

**QED Q's**

**Telephone Call/Contact Centers**

**Service Engineering**

**Queueing Science**

**SNC/CRM**

**July 23, 2004**

**e.mail : [avim@tx.technion.ac.il](mailto:avim@tx.technion.ac.il)**

**Website: <http://ie.technion.ac.il/serveng2004>**

# Contents

1. **Motivation**: "The Right Answer for the Wrong Reason"
2. **Operational Regime (M/M/N)**:
  - **Quality-Driven**
  - **Efficiency-Driven**
  - **The QED (Halfin-Whitt) Regime**
3. Some Intuition

Example from a call center, leading to models with

4. **Impatient** (Abandoning) Customers (M/M/N+G)
5. **Time-Varying** Queues with **Time-Stable** Performance
6. **General Service** Times (G/M/N, G/D/N; G/LN/N)
7. **Heterogeneous** Customers and **Multi-skilled** Agents (SBR)
8. **Forecasting** Parameters

# Supporting Material (Downloadable)

## Background

M. "Call Centers: Research [Bibliography](#) with Abstracts."  
Version 5, July, 2003.

Gans ([U.S.A.](#)), Koole ([Europe](#)), and M. ([Israel](#)):  
"Telephone Call Centers: Tutorial, Review and Research  
Prospects." *MSOM*, 2003.

Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical  
Analysis of a Telephone Call Center: A Queueing-Science  
Perspective." Submitted, 2003.

## M/M/N (Erlang-C); M/G/N

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.

Schwartz: "Simulation Experiments with [M/G/100](#) Queues in the  
Halfin-Whitt (QED) Regime." Project, Technion, July 2002.

Borst, M. and Reiman: "[Dimensioning](#) Large Telephone Call  
Centers." *OR*, 2004.

**M/M/N+M (ErlangA); M/M/N+G; G/G/N+G**

**Garnett, M. and Reiman:** "Designing a Call Center with **Impatient Customers.**" *MSOM*, 2002.

**Zeltyn: Ph.D.** Thesis on the **M/M/N+G Queue (Q/E/QED Asymptotics, Dimensioning, Inference).**

M. and Zeltyn: "The Impact of Customer Patience on Delay and Abandonment: Some **Empirically-Driven Experiments** with the M/M/N+G Queue." *OR Spectrum*, 2004.

M. and Zeltyn: "The Palm/Erlang-A Queue, with Applications to Call Centers." **Teaching Note**, Service Engineering, 2004.

M. and Zeltyn: "The M/M/n+G Queue: **Summary** of Performance Measures." Prepared for Bank of America (BoFA), May 10, 2004.

**In this conference:**

- Whitt: Fluid models; Approximations (of G/G/N+G);  
Uncertainty in arrival rates and staffing levels.

**Time-Dependence (Predictable Variability); Service Networks**

**Jennings, M., Massey and Whitt:** "Server Staffing to Meet Time-Varying Demand." *Management Science*, 1996.

**M., Massey and Reiman:** "Strong Approximations for Markovian Service Networks." *QUESTA*, 1998.

**Feldman:** "Staffing of **Time-Varying** Queues to Achieve **Time-Stable Performance**." Project, Technion, June 2004.

**In this conference:**

- Helber: Profit Maximization.
- Henken; Retrials (**M-Design**).
- Massey: QED Service Networks.

**SBR**

M. and Stolyar: "Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized  $c\mu$ -Rule." OR, 2004. (**ED control** under CRP; No Abandons)

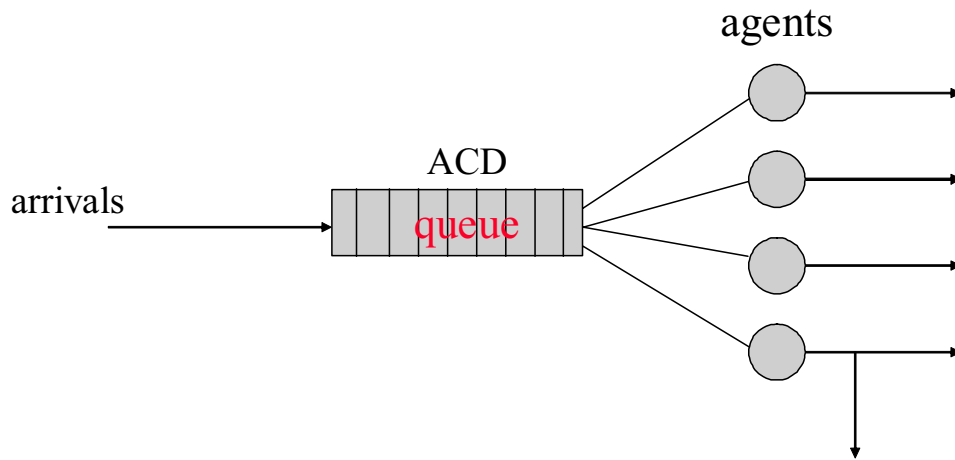
**Gurvich:** "Staffing and Control of the M/M/N Queue with Multi-Class Customers and **Many** Servers." M.Sc. Thesis (**V-Design**)

**Yahalom** and M.: "Optimal Control of Queueing Systems with Multi-Class Customers and Multiple Servers: V- and N-Design." In Preparation.

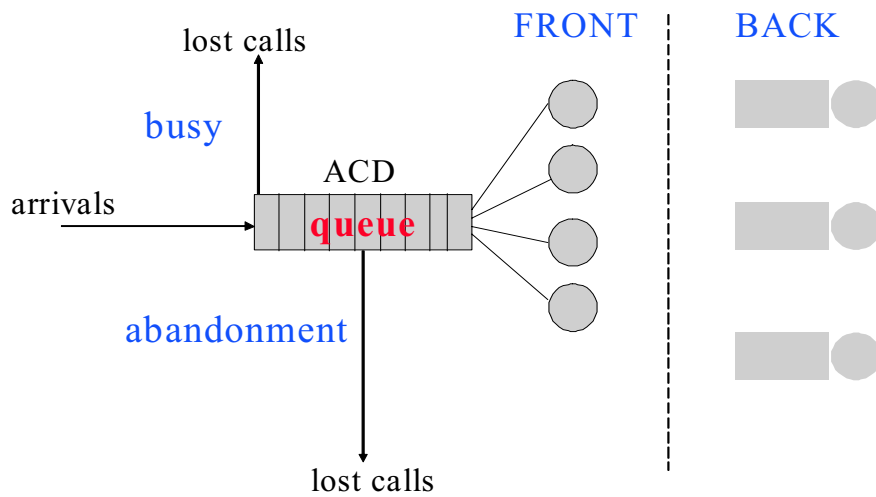
**In this conference:**

- Armony: Control and Staffing (**Reversed-V**).
- Atar, Shaikhet: QED control, with non-basic activities.
- Harrison and Zeevi: QED control under CRP with linear costs; **"ED" staffing under uncertainty**.
- Koole: Approximations, with overflow routing.

# Erlang-C = M/M/N



# Erlang-A <4CallCenters>

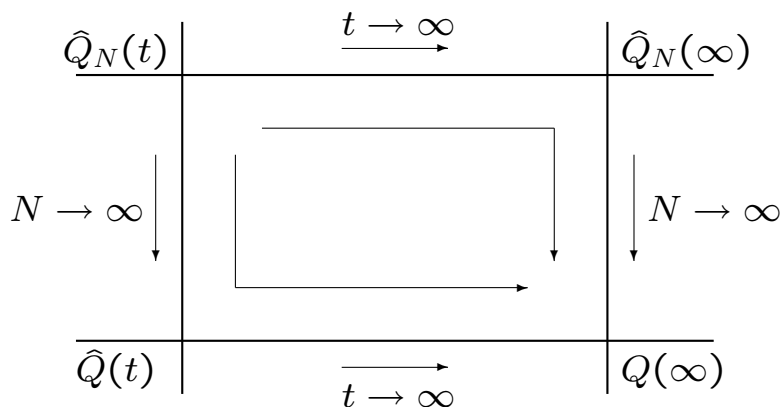


# Approximating Queueing and Waiting

- $Q_N = \{Q_N(t), t \geq 0\} : Q_N(t) =$  **number in system** at  $t \geq 0$ .
- $\hat{Q}_N = \{\hat{Q}_N(t), t \geq 0\} :$  **stochastic process** obtained by centering and rescaling:

$$\hat{Q}_N = \frac{Q_N - N}{\sqrt{N}}$$

- $\hat{Q}_N(\infty) :$  stationary distribution of  $\hat{Q}_N$
- $\hat{Q} = \{\hat{Q}(t), t \geq 0\} :$  process defined by:  $\hat{Q}_N(t) \xrightarrow{d} \hat{Q}(t)$ .



## Approximating (Virtual) **Waiting Time**

$$\hat{V}_N = \sqrt{N} V_N \Rightarrow \hat{V} = \left[ \frac{1}{\mu} \hat{Q} \right]^+ \quad (\text{Puhalskii, 1994})$$

# Staffing Time-Varying Queues:

Two Common Approaches:

**SSA** – Simple Stationary Approximation.

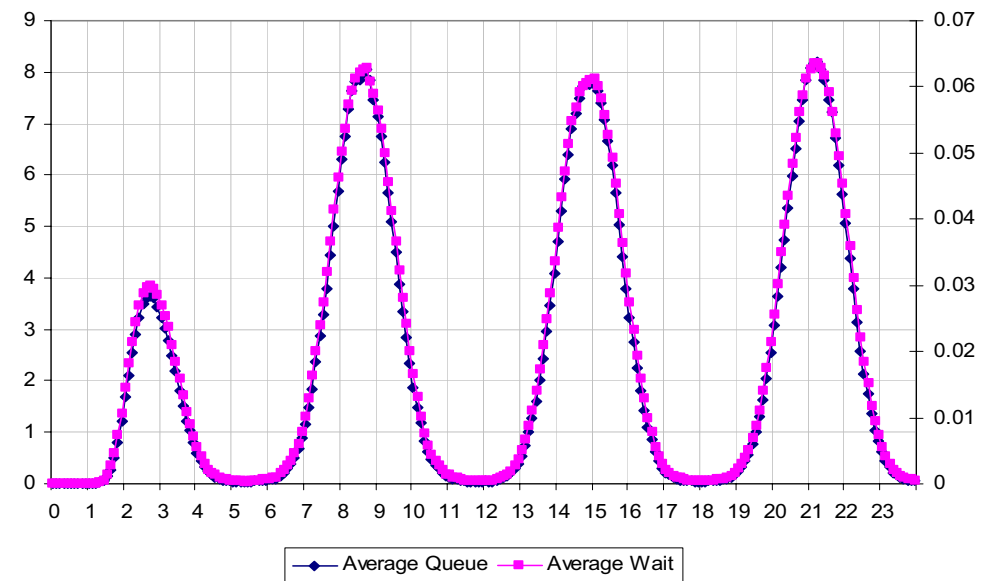
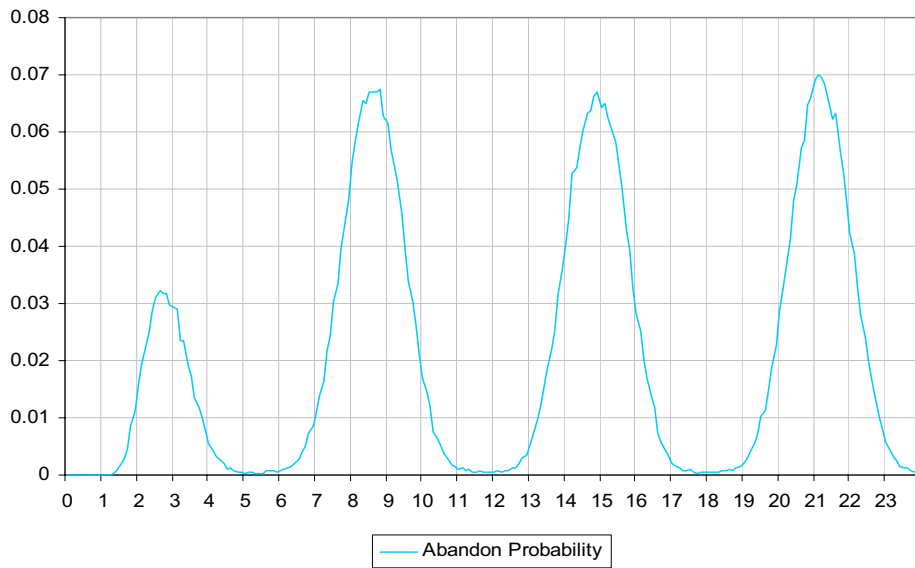
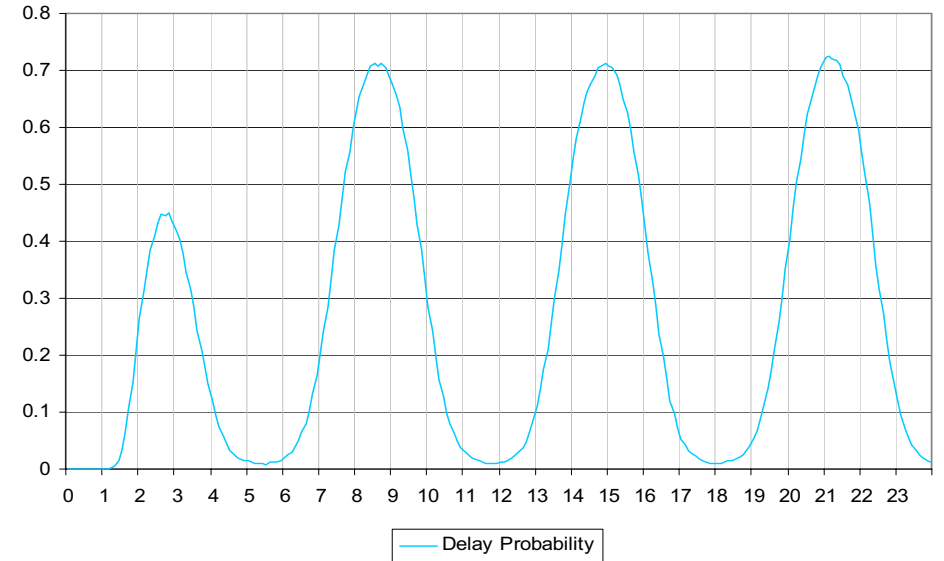
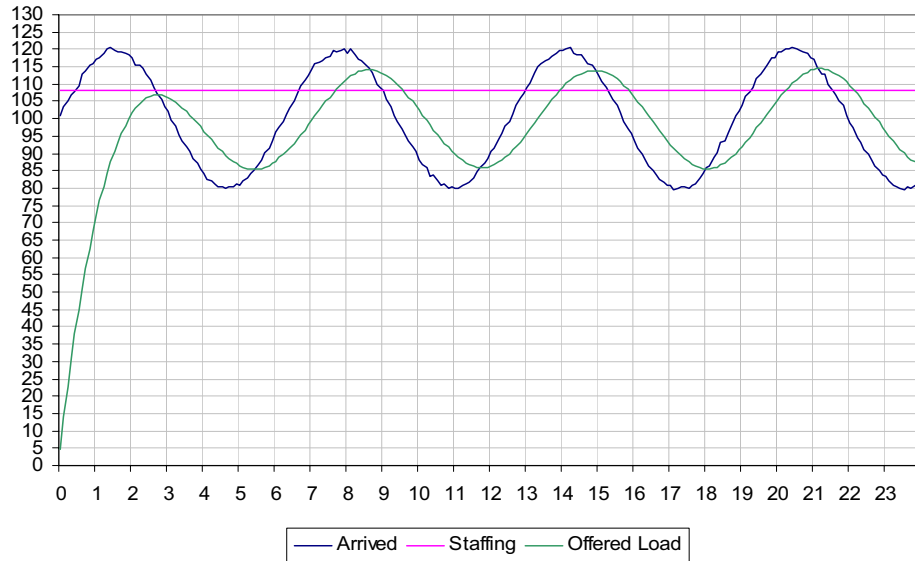
**Constant** staffing levels, based on steady-state M/M/N, with  $\lambda$ =long-run average number of arrivals.

**PSA** – Point-wise Stationary Approximation.

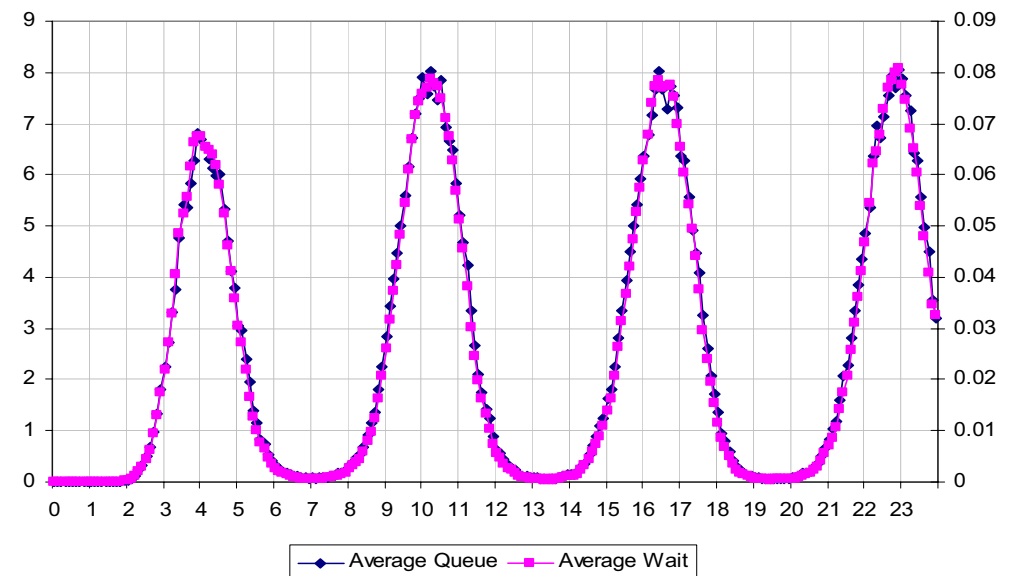
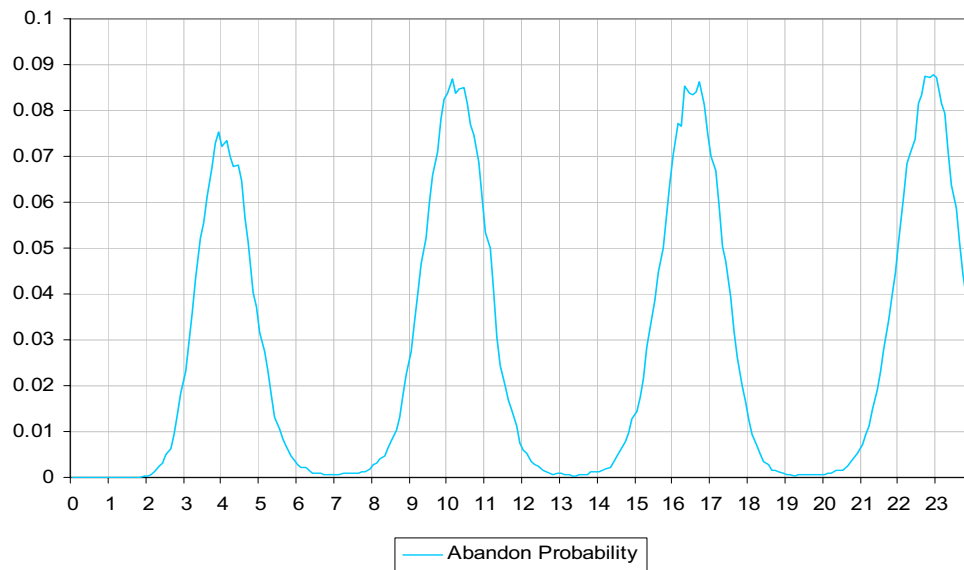
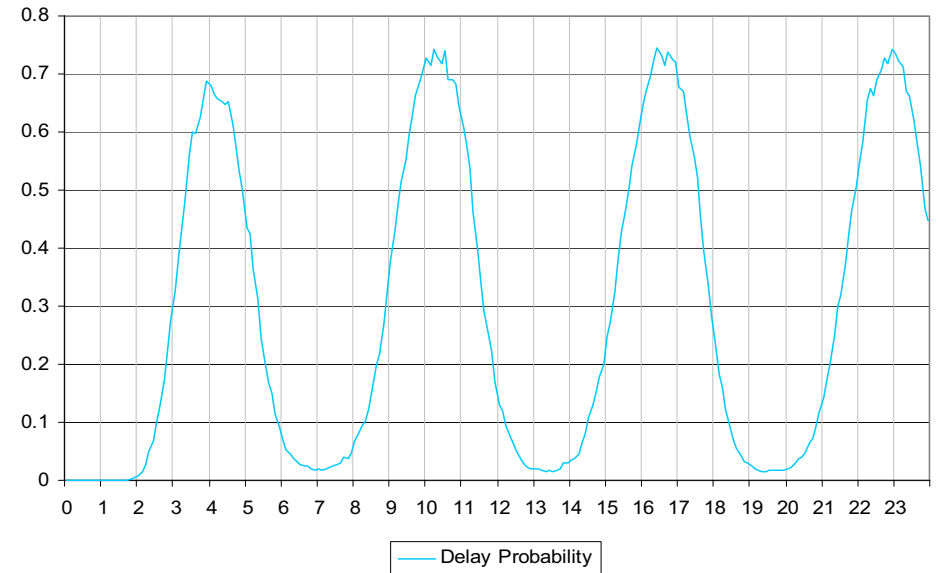
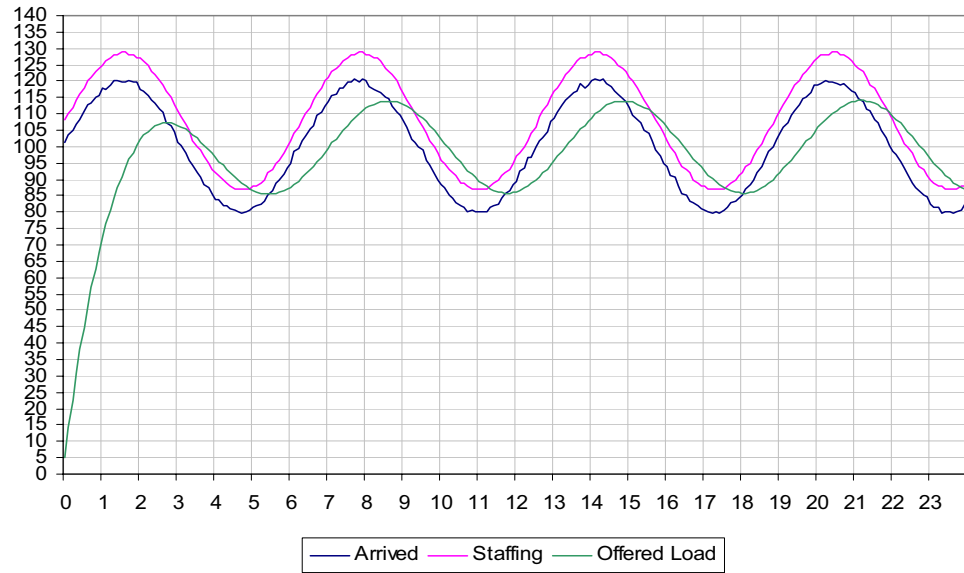
Time-varying staffing levels, based on **steady-state** M/M/N, with  $\lambda = \lambda(t)$  at each time  $t$ .

Could result in time-varying (**highly oscillating**) performance (utilization, service), which is undesirable.

# Simple Stationary Approximation (SSA, $\alpha=0.2$ )



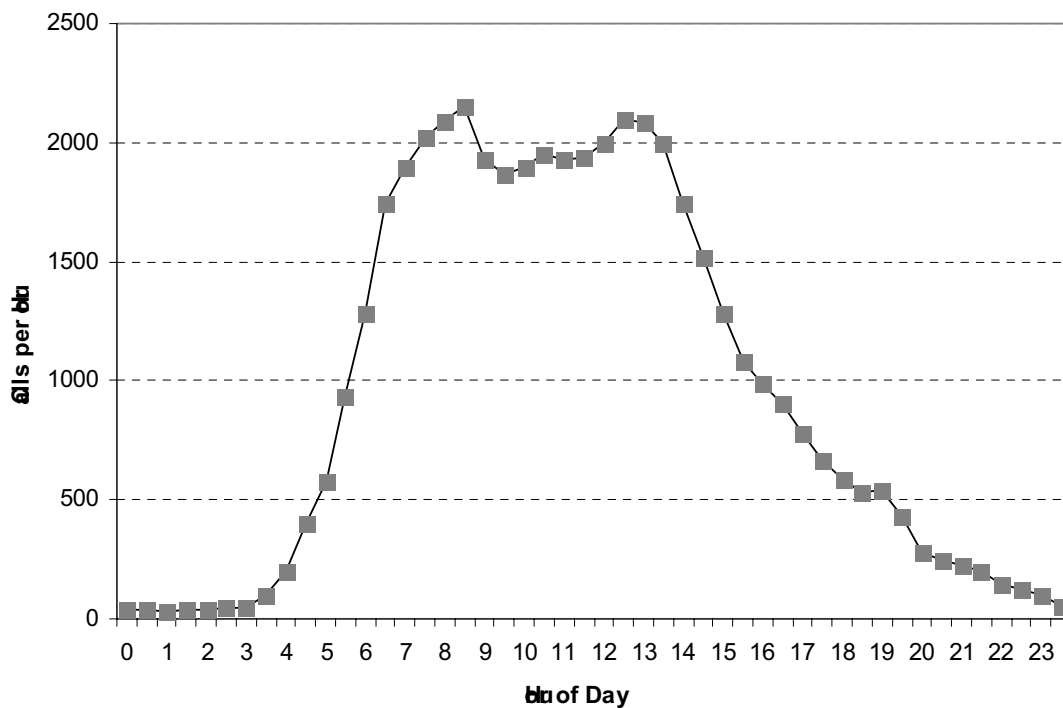
# Point-wise Stationary Approximation (PSA, $\alpha=0.2$ )



# Example: "Real" Call Center

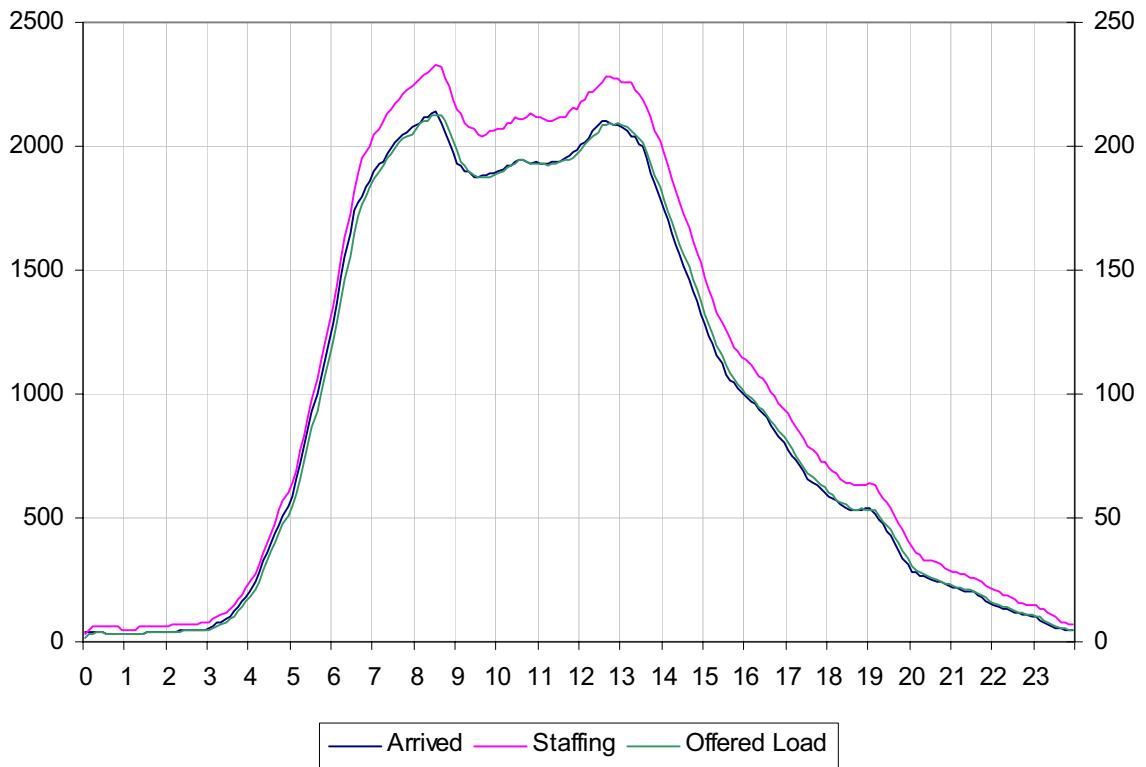
Two-hump arrival functions are common

(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)

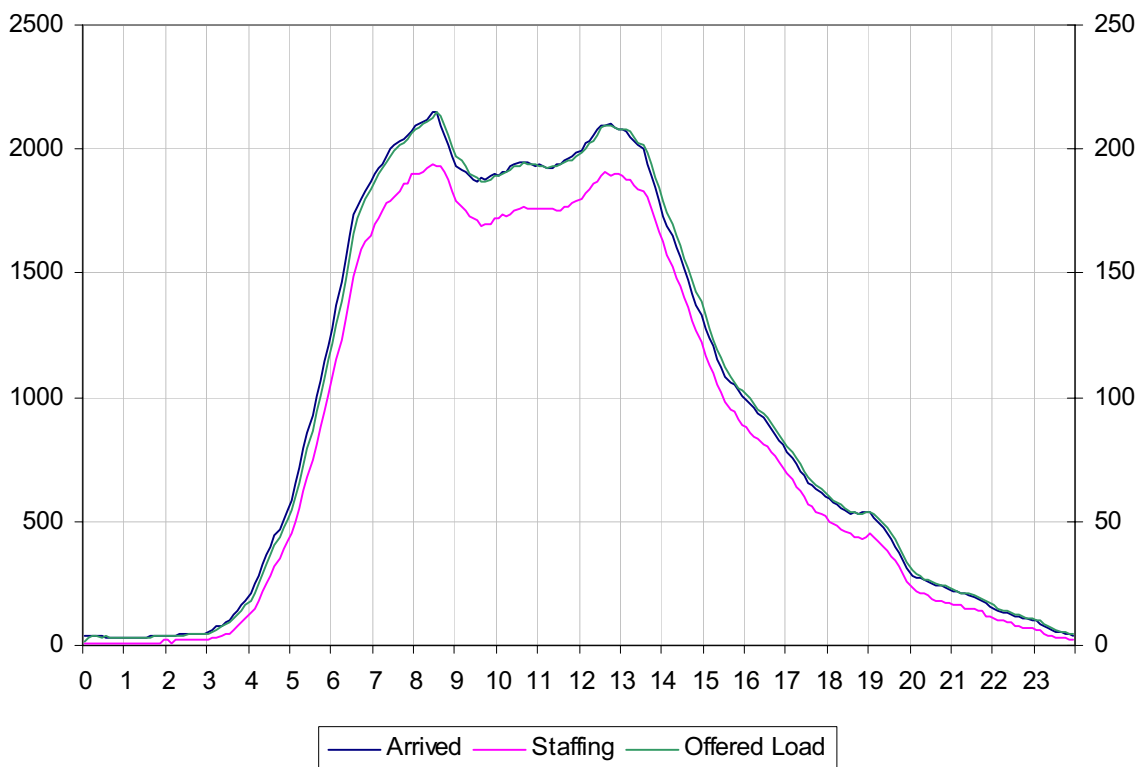


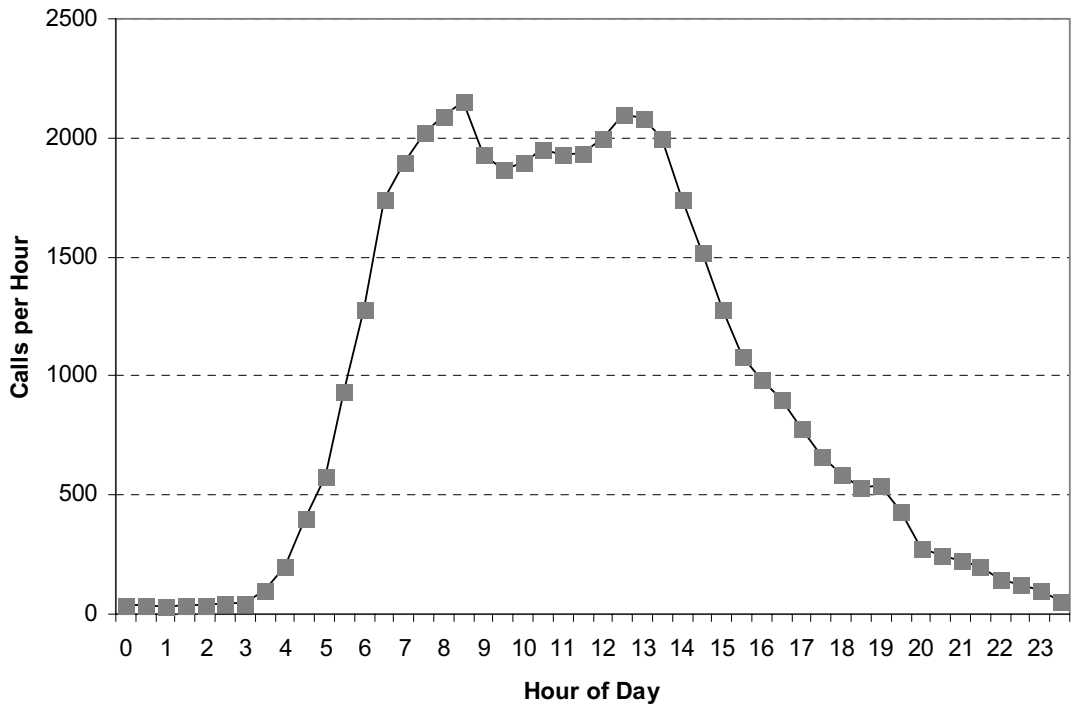
Assume: Service and abandonment rates are **both** exponential having **mean 0.1** (6 min.)

## QD Staffing ( $\alpha=0.1$ )

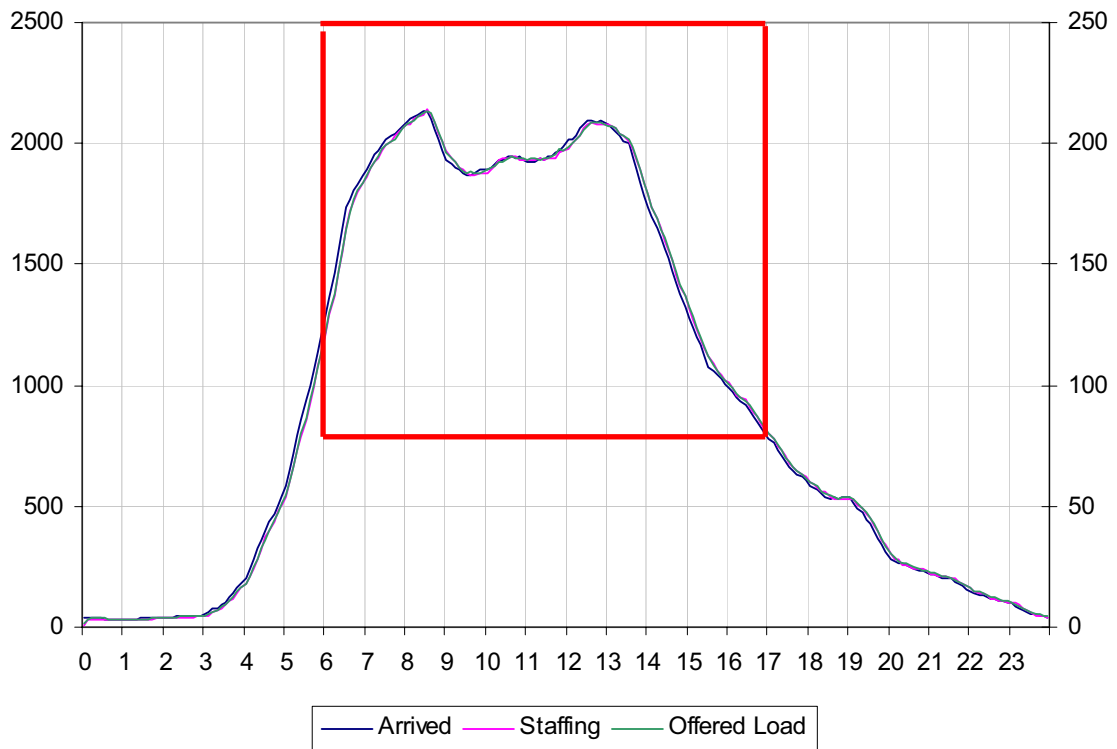


## ED Staffing ( $\alpha=0.9$ )





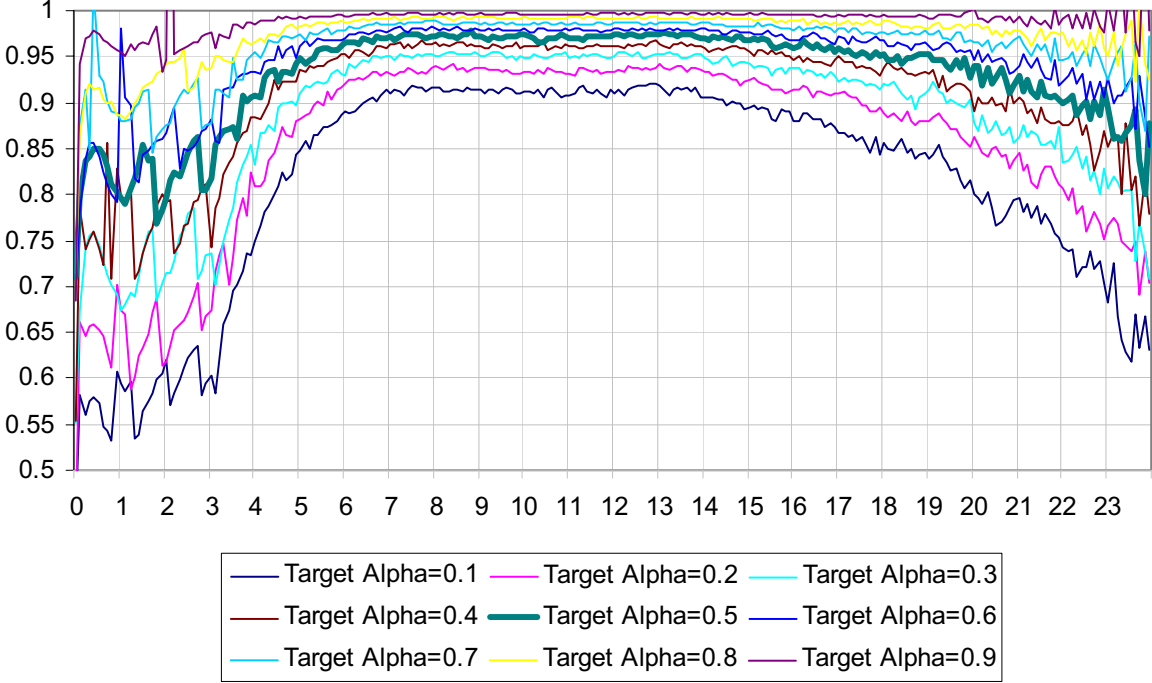
## QED Staffing ( $\alpha=0.5$ )



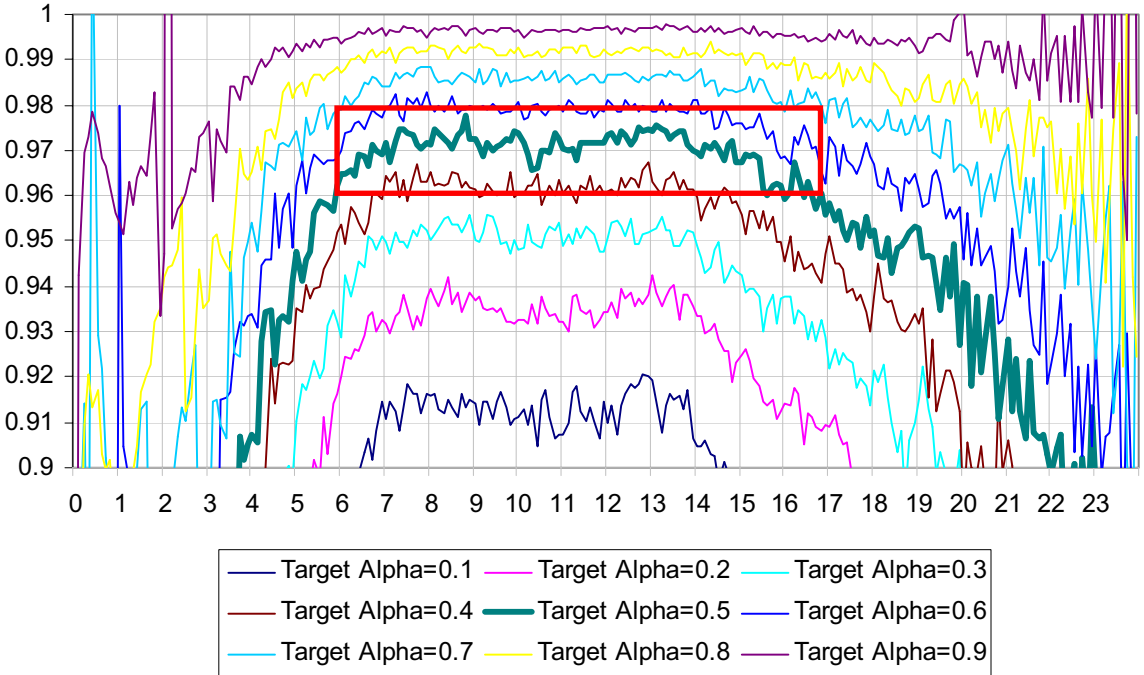


# Utilization

Utilization



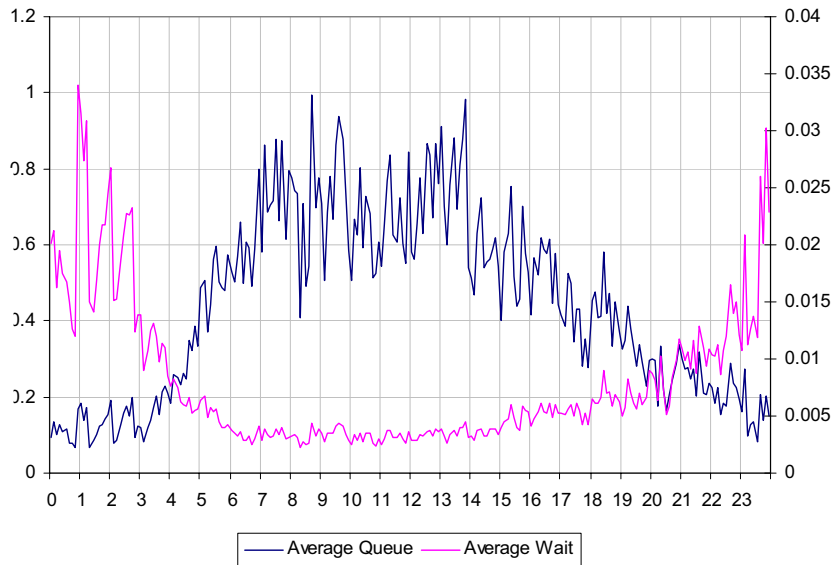
Utilization



# Congestion (Queue, Wait)

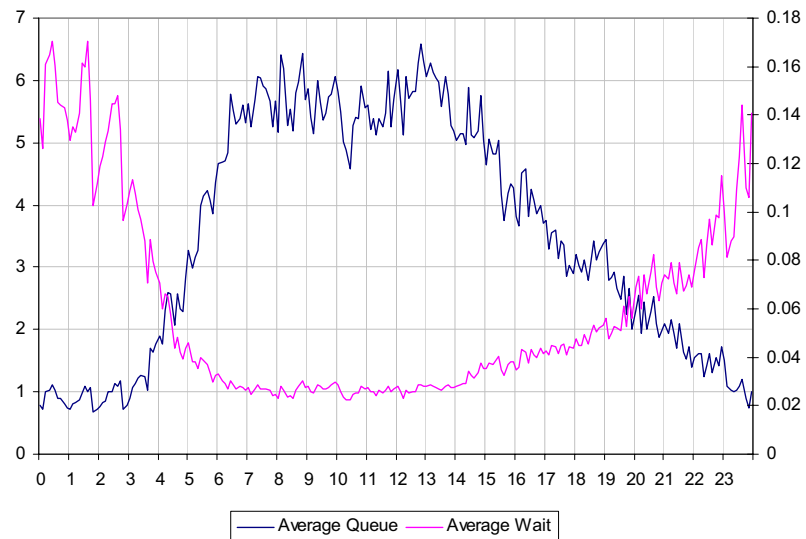
QD  
 $\alpha=0.1$

Negligible



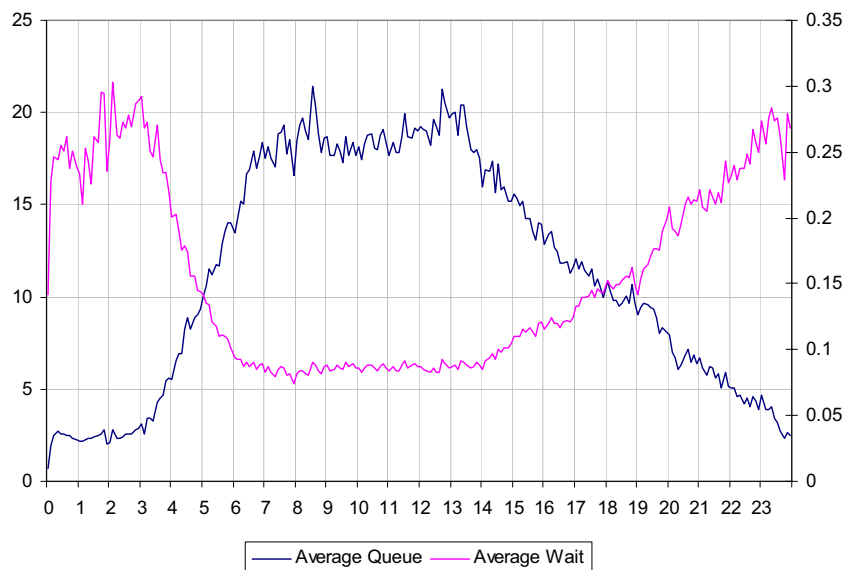
QED  
 $\alpha=0.5$

Seconds



ED  
 $\alpha=0.9$

Minutes

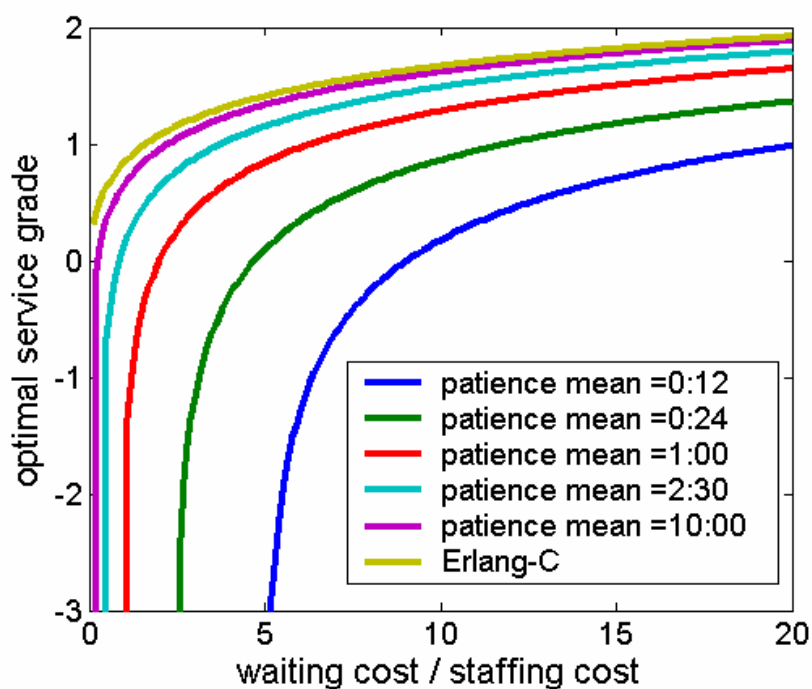


## When QED? Dimensioning

For example, via  $r$  = value of customer-time / agent-salary

Moderately impatient: QED if  $r=2$

Highly impatient:  $r=10$



# What can be achieved

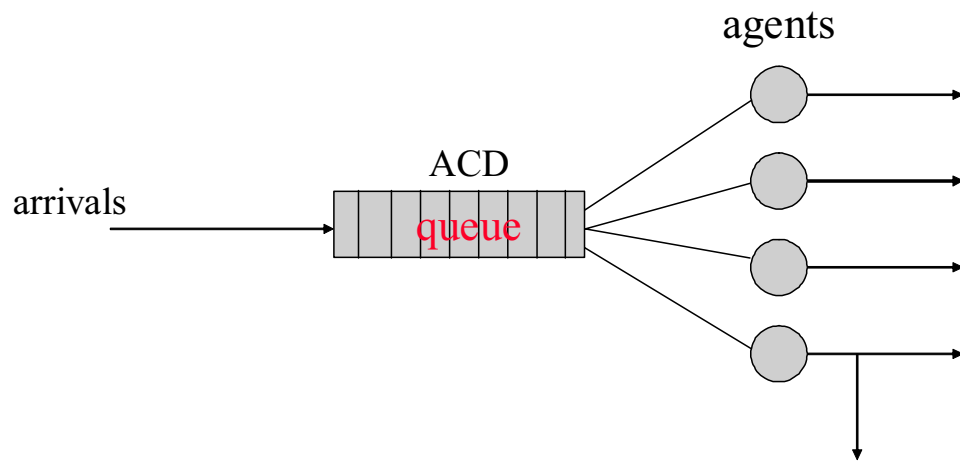
# At what cost

Copy of Summary Interval - Order PK

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

Time	Avg Ans	Speed W	Avg Aban Time	ACD A	Calls 1/M	Avg ACD Time	Avg ACW Time	Aban # Aban	% ACD Time	% Ans	Avg Pos	Calls Staff	Per Pos	% Serv Lev	% Aux Time	% ACW Time	% ACD Time	P
Totals		:00:02	:00:28	10456		:03:47	:00:25	46	53	98	70	149			8			
12:00 AM*		:00:00	:00:00	28		:04:31	:00:02	1	76	51	7	4		51	2	18		61
12:30 AM*		:00:03	:04:10	14		:07:27	:00:33	1	89	52	5	3		48	1	28		63
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		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	:00:08	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	:00:26	1	63	97	94	4		98	5	8		55
10:30 AM*		:00:00		349		:03:35	:00:33	0	62	99	96	4		99	6	8		44
11:00 AM*		:00:00		352		:03:50	:00:27	0	51	100	102	3		100	7	8		45
11:30 AM*		:00:00		348		:03:44	:00:18	0	49	100	97	4		100	8	5		45
12:00 PM*		:00:01		354		:03:59	:00:18	0	52	95	95	4		95	8	5		47
12:30 PM*		:00:00		336		:03:38	:00:21	0	52	99	97	3		99	9	8		46
1:00 PM*		:00:00		347		:03:55	:00:32	0	51	99	98	4		99	11	8		44
1:30 PM*		:00:00		366		:03:52	:00:14	0	56	98	99	4		99	11	7		50
2:00 PM*		:00:01		393		:03:55	:00:17	0	51	100	106	4		100	10	5		46
2:30 PM*		:00:00		403		:03:58	:00:13	0	54	100	112	4		100	10	4		50
3:00 PM*		:00:00	:00:04	410		:04:02	:00:16	1	57	98	110	4		98	8	5		51
3:30 PM*		:00:00		347		:03:59	:00:14	0	60	100	100	3		100	7	5		45
4:00 PM*		:00:00		382		:03:48	:01:37	0	64	100	98	4		100	8	7		47
4:30 PM*		:00:00		379		:03:41	:00:19	0	55	99	97	4		99	8	5		50
5:00 PM*		:00:00		411		:03:53	:00:19	0	53	100	109	4		100	9	5		48
5:30 PM*		:00:01		387		:03:58	:00:19	0	58	99	98	4		99	10	6		51
6:00 PM*		:00:01	:00:21	371		:03:28	:00:25	1	53	98	91	4		98	9	6		47
6:30 PM*		:00:00		280		:03:26	:00:13	0	41	100	90	3		100	8	4		37
7:00 PM*		:00:00		289		:03:24	:00:17	0	42	100	78	3		100	9	5		38

*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  $\rho = R/N$   
 $= 50/100 = 50\%$

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

$\Rightarrow$  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.**

$\lambda$ /hr	$\rho$	$E(W_q) = \text{ASA}$	% 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.**

$\lambda$ /hr	$\rho$	$E(W_q) = \text{ASA}$	% 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.**

$\lambda$ /hr	$\rho$	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

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

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

# Changing N (**Staffing**) in Erlang-C

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

$\lambda$ /hr	<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

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

$\lambda$ /hr	<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

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

$\lambda$ /hr	<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>

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

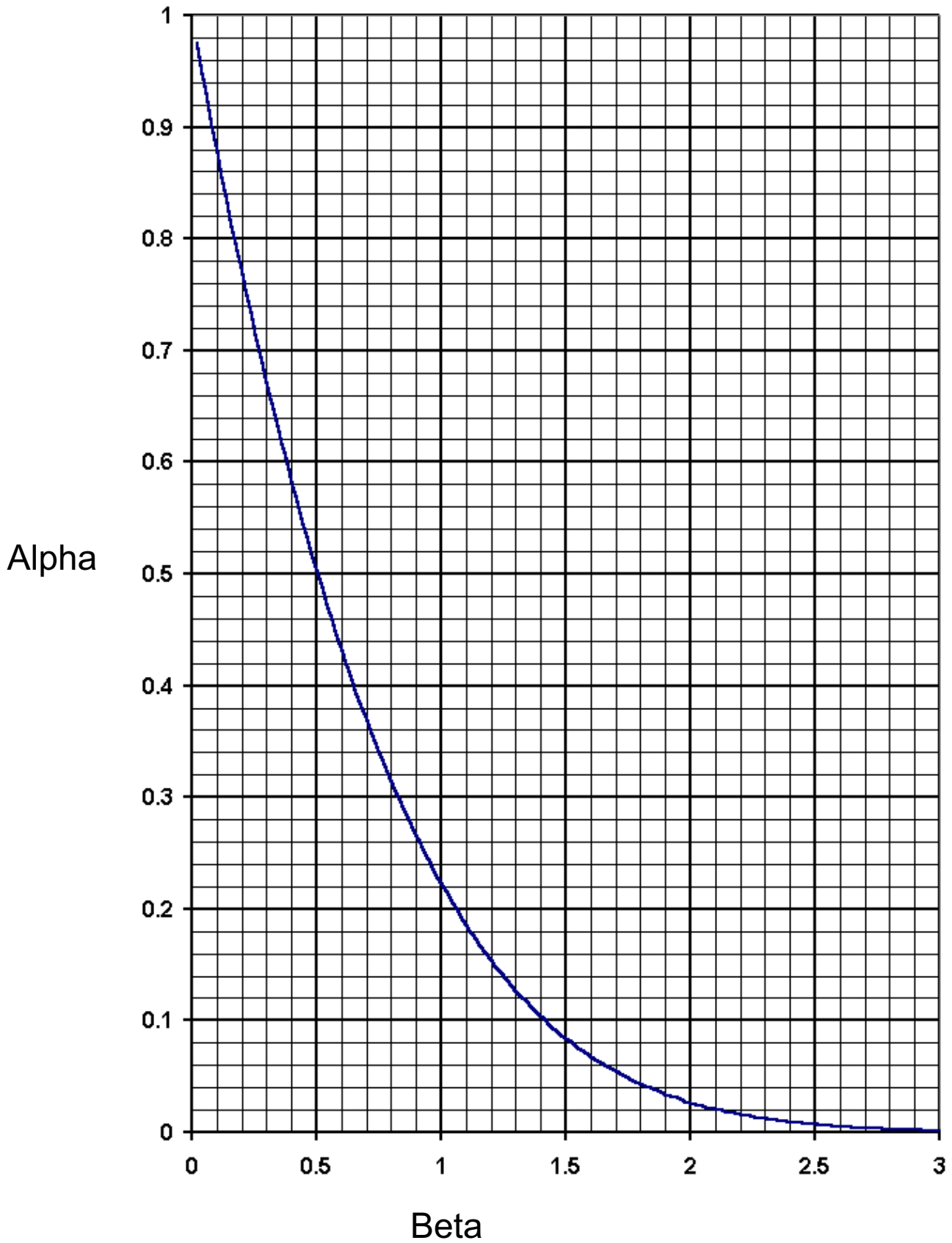
where  $\varphi(\cdot) / \phi(\cdot)$  is the standard normal density/distribution.

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



# 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,  $N \uparrow \infty$

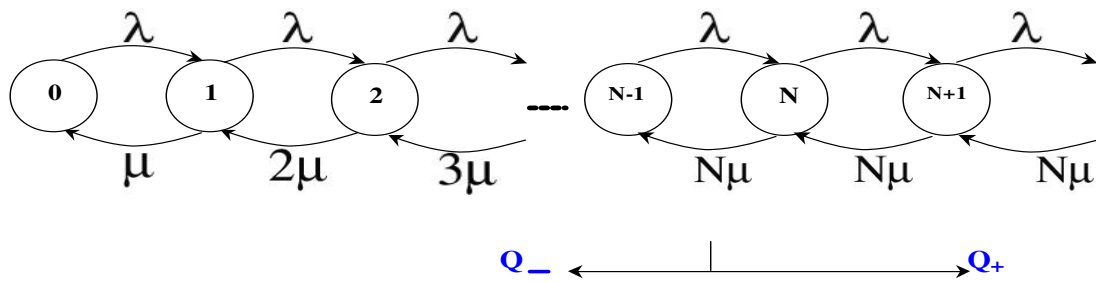
**QED : Some Intuition** (Assume  $\mu = 1$ )

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

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

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

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

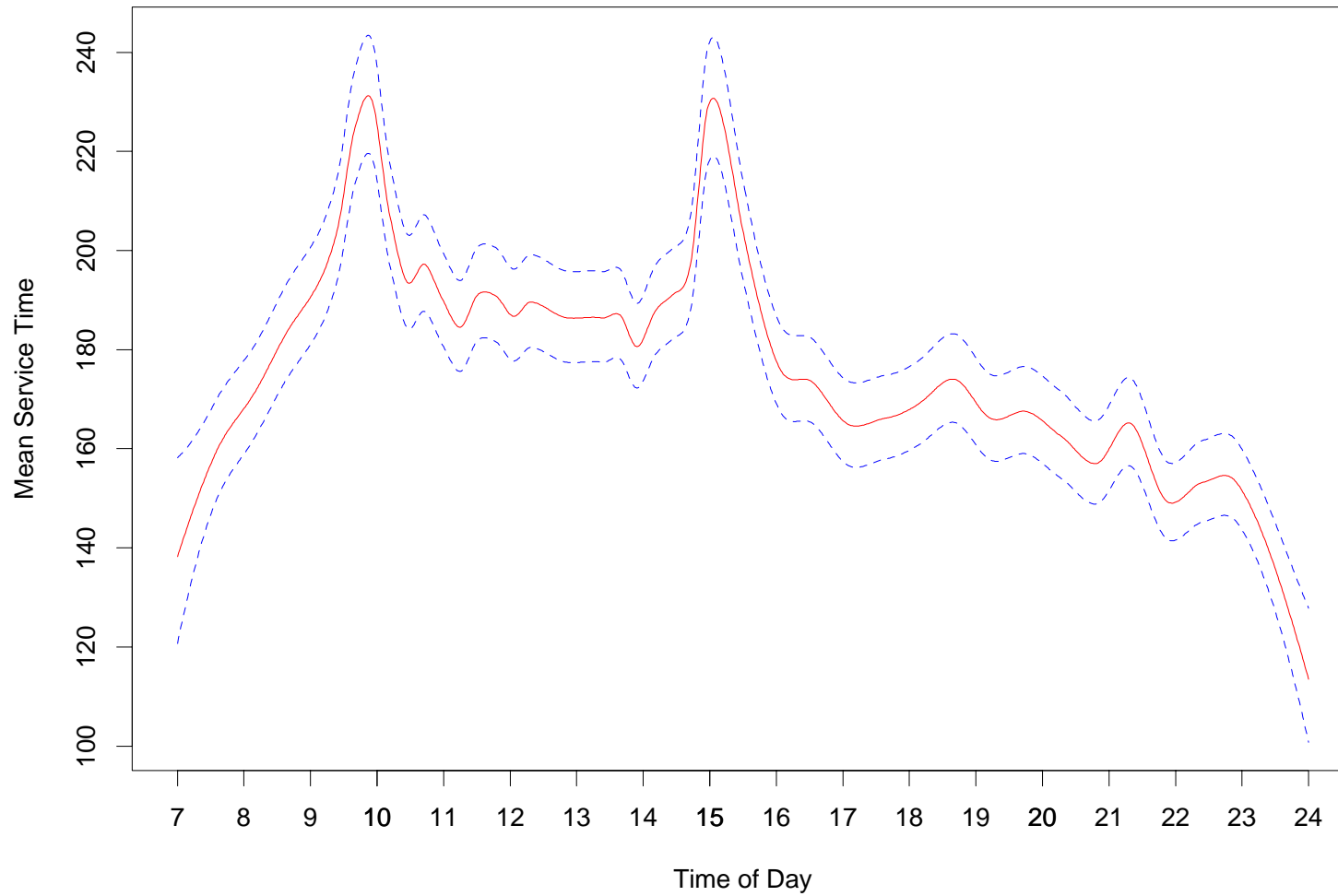
which applies as  $\sqrt{N}(1 - \rho_N) \rightarrow \beta, -\infty < \beta < \infty.$

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  $0 < \beta < \infty.$

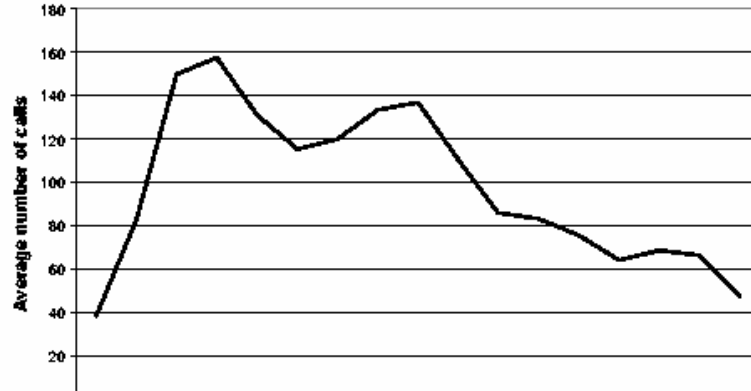
Hence, 
$$E_{2,N} \sim \left[ 1 + \frac{\beta}{h(-\beta)} \right]^{-1}, \text{ assuming } \beta > 0.$$

Figure 12: Mean Service Time (Regular) vs. Time-of-day (95% CI) ( $n = 42613$ )

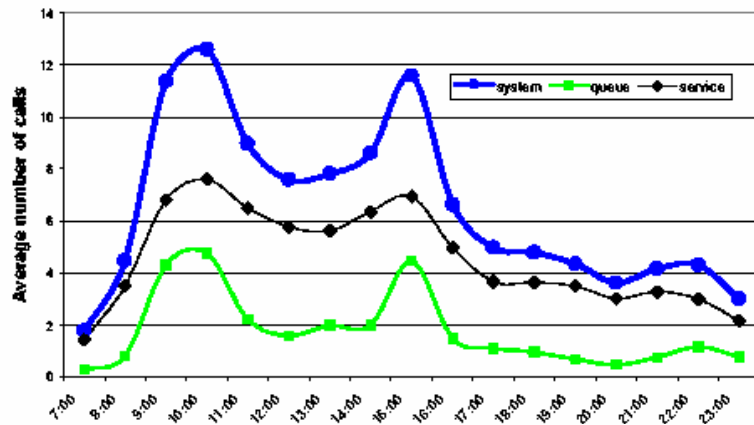


# Predictable Variability

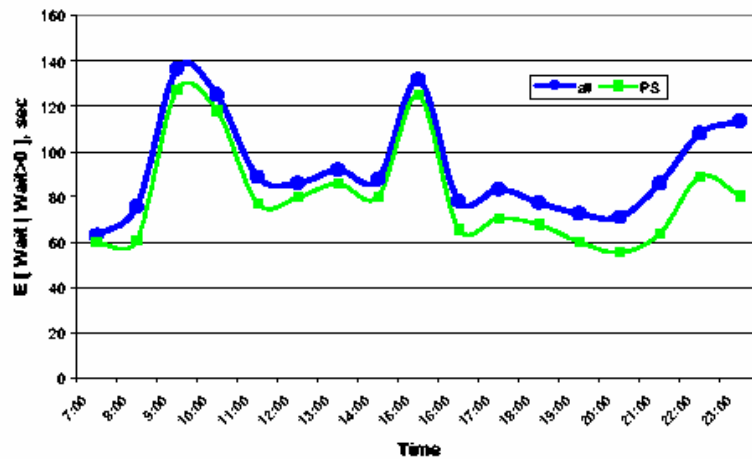
## Arrivals



## Queues

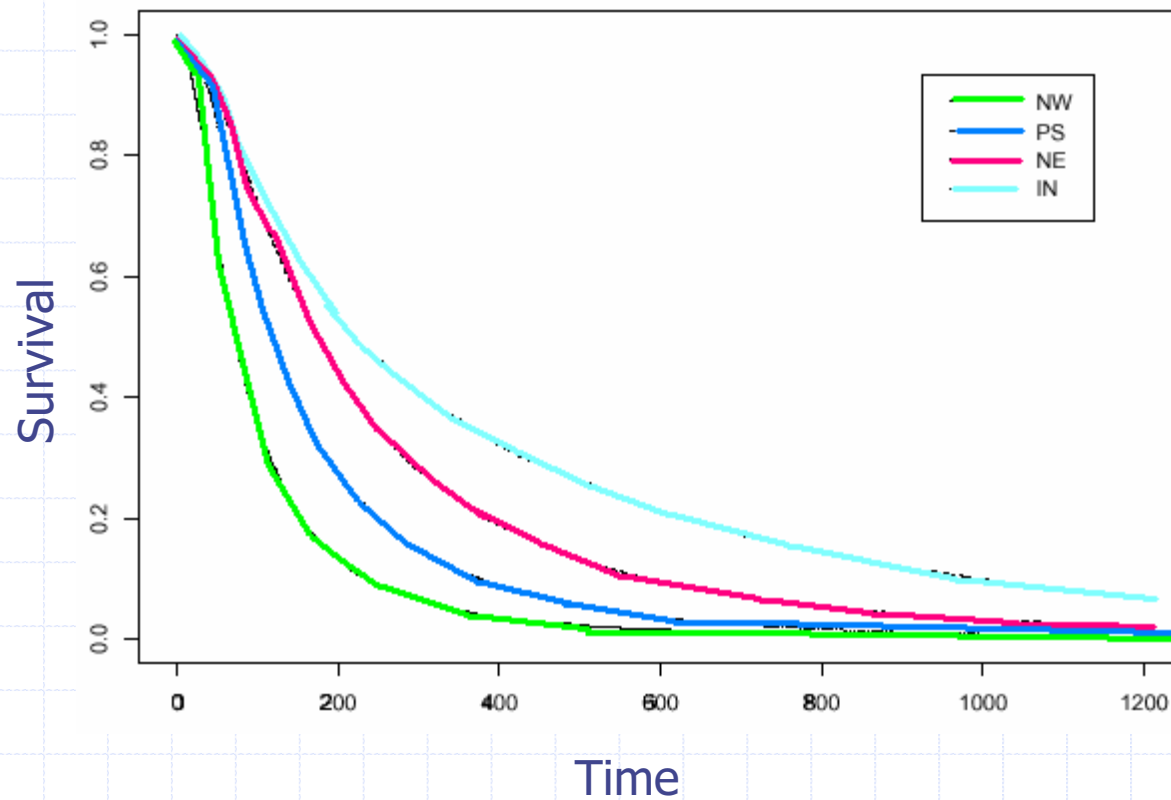


## Waiting



# Service Time

Survival curve, by Types



## Means (In Seconds)

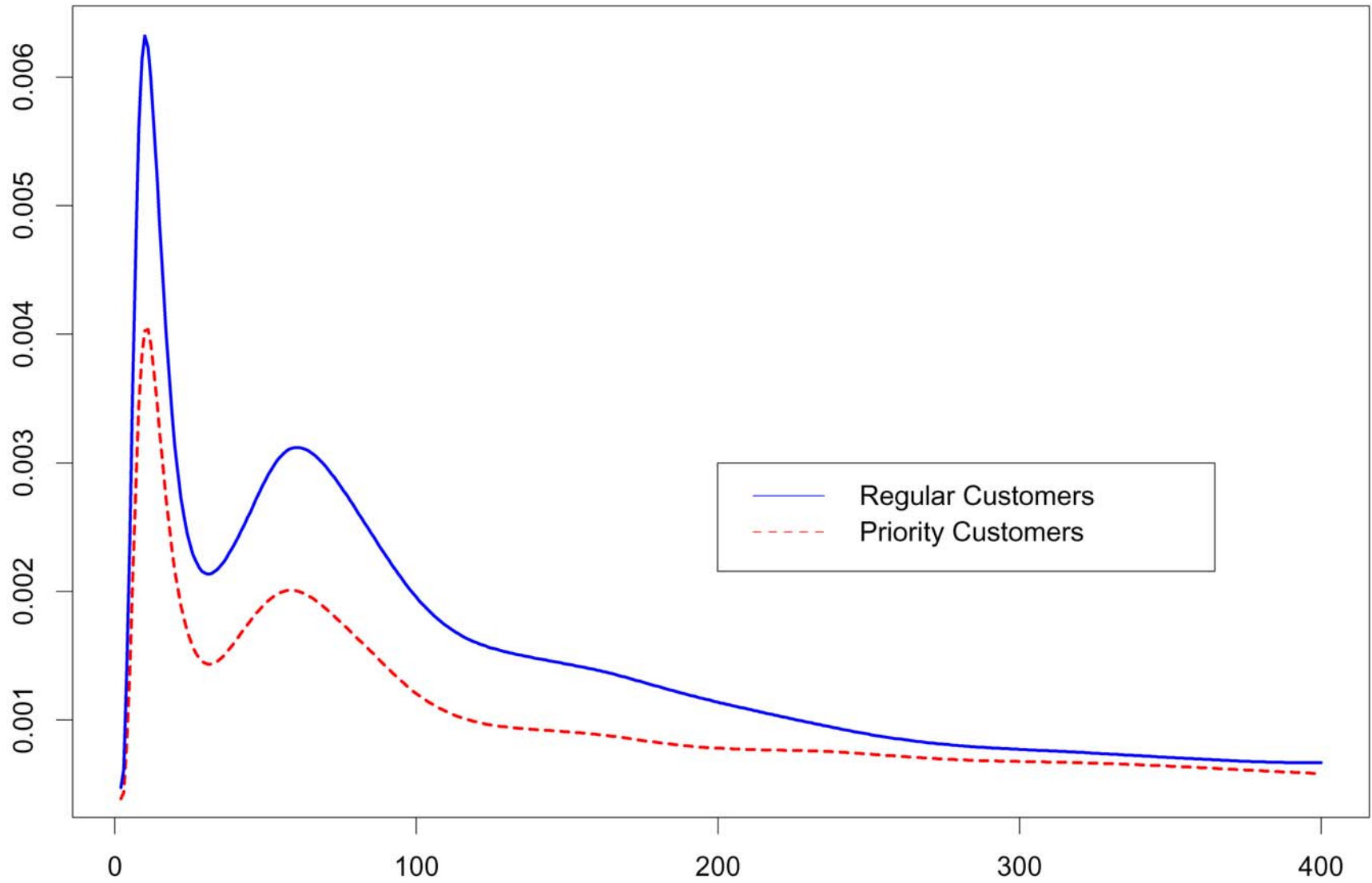
NW (New) = 111

PS (Regular) = 181

NE (Stocks) = 269

IN (Internet) = 381

# Hazard Rate: Empirical (Im)Patience



# Contents

1. **Motivation**: "The Right Answer for the Wrong Reason"
2. **Operational Regime (M/M/N)**:
  - **Quality-Driven**
  - **Efficiency-Driven**
  - **The QED (Halfin-Whitt) Regime**
3. Some Intuition

Example from a call center, leading to models with

4. **Impatient** (Abandoning) Customers (M/M/N+G)
5. **Time-Varying** Queues with **Time-Stable** Performance
6. **General Service** Times (G/M/N, G/D/N; G/LN/N)
7. **Heterogeneous** Customers and **Multi-skilled** Agents (SBR)
8. **Forecasting** Parameters

# Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios ( $E(S) = 3:45$ ):

$\lambda$ /hr	$N$	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%

# Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios ( $E(S) = 3:45$ ):

$\lambda$ /hr	$N$	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
BUT	with	Patience= $E(S)$		
1600	100	96%	0:09	50%

# Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios ( $E(S) = 3:45$ ):

$\lambda$ /hr	$N$	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
BUT	with	Patience= $E(S)$		
1600	100	96%	0:09	50%
AND		could have		<b>%Abandon</b>
1600	100	97.3%	0:23	2.7 %
1600	95	98.4%	0:23	6.5%
1800	105	97.7%	0:23	3.4%

# Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios ( $E(S) = 3:45$ ):

$\lambda$ /hr	<u>N</u>	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
BUT	with	Patience= $E(S)$		
1600	100	96%	0:09	50%
AND		could have		<b>%Abandon</b>
1600	100	97.3%	0:23	2.7 %
1600	95	98.4%	0:23	6.5%
1800	105	97.7%	0:23	3.4%

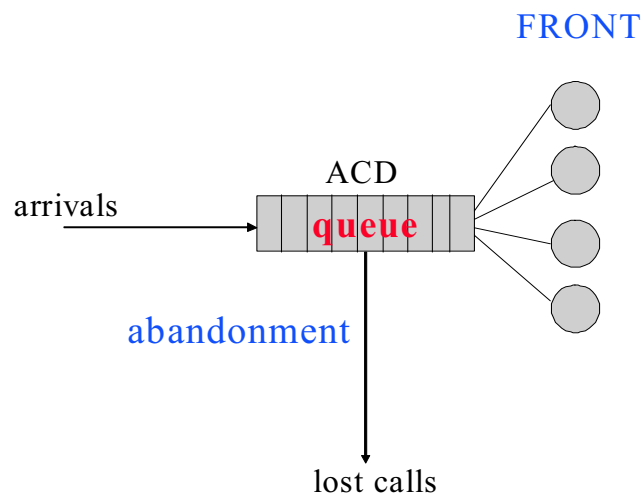
**QED** with **(Im)patient** Customers:

**The "fittest" survive and wait less – much less!**

**Erlang-A**: Erlang-C with Exponential Patience / **A**bandonment

Downloadable implementation: [4CallCenters\(.com\)](http://4CallCenters.com)

# Erlang-A (with G-Patience): M/M/N+G



## QED Theorem (Garnett, M. and Reiman '02; Zeltyn '03)

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

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

- **QED**  $\% \{ \text{Wait} > 0 \} \approx \alpha$ ,  $0 < \alpha < 1$  ;
- **Customers**  $\% \{ \text{Abandon} \} \approx \frac{\gamma}{\sqrt{N}}$ ,  $0 < \gamma$  ;
- **Agents**  $\text{OCC} \approx 1 - \frac{\beta + \gamma}{\sqrt{N}}$   $-\infty < \beta < \infty$  ;
- **Managers**  $N \approx R + \beta \sqrt{R}$ ,  $R = \lambda \times E(S)$  not small;

QED performance (ASA, ...) is easily computable, all in terms of  $\beta$  (the square-root ~~safety~~ staffing level) – see later.

Covers also the Extremes:

$$\alpha = 1 : N = R - \gamma R \quad \text{Efficiency-driven}$$

$$\alpha = 0 : N = R + \gamma R \quad \text{Quality-driven}$$

## QED Approximations (Zeltyn)

$\lambda$  – arrival rate,

$\mu$  – service rate,

$N$  – number of servers,

$G$  – patience distribution,

$g_0$  – patience density at origin ( $g_0 = \theta$ , if  $\exp(\theta)$ ).

$$N = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \quad -\infty < \beta < \infty.$$

$$P\{\text{Ab}\} \approx \frac{1}{\sqrt{N}} \cdot [h(\hat{\beta}) - \hat{\beta}] \cdot \left[ \sqrt{\frac{\mu}{g_0}} + \frac{h(\hat{\beta})}{h(-\beta)} \right]^{-1},$$

$$P\left\{W > \frac{T}{\sqrt{N}}\right\} \approx \left[ 1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)} \right]^{-1} \cdot \frac{\bar{\Phi}(\hat{\beta} + \sqrt{g_0 \mu} \cdot T)}{\bar{\Phi}(\hat{\beta})},$$

$$P\left\{\text{Ab} \mid W > \frac{T}{\sqrt{N}}\right\} \approx \frac{1}{\sqrt{N}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot [h(\hat{\beta} + \sqrt{g_0 \mu} \cdot T) - \hat{\beta}].$$

Here

$$\hat{\beta} = \beta \sqrt{\frac{\mu}{g_0}}$$

$$\bar{\Phi}(x) = 1 - \Phi(x),$$

$$h(x) = \phi(x) / \bar{\Phi}(x), \text{ hazard rate of } N(0, 1).$$

- Generalizing Garnett, M., Reiman (2002) (Palm 1943–53)
- No Process Limits

# Efficiency-Driven Approximations (Zeltyn; Whitt)

$G$  = (Im)Patience distribution

$$N = \frac{\lambda}{\mu} \cdot (1 - \gamma) + o(\lambda), \quad \gamma > 0.$$

Assume the equation

$$G(x) = \gamma$$

has a unique solution  $x^*$ .

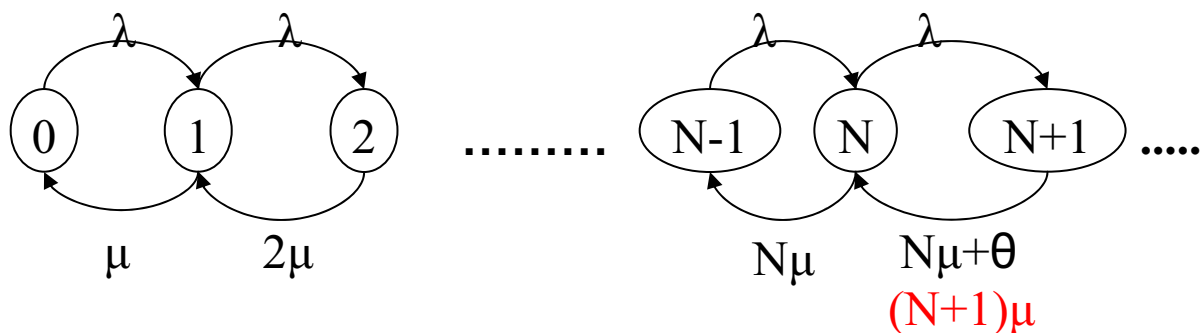
Then

$$\begin{aligned} P\{\text{Ab}\} &\approx \gamma \quad (\text{insensitive to } G) \\ P\{W > T\} &\approx \begin{cases} 1 - G(T), & T < x^* \\ 0, & T > x^* \end{cases}, \\ P\{\text{Ab} \mid W > T\} &\approx \gamma - G(T), \quad 0 \leq T < x^*. \end{aligned}$$

- **Derivation:** Laplace Method, based on Baccelli & Hebuterne (1981)
- Towards Dimensioning (with Borst, Reiman)

# Erlang-A: Moderate (Im)patience

- M/M/N + M queue, with service rate  $\mu$  equals  $\theta$  abandonment rate
- $L_t$ : number-in-system at time t (Birth & Death)
- For **any** N, transition-rates for  $\{L_t, t \geq 0\}$ :



Note: The **same** transition rates as **M/M/ $\infty$**

## Square-Root Staffing: Motivation

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

*PASTA*

$$P\{L(M / M / N + M) \geq N\} =$$

$\theta = \mu$

$$P\{L(M / M / \infty) \geq N\}$$

Fact:  $L(M / M / \infty) \sim \text{Poisson}(R)$ ;  $R = \lambda / \mu$  offered load

# Square-Root Staffing: Motivation

$$P\{W_q(M/M/N+M) > 0\} \stackrel{PASTA}{=}$$

$$P\{L(M/M/N+M) \geq N\} \stackrel{\theta=\mu}{=}$$

$$P\{L(M/M/\infty) \geq N\}$$

Fact:  $L(M/M/\infty) \sim \text{Poisson}(R)$ ;  $R = \lambda/\mu$  offered load

For  $R$  not too small:

$$L(M/M/\infty) \stackrel{d}{\approx} \text{Normal}(R, R) \stackrel{d}{=} R + Z\sqrt{R}$$

$$\Rightarrow P\{W_q > 0\} \approx P\left\{Z \geq \frac{N-R}{\sqrt{R}}\right\} = 1 - \phi\left(\frac{N-R}{\sqrt{R}}\right)$$

## Square-Root Staffing: Motivation

$$\begin{aligned}
 P\{W_q(M/M/N+M) > 0\} &= \\
 & \text{PASTA} \\
 P\{L(M/M/N+M) \geq N\} &= \\
 & \theta = \mu \\
 P\{L(M/M/\infty) \geq N\}
 \end{aligned}$$

Fact:  $L(M/M/\infty) \sim \text{Poisson}(R)$ ;  $R = \lambda / \mu$  offered load

For  $R$  not too small:

$$L(M/M/\infty) \stackrel{d}{\approx} \text{Normal}(R, R) \stackrel{d}{=} R + Z\sqrt{R}$$

$$\Rightarrow P\{W_q > 0\} \approx P\left\{Z \geq \frac{N-R}{\sqrt{R}}\right\} = 1 - \phi\left(\frac{N-R}{\sqrt{R}}\right)$$

Given target delay-probability  $\alpha = 1 - \phi\left(\frac{N-R}{\sqrt{R}}\right)$

$$\Rightarrow N = R + \beta \cdot \sqrt{R}, \quad \text{with} \quad \beta = \phi^{-1}(1 - \alpha)$$

$N$  is the "least integer for which"  $P\{W_q > 0\} \leq \alpha$

# Time-Varying Arrivals

Extension:  $M_t / M / N_t + M$  ( $\mu=\theta$ )

$$N_t = R_t + \beta \cdot \sqrt{R_t} \quad ?$$

Fact:  $L_t \sim \text{Poisson}(R_t)$

$R_t$  – the offered load at time  $t$ , namely:

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

$S_e$  – excess service  $\left( E(S_e) = E(S) \frac{1 + c_s^2}{2} \right)$

## Time-Varying Arrivals

Extension:  $M_t / M / N_t + M$  ( $\mu = \theta$ )

$$N_t = R_t + \beta \cdot \sqrt{R_t} \quad ?$$

Fact:  $L_t \sim \text{Poisson}(R_t)$

$R_t$  – the offered load at time  $t$ , namely:

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

$$S_e - \text{excess service} \left( E(S_e) = E(S) \frac{1 + c_s^2}{2} \right)$$

$L_t \stackrel{d}{\approx} N(R_t, R_t)$  hence, as before:

$$\Rightarrow N_t = \lceil R_t + \beta \cdot \sqrt{R_t} \rceil, \quad \beta = \phi^{-1}(1 - \alpha)$$

hopefully yields time-stable delay probability  $\alpha$ :

Indeed, but in fact **TIME-STABLE PERFORMANCE !**

What if  $\mu \neq \theta$ ?

Use an *Iterative Algorithm* that is *Simulation-Based*

## Performance Measures

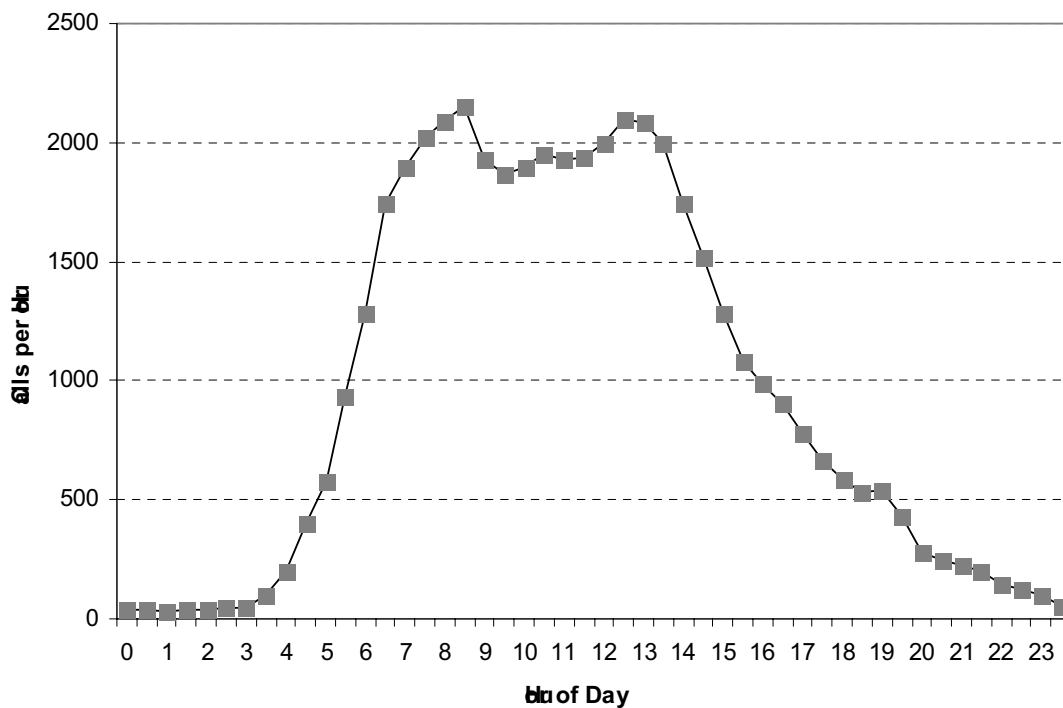
- ***Delay probability in interval  $t$*** , calculated by the fraction of customers who are not served immediately upon arrival, out of all arriving customers during the  $t$  time-interval
- ***Average waiting time in interval  $t$*** , calculated by the average waiting time of all customers arriving during the  $t$  time-interval.
- ***Average queue length in interval  $t$*** , taken constant over the time-interval. The queue length is averaged over all replications
- ***Tail probability in interval  $t$*** , calculated as the probability that queue size equals or exceeds some threshold (e.g. 3 times average queue)
- ***Servers' Utilization in interval  $t$*** , calculated as the fraction of busy-servers during a time-interval (accounting for servers who are busy only a fraction of the interval)
- ***Service grade  $\beta_t$  in interval  $t$*** , which arises from the following "Square-Root Staffing" rule:

$$N_t = R_t + \beta_t \sqrt{R_t}$$

# Example: "Real" Call Center

Two-hump arrival functions are typical

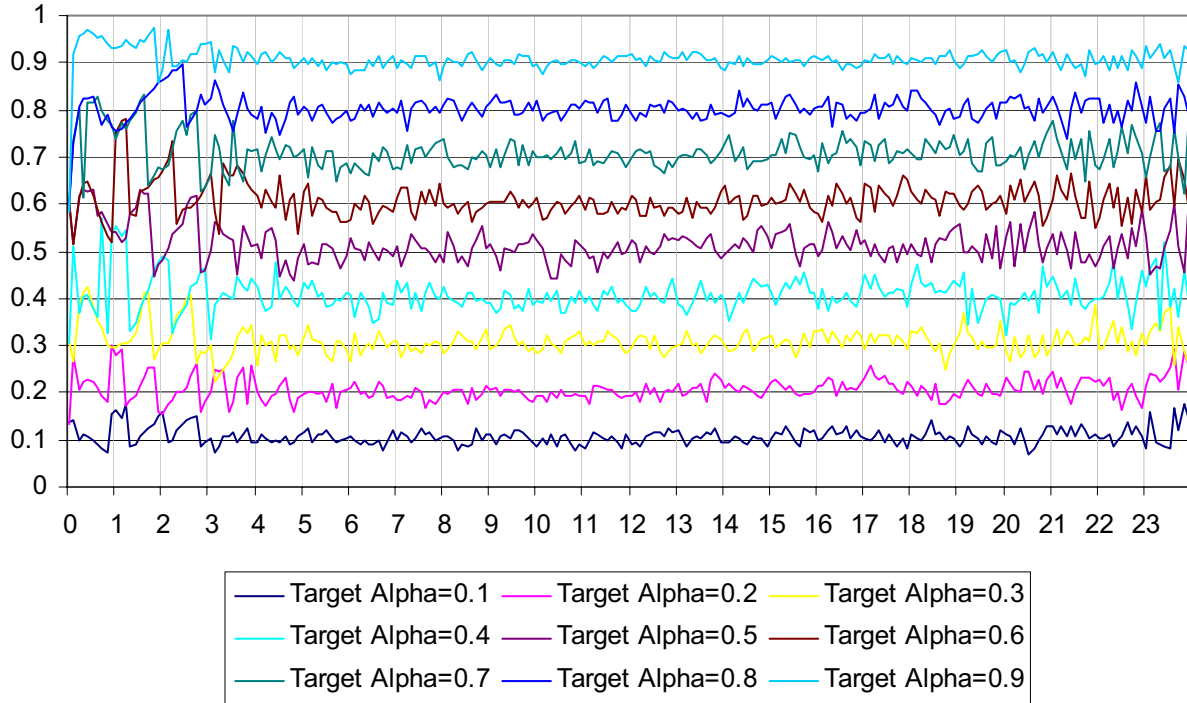
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)



- Service and abandonment rates are **both** exponential having **mean 0.1** (6 min.)

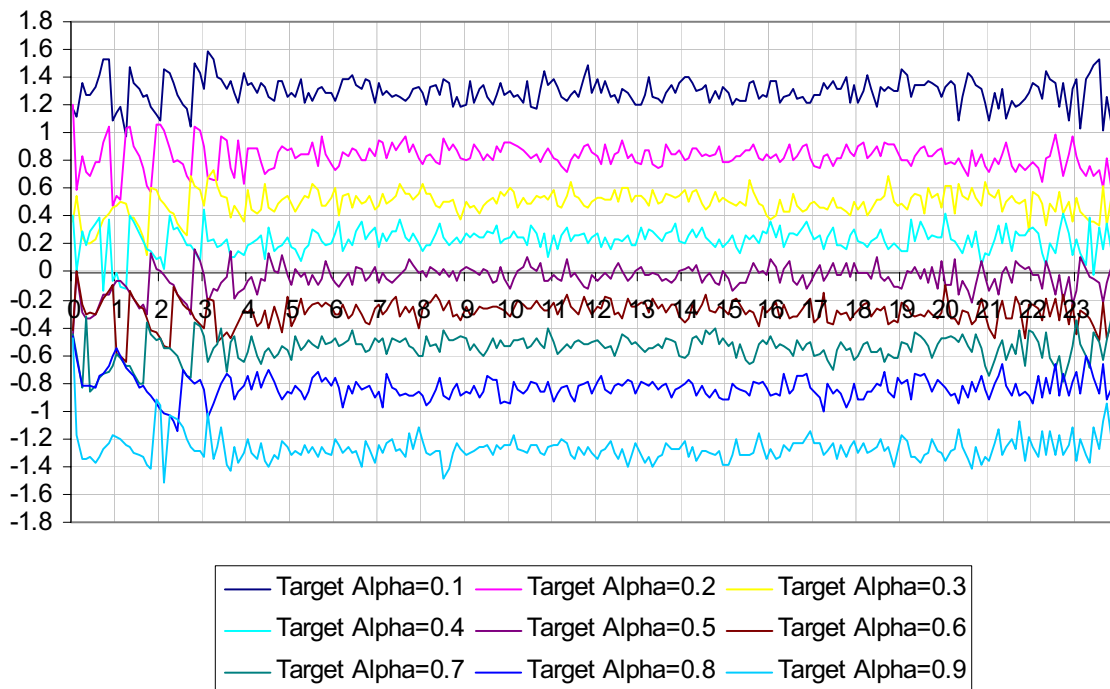
# Delay Probability $\alpha$

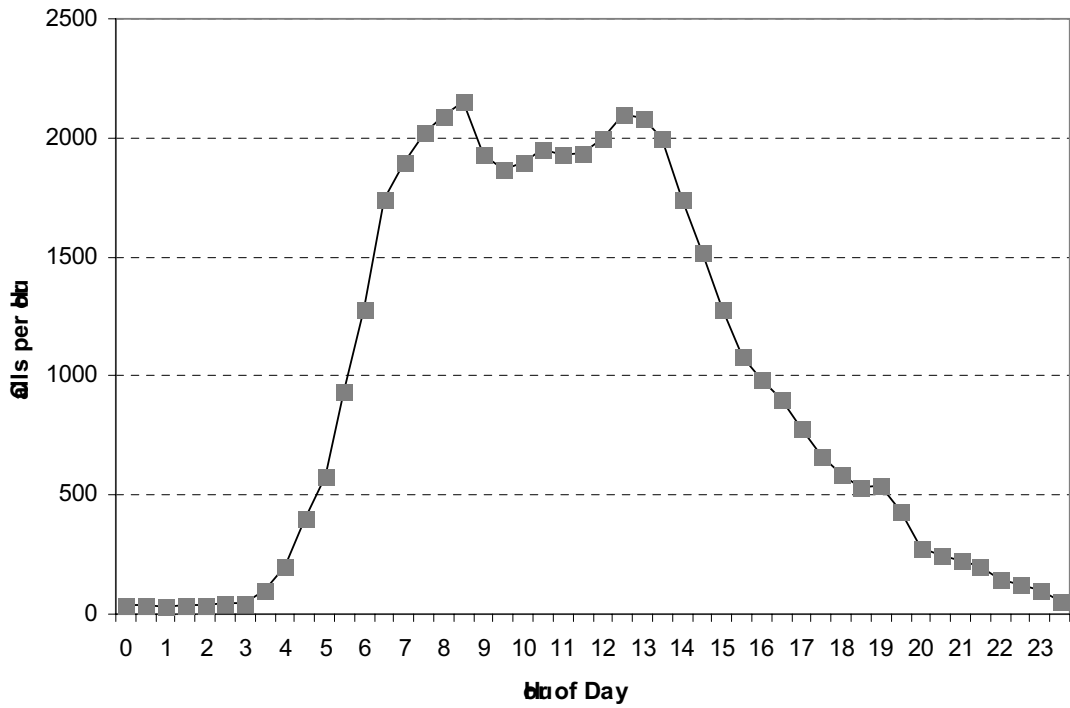
Delay Probability



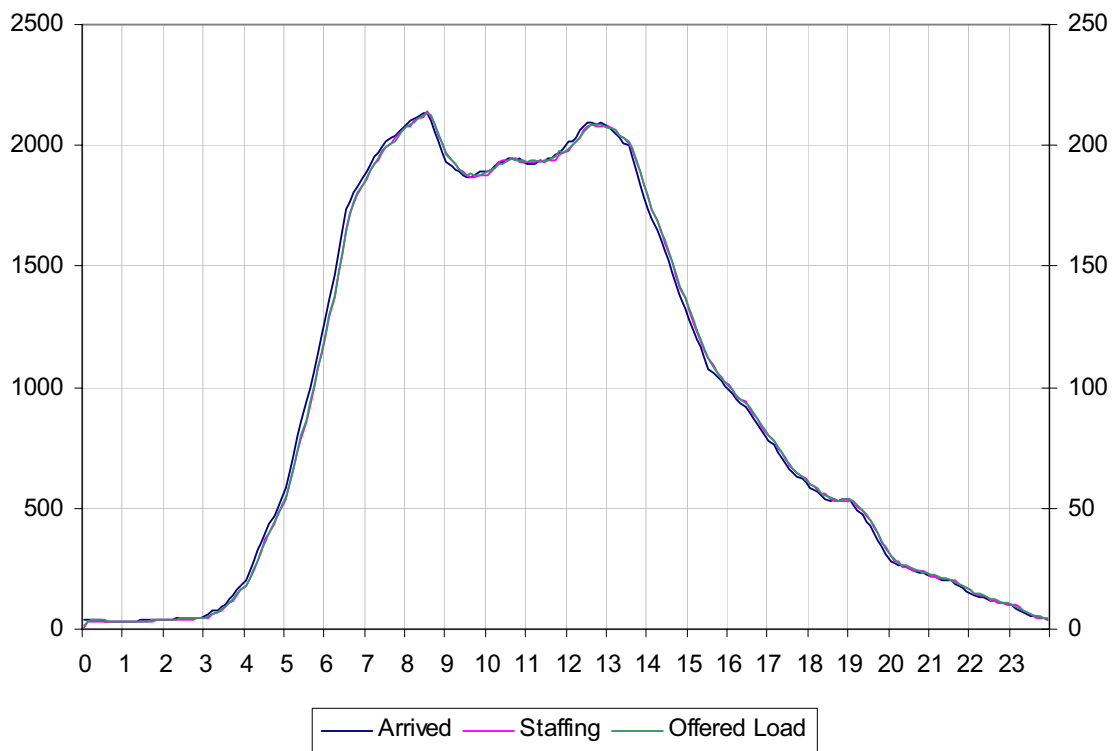
# Service Grade $\beta$

Beta





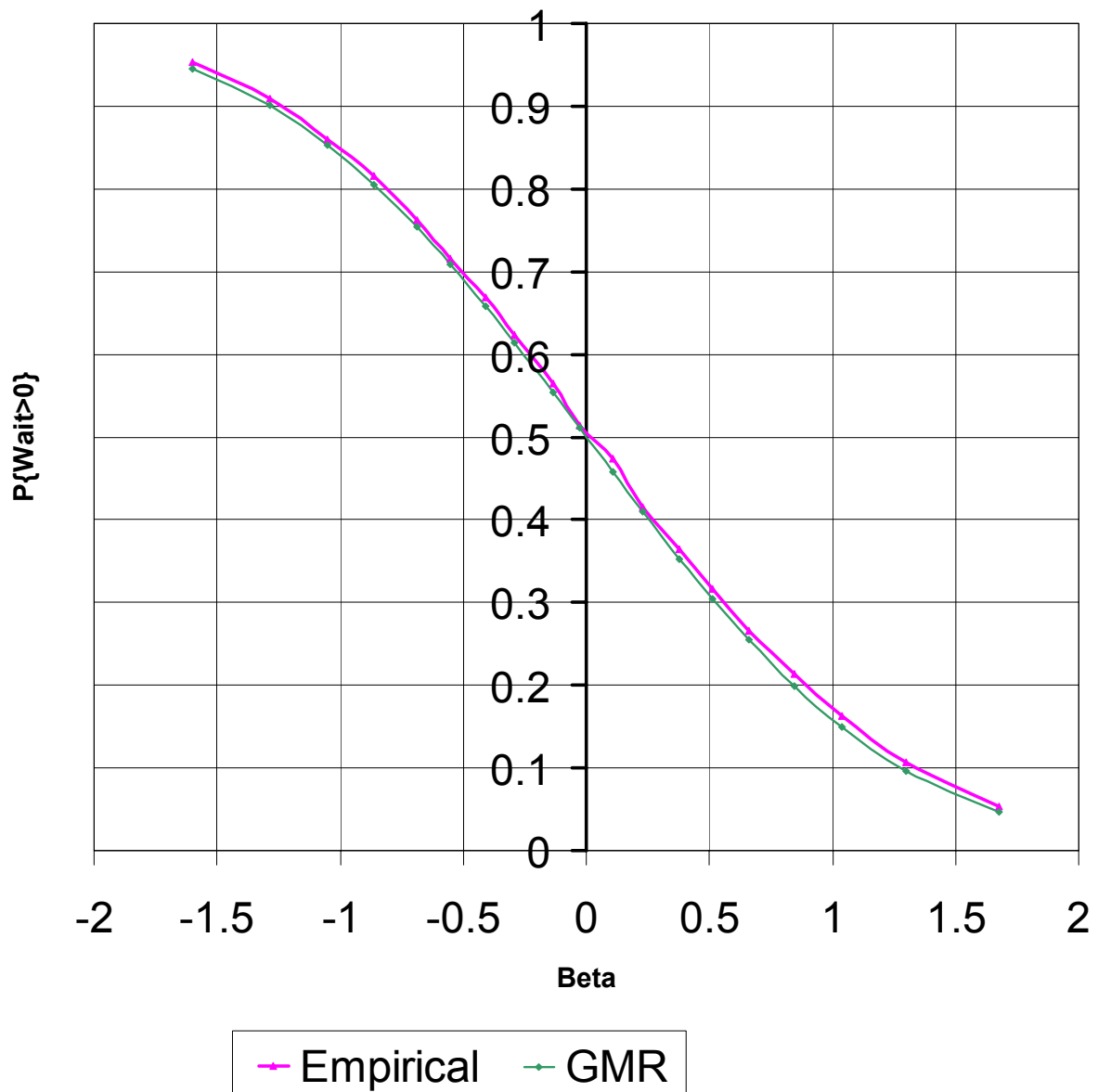
## QED Staffing ( $\alpha=0.5$ )



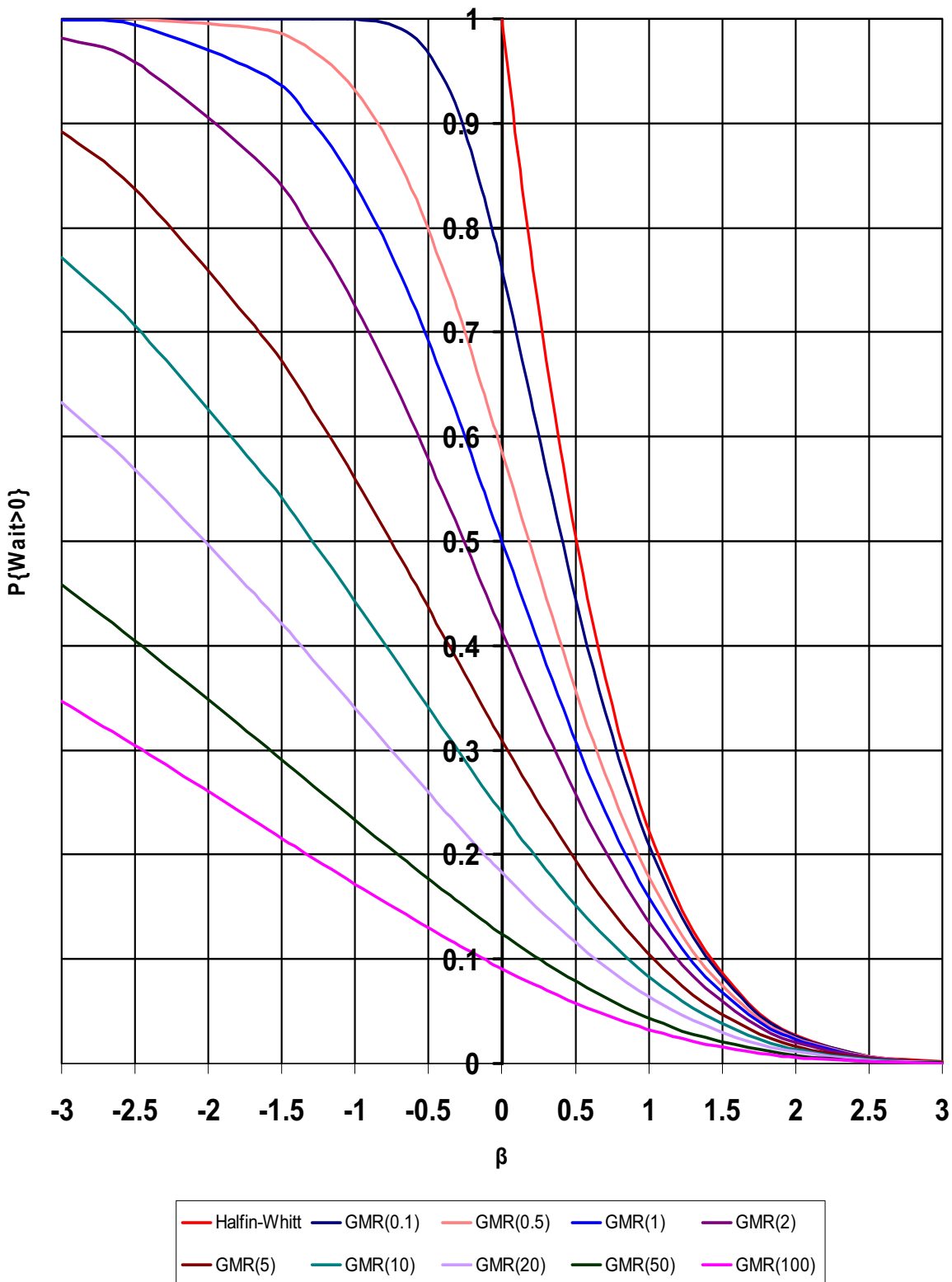
# Erlang-A: Theoretical vs. Empirical

$P\{\text{Wait}>0\}=\alpha$  vs.  $\beta$  ( $N=R+\beta\sqrt{R}$ )

## Moderate Patience

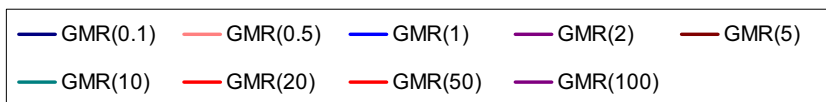
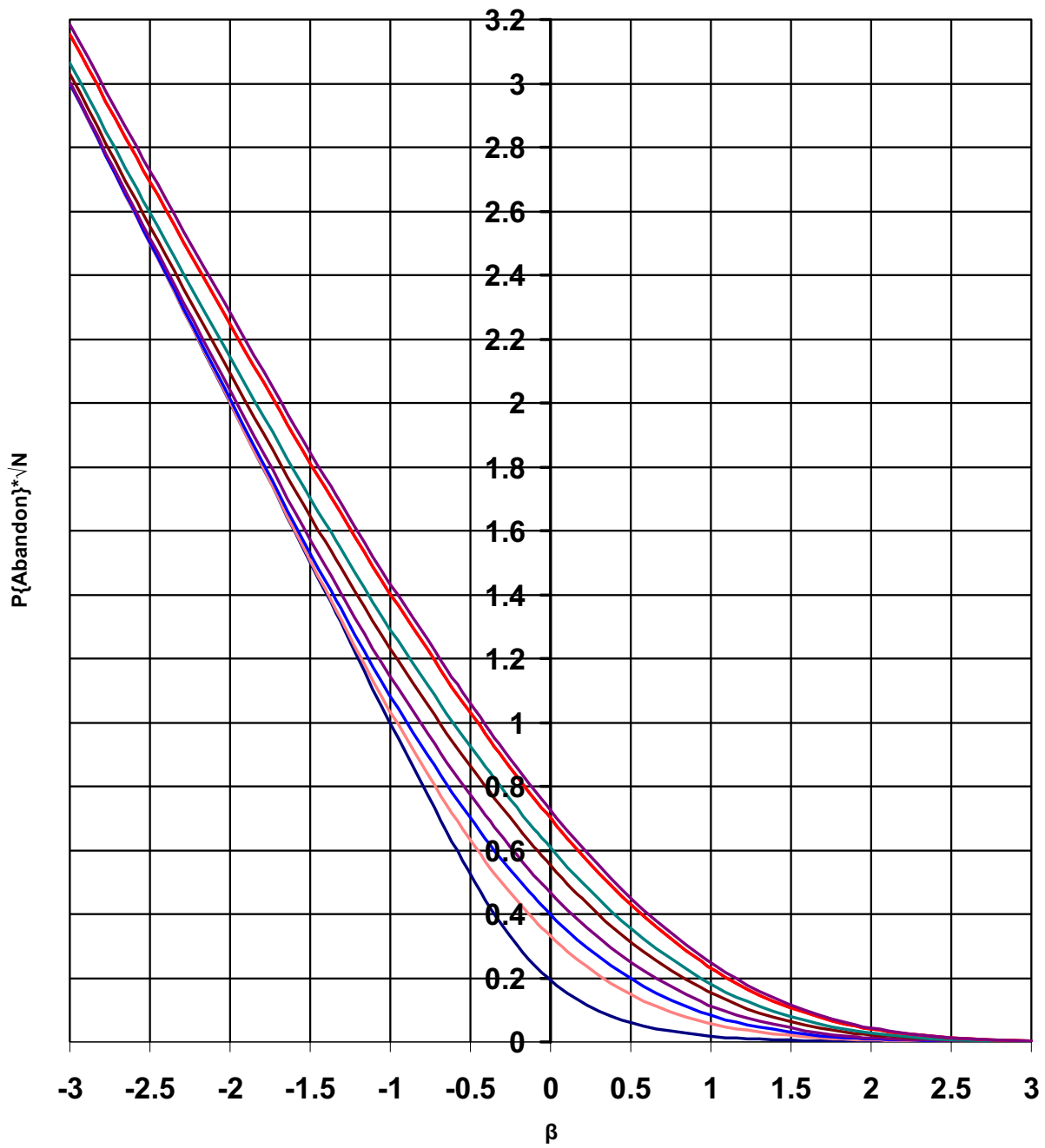


## Erlang-A: $P\{\text{Wait}>0\}=\alpha$ vs. $\beta$ ( $N=R+\beta\sqrt{R}$ )



GMR(x) describes the asymptotic probability of delay as a function of  $\beta$  when  $\frac{\theta}{\mu} = x$ . Here,  $\theta$  and  $\mu$  are the abandonment and service rate, respectively.

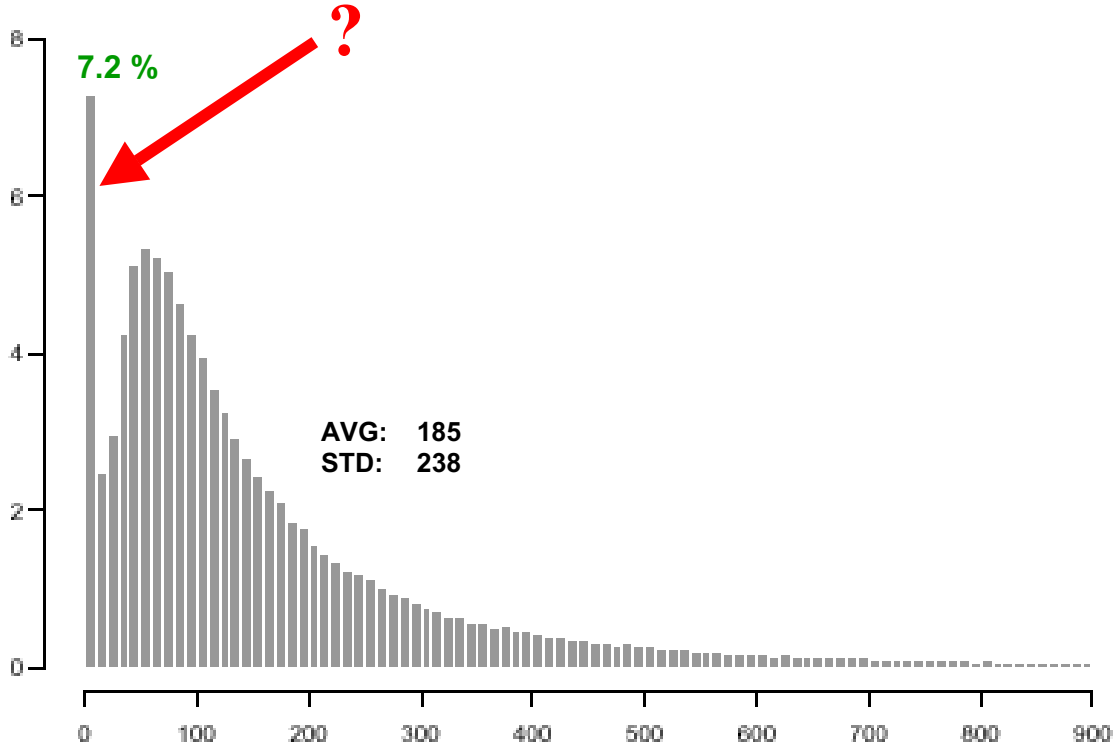
## Erlang-A: $P\{\text{Abandon}\}*\sqrt{N}$ vs. $\beta$



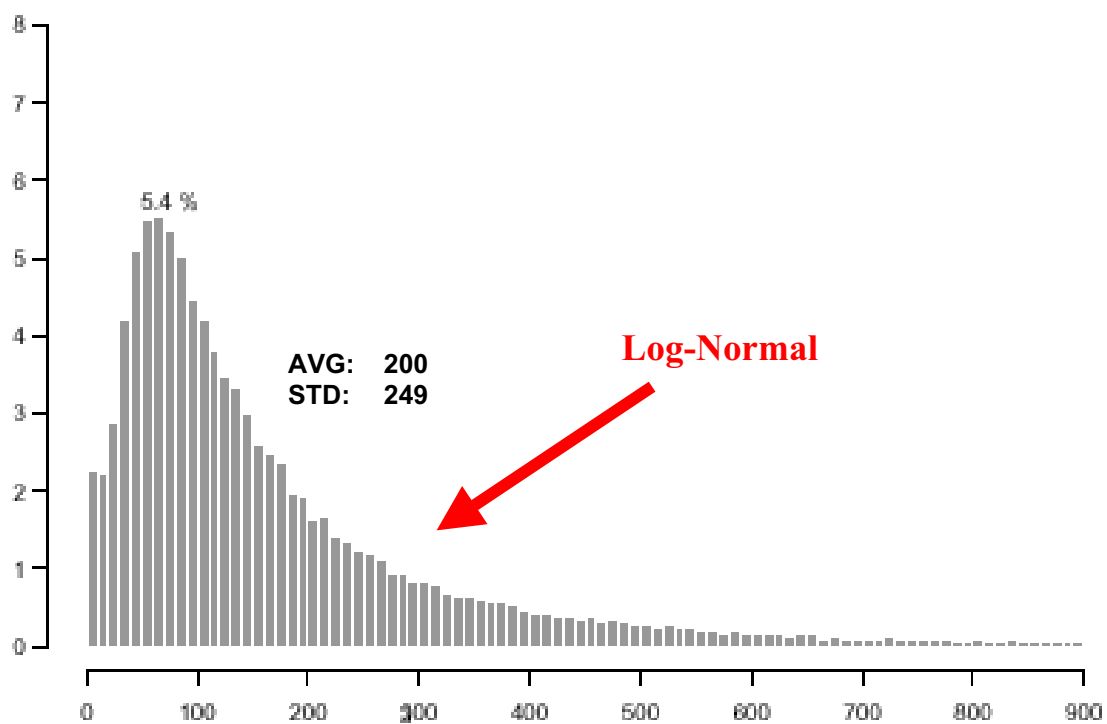
# Beyond Data Averages

## Short Service Times

Jan – Oct:

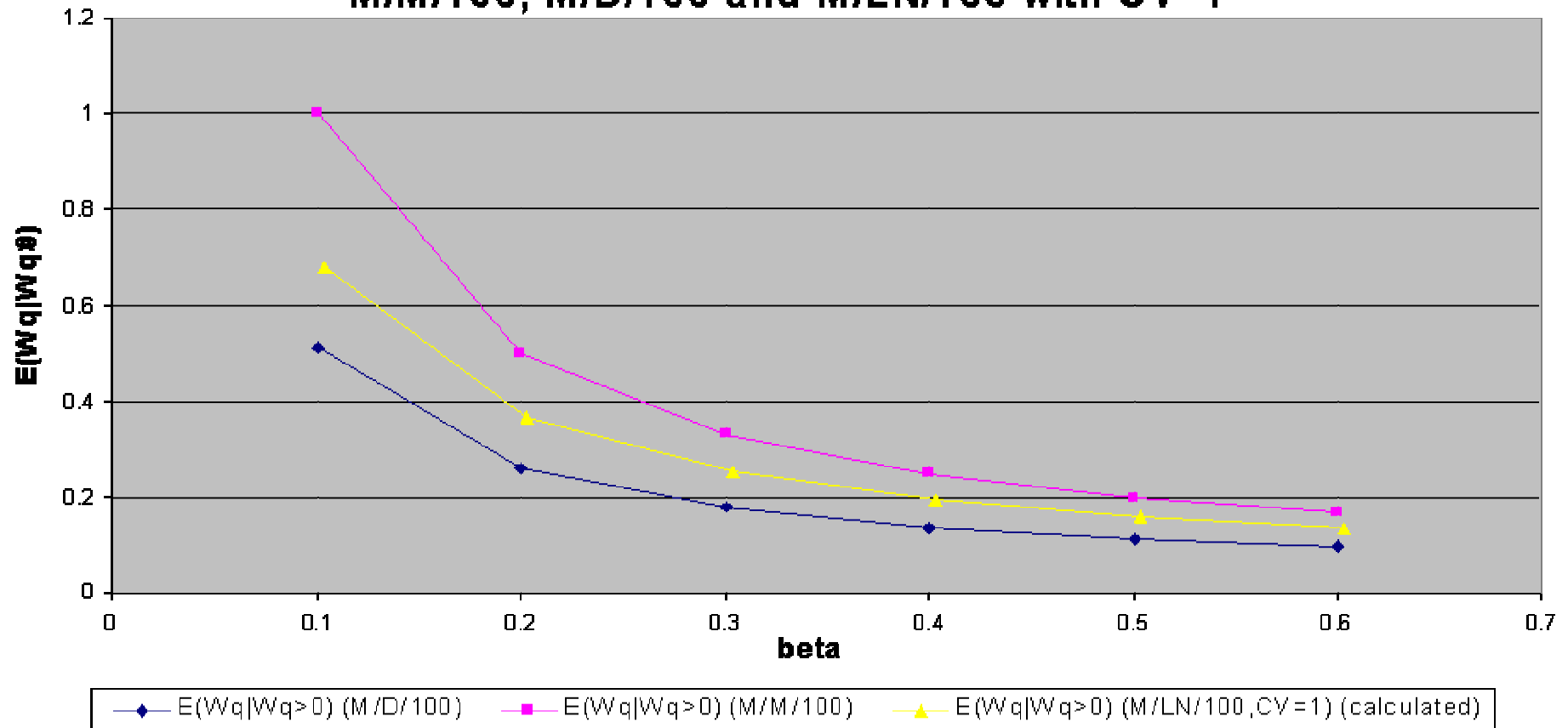


Nov – Dec:



# $E(W_q|W_q>0)$ vs. $\beta$

M/M/100, M/D/100 and M/LN/100 with CV=1



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

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

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

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

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

Observation: Cyclic assignment does not alter waiting times

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

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

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

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

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

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

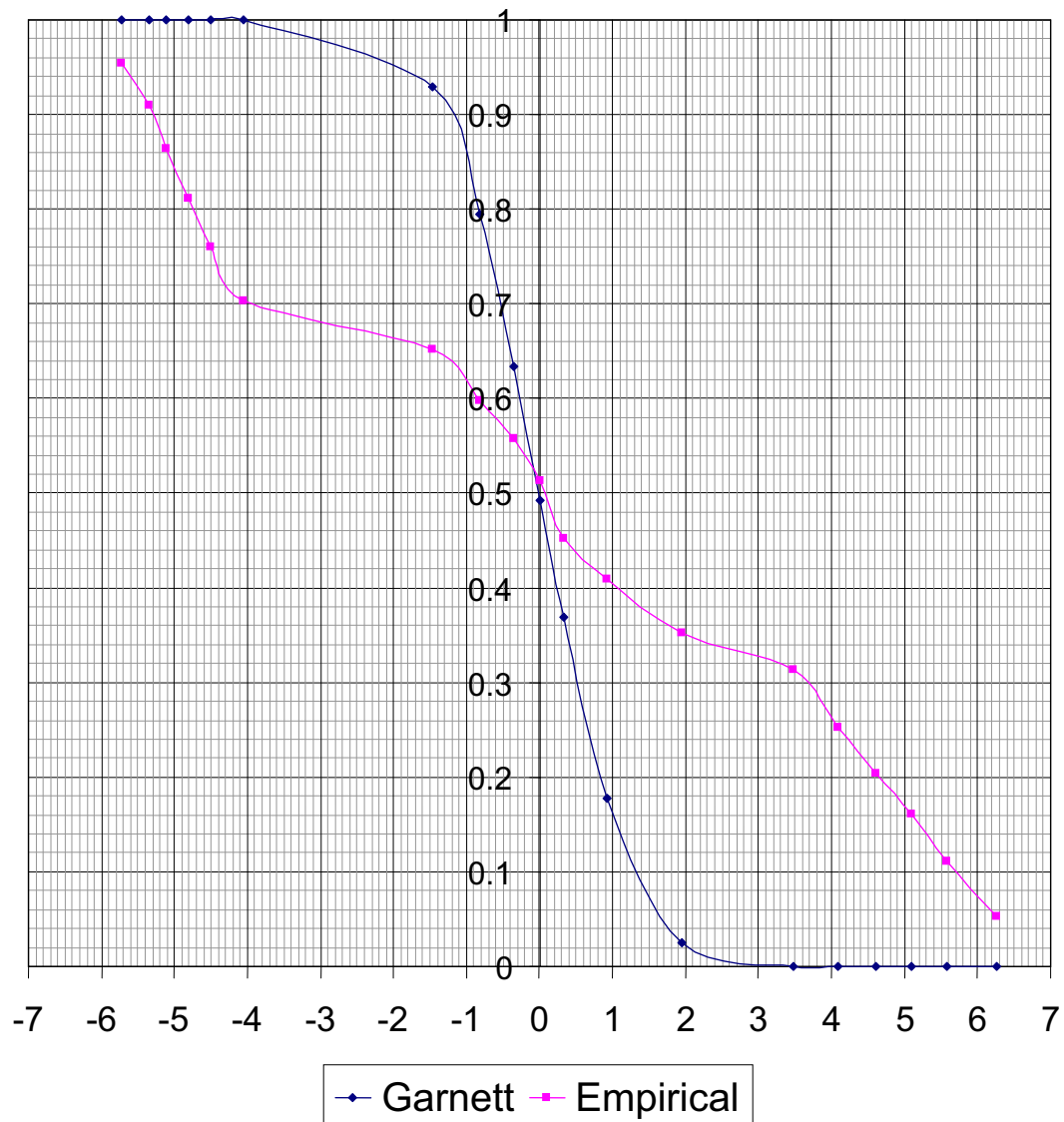
$P(W_N > 0) \rightarrow P(Z < W - \beta) = E\phi(W - \beta) < 1$

( Efficiency:  $N = R+c$  (HT); Quality:  $N = R+bR$  (Stable) )

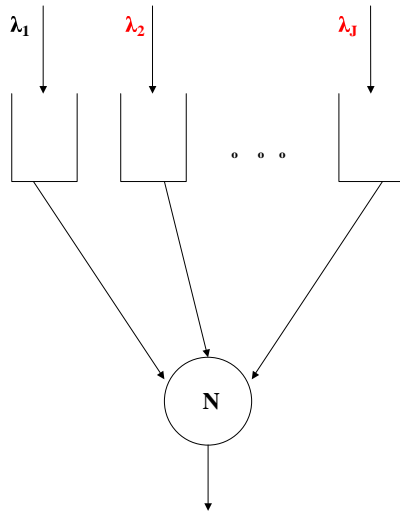
# Forecasting the Arrival Function

## Theoretical & Empirical

### Prob. of Delay vs. $\beta$



## Dimensioning the V-Model



- $J$  customer classes: arrivals  $Poisson(\lambda_j)$ .
- $N$  iid servers: service durations  $Exp(\mu)$ .

### The staffing problem:

Given  $0 < \alpha_1 < \alpha_2 < \dots < \alpha_J < 1$ ,

**Min**  $N$

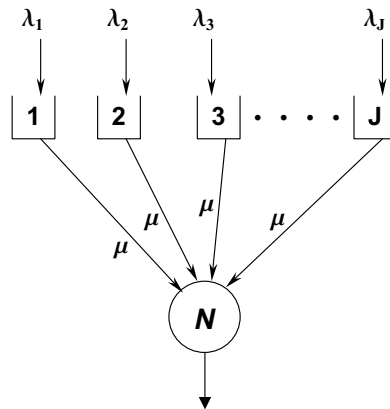
**s.t.**  $P_\pi(W_j(\infty) > 0) \leq \alpha_j, \quad j = 1, \dots, J$

for some scheduling policy  $\pi$

(Could also minimize  $cN + \sum_j d_j \lambda_j EW_j(\infty)$ )

# Multi-Class $M/M/N$ (V-Design)

## Threshold-Based Priorities (Schaack & Larson)



**Static priorities  $1 > 2 > \dots$  with thresholds**

$$0 = K_1 \leq K_2 \leq \dots K_J \leq N$$

i.e. a class- $j$  customer is served when it is of the present highest-priority and the number of idle servers is more than  $K_j$ .

Note:

Optimal Control with State-Dependent Thresholds (Yahalom)

# QED Multi-Class $M/M/N$ (V-Design)

## Threshold-Based Priorities (Gurvich)

Thresholds:  $0 = K_1^N \leq K_2^N \leq \dots \leq K_J^N \leq N$

Service Levels  $0 \leq \alpha_1 \leq \dots \leq \alpha_J \leq 1$ : Delay Probabilities

Consider a sequence indexed by  $N = 1, 2, \dots$

Assume:  $\lambda_j^N / \lambda^N \mu \rightarrow \rho_j > 0, \forall j$  (all classes non-negligible)

Assume:  $K_J^N = o(\sqrt{N})$

Then the following conditions are equivalent:

- **Customer:**  $\lim_{N \rightarrow \infty} P\{W_J^N > 0\} = \alpha_J, \quad 0 < \alpha_J < 1;$
- **Server:**  $\lim_{N \rightarrow \infty} \sqrt{N} (1 - \rho^N) = \beta, \quad 0 < \beta < \infty;$
- **Manager:**  $N \approx R + \beta\sqrt{R}, \quad R = \lambda^N / \mu$  large;

in which case  $\alpha_J = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}$  and, moreover,

the following two conditions are equivalent

- **Customer:**  $\lim_{N \rightarrow \infty} P\{W_j^N > 0\} = \alpha_j, \quad 0 < \alpha_j < 1, \forall j;$
- **Manager:**  $K_{j+1}^N - K_j^N \rightarrow \frac{\ln \alpha_j / \alpha_{j+1}}{\ln \sum_{i=1}^j \rho_i}, \quad 1 \leq j \leq J - 1;$

Solves the Dimensioning Problem

# Iterative Algorithm

## Inputs

- System primitives:  
arrival **function**, service-time distribution,  
patience distribution (when relevant) ;
- Target delay probability  $\alpha$  ;
- Time horizon  $[0, T]$  .

## Outputs

- ✓ Staffing **function**, aiming at  
a delay probability  $\alpha$  is over  $[0, T]$  .

**Starting point:** The *infinite-server heuristics* by

Jennings, M., Massey, Whitt (1996)

## Algorithm (cont.)

**Notation:**  $\forall t \in [0, T]$  (practically  $t=0, \Delta, 2 \cdot \Delta, \dots$ )

$N_i(t)$  – staffing level at time  $t$ ,  
determined in iteration  $i=1, 2, \dots$

$L_i(t)$  – number in the system at  $t$ ,  
under staffing function  $s_i(t)$ .

### Algorithm:

(1)  $i=0$ ;  $N_0(t) \equiv \infty$  (delay probability = 0)

(2) Evaluate the distribution of  $L_i(t)$ , using **simulation**.

(3) Determine  $N_{i+1}(t)$  as follows:

$$N_{i+1}(t) = \arg \min \{c : P\{L_i(t) \geq c\} < \alpha\}, \quad 0 \leq t \leq T.$$

(4) Check stopping condition:

if  $\|N_{i+1}(\cdot) - N_i(\cdot)\|_\infty \leq 1$ , then  $N_{i+1}(\cdot)$  is our staffing level;

else  $i := i+1$ , and go back to (2) .

( $\infty$ ) Last iteration. The algorithm converges to a

Staffing Function  **$N_\infty(\cdot)$  least for which**

$$P\{L_\infty(t) \geq N_\infty(t)\} \leq \alpha, \quad 0 \leq t \leq T.$$