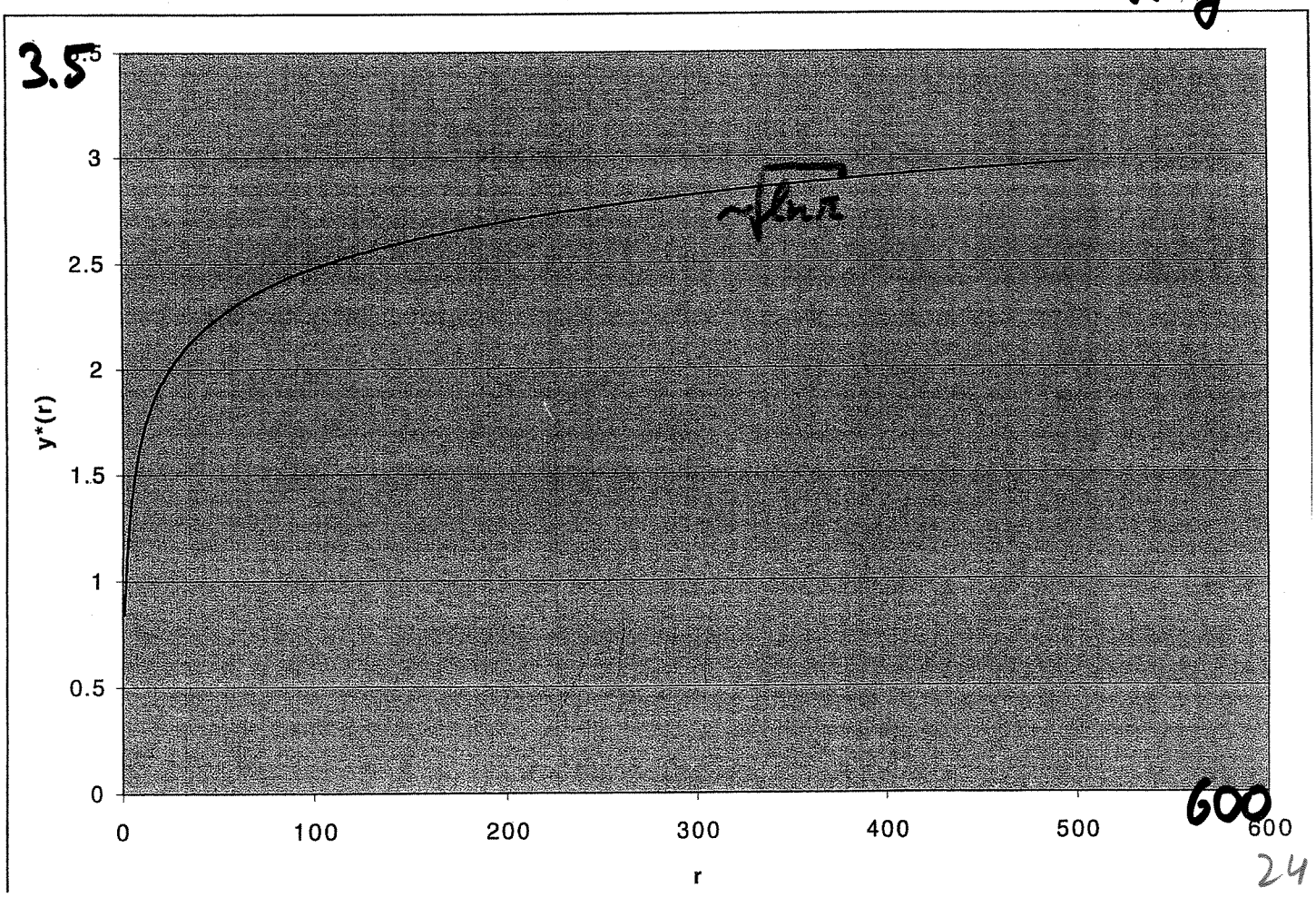
$$\text{Occupancy} = 1-\frac{\Delta}{N} \approx 1-\frac{y^*}{\sqrt{N}}$$

$y^*(r)$



10.0

$r = \frac{\text{waiting}}{\text{staffing}}$



600

√· 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.

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%

Relevant: Large call centers do perform as above.

Numerical Results

Given 'appropriate' y^*, let

$$n^* = \frac{\lambda}{\mu} + y^* \sqrt{\frac{\lambda}{\mu}},$$

and round off n^* to the nearest integer.

In all tests, let $c = 1$, $\mu = 1$

Rationalized:

1) $\lambda = 100$, $\boldsymbol{d} = 0.1, 0.25, 0.5, 1, 2, 4, 10$: exact for all 7 cases

2) $\boldsymbol{d} = 2$, $\lambda =$ all integers from 5 to 100: exact in 83 cases, off by 1 in 13 cases

Efficiency-Driven:

$\boldsymbol{d}_\lambda = \boldsymbol{d}\lambda^{-1/2}$ with $\boldsymbol{d} = 1$, $\lambda =$ all integer multiples of 10 from 10 to 200: exact in 19 cases, off by 1 in 1 case

Quality-Driven:

$\boldsymbol{d}_\lambda = \boldsymbol{d}\sqrt{\lambda}$ with $\boldsymbol{d} = 1$, $\lambda =$ all integer multiples of 10 from 10 to 200: exact in 16 cases, off by 1 in 4 cases

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

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

Formally: $P(\text{Wait} > 20 \text{ sec.}) < 0.2$

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

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

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: "Satisfization" vs. Optimization

Theory: The least N that guarantees $P\{Wait > 0\} < \varepsilon$ 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: $\lambda = 1{,}800$ calls at peak hour (avg)

$E(S) = 4$ min. service time (avg)

$R = 1800 \times \frac{4}{60} = 120$ 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

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

$ASA = E[Wait] = \mathbf{2.7}$ sec. average wait

$ASA \mid Wait > 0 = \mathbf{18}$ sec. average wait of delayed

Scenario Analysis: Reasonable Service Level ?

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

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

$E(S) = 4$ min. service time (avg)

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

Service level constraint: 1 out of 100 delayed (avg), namely 99% answered immediately.

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

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

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

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

Scenario Analysis: on Economies of Scale

"Best Practice" call center

λ = 1,800 calls per peak hour; **E(S) = 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.? $N^* = 140$

$$\Delta = 140 - 120 = 20 \quad \text{(safety staffing)}$$

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

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

- What if E(S) = **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: 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 abandonments.

Peak hour analysis:

$R = \lambda \times E(S) = 1800 \times 4/60 = 120$ Erlangs offered-load

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

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

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

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

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

Performance (via Erlang-C):

- P(Wait > 0) = P(1.22) = 15% delayed
- P(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

Figure 3

Italian data. Beta vs Average Wait

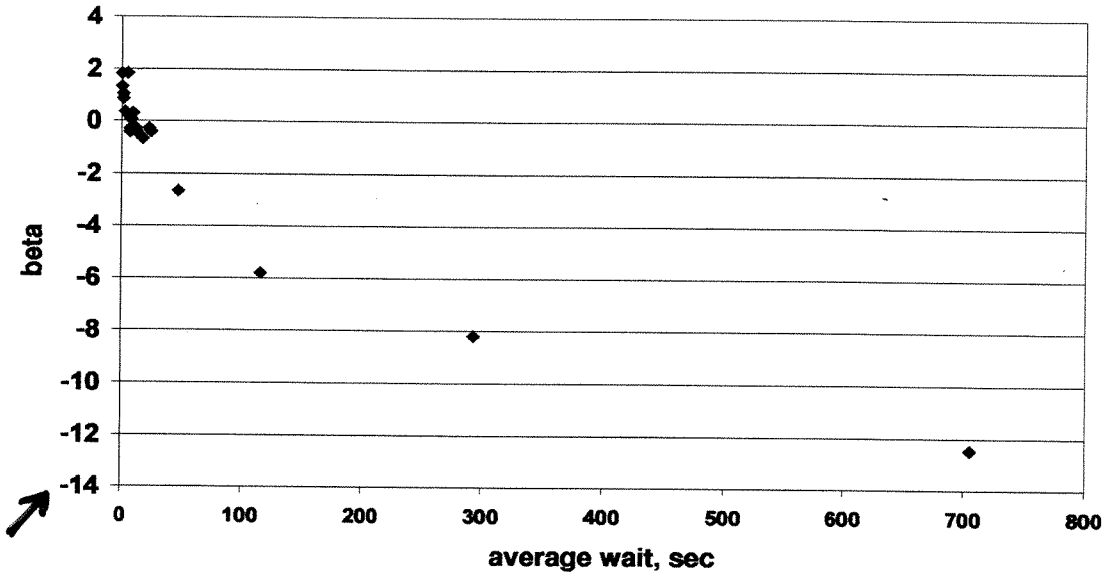


Figure 4

Italian data. Beta vs Average Wait

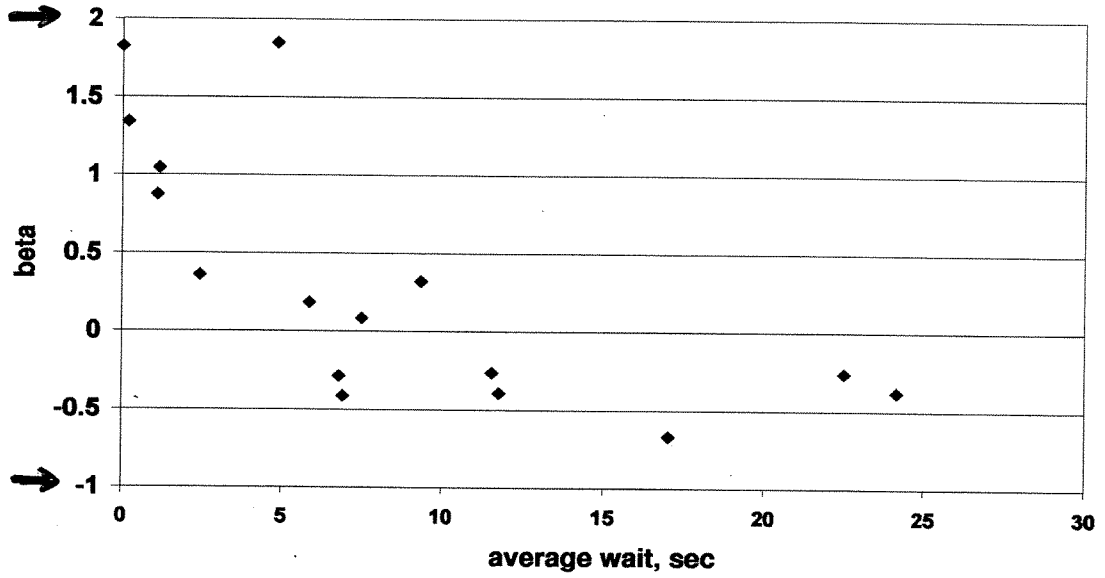


Figure 5

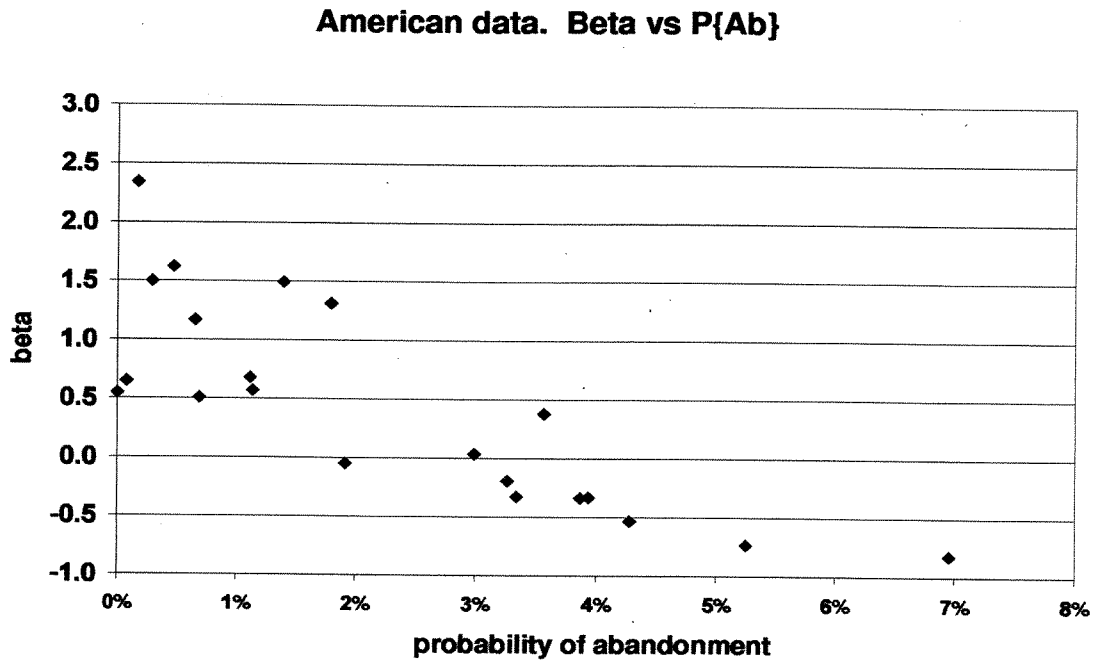
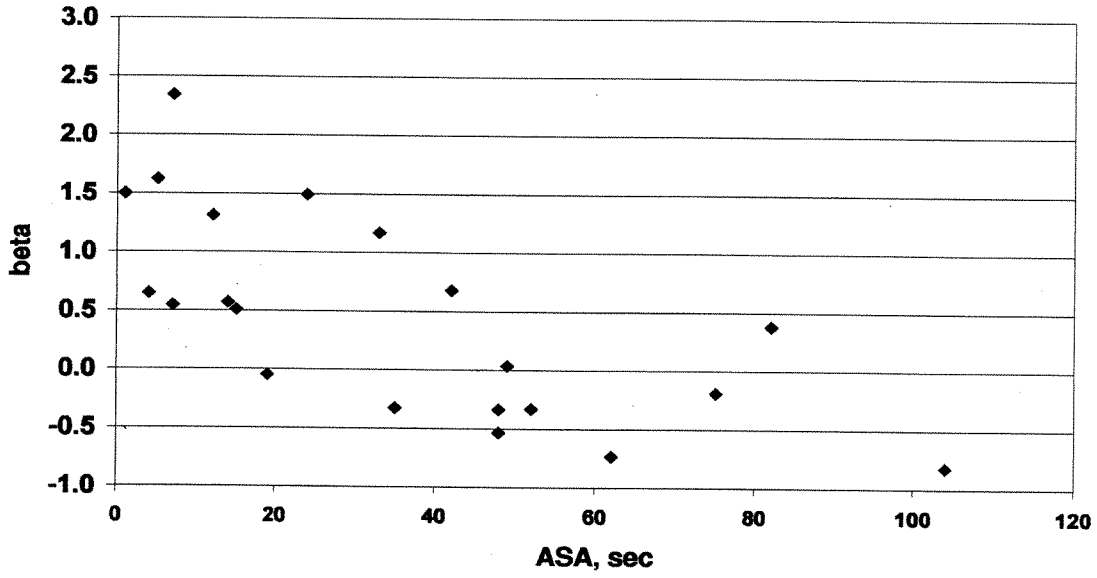


Figure 6



It is significant that one can choose here the same "service regimes" as in Question 2.6. (Note that there are no time intervals that correspond to "regime D"– catastrophic service level).

Rationalized staffing $\Rightarrow$ Abandonments

Abandonments Prevail (10–40%)

Abandonments Matter! Service Level, Economics

E.g. $M/M/N$: $\lambda = 48$, $\mu = 1$, $N = 50$ (Erlang-C)

vs. $M/M/N$ + exponential patience, mean = 2 min.

	$M/M/N$	$M/M/N + M$
Fraction abandoning	–	3%
E[Wait]	20.8 sec.	3.7 sec.
90% percentile	58 sec.	12.5 sec.
E[Queue]	17	3
Agents' utilization	96%	93%

What if $\lambda = 97\%$ of 50 in Erlang-C? E[Wait] = 8.8 sec.

What if $\lambda = 50$? Robustness of Erlang-A

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

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

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

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

$\Rightarrow$ **Serendipity** $\quad P\{\text{Abandon}\} \approx \dfrac{\gamma}{\sqrt{N}}\ , \quad 0 < \gamma < \infty$

Here α and γ are explicitly computable in terms of β, avg. service time, avg. patience.

Extremes:

$\alpha = 1 \Leftrightarrow \beta = -\infty \Leftrightarrow \gamma = \infty$ **Efficiency-driven**

$\alpha = 0 \Leftrightarrow \beta = \infty \Leftrightarrow \gamma = 0$ **Quality-driven**

Designing a Call Center

Approximate Performance Measures

$$N \approx \frac{\lambda}{\mu} + \beta\left(\frac{\lambda}{\mu}\right)^{1/2}$$

$$P\{Wait > 0\} \approx w(-\beta, \sqrt{\mu/\theta})$$

$$P\{Ab|Wait > 0\} \approx 1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}$$

$$P\{Ab\} \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right] \cdot w(-\beta, \sqrt{\mu/\theta})$$

$$E[Wait] \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right] \cdot \frac{w(-\beta, \sqrt{\mu/\theta})}{\theta}$$

$$P\{Wait > t\} \approx w(-\beta, \sqrt{\mu/\theta}) \cdot \frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t)} \cdot e^{-\theta t}$$

$$P\{Ab|Wait > t\} \approx 1 - \frac{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t)}{\Psi(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta}t)}$$

$E[Wait|Wait > t]$ computable

Here

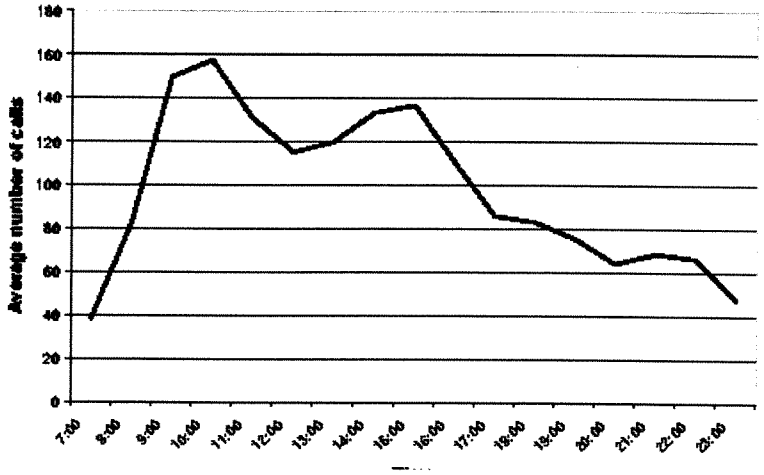
$$w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1},$$

$$h(x) = \frac{\phi(x)}{1 - \Phi(x)},$$

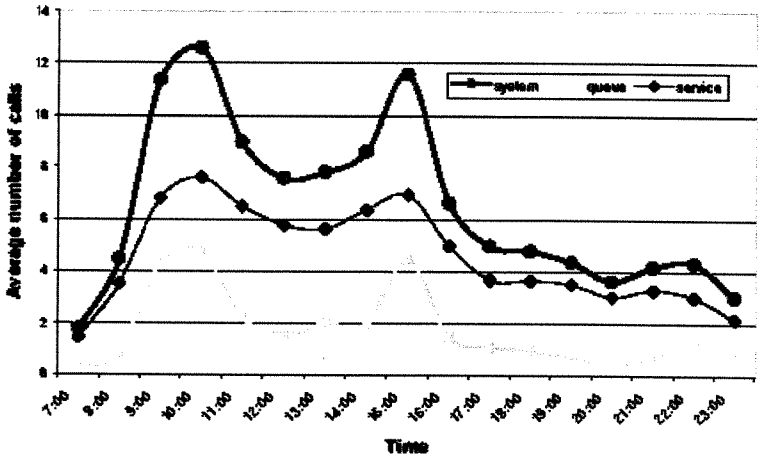
$$\Psi(x, y) = \frac{\phi(x)}{1 - \Phi(x + y)}$$

Time-Varying Queues: Predictable Variability

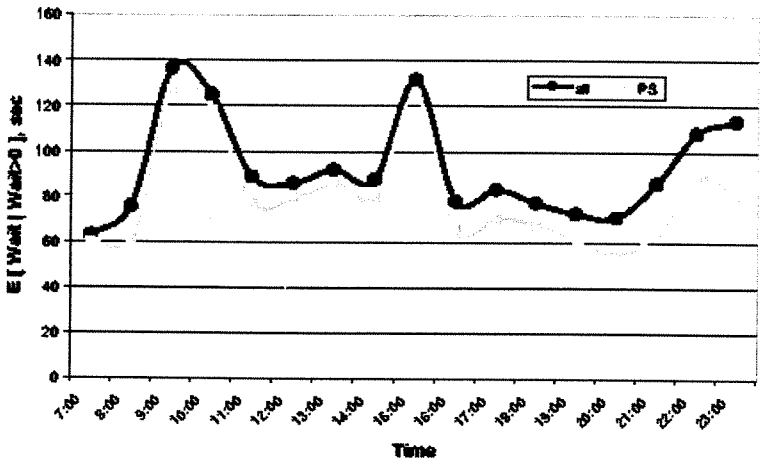
Arrivals



Queues



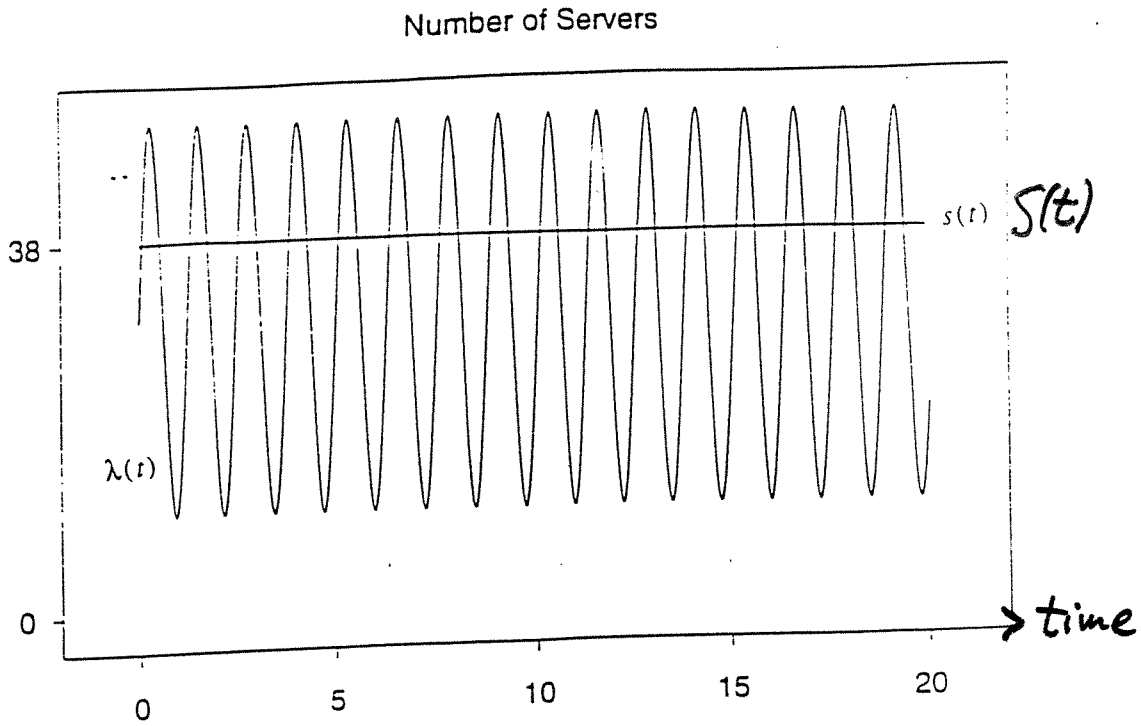
Waiting

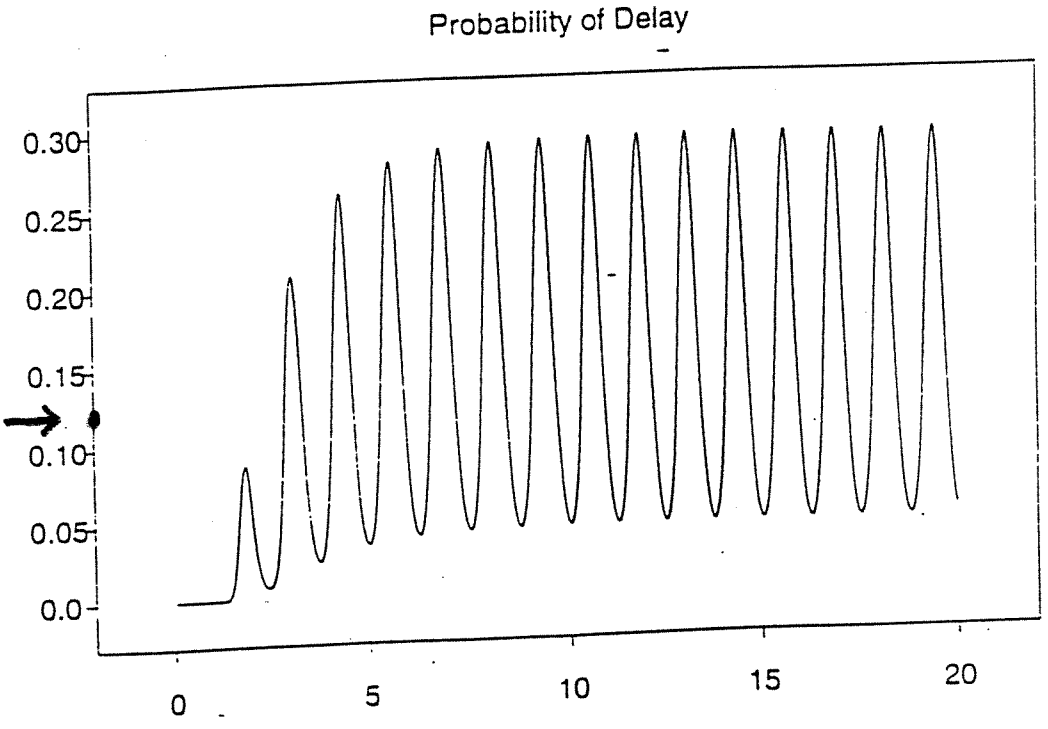


SIMPLE
STATIONARY
APPROX.

$\rho = \frac{\bar{\alpha}}{\mu S}$

$\mu = 1$





Delay Prob.

Target
0.13

Figure 2. The SSA Approximation: The server-staffing levels and delay probabilities as functions of time for the $M_t/M/s_t$ example with rapidly fluctuating sinusoidal arrival-rate function $\lambda(t) = 30 + 20\sin(5t)$ using the simple stationary approximation (SSA) with the average arrival rate 30 and a delay probability target of 0.13. The offered load $\lambda(t)$ is plotted with the constant number of servers.

POINTWISE STATIONARY APPROX.

$$\rho(t) = \frac{\alpha(t)}{\mu S(t)}$$

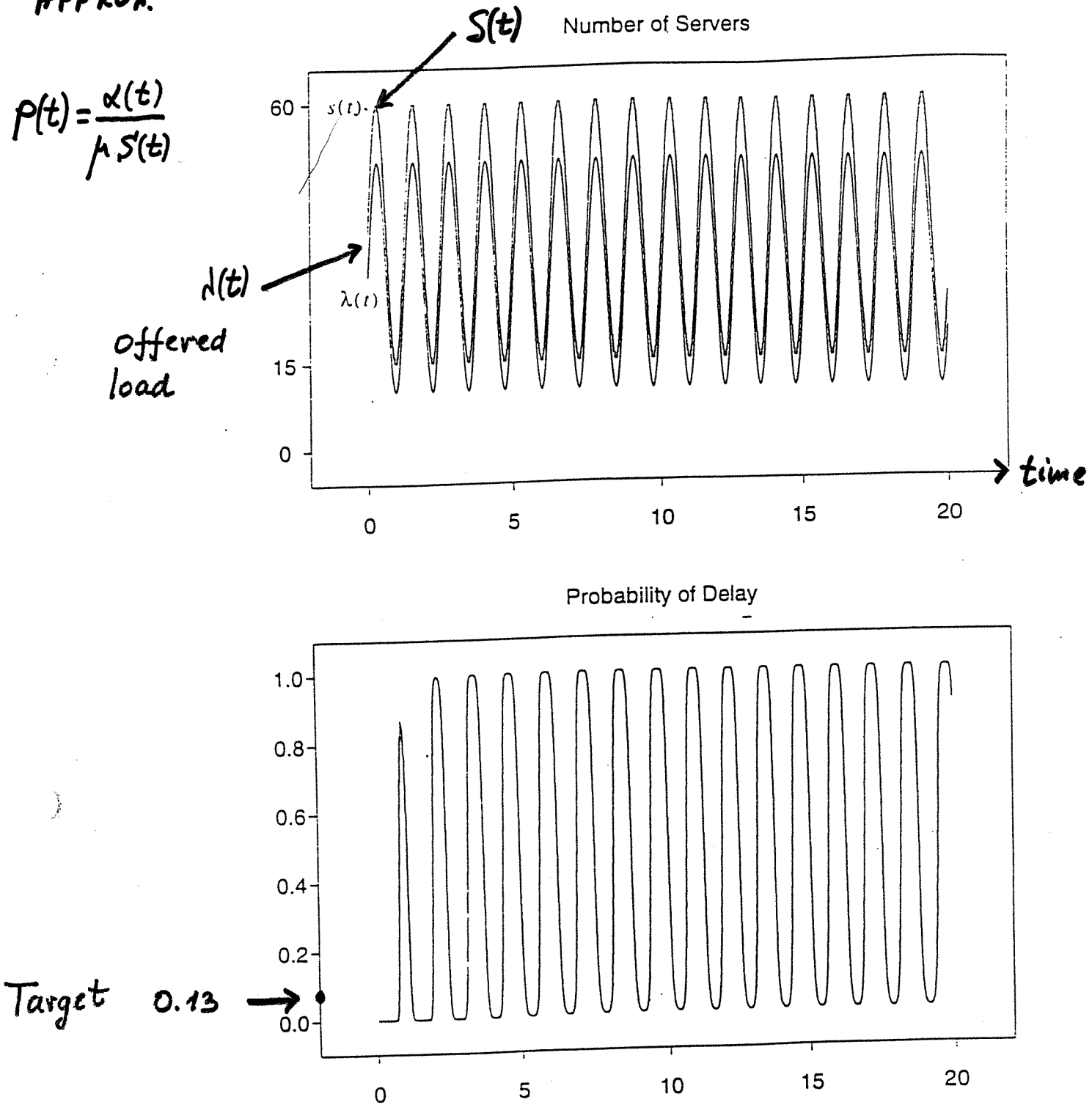


Figure 1. The PSA approximation: The server-staffing levels and delay probabilities as functions of time for the $M_t/M/s_t$ example with rapidly fluctuating sinusoidal arrival-rate function $\lambda(t) = 30 + 20 \sin(5t)$ using the pointwise stationary approximation (PSA) with a delay probability target 0.13. The offered load $\lambda(t)$ is plotted with the number of servers.

$M/G/?$, via ∞-server Heuristics (Folklore, Whitt)

Fact: ∞-server models easy to analyze and approximate.

Eg: Staffing $M/G/N$

Fact: In $M/G/\infty$, $\quad L \sim$ Poisson (R), $\quad R = \lambda \cdot E(S)$

$$P\{W_q(M/G/N) > 0\}$$

$$= P\{L(M/G/N) \geq N\} \qquad \text{PASTA}$$

$$\approx P\{L(M/G/\infty) \geq N\} \qquad \text{Heuristics}$$

$$\approx P\{R + Z\sqrt{R} \geq N\} \qquad \text{Poisson} \approx \text{Normal}$$

$$= 1 - \phi\left(\tfrac{N-R}{\sqrt{R}}\right) \qquad \alpha \text{ delay fraction}$$

$$\leq \alpha \quad \Rightarrow N = \lceil R + 0.5 + \bar{z}_\alpha \sqrt{R} \rceil \qquad (\alpha = P\{Z > \bar{z}_\alpha\})$$

Extension: in $M_t/G/N_t$, staffing of

$$N_t = \lceil R_t + 0.5 + \bar{z}_\alpha \sqrt{R_t} \rceil$$

would maintain delay fraction close to α (constant !).

What is R_t?

∞-server Heuristics (Time-varying)

Fact: In $M_t/G/\infty$, $\quad L_t \sim$ Poisson (R_t)

$R_t =$ offered load: $E\lambda(t - S_e) \cdot E(S) = E\int_{t-S}^{S} \lambda(u)du$

$S_e =$ excess service: $P\{S_e \leq t\} = \frac{1}{E(S)} \int_0^t P\{S > u\}du$

$$ES_e = E(S)\frac{1+C_s^2}{2}$$

Heuristics: Choose N_t such that

$$P\{L_t \geq N_t\} \leq \alpha, \quad P\{L_t \geq N_t - 1\} > \alpha$$

Normal Approximation: $\quad L_t \sim N(R_t, R_t)$

$$\Rightarrow N_t = \lceil R_t + 0.5 + \bar{z}_\alpha \sqrt{R_t} \rceil$$

which performs surprisingly well

(with Jennings, Massey, Whitt (1996)).

∞-server approx.

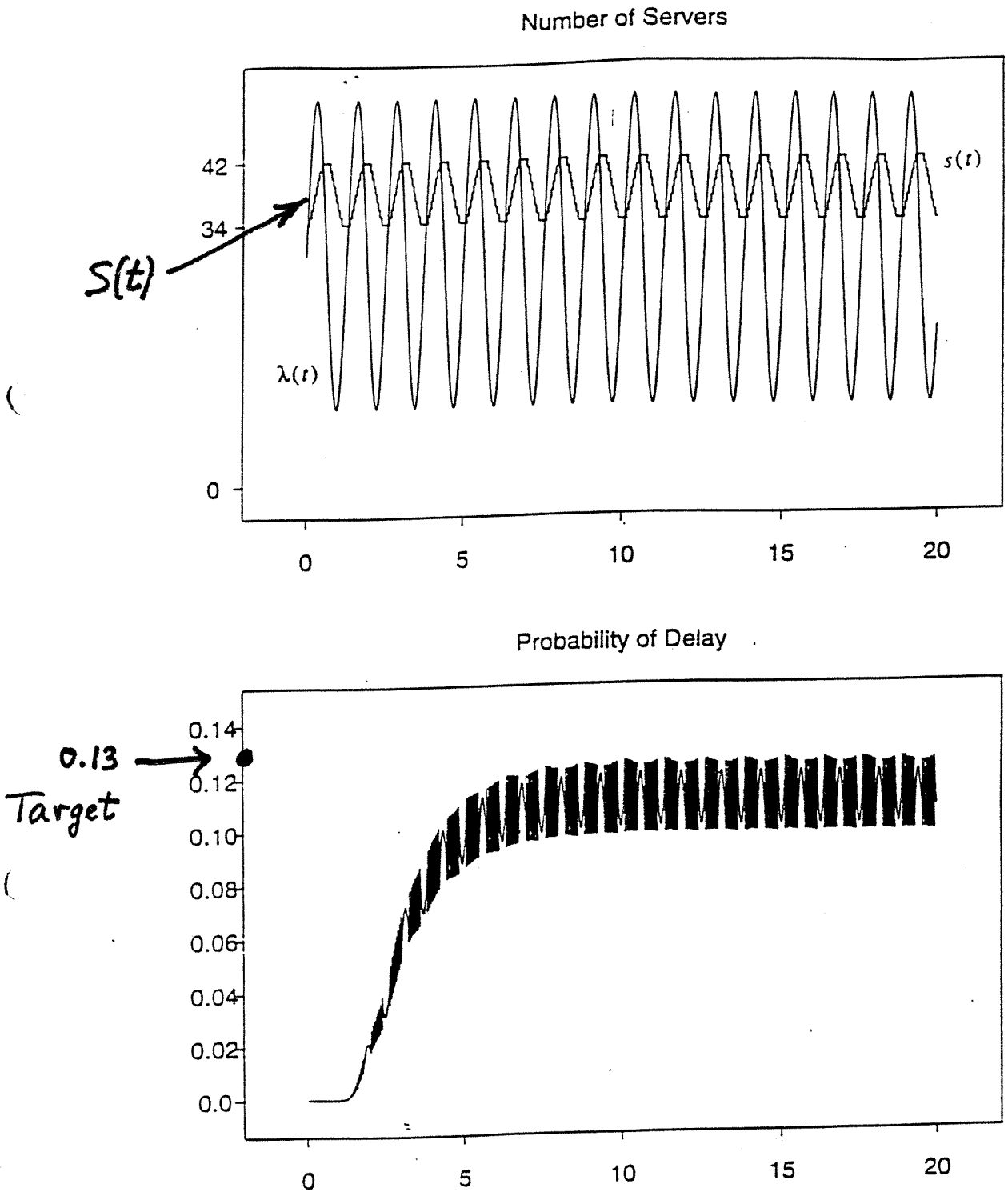


Figure 3. The server-staffing levels and delay probabilities as functions of time for the $M_t/M/s_t$ example with rapidly fluctuating sinusoidal arrival rate $\lambda(t) = 30 + 20 \sin(5t)$ based on $\beta_\varepsilon = 1.282$ ($\varepsilon = 0.1$ and $p_D(\varepsilon) = 0.13$).

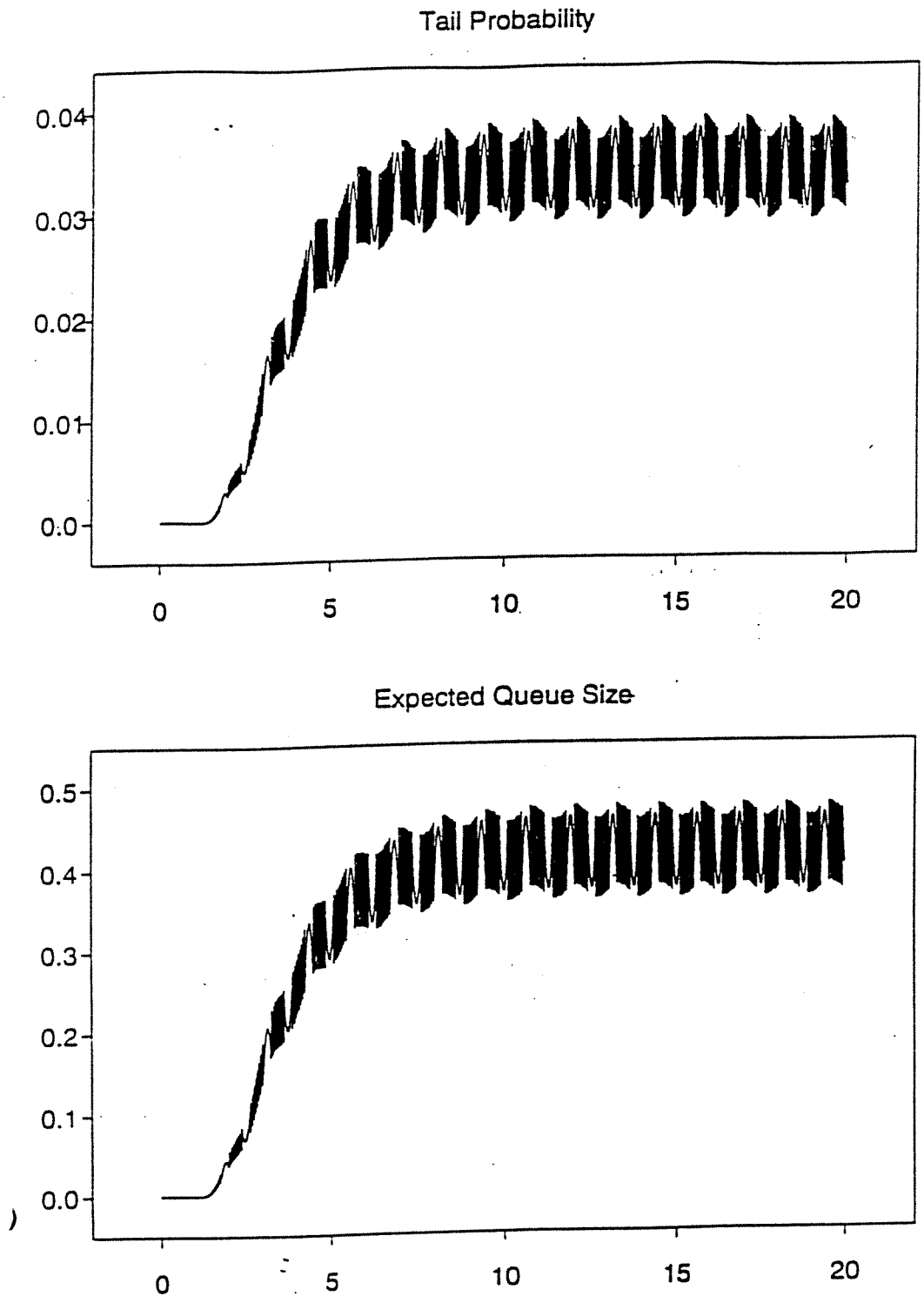
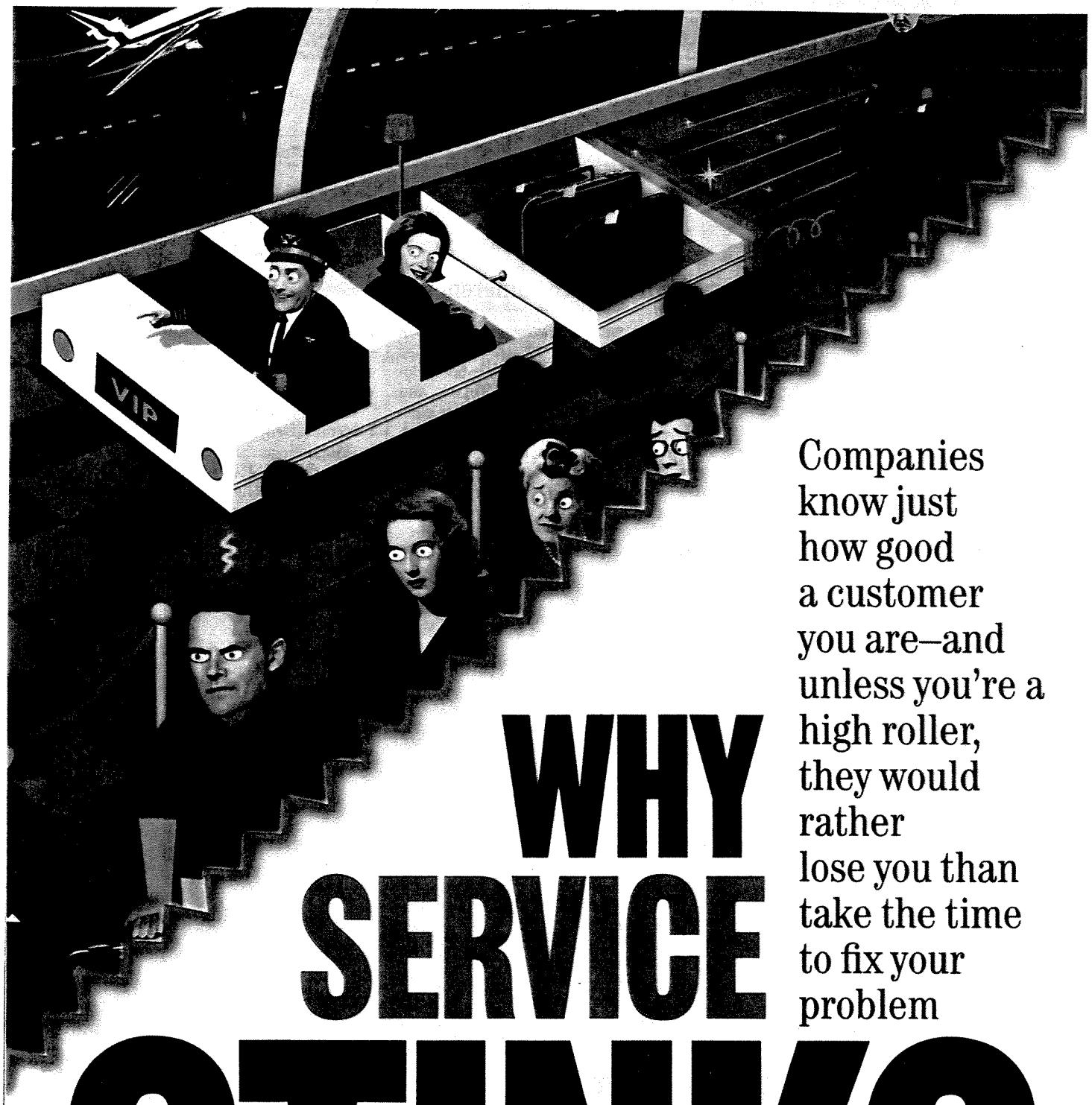


Figure 4. The tail probability $P(Q_s(t) \geq s(t) + 5)$ and expected number in queue (not in service) for the example in Figure 1.



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 take the time to fix your problem

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

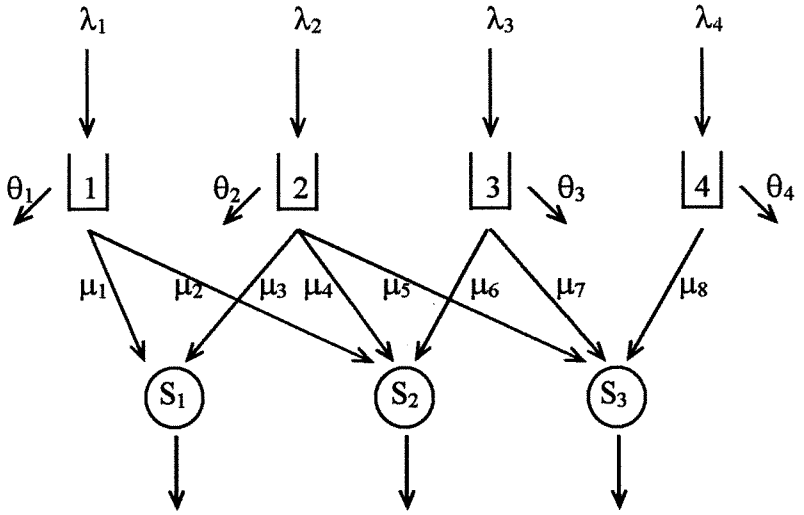
6%

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN

Date: 7:02 pm WED MAR 10, 1999

Introduction

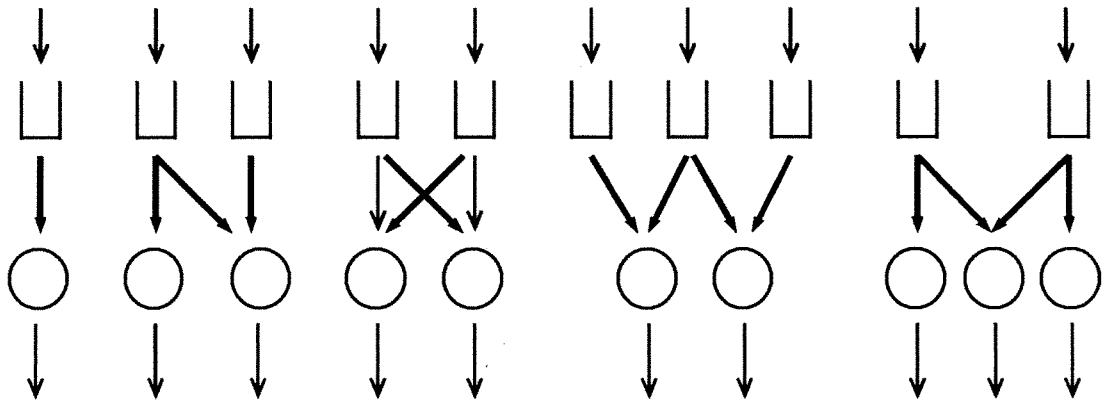
Consider the following multi-queue parallel-server system (animated, for example, by a telephone call-center):



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).



Staffing the "Modern" Basic Call Center

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

   - Conceptual: Halfin-Whitt
   - Dimensioning: Borst, Reiman

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

   - Conceptual: Garnet, Reiman
   - Dimensioning: (Borst, Reiman) in progress

3. Time-Varying (Non-homogenous 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/G/N open and challenging (measure-valued limit)
   - (very) Heavy tails: more than square-root safety?

Staffing: Additional Directions (Features)

- Skills-Based Routing: Current research
- Networks: eg. IVR+ACD, Hierarchical Help Desk
- Adaptive Customers: Shimkin
- Information while Waiting
- Adaptive Agents: Whitt's approach
- Forecasting: Accuracy important