

**QED Q's**

**Telephone Call/Contact Centers**

**Service Engineering**

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

**Website: <http://ie.technion.ac.il/serveng>**

## Supporting Material (Downloadable)

M. "Call Centers: Research **Bibliography** with Abstracts."  
Version 7, December 2006.

**Gans, Koole, and M.**: "Telephone Call Centers: Tutorial, Review and Research Prospects." *MSOM*, 2003. (Sec. 3-4, possibly 2.)

Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." *JASA*, 2005.

**Erlang**: "On the **rational** determination of the number of circuits."  
Written in the 20's; In "*The life and works of A.K. Erlang*," 1948.

Halfin and Whitt: "Heavy Traffic Limits for Queues with Many Exponential Servers." *OR*, 1981.

Jelelnkovic, M. and Momcilovic: "Heavy Traffic Limits for Queues with Many **Deterministic** Servers." *QUESTA*, 2004.

**Borst, M. and Reiman**: "**Dimensioning** Large Telephone Call Centers." *OR*, 2004.

Whitt's website: both the ED and QED regimes.

## Supporting Material (Downloadable)

M. "Call Centers: Research **Bibliography** with Abstracts."  
Version 7, December 2006.

**Gans, Koole, and M.**: "Telephone Call Centers: Tutorial, Review and Research Prospects." *MSOM*, 2003. (Sec. 3-4, possibly 2.)

Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." *JASA*, 2005.

**Erlang**: "On the **rational** determination of the number of circuits."  
Written in the 20's; In "*The life and works of A.K. Erlang*," 1948.

Halfin and Whitt: "Heavy Traffic Limits for Queues with Many Exponential Servers." *OR*, 1981.

Jelelnkovic, M. and Momcilovic: "Heavy Traffic Limits for Queues with Many **Deterministic** Servers." *QUESTA*, 2004.

**Borst, M. and Reiman**: "Dimensioning Large Telephone Call Centers." *OR*, 2004.

Whitt's website: both the ED and QED regimes.

# 4CallCenters: Personal Tool for WFM

4CallCenters v2.01

File Table Settings Help

Performance Profiler Staffing Query Advanced Profiling Advanced Queries What-if Analysis

**Performance Profiler** Performance Profiler allows you to determine and optimize the Performance Level of your Call Center. Enter your call center's parameters below, then press 'Compute'.

Your Call Center's Parameters

Number of Agents Answering Calls: 10  
 Average Time to Handle One Call (mm:ss): 01:00  
 Calls per 60 minute Interval: 100  
 Average Callers' Patience (mm:ss): 01:00  
 Number of Trunks Available: 50

Features: Abandons, Trunks  
 Basic Interval: 60 minutes  
 Target Time: 00:20 (mm:ss)

Change Settings

Compute Add to Table Delete Rows Clear All Export Graph

Results	Basic Interval (minutes)	Target Time to Answer	Number of Agents	Average Handling Time	Calls per Interval	Average Patience	Number of Trunks	Agent's Occupancy	Average Trunks Utilized	%Answer	%Abandon	%Block	Average Speed of Answer	%Answer within Target	Average Queue Length
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															

Settings Parameters Indicators

Ready 19/01/2004 17:01

4CallCenters v2.01

File Table Settings Help

Performance Profiler Staffing Query Advanced Profiling Advanced Queries What-if Analysis

**Advanced Profiling** The Advanced Profiling tool allows you to enter multiple values for each of the input parameters, and produces a performance profile for each combination of parameters.

Compute Add to Table Delete Rows Clear All Export Graph Settings

Input Multi-Value	60	00:20	Range	03:30	40	05:00					
Results	Basic Interval (minutes)	Target Time to Answer	Number of Agents	Average Handling Time	Calls per Interval	Average Patience	Agent's Occupancy	%Answer	%Abandon	Average Speed of Answer	%
226	60.0	00:20.0	8.0	03:30.0	210.0	05:00.0	98.9%	64.6%	35.4%	01:59.4	
227	60.0	00:20.0	9.0	03:30.0	210.0	05:00.0	97.8%	71.9%	20.1%	01:29.0	
228	60.0	00:20.0	10.0	03:30.0	210.0	05:00.0	96.1%	78.4%	21.6%	01:05.8	
229	60.0	00:20.0	11.0	03:30.0	210.0	05:00.0	93.8%	84.1%	15.9%	00:46.8	
230	60.0	00:20.0	12.0	03:30.0	210.0	05:00.0	90.5%	88.7%	11.3%	00:32.4	
231	60.0	00:20.0	13.0	03:30.0	210.0	05:00.0	86.9%	92.3%	7.7%	00:21.7	
232	60.0	00:20.0	14.0	03:30.0	210.0	05:00.0	82.1%	94.9%	5.1%	00:14.1	
233	60.0	00:20.0	15.0	03:30.0	210.0	05:00.0	79.0%	96.8%	3.2%	00:08.9	
234	60.0	00:20.0	16.0	03:30.0	210.0	05:00.0	75.1%	98.0%	2.0%	00:05.4	
235	60.0	00:20.0	4.0	03:30.0	220.0	05:00.0	100%	31.2%	68.8%	05:24.2	
236	60.0	00:20.0	5.0	03:30.0	220.0	05:00.0	100%	39.0%	61.0%	04:22.4	
237	60.0	00:20.0	6.0	03:30.0	220.0	05:00.0	99.9%	46.7%	53.3%	03:31.4	
238	60.0	00:20.0	7.0	03:30.0	220.0	05:00.0	99.7%	54.4%	45.8%	02:48.4	
239	60.0	00:20.0	8.0	03:30.0	220.0	05:00.0	99.3%	61.9%	38.1%	02:11.9	
240	60.0	00:20.0	9.0	03:30.0	220.0	05:00.0	98.5%	69.1%	30.9%	01:41.2	
241	60.0	00:20.0	10.0	03:30.0	220.0	05:00.0	97.1%	75.7%	24.3%	01:15.7	
242	60.0	00:20.0	11.0	03:30.0	220.0	05:00.0	95.1%	81.6%	18.4%	00:55.2	
243	60.0	00:20.0	12.0	03:30.0	220.0	05:00.0	92.5%	86.5%	13.5%	00:39.2	
244	60.0	00:20.0	13.0	03:30.0	220.0	05:00.0	89.3%	90.5%	9.5%	00:27.0	
245	60.0	00:20.0	14.0	03:30.0	220.0	05:00.0	85.8%	93.6%	6.4%	00:18.0	
246	60.0	00:20.0	15.0	03:30.0	220.0	05:00.0	82.7%	95.8%	4.2%	00:11.7	
247	60.0	00:20.0	16.0	03:30.0	220.0	05:00.0	78.1%	97.4%	2.8%	00:07.3	
248	60.0	00:20.0	4.0	03:30.0	230.0	05:00.0	100%	29.8%	70.2%	05:37.6	
249	60.0	00:20.0	5.0	03:30.0	230.0	05:00.0	100%	37.2%	62.7%	04:35.7	
250	60.0	00:20.0	6.0	03:30.0	230.0	05:00.0	99.9%	44.1%	56.8%	03:44.5	
251	60.0	00:20.0	7.0	03:30.0	230.0	05:00.0	99.8%	52.1%	47.9%	03:01.2	
252	60.0	00:20.0	8.0	03:30.0	230.0	05:00.0	99.5%	59.4%	40.6%	02:44.2	
253	60.0	00:20.0	9.0	03:30.0	230.0	05:00.0	99.0%	66.4%	32.6%	01:52.6	
254	60.0	00:20.0	10.0	03:30.0	230.0	05:00.0	97.9%	73.0%	27.0%	01:26.0	
255	60.0	00:20.0	11.0	03:30.0	230.0	05:00.0	96.4%	79.0%	21.0%	01:04.1	
256	60.0	00:20.0	12.0	03:30.0	230.0	05:00.0	94.2%	84.2%	15.8%	00:48.5	
257	60.0	00:20.0	13.0	03:30.0	230.0	05:00.0	91.4%	88.6%	11.4%	00:32.9	
258	60.0	00:20.0	14.0	03:30.0	230.0	05:00.0	88.2%	92.0%	8.0%	00:22.6	
259	60.0	00:20.0	15.0	03:30.0	230.0	05:00.0	84.6%	94.6%	6.4%	00:15.0	
260	60.0	00:20.0	16.0	03:30.0	230.0	05:00.0	80.9%	98.5%	3.5%	00:09.7	
261	60.0	00:20.0	4.0	03:30.0	240.0	05:00.0	100%	28.6%	71.4%	05:50.3	
262	60.0	00:20.0	5.0	03:30.0	240.0	05:00.0	100%	35.7%	64.3%	04:48.4	
263	60.0	00:20.0	6.0	03:30.0	240.0	05:00.0	100%	42.8%	57.2%	03:57.7	
264	60.0	00:20.0	7.0	03:30.0	240.0	05:00.0	99.9%	49.9%	50.1%	03:13.7	
265	60.0	00:20.0	8.0	03:30.0	240.0	05:00.0	99.7%	57.0%	43.0%	02:36.3	
266	60.0	00:20.0	9.0	03:30.0	240.0	05:00.0	99.3%	63.0%	36.2%	02:04.0	
267	60.0	00:20.0	10.0	03:30.0	240.0	05:00.0	98.6%	70.4%	29.6%	01:36.5	
268	60.0	00:20.0	11.0	03:30.0	240.0	05:00.0	97.3%	78.5%	23.5%	01:13.3	
269	60.0	00:20.0	12.0	03:30.0	240.0	05:00.0	95.6%	81.9%	18.1%	00:54.4	
270	60.0	00:20.0	13.0	03:30.0	240.0	05:00.0	93.2%	86.5%	13.5%	00:39.3	
271	60.0	00:20.0	14.0	03:30.0	240.0	05:00.0	90.3%	90.3%	9.7%	00:27.6	
272	60.0	00:20.0	15.0	03:30.0	240.0	05:00.0	87.1%	93.3%	6.7%	00:18.9	
273	60.0	00:20.0	16.0	03:30.0	240.0	05:00.0	83.8%	95.5%	4.5%	00:12.5	

Settings Parameters Indicators

Ready 19/01/2004 10:05

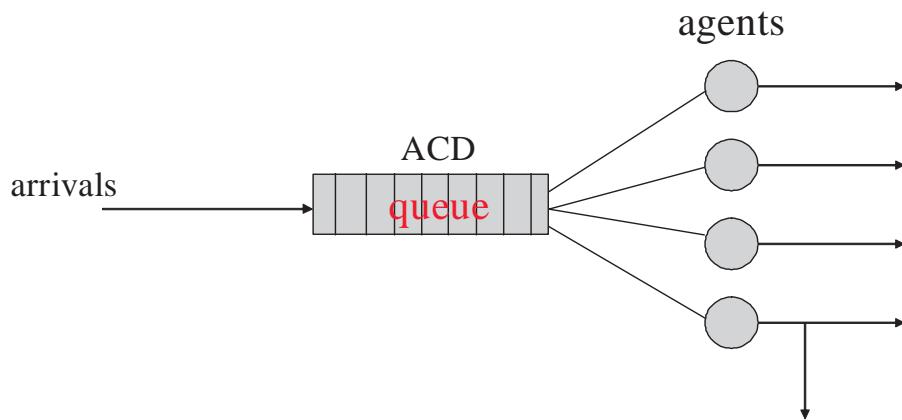
Net abandon vs. Calls per Interval for various Number of Agents

Microsoft Excel - Book1

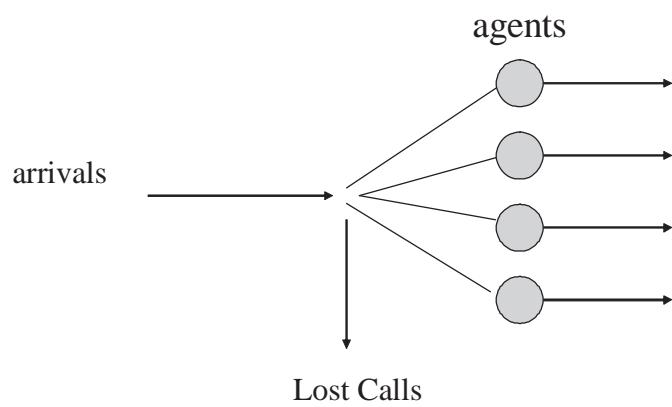
NUM

Start Quick Launch Tera Term - ts-technion... Tera Term - in-technion... Cannot find server - Mic... Microsoft Word - QED\_Q... Microsoft Word - TEMP.doc 4CallCenters - v2.01 Microsoft Excel - Book1 10:05 3 9

## *Erlang-C (M/M/N): # Agents*



## *Erlang-B (M/M/N/N): # Trunks*



# “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 - Classical IE (Erlang-C)

8 transactions per hour  $\Rightarrow$   $E(S) = \underline{\underline{7:30 \text{ minutes}}} (=M)$

<u><math>\lambda</math>/hr</u>	<u>N Agents</u>	<u><math>\rho = OCC</math></u>	<u><math>L_q \equiv Que</math></u>	<u><math>W_q \equiv ASA</math></u>
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

<u><math>\lambda</math>/hr</u>	<u>N</u>	<u><math>\rho = \text{OCC}</math></u>	<u><math>L_q = \text{Que}</math></u>	<u><math>W_q = \text{ASA}</math></u>
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
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
$\infty$	$\infty$	1	$\infty$	7:30 !

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

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

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

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

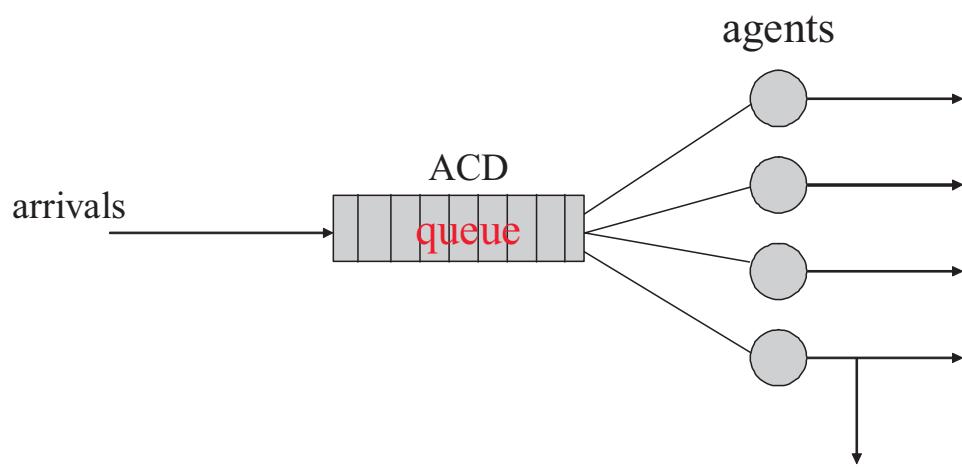
Copy of Summary Interval • Order PK

Date: 7/7/97

Split/Skill: Order PK

Time	Avg Speed	Avg Aban	ACD Calls	Avg ACD	Avg ACW	Aban	% ACD	%	Avg	Calls Per	% Serv	% Aux	% ACW	% ACD	
	Ans	Time	Time	Time	Calls	Time	Ans	Pos	Pos	Lev	Time	Time	Time	Time	
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	98	70	149	8			P	
							#Aban								
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	51	7	4	51	2	18	61	
12:30 AM*	:00:03	:04:10	14	:07:27	:00:33	1	89	52	5	3	48	1	26	83	
1:00 AM*	:00:00		9	:04:54	:11:29	0	91	90	1	7	90	0	28	65	
5:30 AM*			0			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	18	
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	28	
7:00 AM*	:00:00		52	: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	89	75	4	97	9	7	47	
9:30 AM*	:00:02	:00:06	381	:03:45	:00:22	2	60	87	91	4	93	8	8	52	
Peak →	10:00 AM*	:00:02	:00:01	416	:03:49	:00:26	1	63	87	94	4	98	5	8	55
10:30 AM*	:00:00		349	:03:35	:00:33	0	52	99	95	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	6	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		338	: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	8	44	
1:30 PM*	:00:00		368	:03:52	:00:14	0	58	99	99	4	99	11	7	50	
2:00 PM*	:00:01		383	:03:55	:00:17	0	51	100	106	4	100	10	6	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	:00:18	1	57	98	110	4	98	8	5	51	
3:30 PM*	:00:00		347	:03:59	:00:14	0	60	100	100	3	100	7	5	45	
4:00 PM*	:00:00		382	:03:48	:01:37	0	64	100	98	4	100	6	7	47	
4:30 PM*	:00:00		378	:03:41	:00:19	0	65	99	97	4	99	8	5	50	
5:00 PM*	:00:00		411	:03:53	:00:19	0	53	100	109	4	100	8	5	48	
5:30 PM*	:00:01		387	:03:58	:00:19	0	58	99	98	4	99	10	6	51	
6:00 PM*	:00:01	:00:21	371	:03:28	:00:25	1	53	98	81	4	98	8	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	

$$\text{Erlang-}C = M/M/N$$



# 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)

# Rough Performance Analysis

Peak 10:00 – 10:30 a.m., with 100 agents

400 calls

3:45 minutes average service time

2 seconds ASA

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

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

$$= 50 \text{ Erlangs}$$

Occupancy  $\rho = R/N$

$$= 50/100 = 50\%$$

# Rough Performance Analysis

Peak      10:00 – 10:30 a.m., with 100 agents  
400 calls  
3:45 minutes average service time  
2 seconds ASA

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

$$= 400 \times 3:45 = 1500 \text{ min./30 min.}$$
$$= 50 \text{ Erlangs}$$

Occupancy       $\rho = R/N$

$$= 50/100 = 50\%$$

⇒ **Quality-Driven Operation**      (Light-Traffic)  
⇒ Classical Queueing Theory

Above:  $R = 50$ ,       $N = R + 50$ ,       $\approx$  **all served immediately**.

Rule of Thumb:  $N = \lceil R + \delta R \rceil$ ,       $\delta > 0$  service-grade.

**Quality-driven:** 100 agents, 50% utilization

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

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

<u><math>\lambda</math></u> /hr	<u><math>\rho</math></u>	$E(W_q) = ASA$	% Wait = 0
800	50%	0	100%

**Quality-driven:** 100 agents, 50% utilization

⇒ Can increase offered load - by how much?

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

<u><math>\lambda</math></u> /hr	<u><math>\rho</math></u>	$E(W_q) = ASA$	% Wait = 0
800	50%	0	100%
1400	87.5%	0:02 min.	88%
1550	96.9%	0:48 min.	35%
1580	98.8%	2:34 min.	15%
1585	<b>99.1%</b>	<b>3:34 min.</b>	12%

**Quality-driven:** 100 agents, 50% utilization

⇒ Can increase offered load - by how much?

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

<u><math>\lambda</math>/hr</u>	<u><math>\rho</math></u>	$E(W_q) = ASA$	% Wait = 0
800	50%	0	100%
1400	87.5%	0:02 min.	88%
1550	96.9%	0:48 min.	35%
1580	98.8%	2:34 min.	15%
1585	<b>99.1%</b>	<b>3:34 min.</b>	12%

⇒ **Efficiency-driven Operation (Heavy Traffic)**

$$\bar{W}_q \approx \bar{W}_q | W_q > 0 = \frac{1}{N} \cdot \frac{\rho_N}{1 - \rho_N} \cdot E(S) \rightarrow E(S) = 3:45 !$$

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

Above:  $R = 99$ ,  $N = R + 1$ ,  $\approx$  all delayed.

Rule of Thumb:  $N = \lceil R + \gamma \rceil$ ,  $\gamma > 0$  service grade.

# Changing N (Staffing) in Erlang-C

$$E(S) = 3:45$$

<u><math>\lambda</math>/hr</u>	<u>N</u>	OCC	ASA	% Wait = 0
1585	100	99.1%	3:34	12%

# Changing N (Staffing) in Erlang-C

$$E(S) = 3:45$$

<u><math>\lambda</math>/hr</u>	<u>N</u>	OCC	ASA	% Wait = 0
1585	100	99.1%	3:34	12%
1599	<b>100</b>	99.9%	<b>59:33</b>	0%

# Changing N (Staffing) in Erlang-C

$$E(S) = 3:45$$

<u><math>\lambda</math>/hr</u>	<u>N</u>	OCC	ASA	% Wait = 0
1585	100	99.1%	3:34	12%
1599	<b>100</b>	99.9%	<b>59:33</b>	0%
1599	<b>100+1</b>	98.9%	<b>3:06</b>	13%
1599	102	98.0%	1:24	24%
1599	105	<b>95.2%</b>	<b>0:23</b>	<b>50%</b>

# Changing N (Staffing) in Erlang-C

$$E(S) = 3:45$$

<u><math>\lambda</math>/hr</u>	<u>N</u>	OCC	ASA	% Wait = 0
1585	100	99.1%	3:34	12%
1599	100	99.9%	59:33	0%
1599	100+1	98.9%	3:06	13%
1599	102	98.0%	1:24	24%
1599	105	95.2%	0:23	50%

⇒ New Rationalized Operation

Efficiently driven, in the sense that  $OCC > 95\%$ ;

Quality-Driven,  $50\%$  answered immediately

**QED Regime = Quality- and Efficiency-Driven Regime**

Above:  $R = 100$ ,  $N = R + 5$ ,  $50\%$  delayed.

√. Safety-Staffing  $N = \lceil R + \beta \sqrt{R} \rceil$ ,  $\beta > 0$  .

## QED Theorem (Halfin-Whitt, 1981)

Consider a sequence of M/M/N models,  $N=1,2,3,\dots$

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

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

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

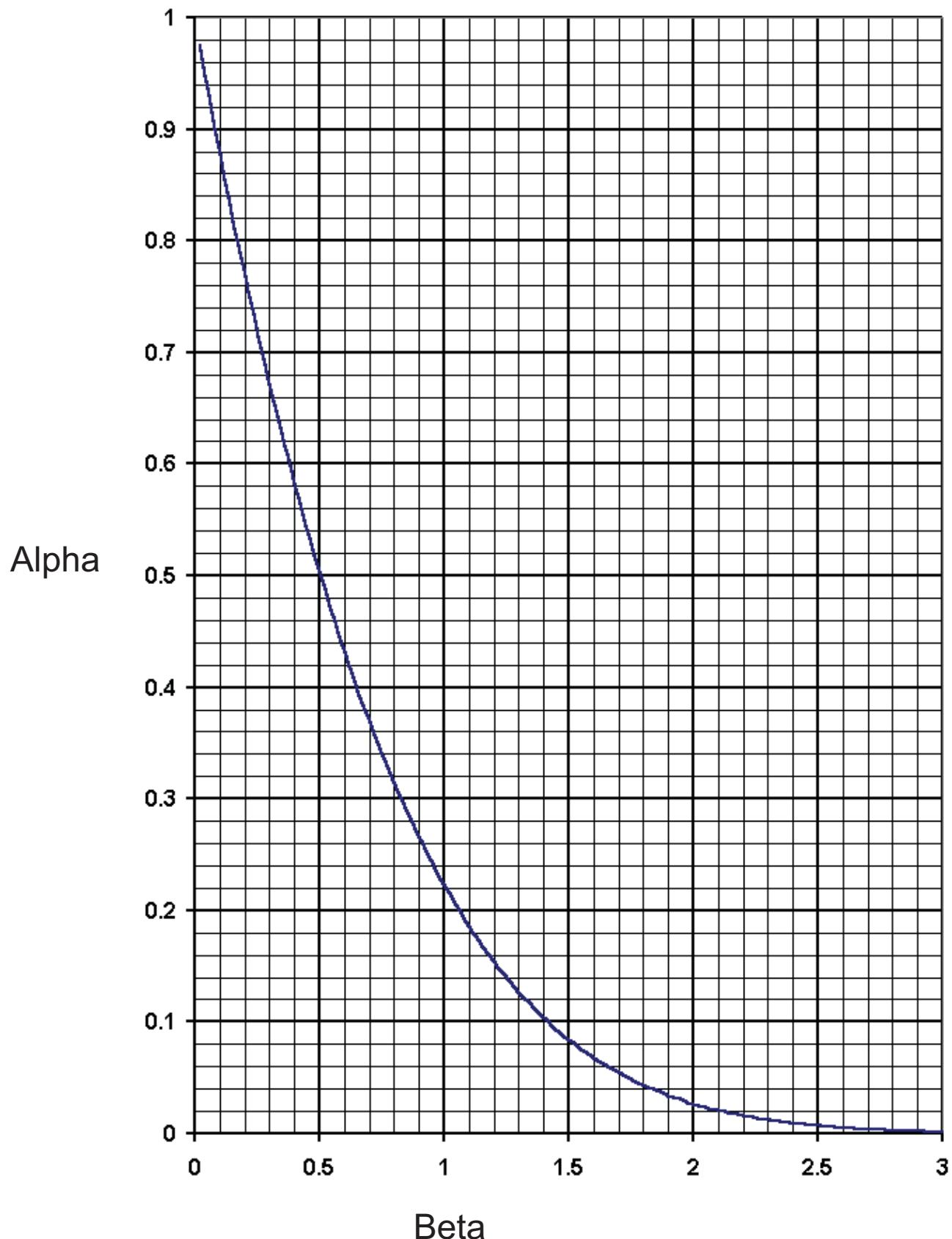
where  $\varphi(\cdot) / \phi(\cdot)$  is the standard normal density/distribution.

Extremes:

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

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

## The Halfin-Whitt Delay Function



## √. Safety-Staffing: Performance

$$R = \lambda \times E(S) \quad \text{Offered load (Erlangs)}$$

$$N = R + \underbrace{\beta \sqrt{R}}_{\beta = \text{“service-grade”} > 0}$$

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

Expected Performance:

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

Erlang-C

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

ASA

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

TSF

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

Occupancy

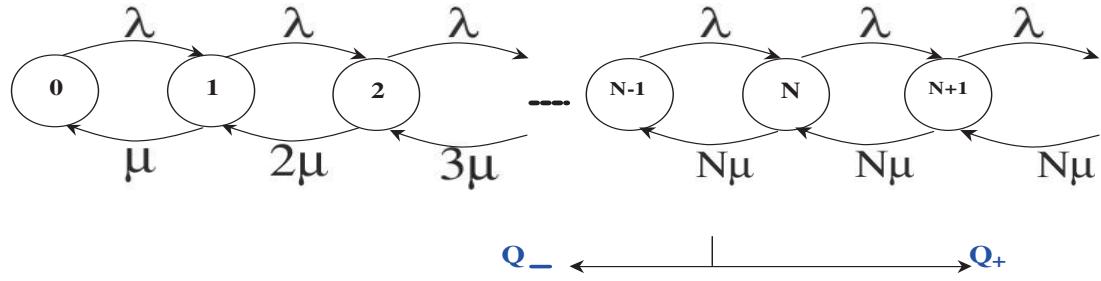
## QED : Some Intuition (Assume $\mu = 1$ )

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

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

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

## M/M/N (Erlang-C) with Many Servers: $N \uparrow \infty$



$Q(0) = N$ : all servers busy, no queue.

Recall 
$$E_{2,N} = \left[ 1 + \frac{T_{N-1,N}}{T_{N,N-1}} \right]^{-1} = \left[ 1 + \frac{1 - \rho_N}{\rho_N E_{1,N-1}} \right]^{-1}.$$

Here 
$$T_{N-1,N} = \frac{1}{\lambda_N E_{1,N-1}} \sim \frac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \frac{1/\mu}{h(-\beta)\sqrt{N}}$$

which applies as  $\sqrt{N}(1 - \rho_N) \rightarrow \beta$ ,  $-\infty < \beta < \infty$ .

Also 
$$T_{N,N-1} = \frac{1}{N\mu(1 - \rho_N)} \sim \frac{1/\mu}{\beta\sqrt{N}}$$

which applies as above, but for  $0 < \beta < \infty$ .

Hence, 
$$E_{2,N} \sim \left[ 1 + \frac{\beta}{h(-\beta)} \right]^{-1}$$
, assuming  $\beta > 0$ .

# Rules of Thumb: Operational Regimes

$$R = \lambda \times E(S) \quad \text{units of work per unit of time (load)}$$

**Efficiency-driven**  $(\% \{ \text{Wait} > 0 \} \rightarrow 100\%)$

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

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

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

**QED Regime**  $(\% \{ \text{Wait} > 0 \} \rightarrow \alpha, \ 0 < \alpha < 1)$

$$N = \lceil R + \beta \sqrt{R} \rceil, \quad \beta > 0 \quad \sqrt{\cdot} \text{ Safety-Staffing}$$

Determine Regimes (Strategy), Parameters (Economics)

**Strategy:** Managers, Agents (Unions), Customers

**Economics:** Minimize agent salaries + waiting cost

## Strategy: Sustain Regime under Pooling

**Base:**  $\lambda = 300/\text{hr}$ ,  $\text{AHT} = 5 \text{ min}$ ,  $N = 30$  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:**  $\lambda = 1200$ ,  $\text{AHT} = 5$ ,  $R = 100$ ;  $N = ?$

**Quality-Driven:** maintain OCC at 83.3%.

$$N = 120, \quad \text{ASA} = .5 \text{ sec}, \quad y = (120 - 100)/10 = 2$$

**Efficiency-Driven:** maintain ASA at 15 sec.

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

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

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

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

$$Q: \quad N = 270$$

$$E: \quad N = 233$$

$$\text{QED: } N = 225 + 1 \cdot \sqrt{225} = 240, \quad \text{OCC} = 94\%, \quad \text{ASA} = 4.7 \text{ sec}$$

## Strategy: Sustain Regime under Pooling

### Economies of Scale

Base case: M/M/N with parameters  $\lambda, \mu, N$

Scenario:  $\lambda \rightarrow m\lambda$  ( $R \rightarrow mR$ )

	Base Case	Efficiency-driven	Quality-driven	Rationalized
Offered load	$R = \frac{\lambda}{\mu}$	$mR$	$mR$	$mR$
Safety staffing	$\Delta$	$\Delta$	$m\Delta$	$\sqrt{m}\Delta$
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 = $P\{\text{Wait}>0\}$	$P(\beta)$	$P\left(\frac{\beta}{\sqrt{m}}\right) \uparrow 1$	$P(\beta\sqrt{m}) \downarrow 0$	$\boxed{P(\beta)}$
Occupancy	$\rho = \frac{R}{R + \Delta}$	$\frac{R}{R + \frac{\Delta}{m}} \uparrow 1$	$\boxed{\rho = \frac{R}{R + \Delta}}$	$\frac{R}{R + \frac{\Delta}{\sqrt{m}}} \uparrow 1$
ASA = $E\left[\frac{\text{Wait}}{E(S)} \mid \text{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \text{ASA}}$	$\frac{1}{m\Delta} = \frac{\text{ASA}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\text{ASA}}{\sqrt{m}}$
TSF = $P\left\{\frac{\text{Wait}}{E(S)} > T \mid \text{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \text{TSF}}$	$e^{-mT\Delta} = (\text{TSF})^m$	$e^{-\sqrt{m}T\Delta} = (\text{TSF})^{\sqrt{m}}$

See: Whitt's "How multi-server queues scale with ...demand",

# Economics: Quality vs. Efficiency

(Dimensioning: with S. Borst and M. Reiman)

Quality  $D(t)$  delay cost ( $t$  = delay time)

Efficiency  $C(N)$  staffing cost ( $N$  = # agents)

**Optimization:**  $N^*$  minimizes Total Costs

- $C \gg D$  : Efficiency-driven
- $C \ll D$  : Quality-driven
- $C \approx D$  : Rationalized - QED

**Satisfization:**  $N^*$  minimal s.t. Service Constraint

Eg. %Delayed <  $\alpha$ .

- $\alpha \approx 1$  : Efficiency-driven
- $\alpha \approx 0$  : Quality-driven
- $0 < \alpha < 1$  : Rationalized - QED

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

## Economics: √· Safety-Staffing

Optimal

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

where

**d** = delay/waiting costs

**c** = staffing costs

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

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

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

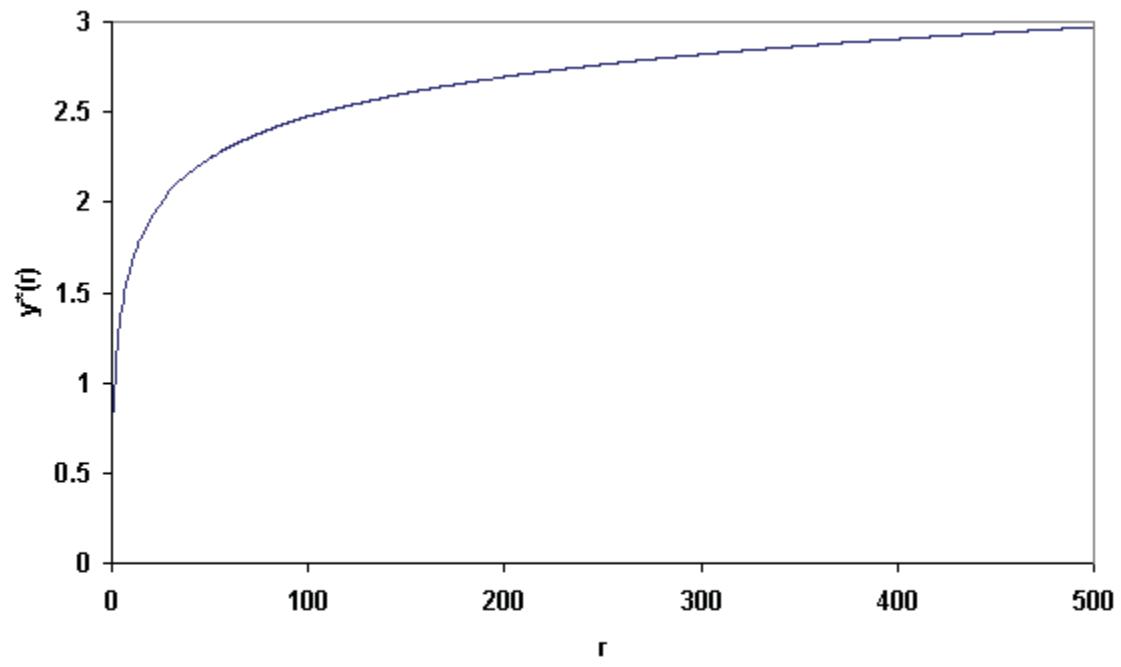
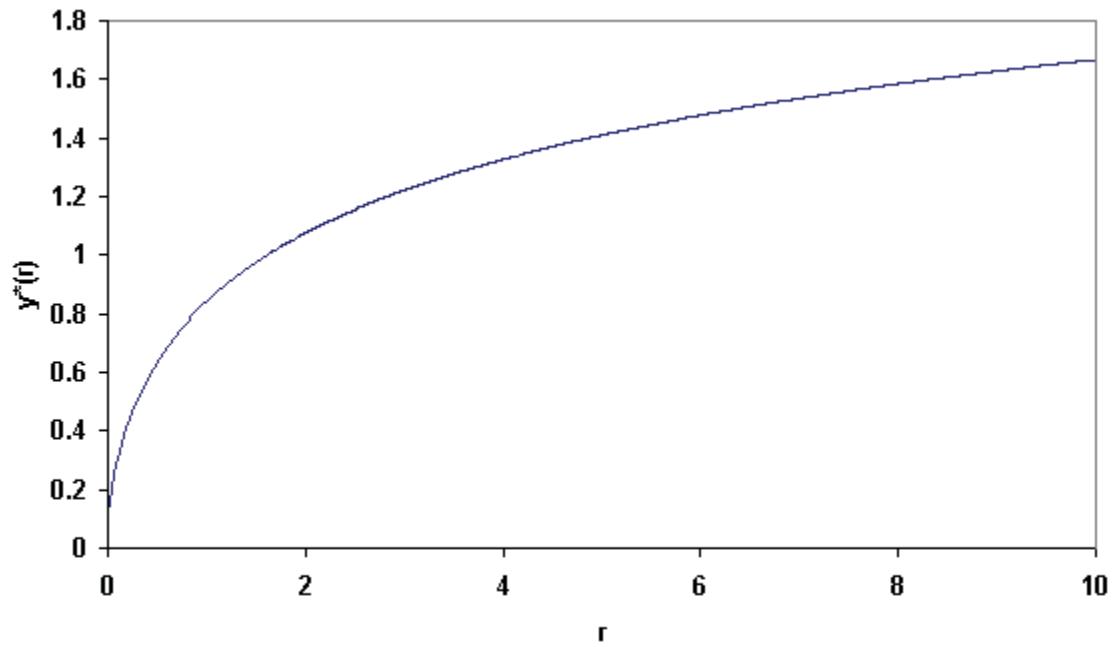
$$P\{\text{Wait} > 0\} \approx P(y^*) = \left[ 1 + \frac{y^* \phi(y^*)}{\varphi(y^*)} \right]^{-1} \quad \text{Erlang-C}$$

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

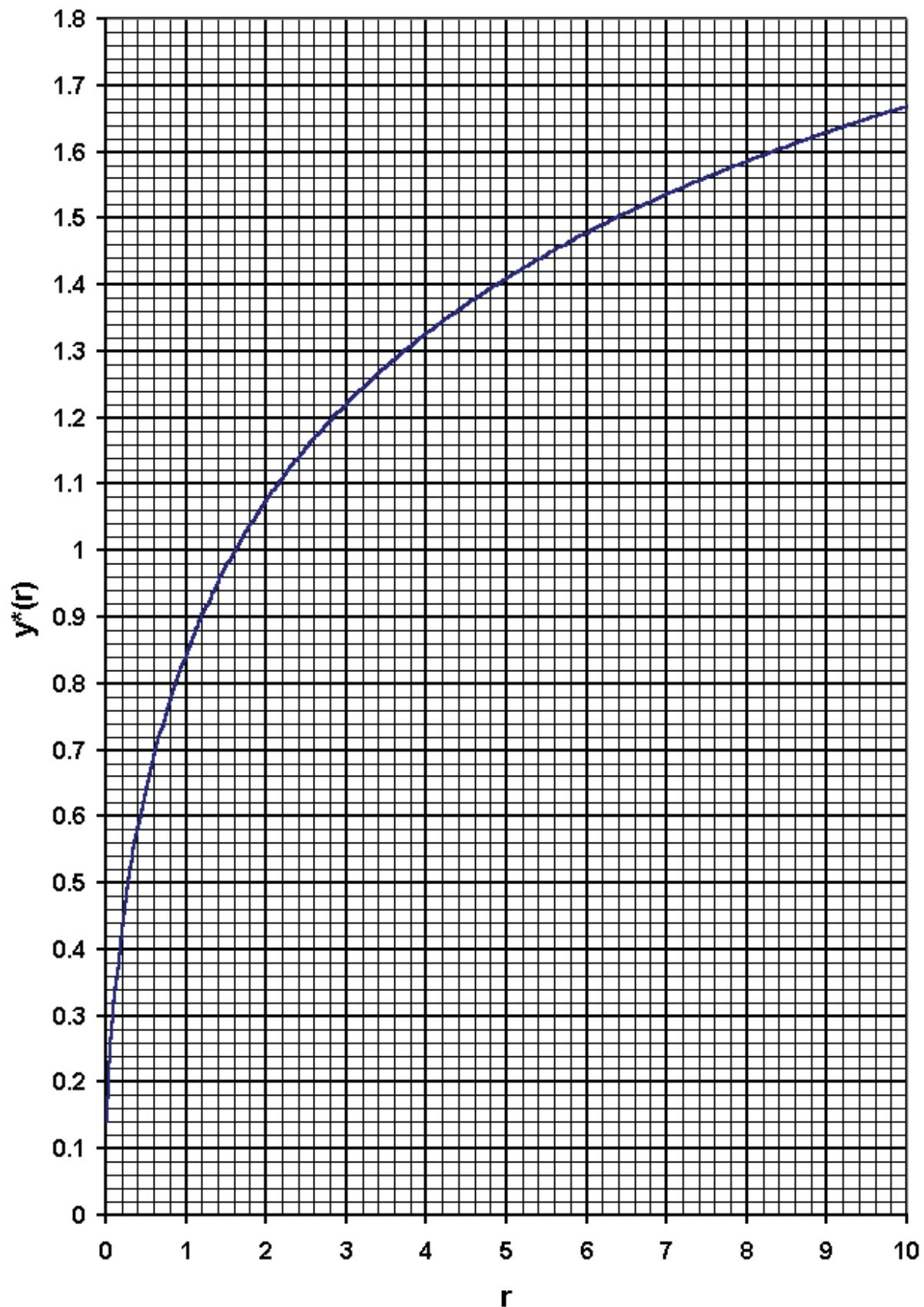
$$\text{ASA} = E\left[ \frac{\text{Wait}}{E(S)} \mid \text{Wait} > 0 \right] = \frac{1}{\Delta}$$

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

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



$$y^*(r), \quad r = \text{cost of delay} / \text{cost of staffing}$$



# Rules-of-Thumb in an "Erlang-C World"

**R = Offered Load (not small)**

Efficiency-Driven:  $N = R + 2$  (or 3, or...);

Expect that essentially **all** customers are delayed in queue, that average delay is about 1/2 (or 1/3, or...) average service-time, and that agents utilization is extremely high (close to 100%).

Quality-Driven:  $N = R + (10\% - 20\%) R$

Expect essentially **no** delays of customers.

**QED:**  $N = R + 0.5\sqrt{R}$

Expect that about **half** of the customers are not delayed in queue, that average delay is about one-order less than average service-time (seconds vs. minutes), and that agents utilization is high (90-95%).

Can determine regime scientifically:

**Strategy:** Retain performance levels under Pooling (4CC demo)

**Economics:** Minimize agent salaries + congestion cost, or

**Satisfization:** Least Number of Agents s.t. Constraints

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

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

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

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

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

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 (idleness) = 7% (vs. 3% before)

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

## Scenario Analysis: “Reasonable” Service Level ?

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).

Example:  $\lambda = 1,800$  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: 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)

Note: **Satisfization easier to model but Costs easier to grasp.**