Service Engineering and Management

Designing Call Centers with Impatient Customers

1. Massey, Reiman, Rider, Stolyar
2. Garnet, Reiman
3. Shimkin, Zohar
4. Carmon; Ritov, Sakov, Zeltyn
5. Borst, Reiman

Avi Mandelbaum
Industrial Engineering and Management
TECHNION, ISRAEL
avim@tx.technion.ac.il

Designing Call/Contact Centers with Impatient Customers: 10 Years History, or A Modelling Spectra

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

   as the standard Steady-state model
4. Massey, Reiman, Rider, Stolyar: Predictable variability ⇒

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

   Online Models and Simulation
9. Garnett: Practice ⇒ 4CallCenters.com

אגף שרות ותמי

שירות לקוחות (1013)

06/11/99 יום שבת

% answered

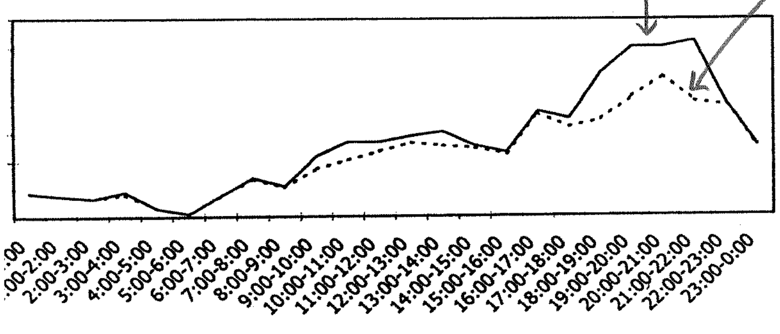
פניות והתפלגות על פני היום

שעה	פניות שנענו	פניות שננטשו	סה"כ פניות	אחוז פניות	אחוז מענה	זמן המתנה מקסימלי (שניות)	זמן המתנה ממוצע למענה (שניות)
0:00-1	17	0	17	1%	100%	2	0
1:00-2	15	0	15	1%	100%	2	1
2:00-3	13	0	13	1%	100%	2	1
3:00-4	16	2	18	1%	89%	74	6
4:00-5	6	0	6	0%	100%	2	0
5:00-6	2	0	2	0%	100%	2	1
6:00-7	15	0	15	1%	100%	2	0
7:00-8	27	1	28	2%	96%	66	9
8:00-9	21	1	22	2%	95%	2	0
9:00-10	35	9	44	4%	80%	292	46
10:00-1	41	13	54	4%	76%	260	75
11:00-1	47	7	54	4%	87%	340	35
12:00-1	53	5	58	5%	91%	222	40
13:00-1	51	10	61	5%	84%	396	73
14:00-1	49	2	51	4%	96%	192	17
15:00-1	45	1	46	4%	98%	120	10
16:00-1	73	2	75	6%	97%	118	18
17:00-1	64	6	70	6%	91%	132	23
18:00-1	69	33	102	8%	68%	532	130
19:00-2	85	35	120	10%	71%	370	131
20:00-2	100	20	120	10%	83%	214	58
21:00-2	82	42	124	10%	66%	340	106
22:00-2	79	2	81	7%	98%	94	9
23:00-0	49	1	50	4%	98%	42	5
	1054	192	1246	100%	90%	532	33

התפלגות הפניות ע"פ היום
(סה"כ פניות מול פניות שנענו)

Called

answered



08:00-22:..

..גה ממוצע ..ניות)	72
..תנה מקס'	532
..יחות נענות	81%

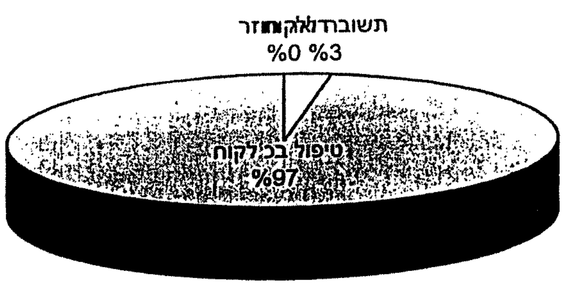
% answered

סה"כ נציגים	22
זמן עבודה נטו (דקות)	7260

התפלגות חזרה ללקוח

שפות זרות	0
דואר חוזר	0
תשובה ללקוח	15
טיפול בכ.לקוח	515
סה"כ	530

התפלגות חזרה ללקוח
(באחוזים)



..דה במיוחדים מתוך סה"כ זמן עבודה כולל	13%
האיפיונים ביחס למספר השיחות שנענו	113%

3

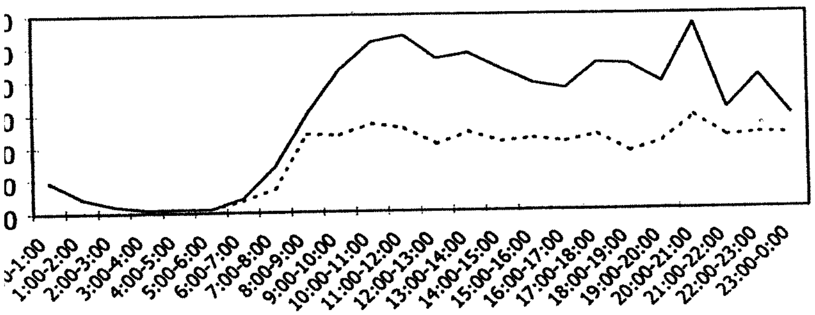
אגף שירות ותמיכה

שירות לקוחות (1013) יום ה' 25/11/99

הפניות והתפלגות על פני היום

שעה	פניות שנענו	פניות שננטשו	סה"כ פניות	אחוז פניות	אחוז מענה	זמן המתנה מקס' (שניות)	זמן המתנה ממוצע למענה (שניות)
0:00	47	0	47	1%	100%	44	2
1:00	21	0	21	1%	100%	2	0
2:00	10	0	10	0%	100%	148	15
3:00	4	0	4	0%	100%	2	1
4:00	5	0	5	0%	100%	2	0
5:00	5	0	5	0%	100%	86	18
6:00	17	4	21	1%	81%	198	50
7:00	36	34	70	2%	51%	528	268
8:00	120	31	151	4%	79%	482	100
9:00	118	98	216	6%	55%	682	255
10:00	135	123	258	7%	52%	478	283
11:00	128	140	268	8%	48%	494	307
12:00	102	130	232	6%	44%	738	350
13:00	121	119	240	7%	50%	634	292
14:00	106	110	216	6%	49%	464	250
15:00	111	83	194	5%	57%	416	188
16:00	105	81	186	5%	56%	600	253
17:00	116	108	224	6%	52%	514	226
18:00	89	133	222	6%	40%	636	461
19:00	105	89	194	5%	54%	548	254
20:00	144	140	284	8%	51%	402	197
21:00	112	42	154	4%	73%	318	109
22:00	116	88	204	6%	57%	638	191
23:00	115	29	144	4%	80%	334	46
	1988	1582	3570	100%	68%	738	172

התפלגות הפניות ע"פ היום
(סה"כ פניות מול פניות שנענו)



08:00-

ממוצע (...)	217
מקס'	738
נענות	53%

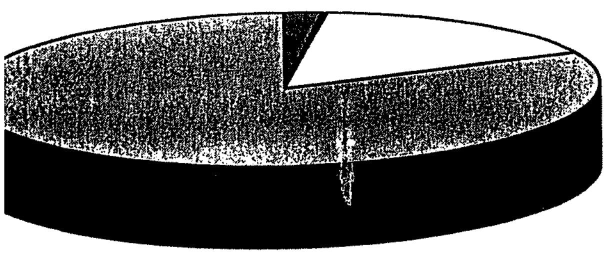
סה"כ נציגים	55
זמן עבודה נטו (דקות)	18150

התפלגות חזרה ללקוח

שפות זרות	50
דואר חוזר	10
תשובה ללקוח	439
טיפול בכ.לקוח	2200
סה"כ	2699

כמות ה... מתוך סה"כ זמן עבודה	19%
טיפולים ביחס למספר השיחות שנענו	124%

התפלגות חזרה ללקוח
(באחוזים)



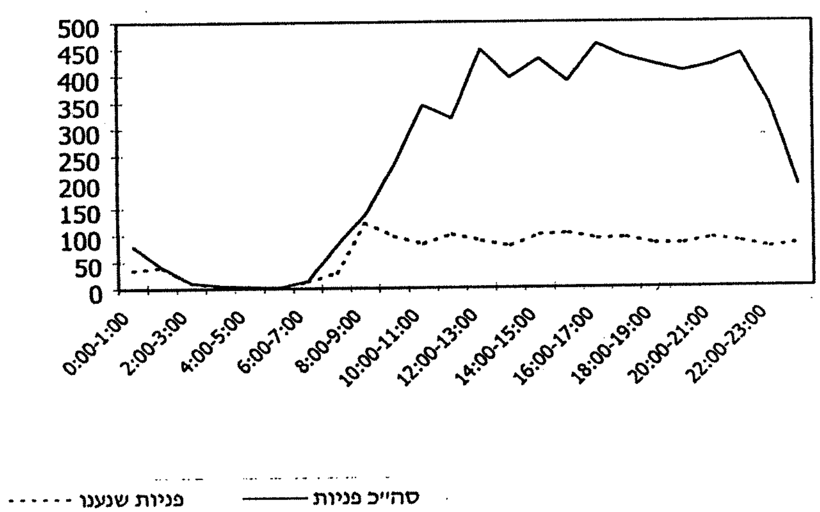
אגף שרות ותמיכה

שרות לקוחות (1013) יום א' 21/11/99

הפניות והתפלגות על פני היום

שעה	פניות שנענו	פניות שננטשו	סה"כ פניות	אחוז פניות	אחוז מענה	זמן המתנה מקסימלי (שניות)	זמן המתנה ממוצע למענה (שניות)
0:00	34	46	80	1%	43%	656	230
1:00	40	0	40	1%	100%	2	0
2:00	10	1	11	0%	91%	2	1
3:00	3	2	5	0%	60%	110	37
4:00	3	0	3	0%	100%	0	0
5:00	1	0	1	0%	100%	0	0
6:00	12	2	14	0%	86%	104	27
7:00	29	50	79	1%	37%	564	312
8:00	121	16	137	2%	88%	338	66
9:00	96	135	231	4%	42%	702	377
10:00	82	263	345	6%	24%	866	657
11:00	99	220	319	5%	31%	736	528
12:00	87	361	448	7%	19%	1022	681
13:00	78	317	395	7%	20%	1214	896
14:00	98	332	430	7%	23%	944	718
15:00	101	288	389	6%	26%	864	624
16:00	90	368	458	8%	20%	994	761
17:00	92	342	434	7%	21%	812	596
18:00	81	338	419	7%	19%	1100	851
19:00	81	326	407	7%	20%	860	682
20:00	91	327	418	7%	22%	834	642
21:00	84	355	439	7%	19%	922	651
22:00	73	273	346	6%	21%	882	694
23:00	80	110	190	3%	42%	678	412
	1566	4472	6038	100%	45%	1214	435

התפלגות הפניות ע"פ היום
(סה"כ פניות מול פניות שנענו)



08:00-22

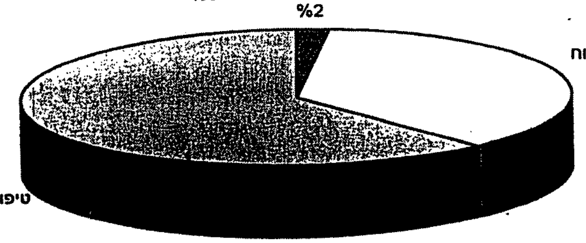
ממוצע	376
מקס'	1214
נענות	24%

סה"כ נציגים	47
זמן עבודה נטו (דקות)	15510

התפלגות חזרה ללקוח

שפות זרות	20
דואר חוזר	0
תשובה ללקוח	395
טיפול בכ.לקוח	662
סה"כ	1077

התפלגות חזרה ללקוח
(באחוזים)



5

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

~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

~35%

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

~5%

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN   Date: 7:02 pm WED MAR 10, 1999

6/13/00 - Tue

Charlotte - 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%

Rationalized staffing $\Rightarrow$ Abandonments

Abandonments Prevail (10–40%)

Abandonments Matter! Service Level

Economics

E.g. $M/M/N$: $\lambda = 48, \quad \mu = 1, \quad N = 50$

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

	$M/M/N$	$M/M/N+M$
Fraction abandoning	–	3.1%
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 $\lambda = 50$? Robustness

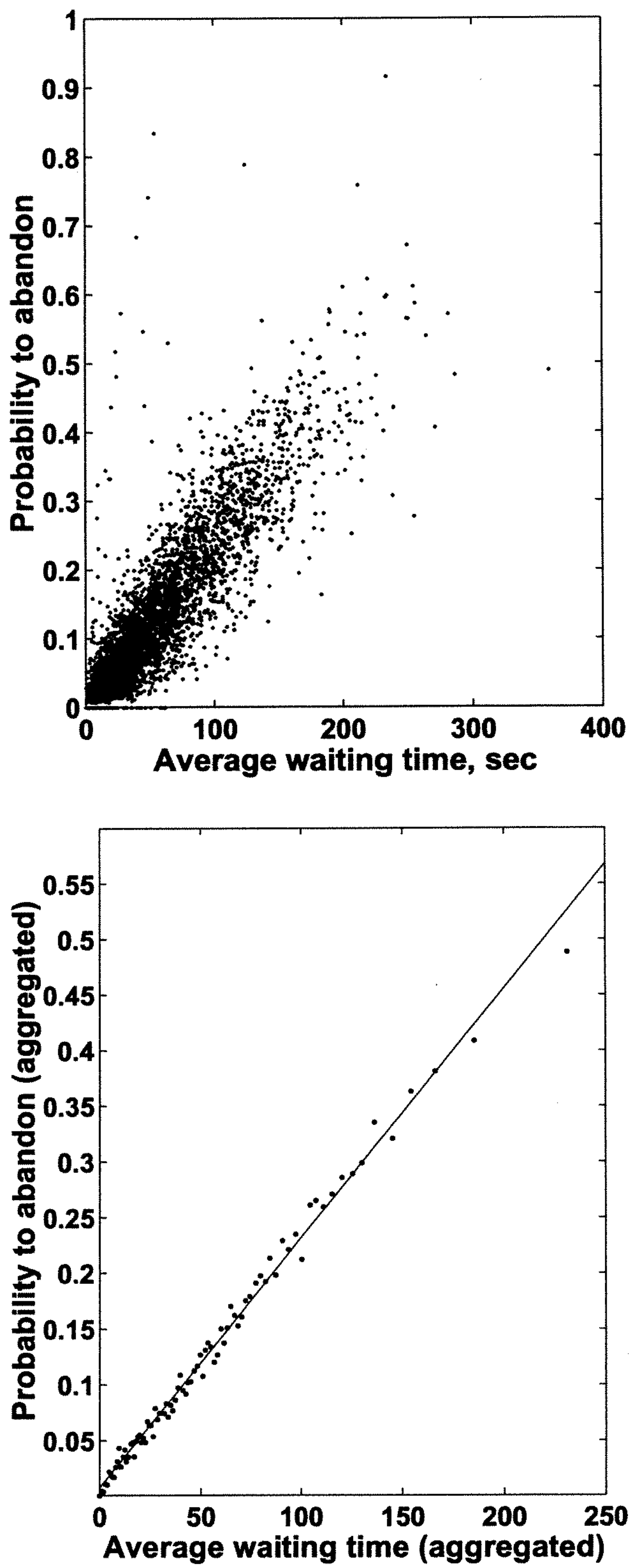
Abandonments Important

- 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)

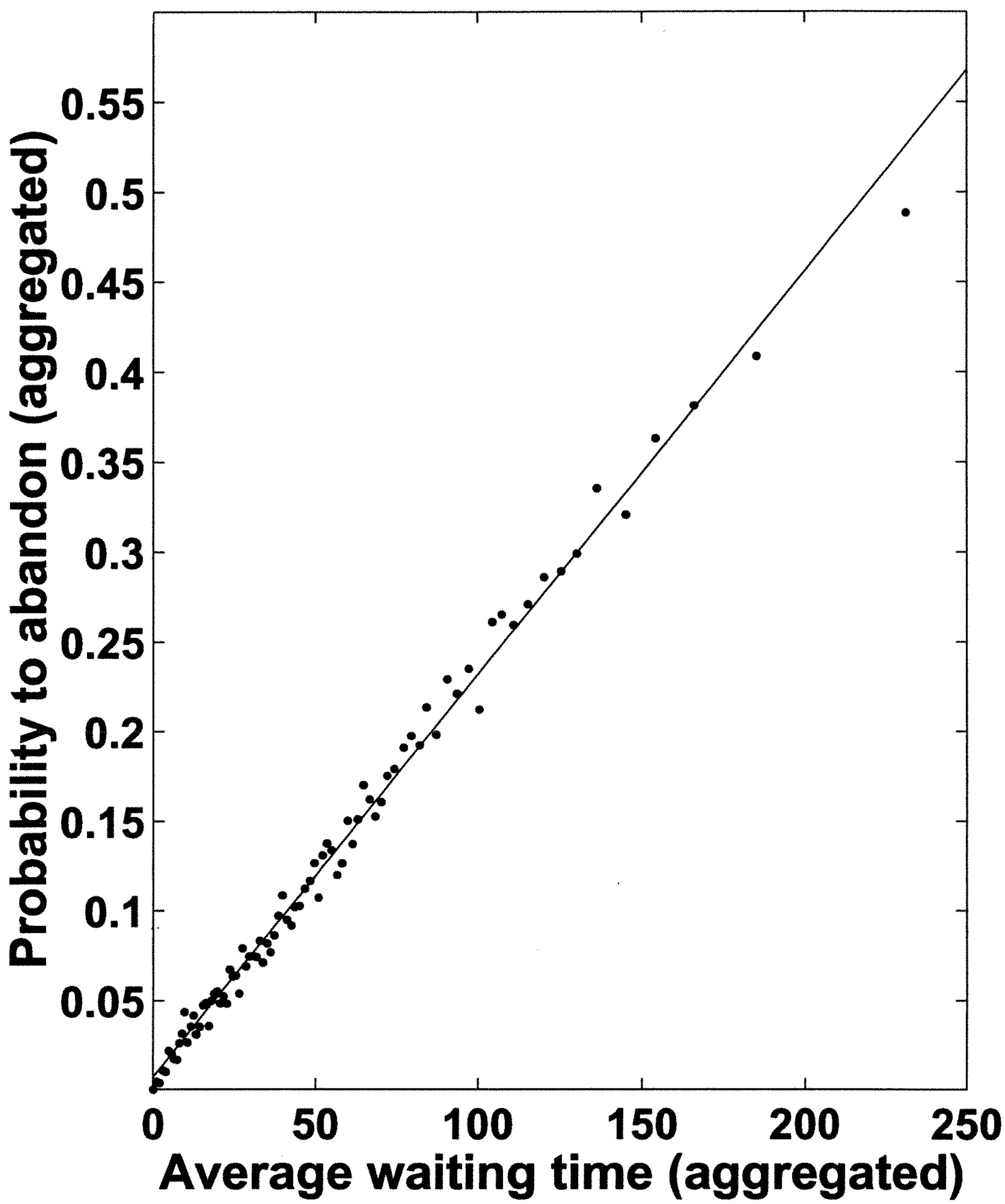
But Abandonment also Interesting

- 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

Fraction Abandoning vs. Average Waiting Time



Fraction Abandoning vs. Average Waiting Time

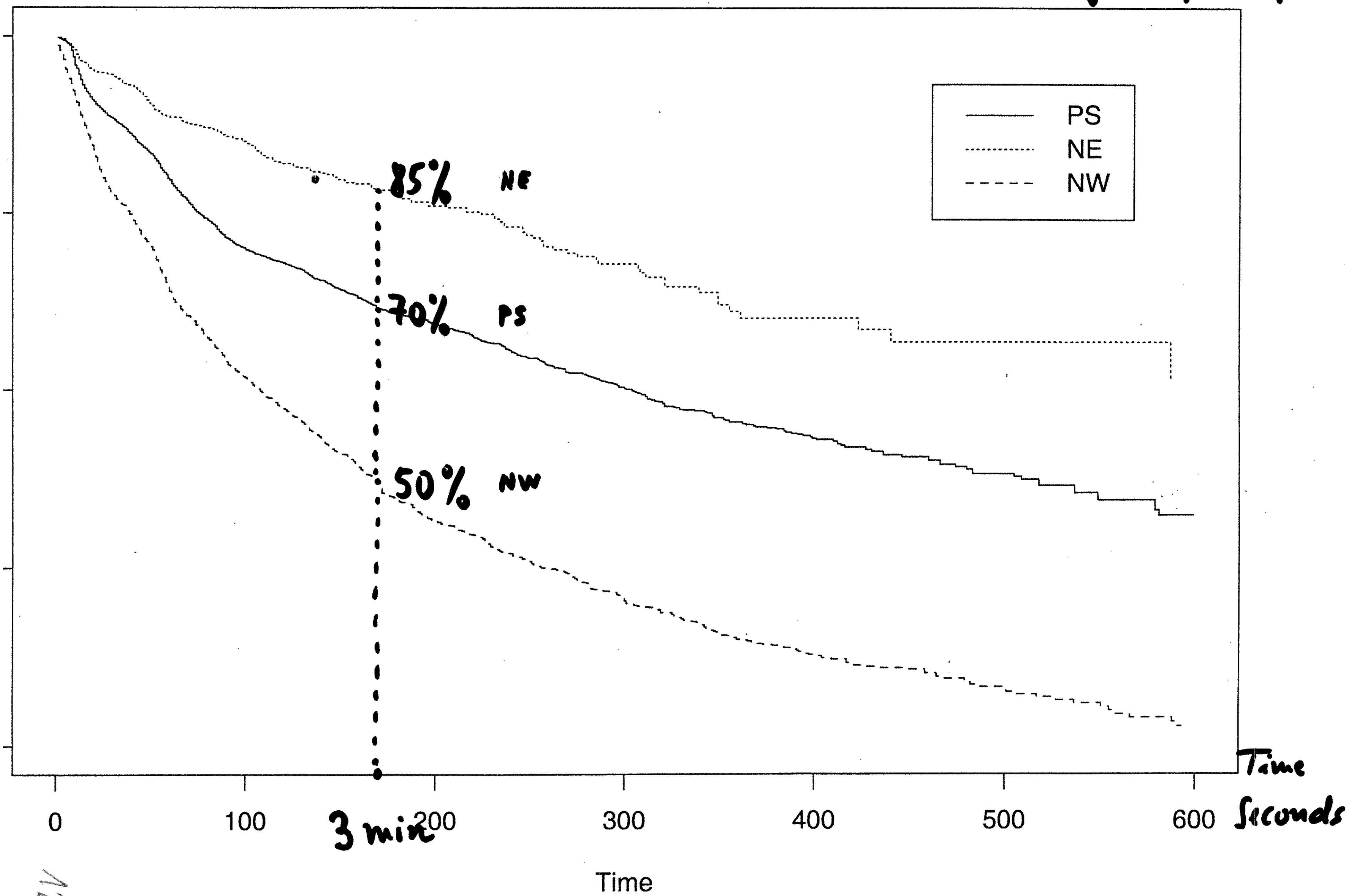


$$P_{ab} = \theta \overline{W_q}$$

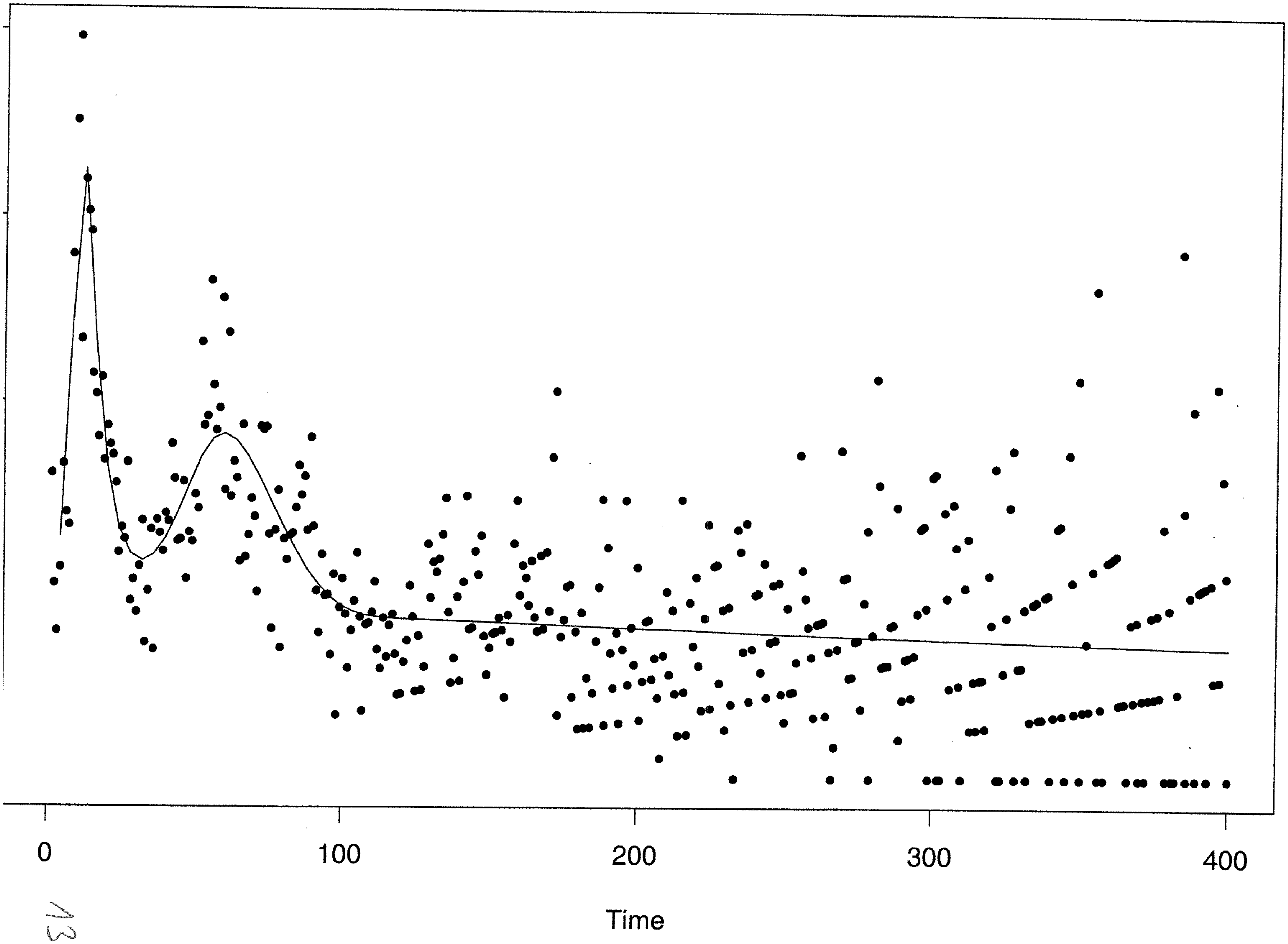
$$\frac{1}{\theta} = \frac{\overline{W_q}}{P_{ab}} = \frac{TTT}{\# \text{abandon}} \Rightarrow \text{censoring}$$

Aug - Patience Survival curve : How long willing to wait before hanging up the phone?

ival = still waiting

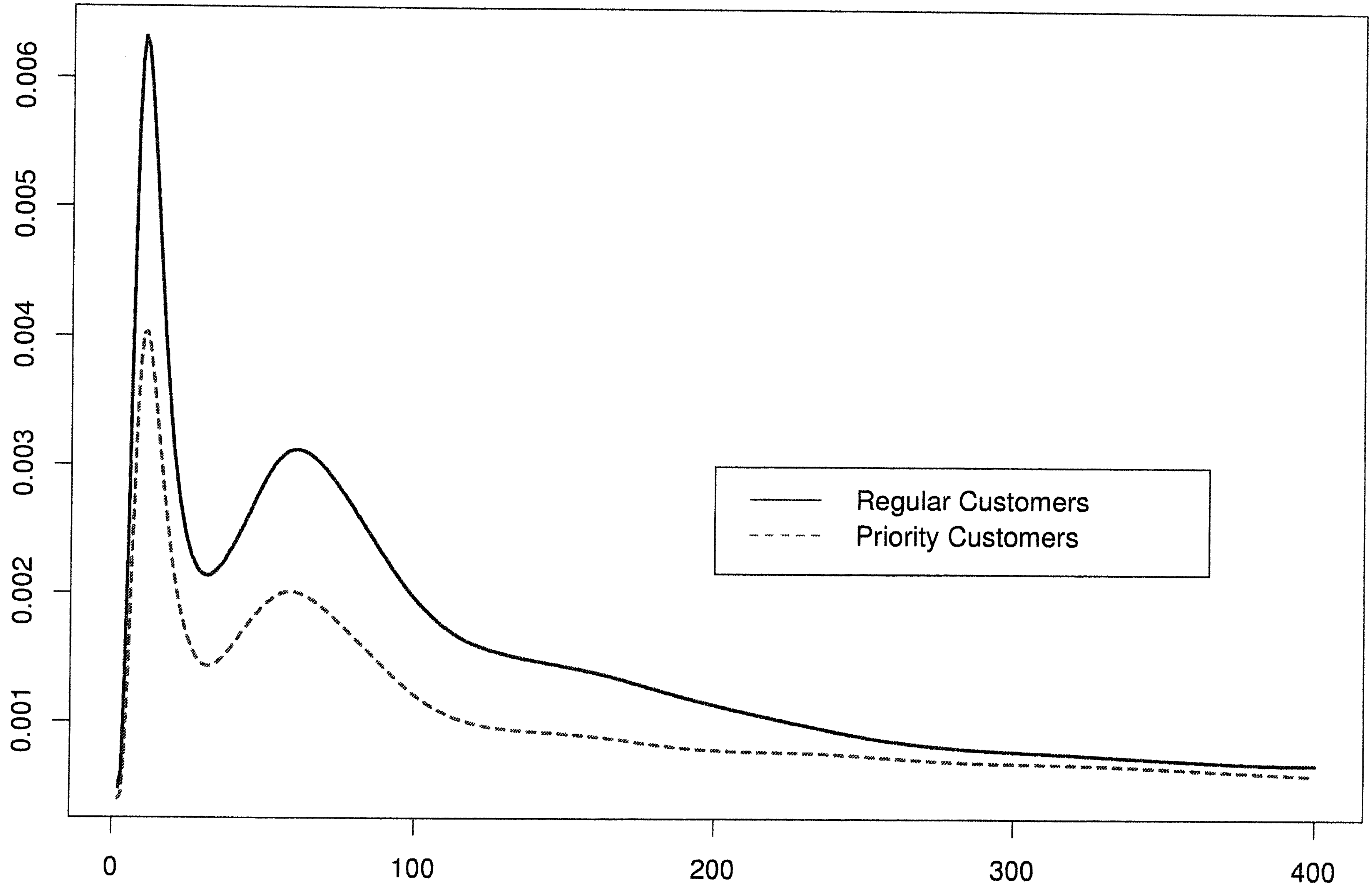


Aug - Patience - Hazard Rates



Hazard Rate: Empirical (Im)Patience

Palm's 4th
Law of Irritation



PATIENCE INDEX

- How to define? measure? manage?

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 **censored** by "wait".

"Uncensoring" (simplified)

$$\text{Willing to wait} \quad 1 + 2 \times \frac{360K}{90K} = 1 + 2 \times 4 = \mathbf{9} \text{ min.}$$

$$\text{Expect to wait} \quad 2 + 1 \times \frac{90K}{360K} = 2 + 1 \times \frac{1}{4} = \mathbf{2.25} \text{ min.}$$

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

- Strongly supported by ongoing research (Wharton).

The 6th Law of Congestion

Customer-Focused Queueing Theory

– 200 abandonments in Direct-Banking

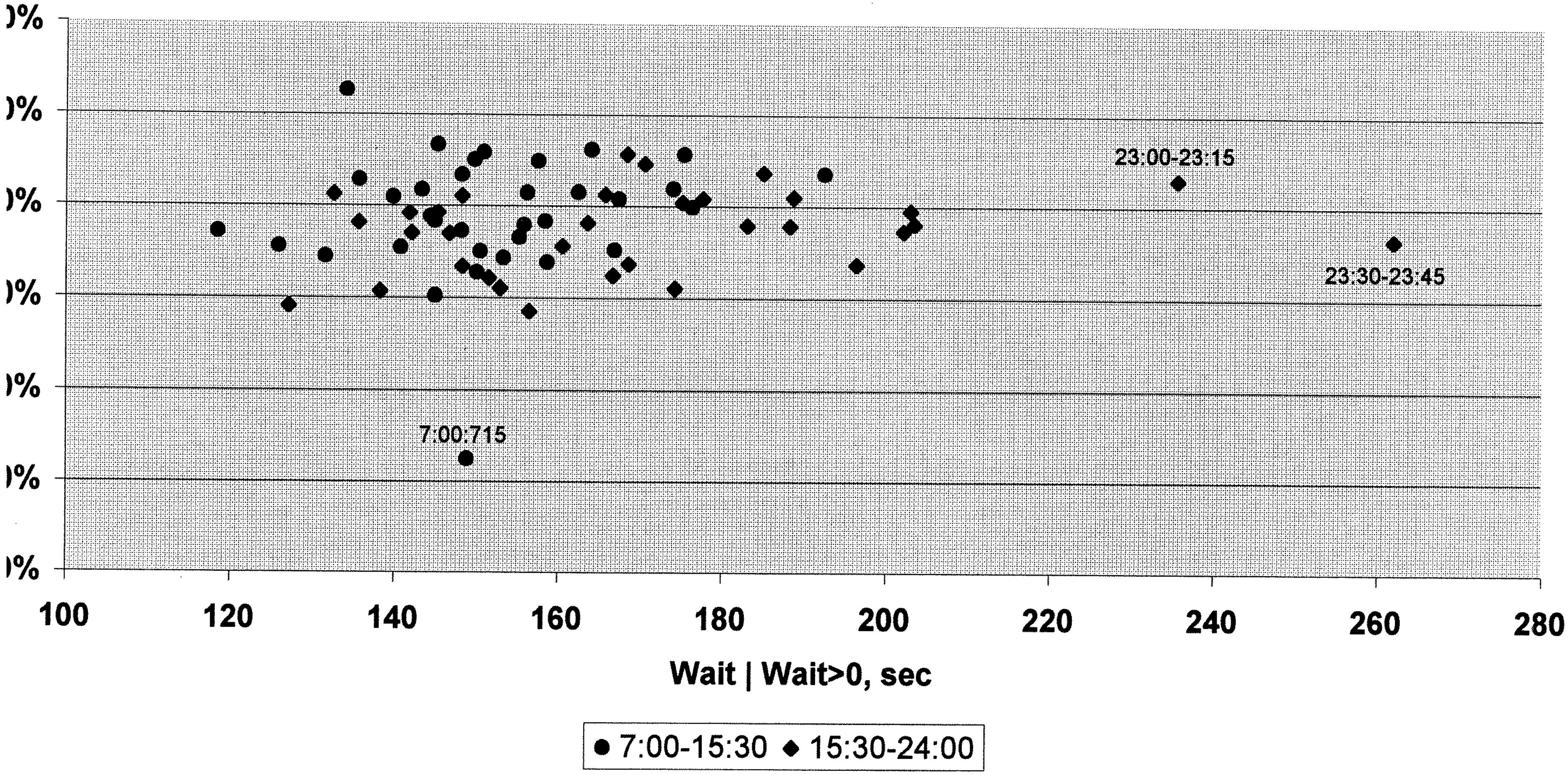
– unscientific

Reason to Abandon	Actual Abandon Time (sec)	Perceived Abandon Time (sec)	Perception Ratio
Fed up waiting (77%)	70	164	2.34
Not urgent (10%)	81	128	1.6
Forced to (4%)	31	35	1.1
Something came up (6%)	56	53	0.95
Expected call-back (3%)	13	25	1.9

The Expectation → Experience → Perception Cycle

Adaptivity ?

IN customers, 1999.
Percent {abandonments | Wait>0} vs. {Wait | Wait>0}.



17

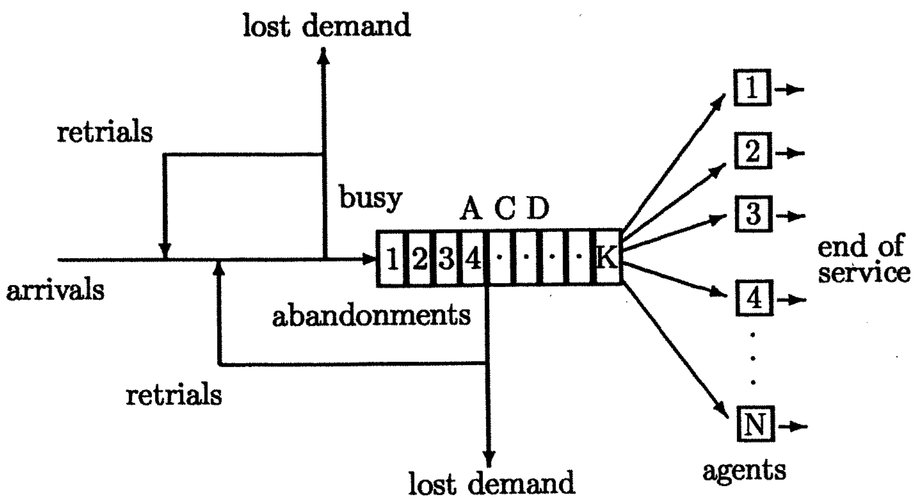
Designing Call Centers with Impatient Customers

Palm's Model: Erlang-C with Impatience

Joint Research with
Ofer Garnet, Marty Reiman

Avi Mandelbaum
Industrial Engineering and Management
TECHNION, ISRAEL

Schematic Model of a Telephone Call Center



Abandonments (no retrials):

- Palm (1937, 1953)
- Barrer (1957)
- Riordan (1962)
- Baccelli and Hebuterne (1981)

$M/M/N/B+\underline{M}$ System (Garnet, Reiman)

- Poisson arrivals-rate λ
- Service times – $\exp(\mu)$
- N statistically identical agents attending to single queue
- Service policy FCFS
- System capacity: B $\quad (B < \infty \Rightarrow$ Busy Signal)
- Customers' <u>patience</u> – $\exp(\theta)$

$\{Q(t), t \geq 0\}$ - number of customers in the system:

<u>Birth</u> & <u>Death</u>

$\{Q(t), t \geq 0\}$ - transition diagram

$$
\begin{tikzcd}
\boxed{0} \arrow[r, shift left, "\lambda"] & \boxed{1} \arrow[l, shift left, "\mu"] \cdots & \boxed{N-1} \arrow[r, shift left, "\lambda"] & \boxed{N} \arrow[l, shift left, "N\mu"] \arrow[r, shift left, "\lambda"] & \boxed{N+1} \arrow[l, shift left, "N\mu+\theta"] \cdots & \boxed{B-1} \arrow[r, shift left, "\lambda"] & \boxed{B} \arrow[l, shift left, "N\mu+(B-N)\theta"]
\end{tikzcd}
$$

"Everything" calculable via stationary distribution $(B \leq \infty)$

$$
\pi_k = \begin{cases} \dfrac{(\lambda/\mu)^k}{k!}\pi_0 & 0 \leq k \leq N \\ \displaystyle\prod_{j=N+1}^{k} \left(\frac{\lambda}{N\mu + (j-N)\theta}\right) \frac{(\lambda/\mu)^N}{N!}\pi_0 & N < k \leq B \end{cases}
$$

where

$$
\pi_0 = \left[\Sigma_{k=0}^{N} \frac{(\lambda/\mu)^k}{k!} + \Sigma_{k=N+1}^{B} \Pi_{j=N+1}^{k} \left(\frac{\lambda}{N\mu+(j-N)\theta}\right) \frac{(\lambda/\mu)^N}{N!}\right]^{-1}
$$

$M/M/N/B+M$ - Characteristics

Notation:

- $P\{Ab\}$ - abandonment probability (fraction)
- $P\{Wait > 0\}$ - waiting probability (fraction)
- $P\{Block\}$ - blocking probability in an $M/M/N/N$ system

1. $P\{Ab\}$ increases monotonically in θ, λ
   $P\{Ab\}$ decreases monotonically in N, μ
   (Bhattacharya and Ephremides (1991))

2. $P\{Ab\} \leq P\{Block\}$
   (Boxma and de Waal (1994))

3. $P\{Ab\} = \theta \cdot E[Wait]$

   ($\lambda P\{Ab\} = \theta \cdot E[No.\ in\ queue]$ & Little)

4. Stationary distribution always exists (in particular for $B = \infty$)
   (Sandwiched between infinite-server models)

Calculation of Exact Measures

- V - virtual waiting time = waiting time of a customer with infinite patience (test customer).
- X - customer's patience ($X \sim \exp(\theta)$, independent of V).
- $Wait \equiv V \wedge X$ - actual waiting time.

Performance measures of the form $E[f(V,X)]$:

$f(v,x)$	$E[f(V,X)]$
$1_{\{v>x\}}$	$P\{V > X\} = P\{Ab\}$
$1_{(t,\infty)}(v \wedge x)$	$P\{Wait > t\}$
$1_{(t,\infty)}(v \wedge x)1_{\{v>x\}}$	$P\{Wait > t; Ab\}$
$(v \wedge x)1_{\{v>x\}}$	$E[Wait; Ab]$
$(v \wedge x)1_{(t,\infty)}(v \wedge x)1_{\{v>x\}}$	$E[Wait; Wait > t; Ab]$
$g(v \wedge x)$	$E[g(Wait)]$

From these obtain more "natural" measures, for example

$$P\{Ab|Wait > t\} = \frac{P\{Wait > t; Ab\}}{P\{Wait > t\}}$$

22

Exact Calculation of Performance Measures - continued

$$E[f(V,X)] = E[f(V,X)\cdot 1_{\{V>0\}}] + E[f(0,X)]\cdot (\sum_{k=0}^{N-1} \pi_k + \pi_B)$$

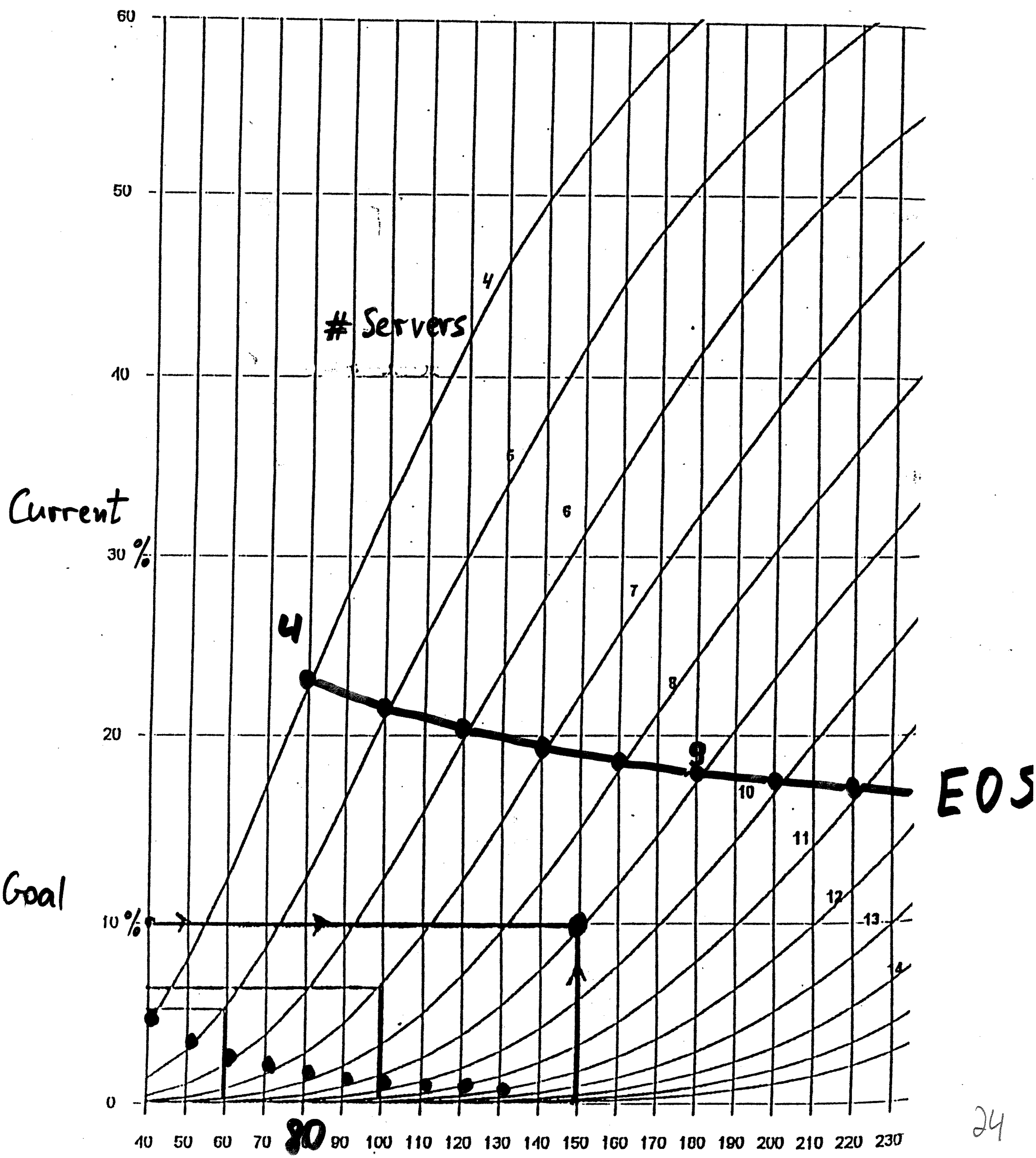
$$E[f(V,X)\cdot 1_{\{V>0\}}] = \sum_{n=0}^{B-N-1} \pi_{N+n} \int_0^\infty \int_0^\infty f(t,x)\theta e^{-x\theta} dx dF_{V_n}(t)$$

(V_n - virtual waiting time given $N+n$ in system on arrival
$\stackrel{d}{=}$ $\exp(N\mu) + \exp(N\mu+\theta) + \ldots + \exp(N\mu+n\theta)$)

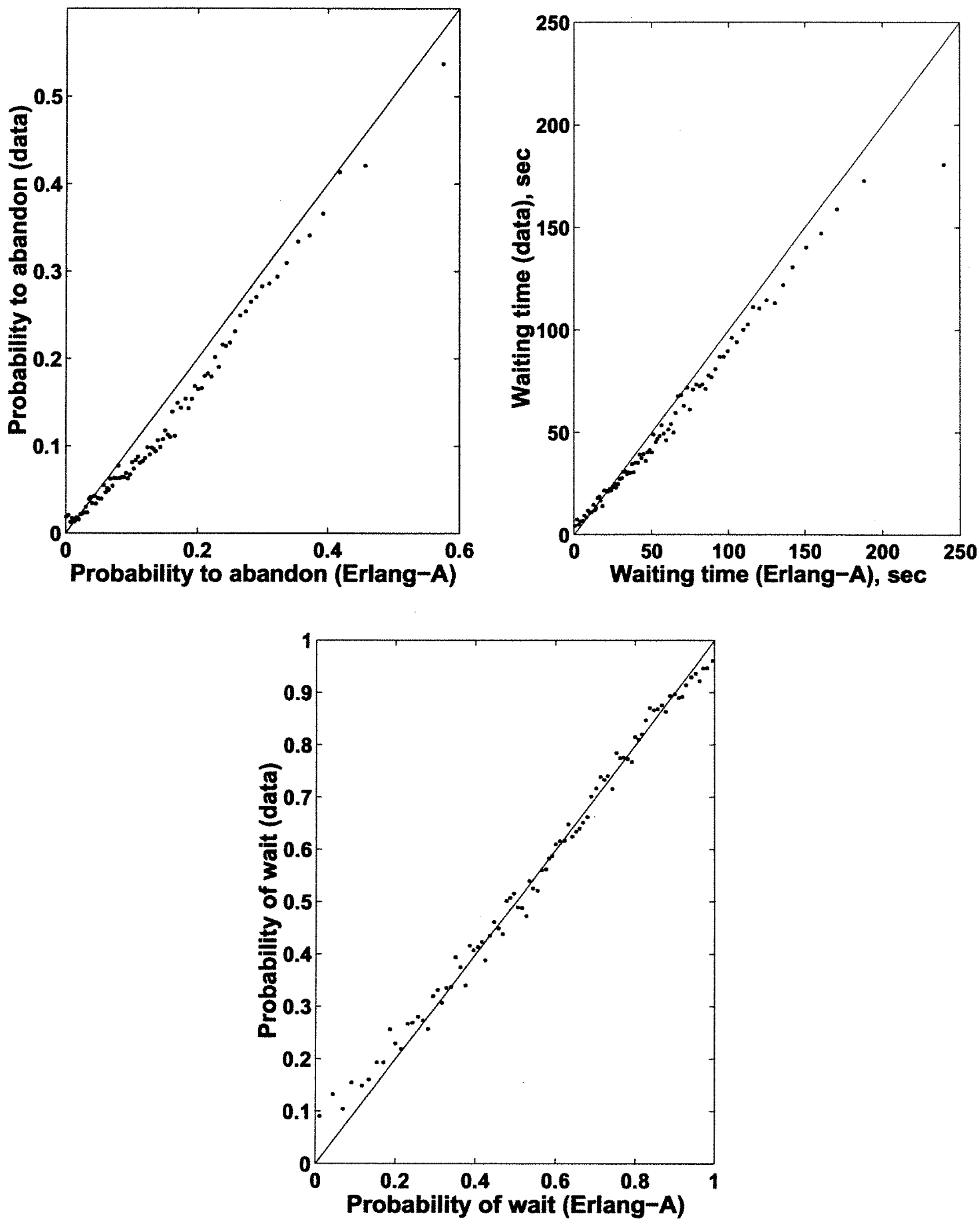
Abandonment Rate (% Arrivals = Enter)

@ given load (efficiency) : 20 cust's / server / hr

service level (abandons) improves with size



Erlang-A Formulae vs. Data Averages



4CallCenters.com

Motivation: Insight, Numerical stability

Focus: Large Call Centers — λ, N large; $B = \infty$

Framework: Sequence of $M/M/N+M$ systems, indexed by N

- $Q_N = \{Q_N(t),\ t \geq 0\}$
- $V_N = \{V_N(t),\ t \geq 0\}$, virtual waiting time of a test customer (infinite-patience)
- Parameters λ_N, μ, θ_N

$\theta_N \uparrow \infty$	impatient
$\theta_N \downarrow 0$	patient
$\theta_N \equiv \theta$	rational

Offered load $R_N = \dfrac{\lambda_N}{\mu}$ $(\rho_N = R_N/N)$

Approximations of Process and Stationary Distribution

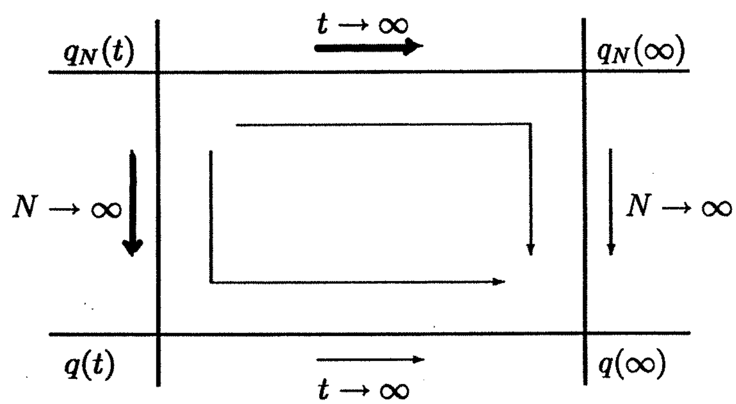
$$q_N(t) = \frac{1}{\sqrt{N}} [Q_N(t) - N], \quad 0 \leq t \leq \infty$$

$$v_N(t) = \sqrt{N}\, V_N(t)$$

Approximating the Process $\{Q_N(t), t \geq 0\}$

Notation:

- $q_N = \{q_N(t), t \geq 0\}$ - stochastic process obtained by the rescaling:
$$q_N = \frac{Q_N - N}{\sqrt{N}}$$
- $q_N(\infty)$ - stationary distribution of q_N
- $q = \{q(t), t \geq 0\}$ - process defined by: $q_N \xrightarrow{d} q$



27

Towards Rules of Thumb ($\theta_N \equiv \theta$)

Following Halfin-Whitt and Flemming-Simon-Stolyar.

Theorem: *Let* $N \sim R + \beta\sqrt{R}$

$$\alpha \equiv \lim_{N\to\infty} P_N\{Wait > 0\}$$

$$\Updownarrow$$

$$\beta \equiv \lim_{N\to\infty} \sqrt{N}\,(1-\rho_N) \qquad \Rightarrow \rho \sim 1 - \frac{\beta}{\sqrt{N}} \text{ !}$$

$$\Delta \equiv \lim_{N\to\infty} \sqrt{N} P_N\{Ab\} \qquad \Rightarrow P\{Ab\} \sim \frac{\Delta}{\sqrt{N}} \text{ !}$$

Then

$$1 > \alpha > 0 \text{ iff } \infty > \beta > -\infty$$
$$\text{iff } \infty > \Delta > 0$$

in which case

$$\alpha = w(-\beta, \sqrt{\mu/\theta})$$
$$\Delta = [\sqrt{\theta/\mu} \cdot h(\beta\sqrt{\mu/\theta}) - \beta] \cdot \alpha$$

Here

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

$$h(x) = \frac{\phi(x)}{1-\Phi(x)}, \quad \text{hazard rate of std. normal}$$

Designing a Call Center with Impatient Customers

Rules of Thumb - continued

$$(\boldsymbol{\theta_N \equiv \theta})$$

Extremes:

$\boldsymbol{\alpha = 1} \Leftrightarrow \beta = -\infty \Leftrightarrow \Delta = \infty$ **Efficiency-driven**

$\boldsymbol{\alpha = 0} \Leftrightarrow \beta = \infty \Leftrightarrow \Delta = 0$ **Quality-driven**.

Ultimate service-quality and efficiency via scale ($N \uparrow \infty$):

Waiting	$\alpha \leq \epsilon$
Utilization	$\rho = \frac{R}{N} \geq 1 - \epsilon$
Abandonments	$\Delta \leq \epsilon$.

New "Rules of Thumb" $(M/M/N+N)$

	N	ρ	$P\{\text{Wait} > 0\}$	$P\{Ab\}$
Efficiency driven	$R - \epsilon R$	$1+\epsilon$	1	ϵ
Quality driven	$R + \epsilon R$	$1-\epsilon$	0	$\frac{\Delta}{\sqrt{N}}$
Rationalized	$R + \beta\sqrt{R}$ $-\infty < \beta < \infty$	$1 - \frac{\beta}{\sqrt{N}}$	$\alpha(\beta)$ $0 < \alpha < 1$	$\frac{\Delta(\beta)}{\sqrt{N}}$

Compare with Previous ($M/M/N$, Halfin-Whitt)

E	$R+\epsilon$	$1 - \frac{\epsilon}{N}$	1	–
Q	$R+\epsilon R$	$\frac{1}{1+\epsilon}$	0	–
R	$R + \beta\sqrt{R}$ $\beta > 0$	$1 - \frac{\beta}{\sqrt{N}}$	$\alpha(\beta)$ $0 < \alpha < 1$	–

Designing a "Large" Call Center – Rationalized Service Level

Presently $$N \approx R + \beta\sqrt{R} \qquad \left(R = \frac{\lambda}{\mu}\right)$$

β = service grade

with

$$P\{\text{Wait} > 0\} \sim \alpha(\beta) \qquad \left(= w\left(-\beta, \sqrt{\mu/\theta}\right)\right)$$

$$P\{Ab\} \sim \frac{\Delta(\beta)}{\sqrt{N}} .$$

Forecast $\lambda \uparrow \hat{\lambda}$ $\quad \left(R \uparrow \hat{R} = \frac{\hat{\lambda}}{\mu}\right)$

To maintain present service level, recommend staffing

$$\hat{N} \approx \hat{R} + \beta\sqrt{\hat{R}} ,$$

and expect performance as follows:

$P\{\text{Wait} > 0\}$ unchanged

$P\{Ab\}$ reduced by $\sqrt{\frac{N}{\hat{N}}} \sim \sqrt{\frac{\lambda}{\hat{\lambda}}}$ EOS

β?

Designing a Call Center - Service-Level

$P\{Wait > T_1\ ;\ \neg Ab\}$ - "loyal" customers with long waits.

$P\{Wait \le T_2\ ;\ Ab\}$ (T_2 small) - "unworthy" customers, not willing to wait even a short period.

Proposed service-level measure: $(0 \le a \le 1)$

$$
\begin{aligned}
\neg SL &= aP\{Wait > T_1\ ;\ \neg Ab\} + (1-a)P\{Wait > T_2; Ab\} \\
&\approx w(-\beta, \sqrt{\mu/\theta}) \cdot h(\beta\sqrt{\mu/\theta}) \\
&\quad \cdot \Big[a\Psi(\mu\sqrt{\theta\mu N}T_1, \beta\sqrt{\mu/\theta} + \sqrt{N\theta}/\mu) \\
&\quad +(1-a)\cdot e^{-\theta T_2}\Psi(\mu\sqrt{\theta/\lambda}T_2, \beta\sqrt{\mu/\theta}) \\
&\quad -(1-a)\cdot \Psi(\mu\sqrt{\theta\mu N}T_2, \beta\sqrt{\mu/\theta} + \sqrt{N\theta/\mu}))\Big] \ .
\end{aligned}
$$

Here

$$
\begin{aligned}
w(x,y) &= \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}, \\
h(x) &= \frac{\phi(x)}{1-\Phi(x)}, \\
\Psi(x,y) &= \frac{\phi(x)}{1-\Phi(x+y)}
\end{aligned}
$$

Approximations - Application

Estimating the parameters :

- Service rate (μ) : from ACD data

- Arrival rate (λ) : from ACD data

- Abandonment rate (θ) :

$$\widehat{\frac{1}{\theta}} = \frac{\binom{No.}{served} \times \binom{Average\ wait}{of\ served} + \binom{No.}{abandoning} \times \binom{Average\ time}{to\ abandon}}{No.\ abandoning}$$

or

$$P\{Ab\} = \theta \cdot E[Wait] \quad \text{(linear regression)}$$

- Parameter β :

$$\beta = \frac{N\mu - \lambda}{\sqrt{N\mu}}$$

- Parameter γ : ???

Designing a Call Center

Approximate Performance Measures

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

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

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

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

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

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

$E[Wait|Wait > t]$ computable

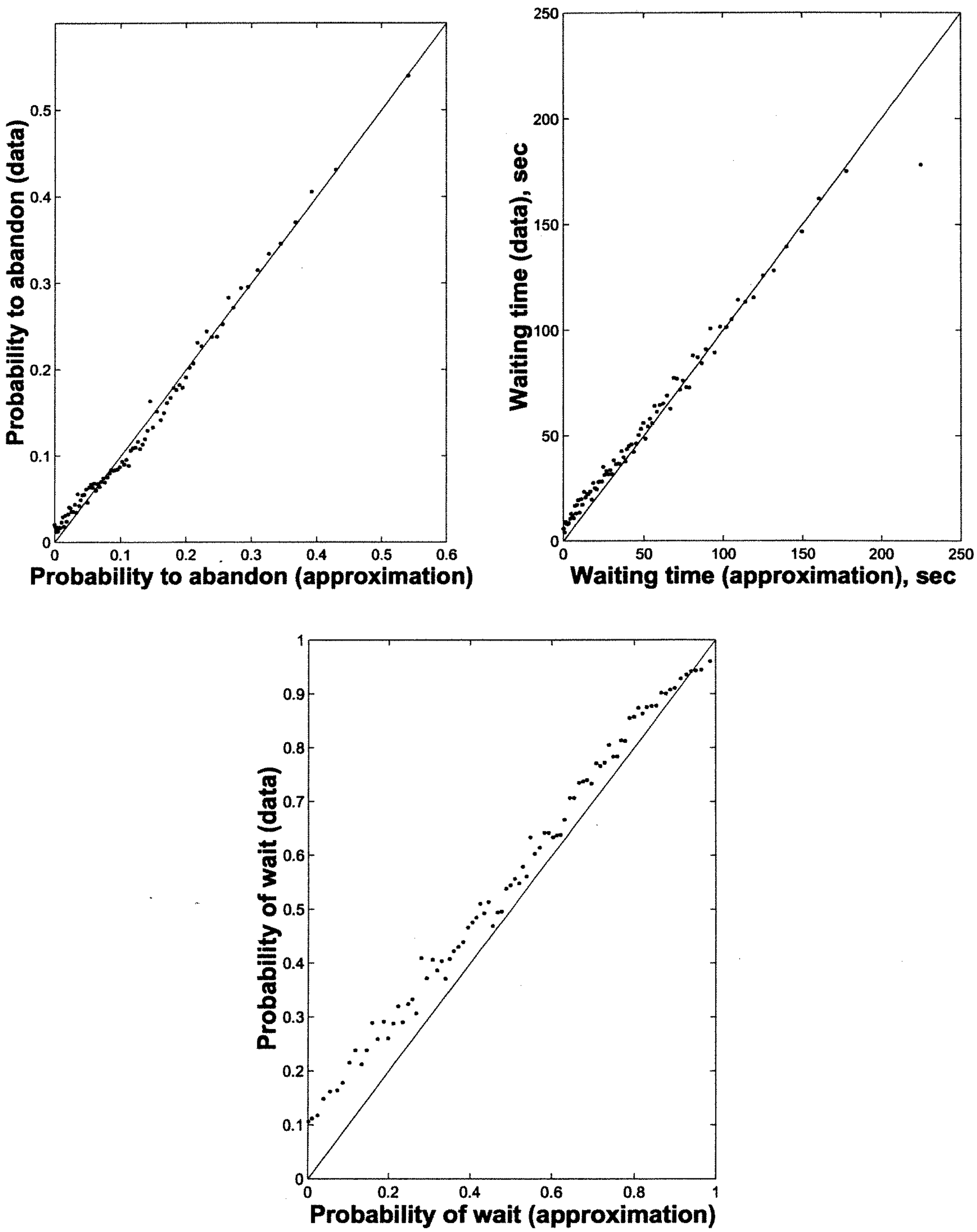
Here

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

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

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

Erlang-A Aproximations vs. Data Averages



Patience Characteristics

$1/\theta_N$ average time to abandon

Model $\theta_N = \delta N^{\gamma}$, $\delta > 0$

$\gamma > 0$ impatient

$\gamma < 0$ very patient

$\gamma = 0$ $\theta_N \equiv \theta$

Weak Convergence
(Process + Stationary Distribution)

Theorem: $\underline{q_N} \stackrel{d}{\Longrightarrow} q \quad \textit{if} \;\; q_N(0) \stackrel{d}{\Longrightarrow} q(0).$

Theorem: $\underline{v_N} \stackrel{d}{\approx} \frac{1}{\sqrt{N}} \left(\frac{q}{\mu}\right)^{+}.$ Puhalski

q :

$\theta \downarrow 0$
Patient

$$\begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}\, dW(t) \\ f(x) = \begin{cases} -\mu(\beta + x) & x \le 0 \\ -\mu\beta & x > 0 \end{cases} \end{cases} \qquad \begin{pmatrix} OU & q \le 0 \\ BM & q > 0 \end{pmatrix}$$

M/M/N
Erlang C — Halfin-Whitt

$\theta \uparrow \infty$
Impatient

$$\begin{cases} dq(t) = -\mu(\beta + q(t))dt \\ \qquad\qquad + \sqrt{2\mu}\, dW(t) - dY(t) \qquad (ROU\ q \le 0) \\ Y(0) = 0\,,\ Y \uparrow 0\,,\ q\, dY = 0 \end{cases}$$

M/M/N/N
Erlang B — Jagerman

θ fixed
Rational

$$\begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}\, dW(t) \\ f(x) = \begin{cases} -\mu(\beta + x) & x \le 0 \\ -(\mu\beta + \theta x) & x > 0 \end{cases} \end{cases} \qquad \begin{pmatrix} OU & q \le 0 \\ OU & q > 0 \end{pmatrix}$$

Fleming-Simon-Stolyar
Erlang A

(W - standard Brownian Motion)

(β - as before; will be re-introduced momentarily)

Designing a Call Center - Selecting a Model

- Very impatient customers - $M/M/N/N$ model
- Very patient customers - $M/M/N$ model
- "Balanced" abandoning - sub-model $\gamma = 0$

A Service Center in RUSH HOUR

amazon.com

We're sorry!

Our store is closed temporarily for scheduled maintenance. If you enter your e-mail address, we'll notify you as soon as we reopen. Again, our apologies for the inconvenience.

Thanks for your patience,

Your friends at Amazon.com

Please enter your e-mail address: Submit amazon.co.uk™

We'll be right back!

We're sorry, but our store is closed temporarily. We expect to be back soon.
If you would like to be notified when we reopen, please leave your
e-mail address below and we will be happy to let you know.

Keep shopping...

Feel free to visit our partner site Amazon.com, which remains open.
Again, we apologise for the inconvenience, and thank you for your patience.

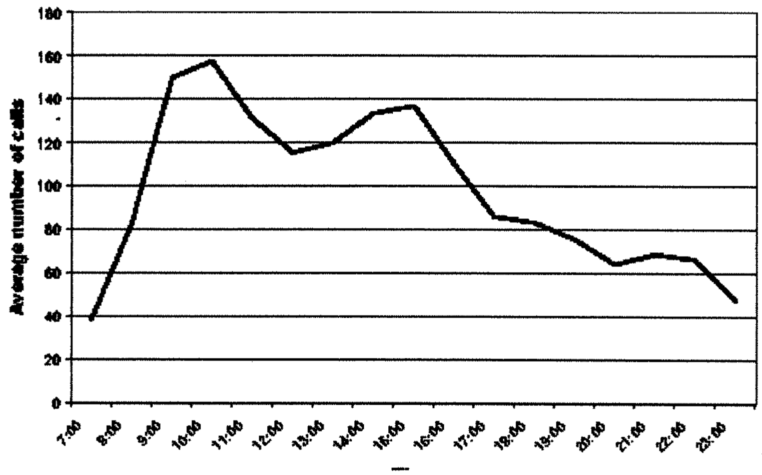
Your friends at Amazon.co.uk

Please enter your e-mail address: [ ] Submit

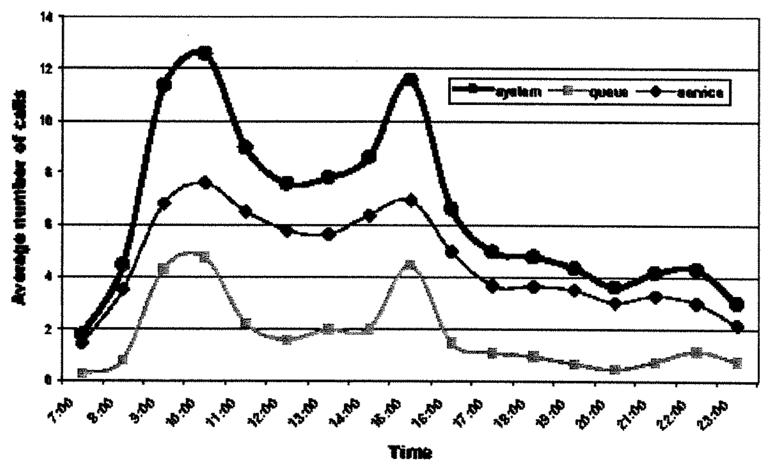
39.

Time-Varying Queues: Predictable Variability

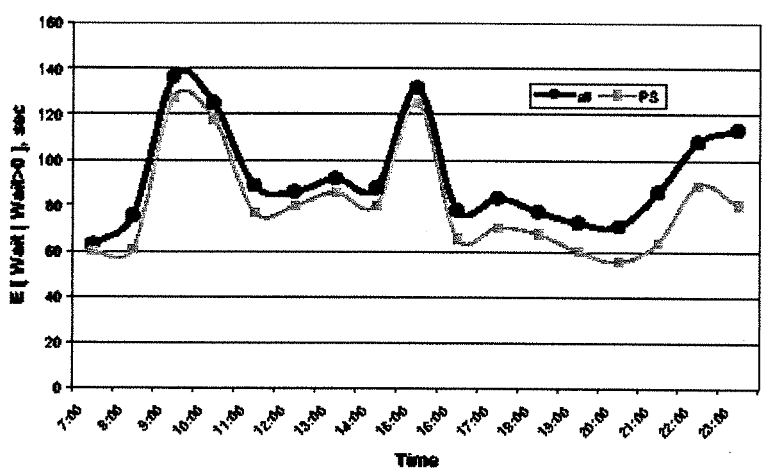
Arrivals



Queues

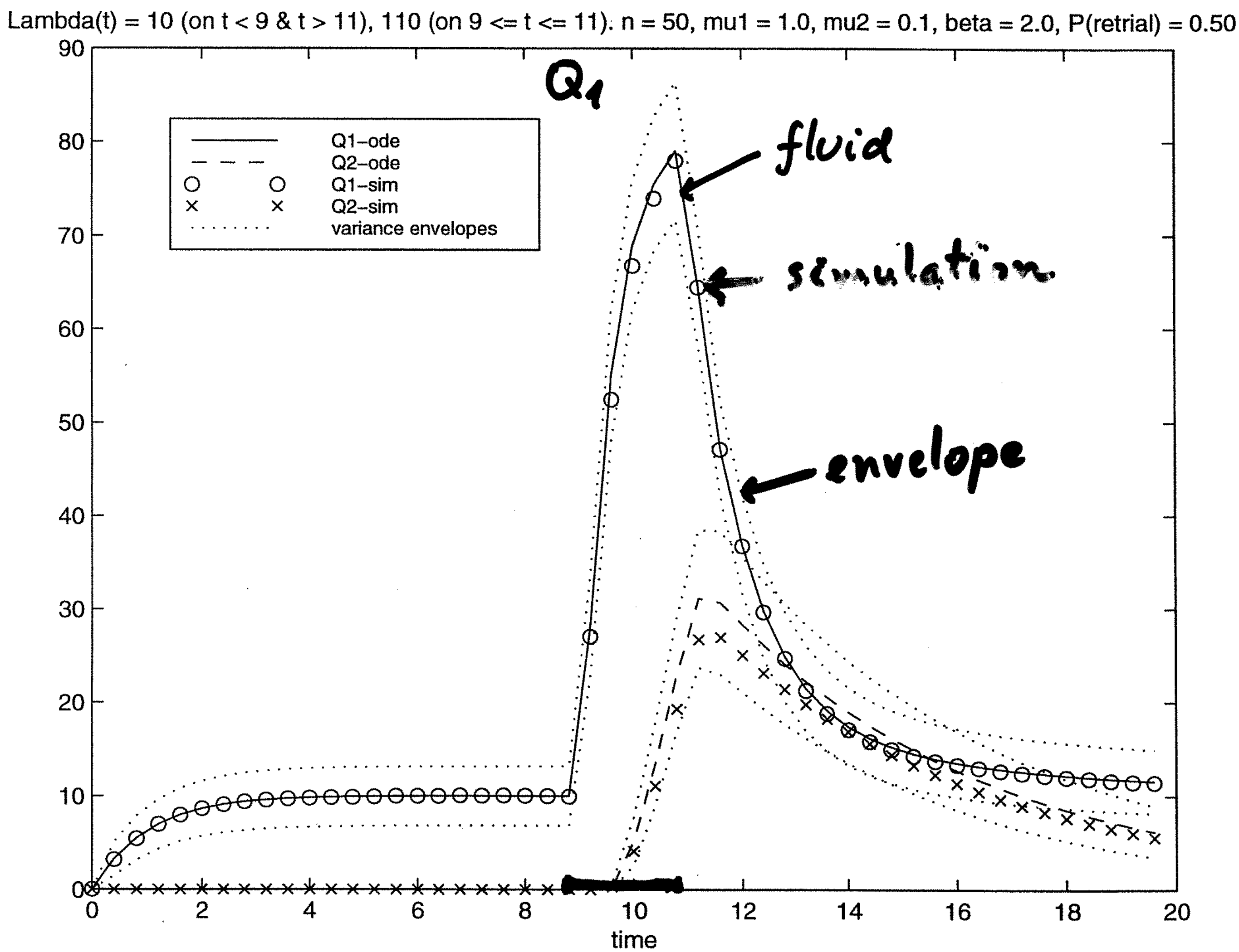


Waiting



Rush Hour

Deterministic Time-Varying Motion
Predictable Variability



$\lambda_t \equiv 10$ Peak = 110

Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials[1]

Avi Mandelbaum
Technion Institute
Haifa, 32000, ISRAEL
avim@tx.technion.ac.il

William A. Massey
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
will@research.bell-labs.com

Martin I. Reiman
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
marty@research.bell-labs.com

Brian Rider
Courant Institute
New York, NY 10012-1185, U.S.A.
riderb@cims.nyu.edu

Alexander Stolyar
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
stolyar@research.bell-labs.com

April 7, 2000

Abstract

We consider a Markovian multiserver queueing model with time dependent parameters where waiting customers may abandon and subsequently retry. We provide simple fluid and diffusion approximations for both the queue length and virtual waiting time processes arising in this model.

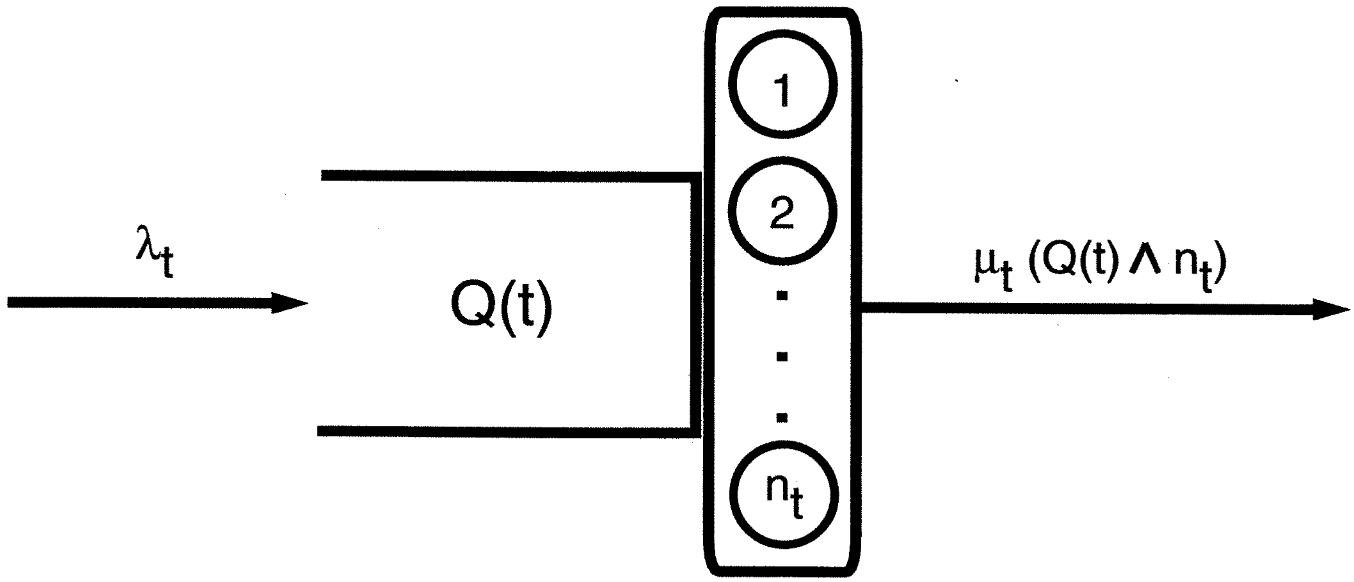
These approximations, which are justified by limit theorems where the arrival rate and number of servers grow large, are compared to simulations, and perform extremely well.

Keywords: Call Centers, Fluid Approximations, Diffusion Approximations, Multiserver Queues, Queues with Abandonment, Virtual Waiting Time, Queues with Retrials, Nonstationary Queues.

[1]Submitted to the Selected Proceedings of the Fifth INFORMS Telecommunications Conference.

Fluid ~~Approximation~~ Model

$M_t / M_t / n_t$

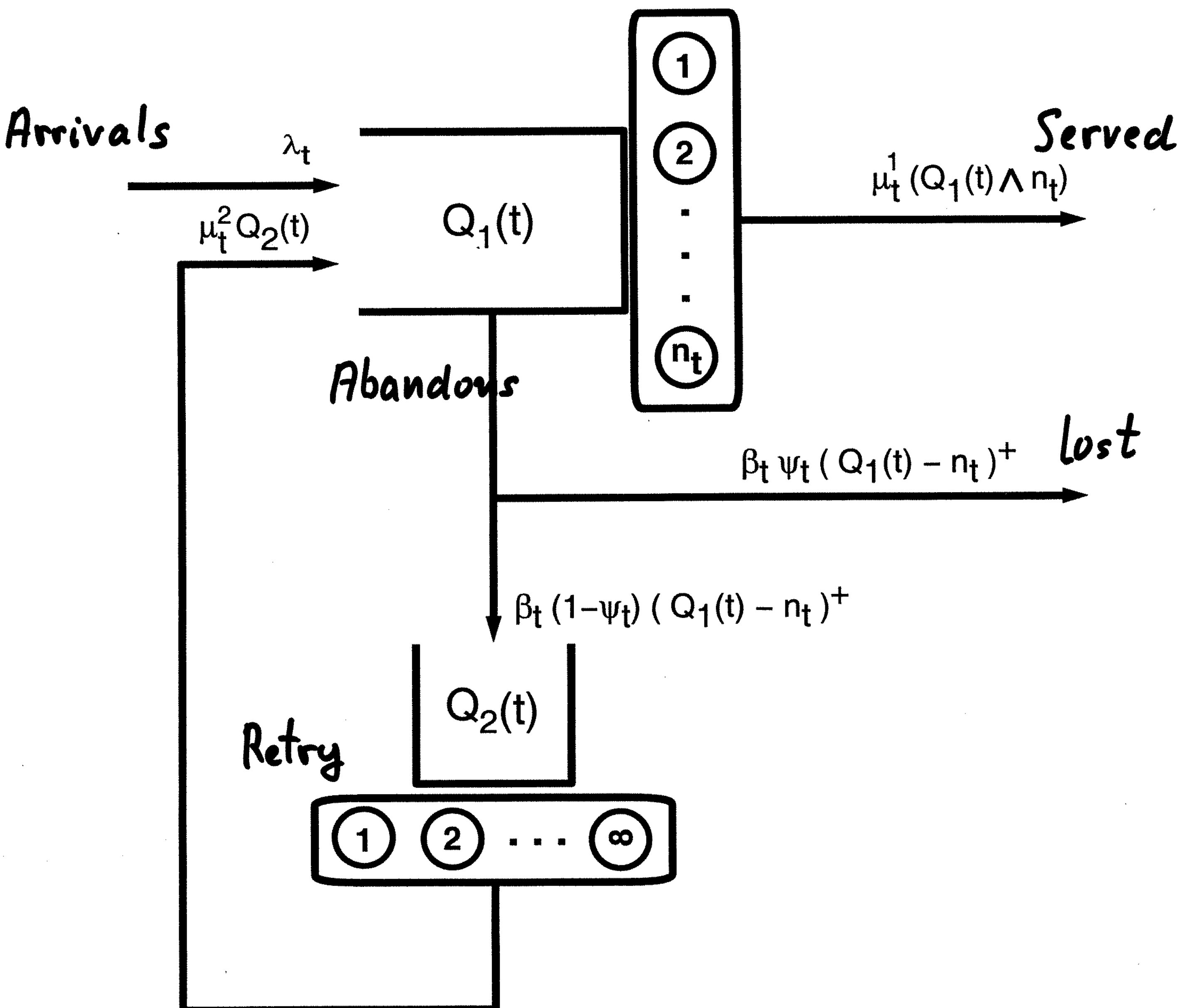


Fluid: $\frac{d}{dt} Q^{\circ}(t) = \lambda_t - \mu_t \left(Q^{\circ}(t) \wedge n_t \right)$

Time Varying Multiserver Queues ...

Massey, Reiman, Rider, Stolyar

Call Center: A Multiserver Queue with Abandonment and Retrials



Primitives

λ_t external arrival rate to calling node 1

e.g., continuously changing, sudden peak

μ_t^1 service rate at 1

e.g., change in nature of work, fatigue

n_t number of servers

e.g., in response to offered load

β_t abandonment rate from 1

e.g., in response to IVR discouragement at predictable overloading

ψ_t prob. of no retrial

$1/\mu_t^2$ average time to retry

Large system: $\eta \uparrow \infty$ scaling parameter

$$\lambda_t \to \eta\,\lambda_t$$

$$n_t \to \eta\,n_t$$

Model

Fluid Appr~~ox~~imation

$Q^0(t) = (Q_1^0(t),\ Q_2^0(t))$

Solution to nonlinear differential balance equations

$$
\begin{aligned}
\frac{d}{dt}\, Q_1^0(t) \;&=\; \lambda_t + \mu_t^2\, Q_2^0(t) \\
&\qquad - \mu_t^1\, (Q_1^0(t) \wedge n_t) - \beta_t\, (Q_1^0(t) - n_t)^+ \\
\frac{d}{dt}\, Q_2^0(t) \;&=\; \beta_1 (1 - \psi_t)(Q_1^0(t) - n_t)^+ \\
&\qquad - \mu_t^2\, Q_2^0(t)
\end{aligned}
$$

Justification: **Functional Strong Law of Large Numbers**, with $\lambda_t \to \eta\lambda_t,\ n_t \to \eta n_t$.

As $\eta \uparrow \infty$,

$$\frac{1}{\eta}\, Q^{\eta}(t) \to Q^0(t)\,, \quad \text{uniformly on compacts.}$$

Diffusion Refinement

$$Q^\eta(t) \stackrel{d}{=} \eta\, Q^0(t) + \sqrt{\eta}\; Q^1(t) + o(\sqrt{\eta})$$

Justification: **Functional Central Limit Theorem**

$$\sqrt{\eta}\left[\frac{1}{\eta}\, Q^\eta(t) - Q^0(t)\right] \Rightarrow Q^1(t), \qquad \text{in } D[0,\infty)$$

Q^1 solution to stochastic differential equation.

Deduce differential equations for

$EQ^1(t)$, Var $Q^1(t)$: confidence envelopes

Can be solved numerically.

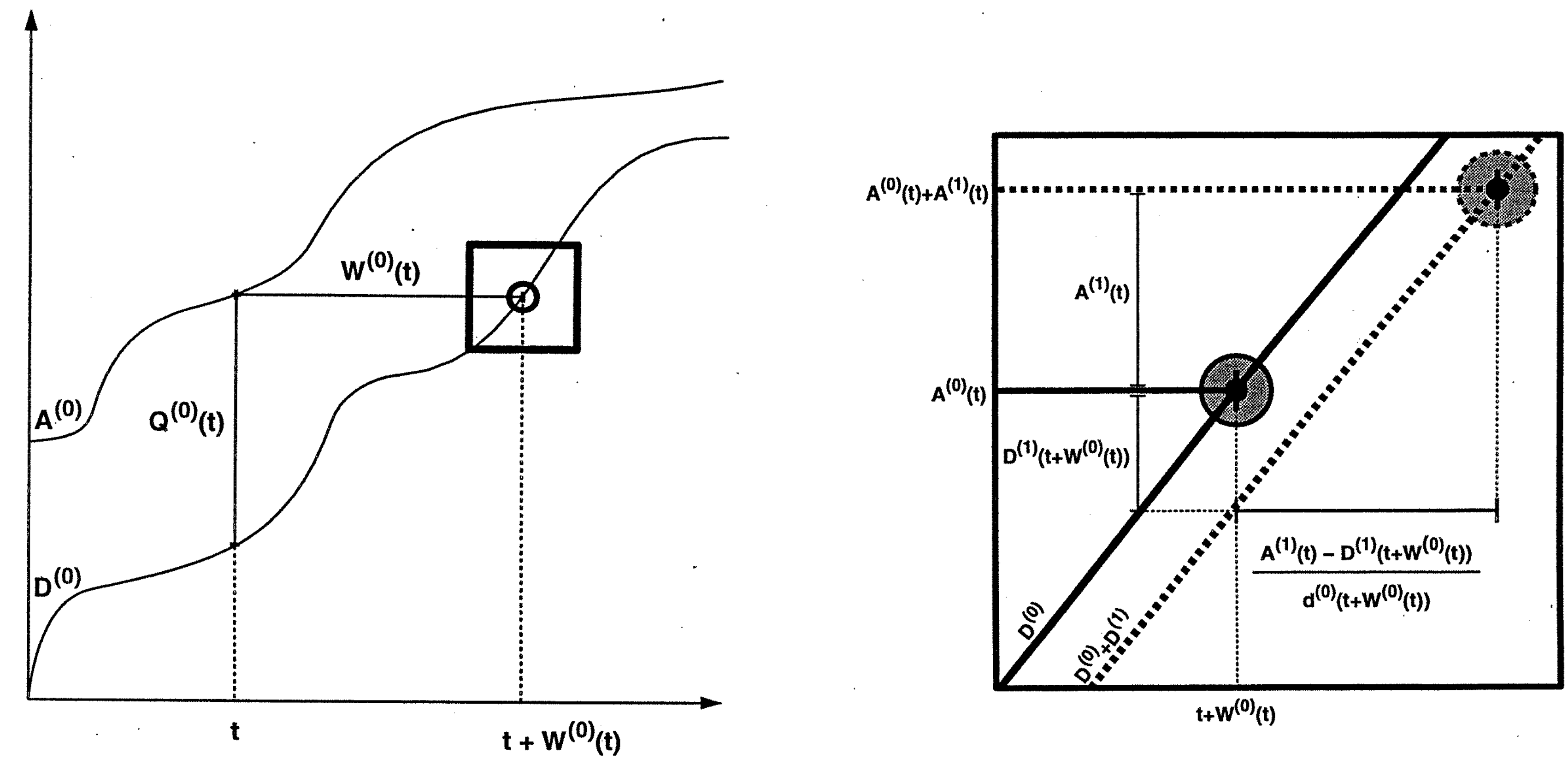
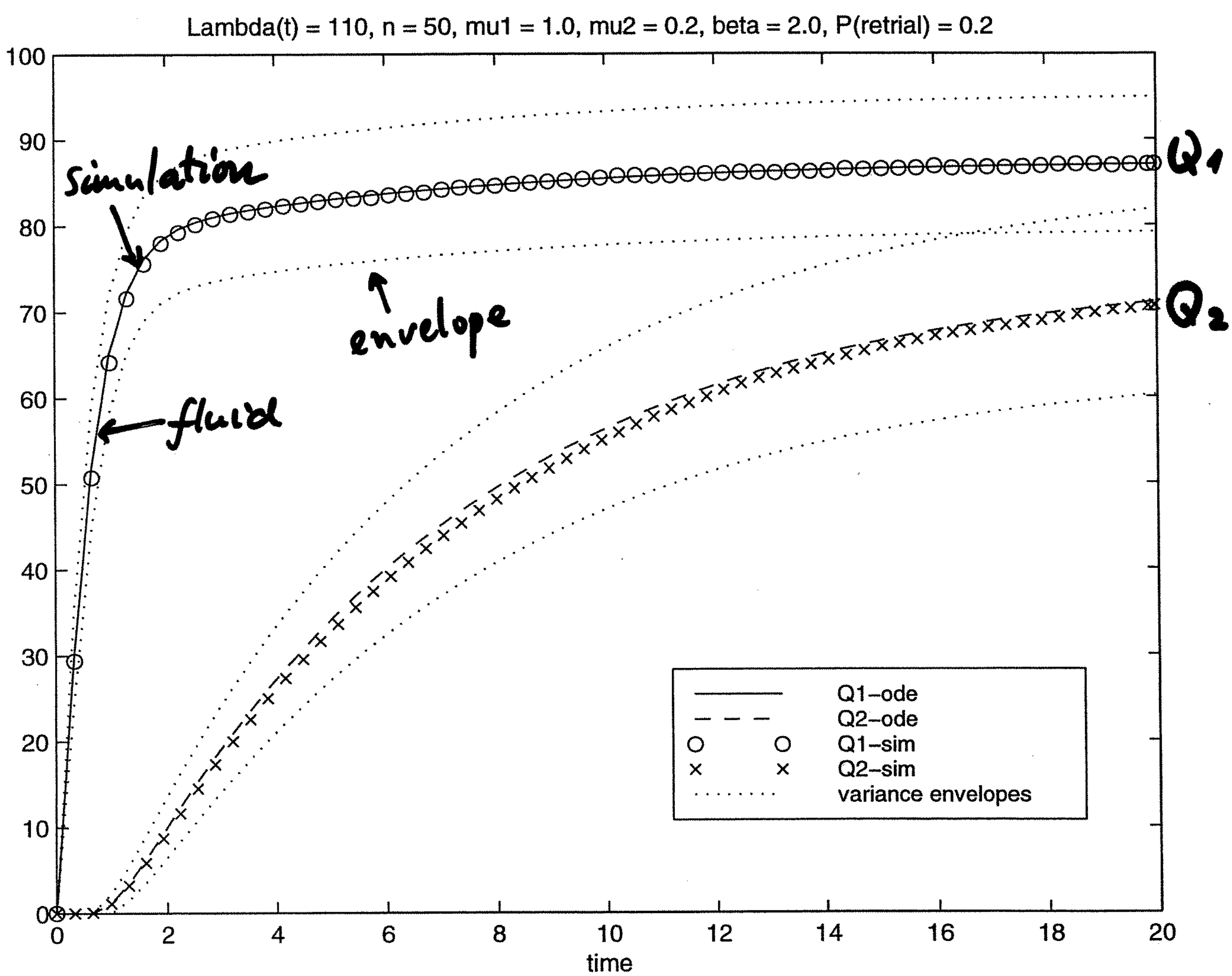


Figure 2: The diffusion term for the virtual waiting time

$\lambda_t \equiv 110$

$n_t \equiv 50$



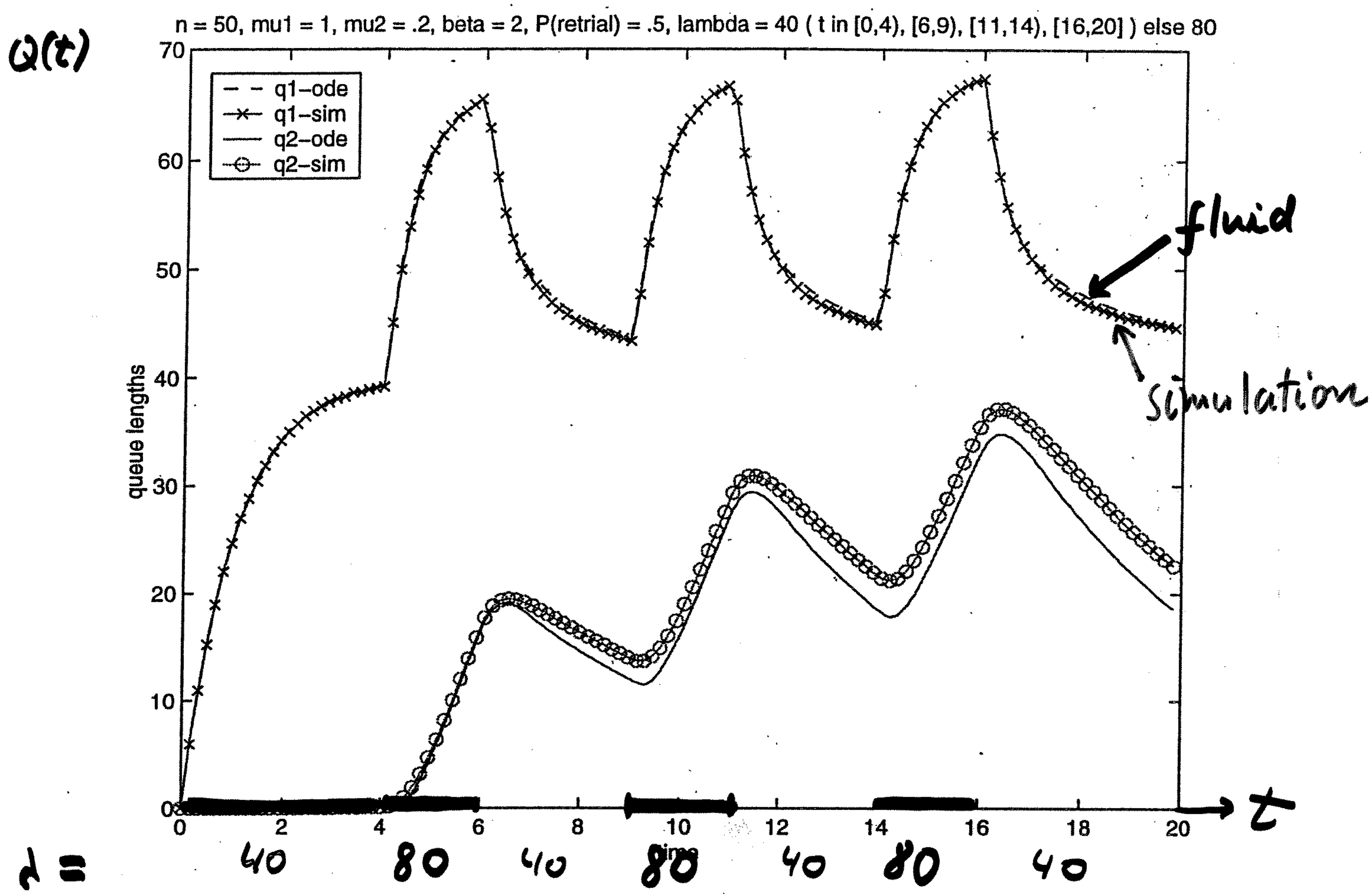


Figure 2: Numerical example: Empirical averages of $Q_1(t)$ and $Q_2(t)$ versus their fluid approximations for square wave case.

Prob.

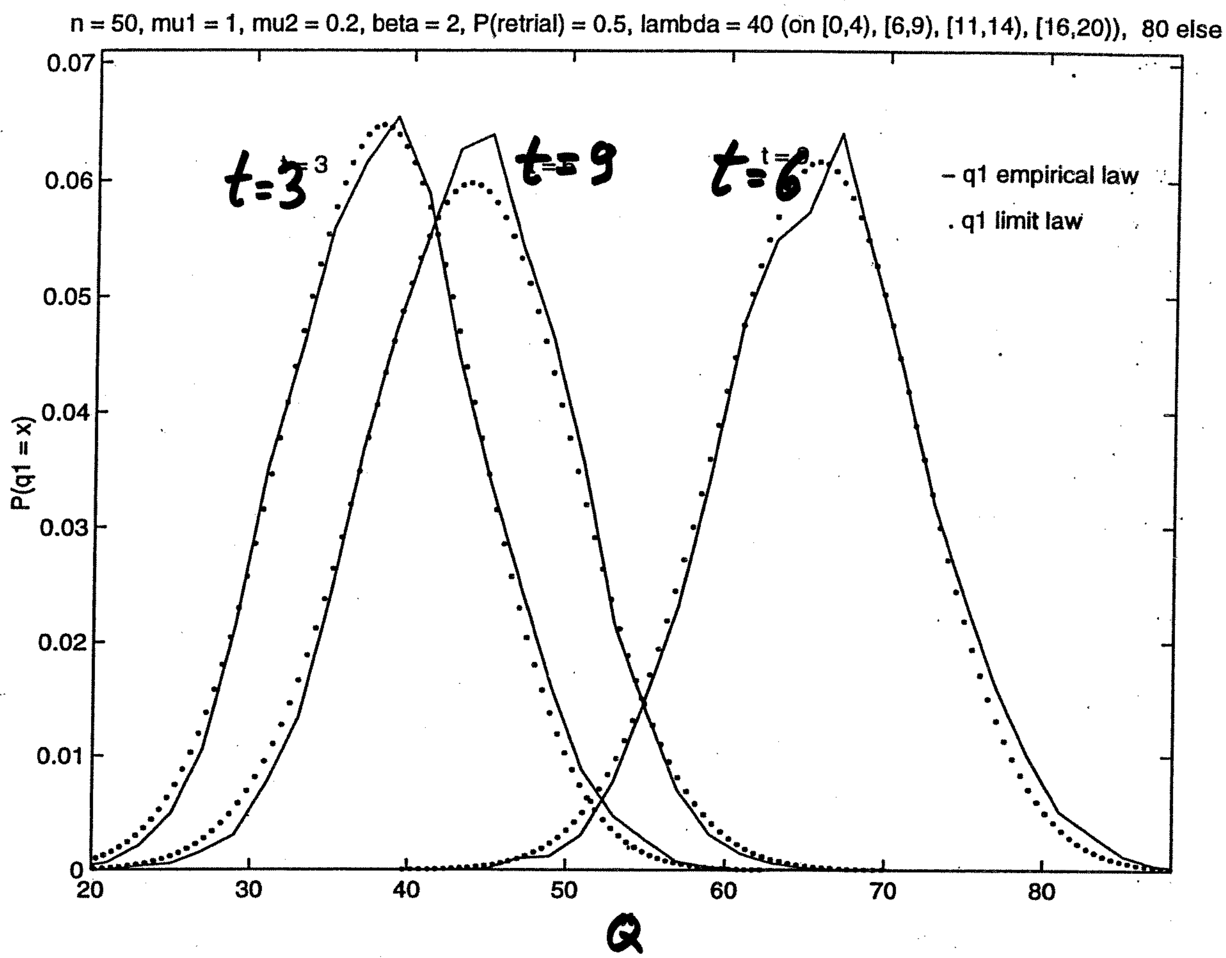


Figure 5: Numerical example: Empirical distribution of Q_1 versus the same from its diffusion approximation.

The Fluid View

- Predictable variability is dominant (Std << Mean)
- The value of the fluid-view increases with the complexity of the system from which it originates

- Legitimate models of flow systems
  - Often simple and sufficient; empirical, predictive
    - Capacity analysis
    - Inventory build-up diagrams
    - Mean-value analysis
- Approximations
  - First-order fluid approx. of stochastic systems
    - Strong Laws of Large Numbers

    (vs. 2nd-order diffusion approx., Central Limits)
  - Long-run
    - Long horizon, smooth-out variability (strategic)
  - Short-run
    - Short horizon, deterministic (operational)
- Technical tools
  - Lyapunov functions to establish stability (Long-run)
  - Building blocks for stochastic models (M(t)/M(t)/1)

Server Staffing To Meet Time-Varying Demand

Otis B. Jennings • Avishai Mandelbaum • William A. Massey • Ward Whitt
School of Industrial and System Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
Industrial and Management Engineering, The Technion, Haifa 32000, Israel
Bell Laboratories, Lucent Technologies, Room 2C-120, Murray Hill, New Jersey 07974-0636
AT&T Laboratories, Room 2C-178, Murray Hill, New Jersey 07974-0636

We consider a multiserver service system with general nonstationary arrival and service-time processes in which $s(t)$, the number of servers as a function of time, needs to be selected to meet projected loads. We try to choose $s(t)$ so that the probability of a delay (before beginning service) hits or falls just below a target probability at all times. We develop an approximate procedure based on a time-dependent normal distribution, where the mean and variance are determined by infinite-server approximations. We demonstrate that this approximation is effective by making comparisons with the exact numerical solution of the Markovian $M_t/M/s_t$ model.

(*Operator Staffing; Queues; Nonstationary Queues; Queues with Time-dependent Arrival Rates; Multiserver Queues; Infinite-server Queues; Capacity Planning*)

1. Introduction

We propose a procedure to determine how many servers are needed, as a function of time, in a *nonstationary* stochastic service system. We assume that any number of servers can be assigned as a function of time in response to *projected* loads, but that the server assignment cannot be changed adaptively in real time in response to *observed* loads. This problem is often referred to as the *operator staffing problem*; e.g., see Andrews and Parsons (1989, 1993), Brigandi et al. (1994), Buffa et al. (1976), Gaballa and Pearce (1979), Grassmann (1986, 1988), Hall (1991, Chapter 7), Holloran and Byrn (1986), Kolesar (1986), Kolesar et al. (1975), Larson (1972), Quinn et al. (1991), Segal (1974), Sze (1984), and Thompson (1993, 1994).

We investigate the operator staffing problem in the context of the $G_t/GI_t/s_t$ queueing model. There is unlimited waiting room and the service discipline is first-come first-served (FCFS). The arrival process is partially characterized by a time-dependent arrival-rate function $\lambda(t)$ and a time-dependent variability function $c_a^2(t)$ (to be defined below); the service times are mutually independent and independent of the arrival process; the cumulative distribution function (cdf) of the service time of an arrival at time t is $G_t(x)$; and the number of servers as a function of time is $s(t)$, which is subject to control, given the functions $\lambda(t)$, $c_a^2(t)$ and $G_t(x)$. This model contains as an important special case the fully specified $M_t/GI/s_t$ model with nonhomogeneous Poisson arrival process.

We do not consider the important problem of *forecasting uncertainty*, which means uncertainty about the arrival-rate function and other elements of the model. We also do not consider customer abandonments and retrials. These phenomena can be important, but we are primarily interested in providing a sufficiently high grade of service that these phenomena will be negligible.

We develop a procedure for determining the required number of servers, $s(t)$, as a function of time t. We do not discuss the subsequent problems of determining the actual work schedules to provide these servers. *We choose $s(t)$ so that the probability of delay* (the probability that an arrival at time t would have to wait before

53

0025-1909/96/4210/1383$01.25

Designing Call Centers with Impatient Customers

Rational Abandonments from Phantom Queues

Joint Research with
Nahum Shimkin, Ety Zohar

Avi Mandelbaum
Industrial Engineering and Management
TECHNION, ISRAEL

Rational Consistent Equilibrium

Rationality Each customer optimizes own utility

Consistency Perceived virtual waiting time dist
= Actual

Assumptions: multi-types, continuously distributed
linear waiting cost, fixed service reward

Theorem $\exists\,!$ rational consistent equilibrium

Fixed point of

- F perceived dist of virtual wait, by all types
  - abandonment time optimized, per type
  - G patience distribution
- F actual dist of virtual wait in $M/M/N+G$

Notes:

- Equilibrium explicitly computable, up to a scalar fixed-point eq.
- Equilibrium hazard-rate dist is DFR - F

Rational Abandonments from Invisible (Phantom) Queues (Shimkin)

Goal: Model **Patience** (time to abandon) while waiting, capturing interplay with system performance

Tradeoffs: Costly waits lead to frequent abandons, which lead to less waits, ... leading to a (surprisingly) unique

Nash Equilibrium

Framework: $M/M/N+G$

Baccelli and Hebuterne '81, ...

Observation: The virtual waiting time has Increasing Failure Rate (IFR). Hence, either

renege or **join to get served.**

$\Rightarrow$ Modify $M/M/m$

The M/M/N(q) model

Palm ('53): "relatively often,... through errors in dialing,... (subscribers) would not receive any ringing tone, so that they were presented with a delay time of unlimited duration"...

leading to

Arriving customer: either

joins a standard M/M/N queue w.p. q; or

w.p. $1 - q$ routed to a "black hole", where never served, without being notified.

Interpretations: proxy for

- system fault – occasional unusual congestion
- individual fault – customer "forgotten"

Rational Consistent Equilibrium

Type z customer, distributed P_Z

C_z – cost rate, R_z – reward (constant) ($U_z(t)$)

Complete system description requires

F_z – belief of each z-type customer for

F – the virtual waiting time distribution.

With F_z exogeneously specified:

Compute T_z = optimal (rational) abandon time;

Deduce patience distribution $G(t) = P_Z\{z : T_z \le t\}$.

Then resulting $M/M/N + G$ queue yields F.

However, F_z should be related to F: Learning

Consistency Assumption: $F_z \equiv F, \quad \forall z$.

Consistent equilibrium: decision profile $\{T_z\}$,

which is a fixed point of the map:

$$\{T_z\} \xrightarrow{P_Z} G \xrightarrow{M/M/N+G} F \xrightarrow{F_z \equiv F,\ U_z(\cdot)} \{T_z\}$$

$$\underset{\text{expectation } n}{F} \xrightarrow{U_z} \{T_z\} \xrightarrow{P_z} G \xrightarrow{M/M/N+G} \underset{\text{perception } n+1}{F}$$

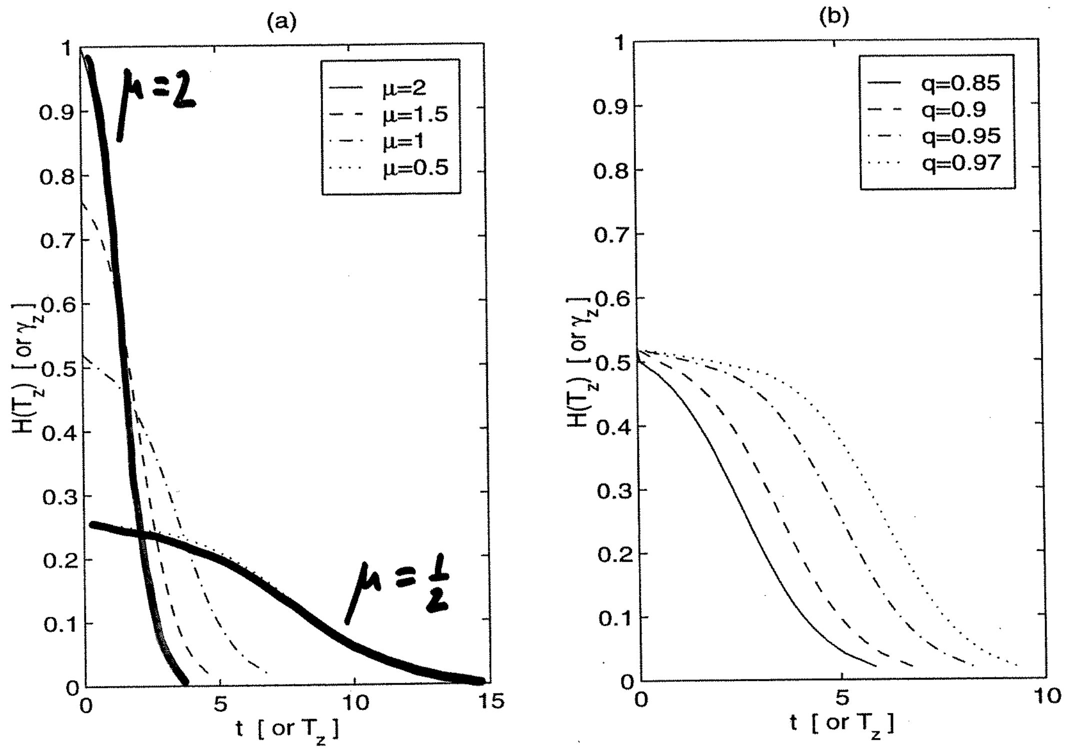
Main Results

Theorem: M/M/N(q) with rational abandonment in equilibrium

- *An equilibrium profile exists, is unique, and is non-trivial in general.*
- *The equilibrium profile can be computed explicitly, up to a scalar fixed-point equation.*
- *The equilibrium hazard rate of the virtual waiting time is nonincreasing.*

Interplay with Performance

$z \sim U$
μ varies



Equilibrium profiles for an M/M/m(q) system with uniformly distributed γ_z.

(a) $q = 0.9$ while μ is varried

(b) $\mu = 1$ while q is varried

As $\mu \uparrow$, more enter system $\quad \theta \uparrow$

but they are more impatient $\quad T \downarrow$

Learning (Simulation)

- M/M/1 queue (overloaded, to be interesting).

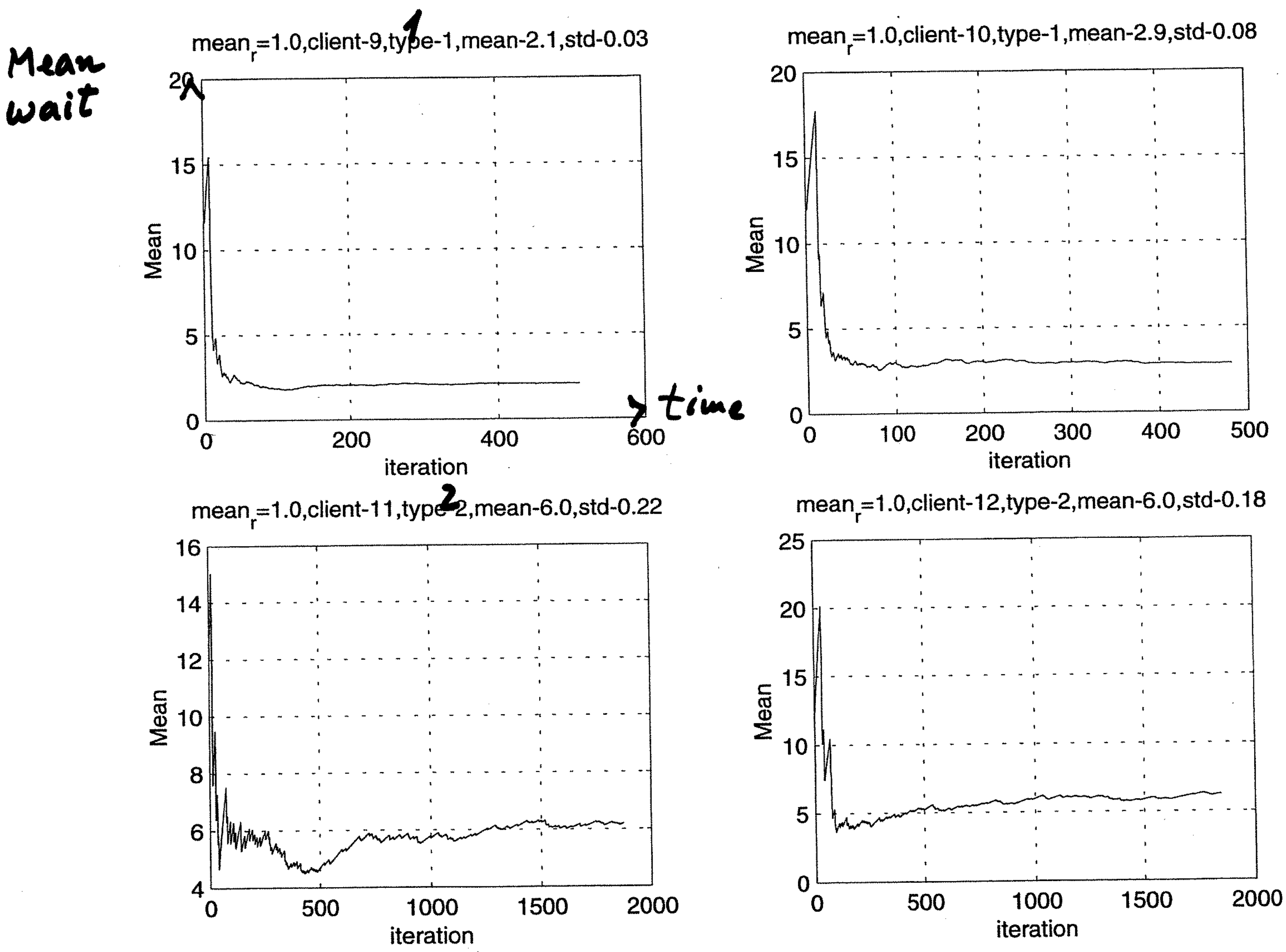
- **Partial** consistency — subjective waiting time is influenced by objective **mean** waiting time.

- **Customers assume**: i.i.d. **exponential** (virtual) waits (constant hazard rate).

- Estimation of mean waiting time is performed by some statistical procedure, for example (**censored**) MLE, based on individual experience.

- Equilibrium ?

Partial Consistency : "Exponential" Lenses



Type-dependent learning

Censored estimation

Further Research

- Extensions to the basic model:
  - Nonlinear (convex) waiting costs
  - More general "failure" models – priorities, variable # of servers, server vacations (outbound calls)
- Partially consistent equilibrium – parametric learning (e.g., F_z exponential mean m_z – empirical backup)
- Retrials
- Status information (instead of "music")
- Dynamic learning schemes (simulations)
- Empirical verification, tuning and refinements; in particular model the cost of waiting (psychological, economical), and the triggers for abandons and retrials.
- Customer-focused design: Pooling types