

Service Engineering

QED Q's

Quality & Efficiency Driven

Telephone Call/Contact Centers

Service Engineering

QED Q's

**Quality & Efficiency Driven
Telephone Call/Contact Centers**

INFORMS San-Francisco

November 15, 2005

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

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

Contents

1. Background
2. **Operational Regime:** Quality-Driven, Efficiency-Driven

QED Q's = Quality & Efficiency Driven

in an M/M/N (Erlang-C) world

3. Intuition; Dimensioning

Contents

1. Background
2. **Operational Regime:** Quality-Driven, Efficiency-Driven

QED Q's = Quality & Efficiency Driven

in an M/M/N (Erlang-C) world

3. Intuition; Dimensioning

Leading to **models**, **inference** and **tools**:

4. **Impatient** (Abandoning) Customers (Erlang-A)
5. Predictably **(Time) Varying** Queues

Contents

1. Background
2. **Operational Regime:** Quality-Driven, Efficiency-Driven

QED Q's = Quality & Efficiency Driven

in an M/M/N (Erlang-C) world

3. Intuition; Dimensioning

Leading to **models**, **inference** and **tools**:

4. **Impatient** (Abandoning) Customers (Erlang-A)
5. Predictably **(Time) Varying** Queues
6. **General Service** Times (M/D/N, M/LN/N, ...)
7. **Human**/Smart Customers and Smart Systems
8. **Heterogeneous** Customer Types and
Partially Overlapping Server Skills (SBR)

Supporting Material (Downloadable)

Gans, Koole, M:

“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." *JASA*, 2005.

Feigin, M., Trofimov.:

"[Data MOCCA](#): Models for Call Center Analysis."
Ongoing (available for use: 7GB or 20GB memory).

Research-Partners

QED/QD/ED Q's:

Garnett, Reiman: M/M/N+M-Patience (Erlang-A)

Zeltyn: M/M/N+G-patience

Jelesnkovic, Momcilovic: G/G/N w/ $G=D$, finite-support

Kaspi, Ramanan: G/G/N+G

Massey, Reiman, Rider, Stolyar: Service Networks

Dimensioning:

Borst, Reiman; Zeltyn: Erlang-C and A

Armony, Gurvich: V- and Reversed-V

Massey, Whitt; Jennings, Feldman, Rozenhmidt:

Stablizing Time-Varying Q's

SBR:

Atar, Reiman; Stolyar: Control

Atar, Shaikhet: Null-Controllability

History

"The Life of Work of A. A. Markov," by Basharin et al, Linear Algebra and its Applications, 2004.

Erlang Models (B / C):

Erlang, A.K. "Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges", Elektroteknikeren, 1917; English translation in the "Life and Work of A.K. Erlang" 1948, by Brokemeyer et al.

Square-Root Staffing / Dimensioning:

Erlang, A.K.: "On the Rational Determination of the Number of Circuits", 1924; first published in the "Life....," 1948; proofs on page 120-6.

Impatient Customers (Erlang-A / Irritation):

Palm, C. "Etude des Delais D'Attente", Ericsson Technics, 1937.

Palm, C.: "Methods of Judging the Annoyance Caused by Congestion", Tele, 1953.

Tele-Nets: Call/Contact Centers

Scope

Examples

Information
(uni, bi-dir)

#411, Tele-pay, Help Desks

Business

Tele-Banks, #800-Retail

Emergency

Police #911

Mixed

Info + Emerg.

Utility, City Halls

Info + Bus.

Airlines, Insurance, Cellular

Scale

- 10s to 1000s of agents in a “single” Call Center
- X% of work force in call centers (up to several millions)
- 70% of total business transactions in call centers
- 20% growth rate of the call center industry
- Leading-edge technology, but 70% costs for “people”

What can be achieved

At what cost

Copy of Summary Interval - Order PK

Date: 7/7/97

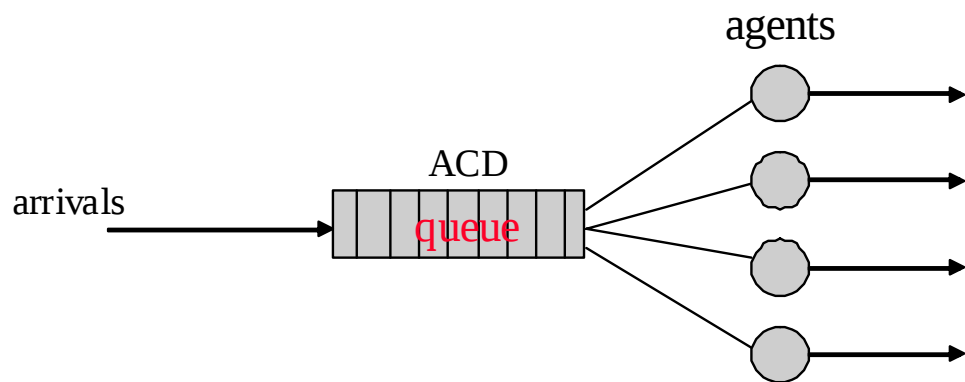
Split/Skill: Order PK

Time	Avg Speed Ans	Avg Aban Time	ACD Time	Avg ACD Time	Avg ACW Time	Aban Calls	% ACD Time	% Ans	Avg Calls Per Staff	% Serv Lev	% Aux Time	% ACW Time	% ACD Time	
	W		A	1/M		# Aban			N				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	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	0
6:00 AM*	:00:00		12	:03:21	:00:18	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
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	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		336	:03:38	:00:21	0	52	99	97	3	99	9	8	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	8	44
1:30 PM*	:00:00		366	:03:52	:00:14	0	56	99	99	4	99	11	7	50
2:00 PM*	:00:01		393	:03:55	:00:17	0	51	100	106	4	100	10	5	46
2:30 PM*	:00:00		403	:03:58	:00:13	0	54	100	112	4	100	10	4	50
3:00 PM*	:00:00	:00:04	410	:04:02	:00:16	1	57	98	110	4	98	8	5	51
3:30 PM*	:00:00		347	:03:59	:00:14	0	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

Peak



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.

λ/hr	ρ	$E(W_q) = \text{ASA}$	% 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.**

λ /hr	ρ	$E(W_q) = \text{ASA}$	% 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	99.1%	3:34 min.	12%

Quality-driven: 100 agents, 50% utilization

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

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

λ/hr	ρ	$E(W_q) = \text{ASA}$	% 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	99.1%	3:34 min.	12%

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

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

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

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

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

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

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

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

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

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

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

λ /hr	<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%

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

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

λ /hr	<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
$$\alpha = \left[1 + \frac{\beta}{h(-\beta)} \right]^{-1},$$

where $h(\cdot)$ is the hazard-rate of standard-normal.

Extremes:

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

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

√. Safety-Staffing: Performance

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

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

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

Expected Performance:

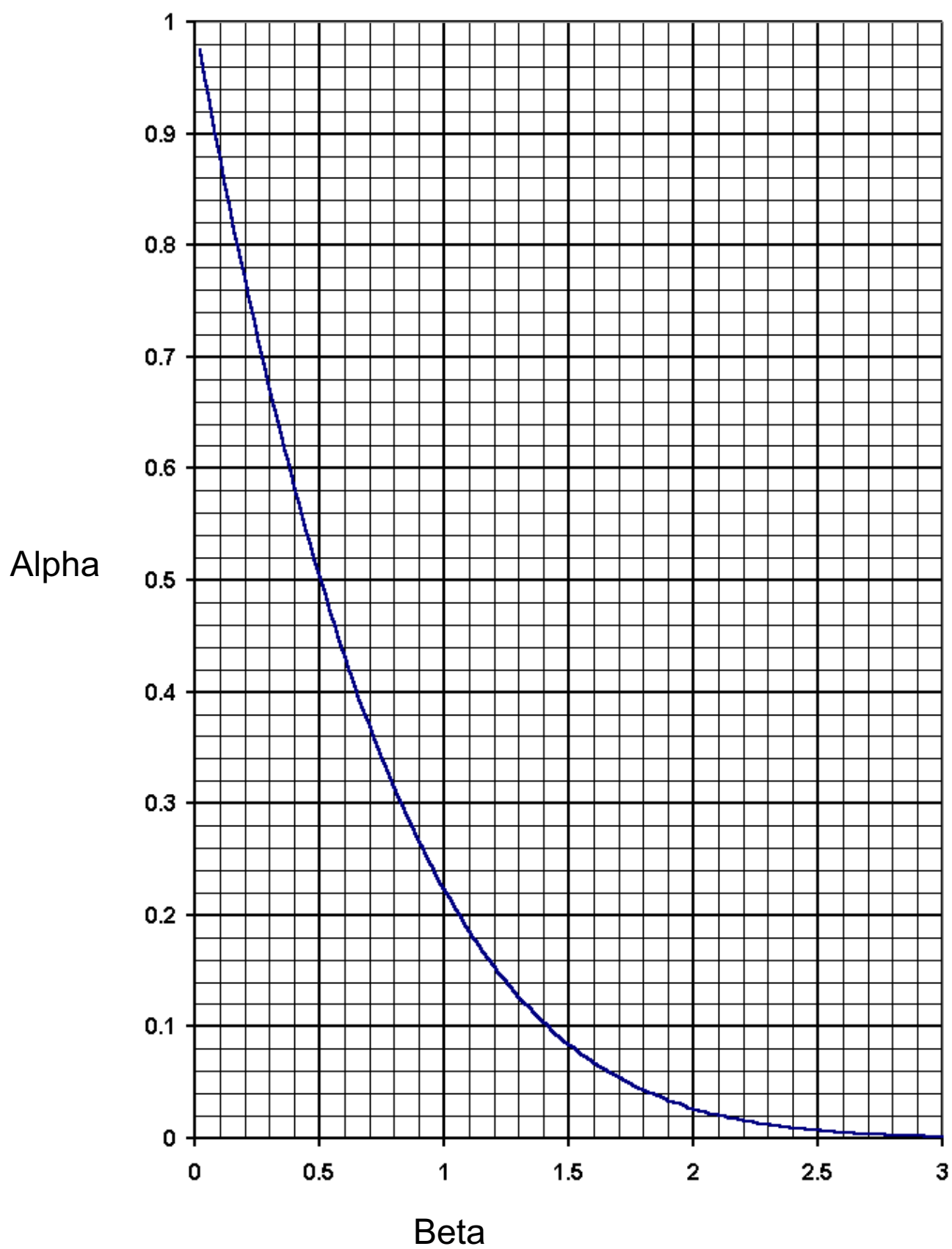
$$\% \text{ Delayed} \approx P(\beta) = \left[1 + \frac{\beta}{h(-\beta)} \right]^{-1}, \quad \beta > 0 \quad \text{Erlang-C}$$

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

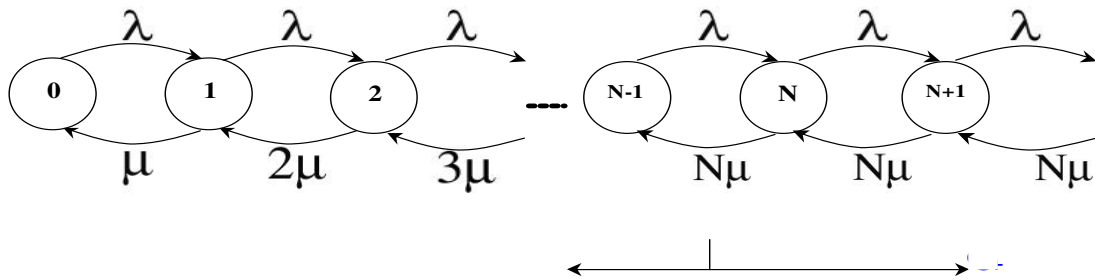
$$\% \left\{ \frac{\text{Wait}}{E(S)} > T \mid \text{Wait} > 0 \right\} = e^{-T\Delta} \quad \text{TSF}$$

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

The Halfin-Whitt Delay Function $P(\beta)$



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



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

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

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 $0 < \beta < \infty$.

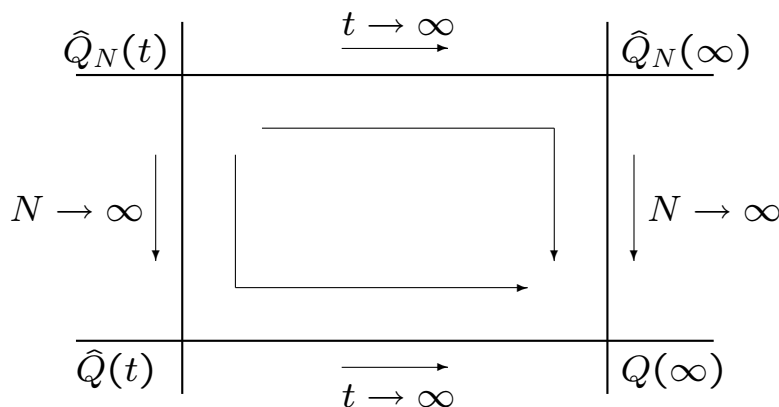
Hence, $E_{2,N} \sim \left[1 + \frac{\beta}{h(-\beta)} \right]^{-1}$, assuming $\beta > 0$.

Approximating Queueing and Waiting

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

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

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



Approximating (Virtual) Waiting Time

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

Economics: Quality vs. Efficiency

Dimensioning: with Borst and 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

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

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

Costs: $\mathbf{d}$ = delay/waiting

$\mathbf{c}$ = staffing

Min Total Costs

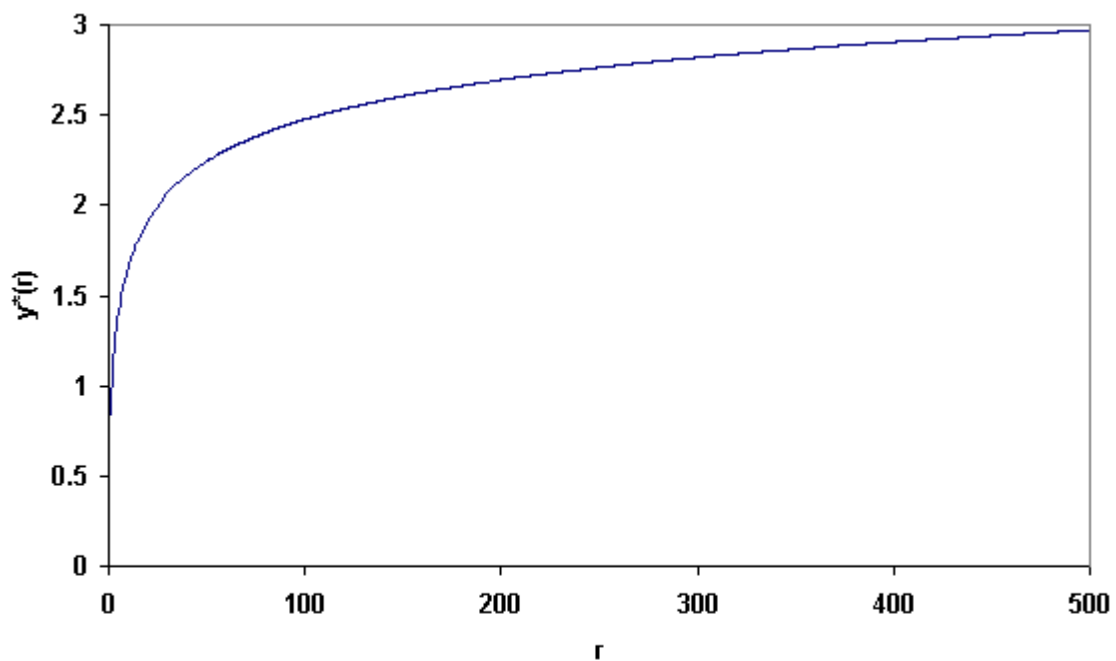
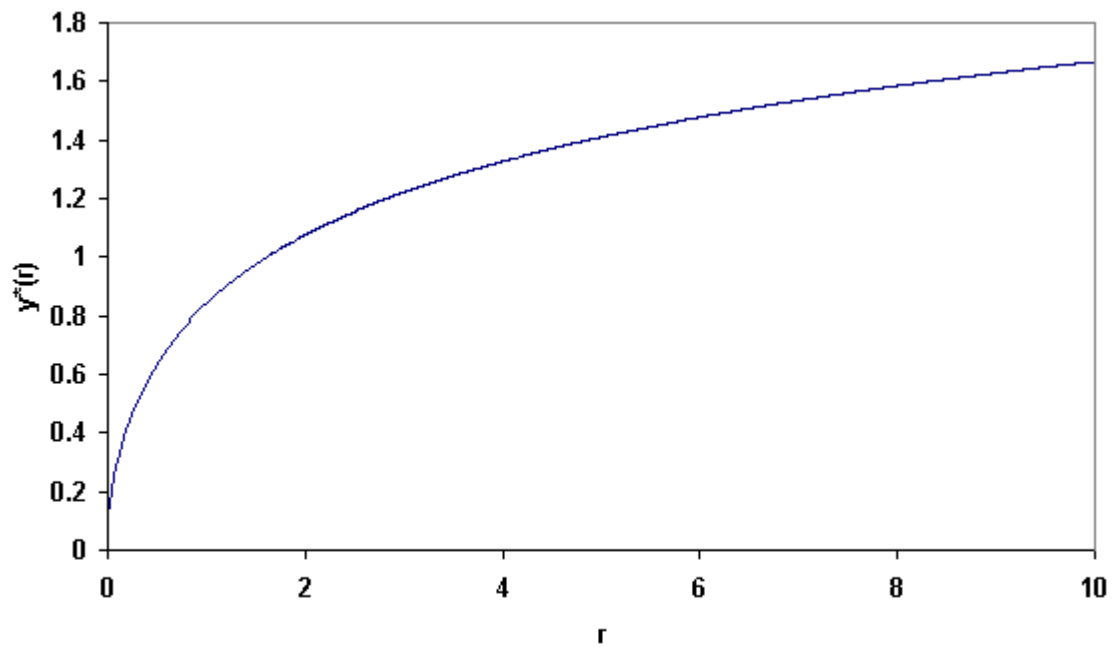
Asymptotically Optimal

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

$$\begin{aligned} \text{Here } y^*(\mathbf{r}) &\approx \left(\frac{r}{1 + r(\sqrt{\pi/2} - 1)} \right)^{1/2}, & 0 < r < 10 \\ &\approx \left(2 \ln \frac{r}{\sqrt{2\pi}} \right)^{1/2}, & r \text{ large.} \end{aligned}$$

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

r = cost of delay / cost of staffing



√. Safety-Staffing: Overview

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

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

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

Relevant: Medium to Large CC do perform as above.

Instructive: In large call centers, high resource utilization and service levels could **coexist**, which is enabled by **economies of scale** that dominate stochastic variability.

Arrivals: Inhomogeneous Poisson

Figure 1: Arrivals (to queue or service) – “Regular” Calls

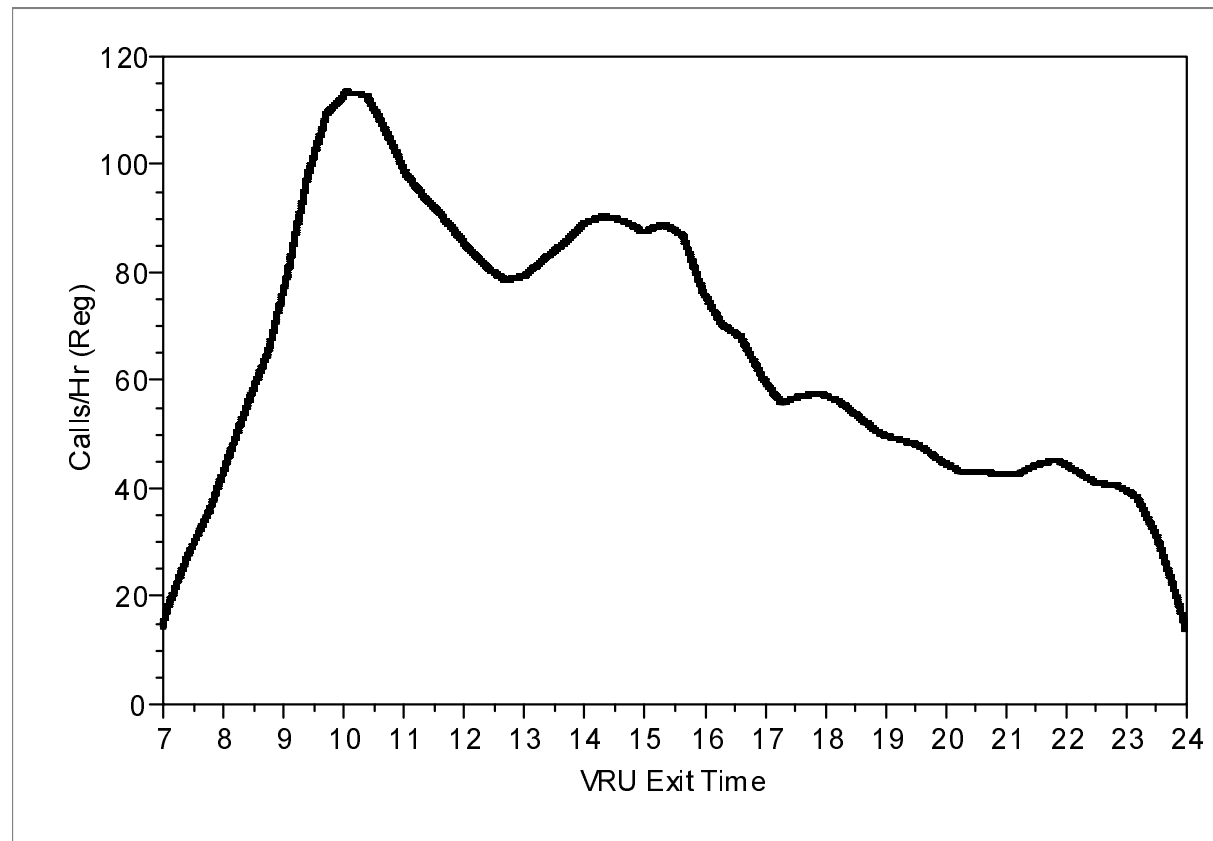
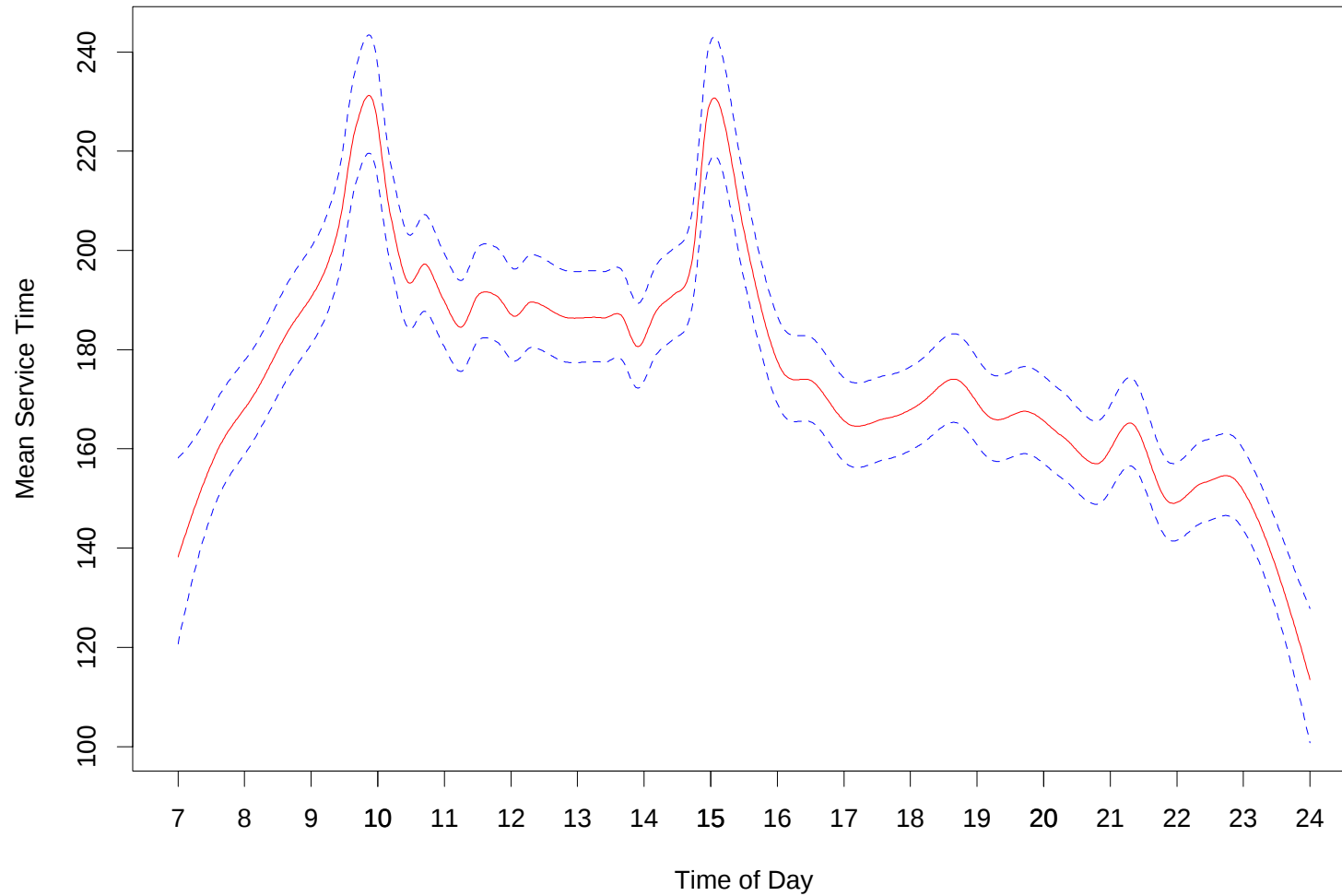
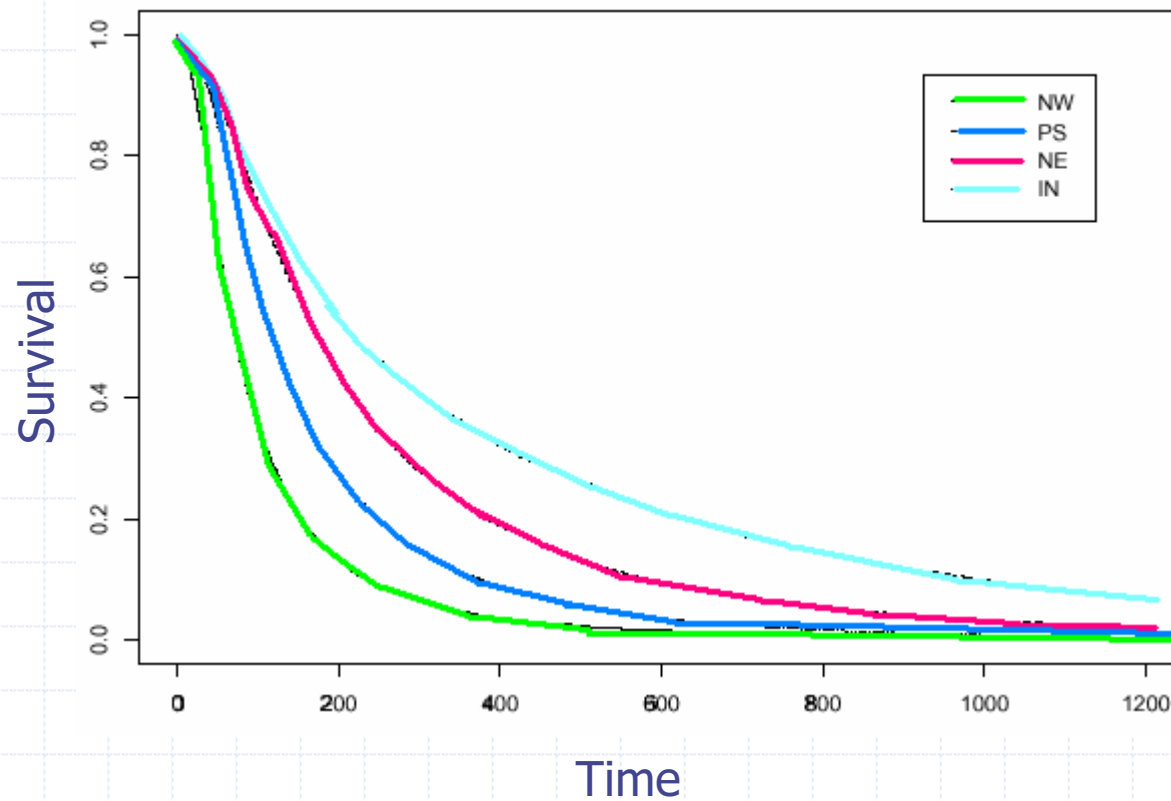


Figure 12: Mean **Service Time** (Regular) vs. Time-of-day (95% CI) ($n = 42613$)



Service Time

Survival curve, by Types



Means (In Seconds)

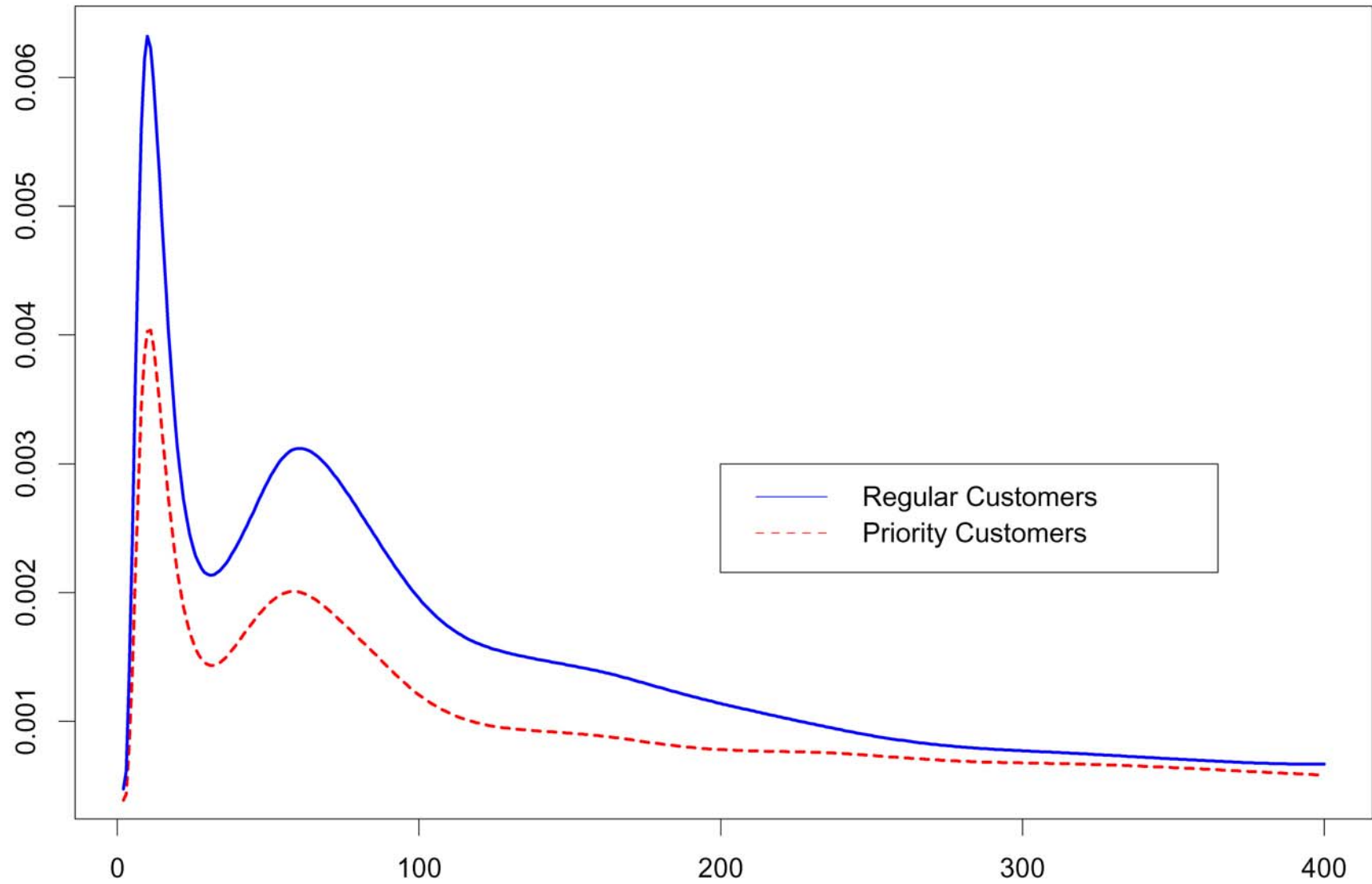
NW (New) = 111

PS (Regular) = 181

NE (Stocks) = 269

IN (Internet) = 381

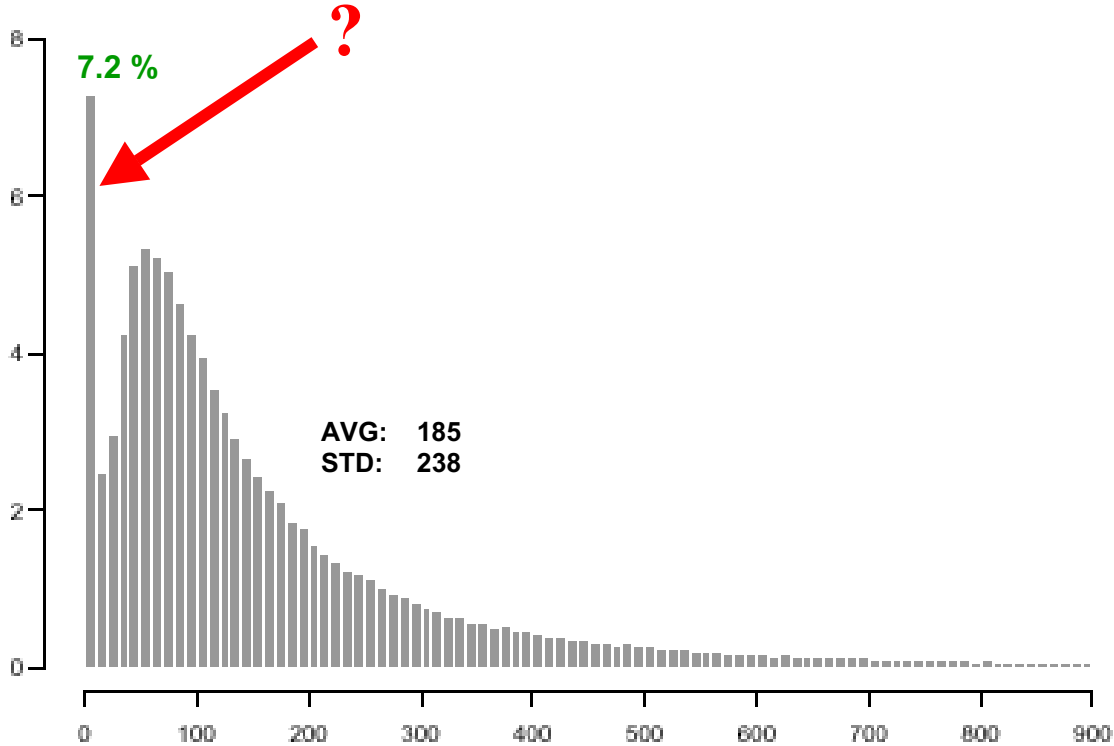
Hazard Rate: Empirical (Im)Patience



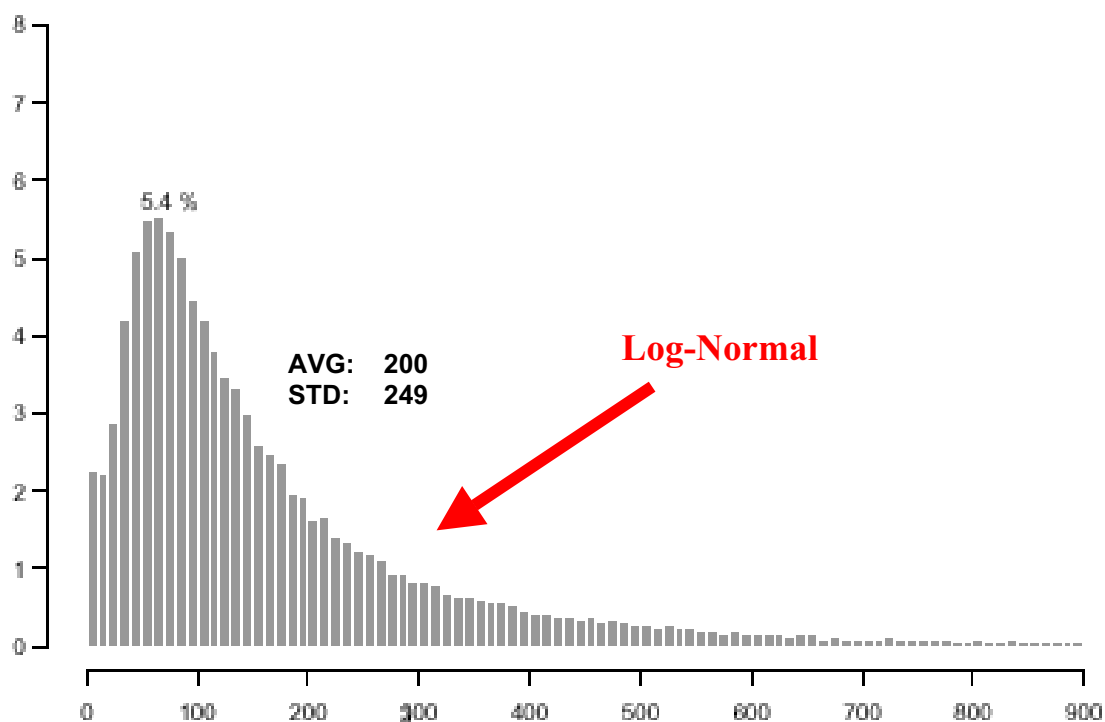
Beyond Data Averages

Short Service Times

Jan – Oct:



Nov – Dec:



Contents

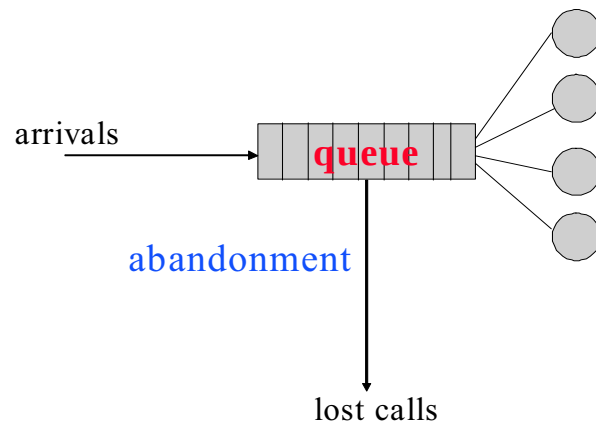
Operational Regime: Quality-Driven, Efficiency-Driven

QED Q's = Quality & Efficiency Driven

models, inference and tools:

4. **Impatient** (Abandoning) Customers (Erlang-A)
5. Predictably **(Time) Varying** Queues
6. **General Service** Times (M/D/N, M/LN/N, ...)
7. **Human**/Smart Customers and Smart Systems
8. **Heterogeneous** Customer Types and
Partially Overlapping Server Skills (SBR)
9. **Data MOCCA**: Models for Call Center Aalysis
10. **4CallCenters**: (Personal) **Tool** for WFM

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



QED Theorem Garnett, M. and Reiman '02; Zeltyn

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

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

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

QED performance (ASA, ...) is easily computable, all in terms of β (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

λ – arrival rate,

μ – service rate,

N – number of servers,

G – patience distribution,

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

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

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

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

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

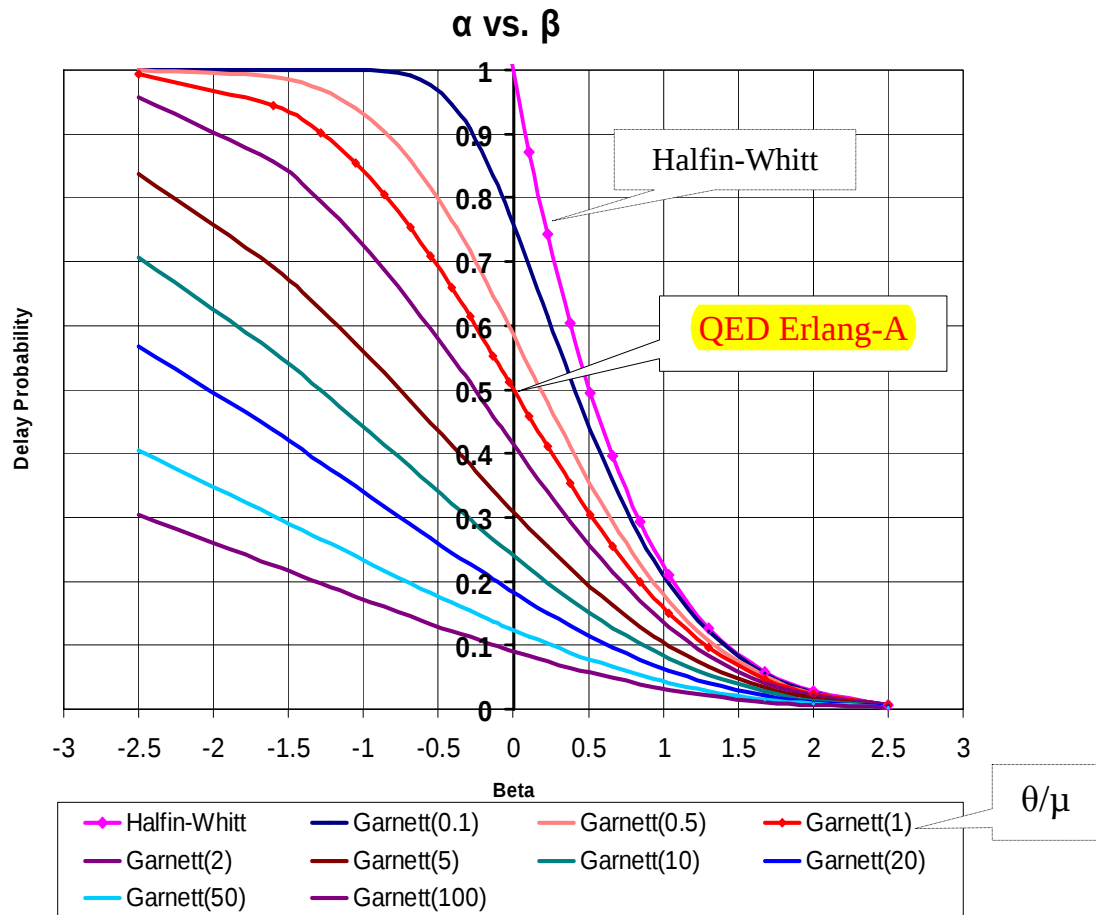
Here

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

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

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

HW/GMR Delay Functions

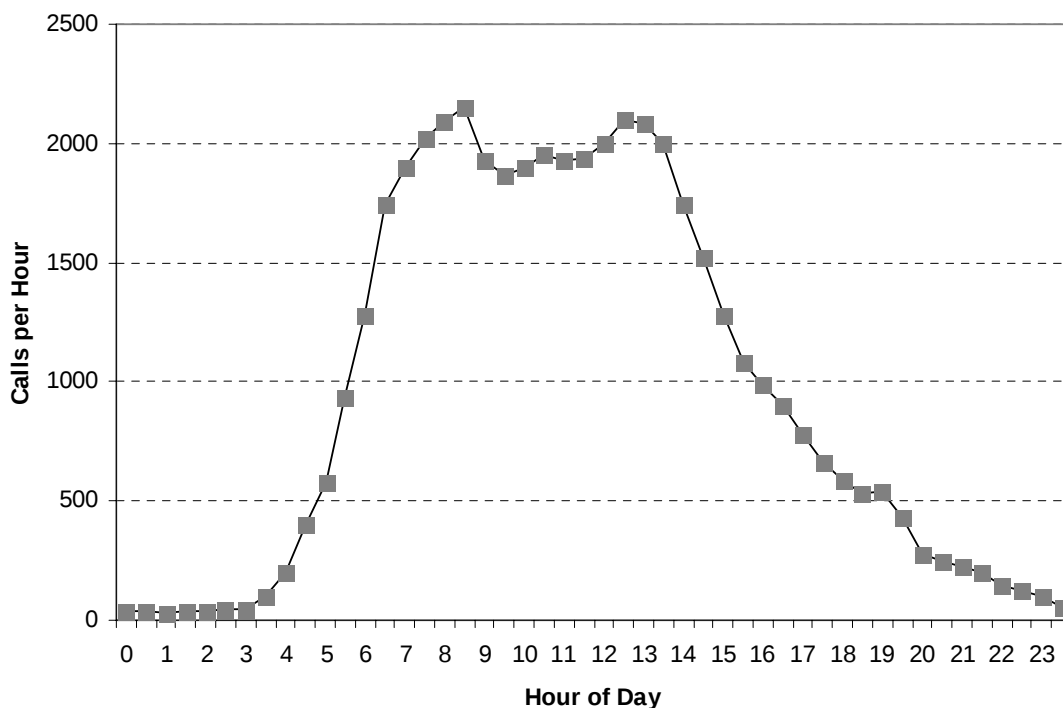


Example: "Real" Call Center

(The "Right Answer" for the "Wrong Reasons")

Time-Varying (two-hump) arrival functions common

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



Assume: Service and abandonment times are **both**
Exponential, with **mean 0.1** (6 min.)

Time-Varying Arrivals

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ **?** θ

?

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

Time-Varying Arrivals

Model $M_t / M / N_t + M$

Parameters $\lambda(t) \quad \mu \quad ? \quad \theta$

?

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

Offered Load:

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

Average # in $M_t / M / \infty$

Time-Varying Arrivals

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ $?$ θ

$?$

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

Offered Load:

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

Average # in $M_t / M / \infty$

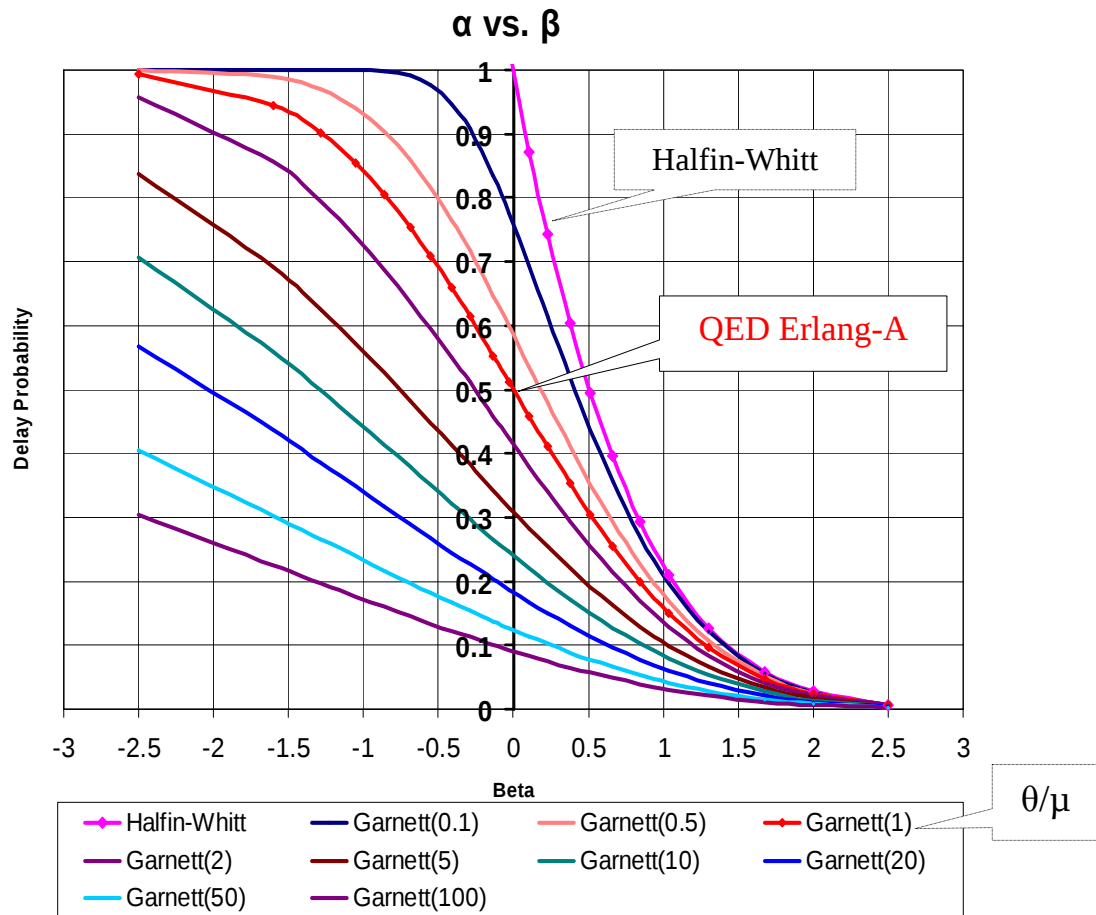
Gives rise to **TIME-STABLE PERFORMANCE**

(Why? Think $M_t / M / N_t + M$ with $\mu = \theta$;

And if $\mu \neq \theta$, or generally:

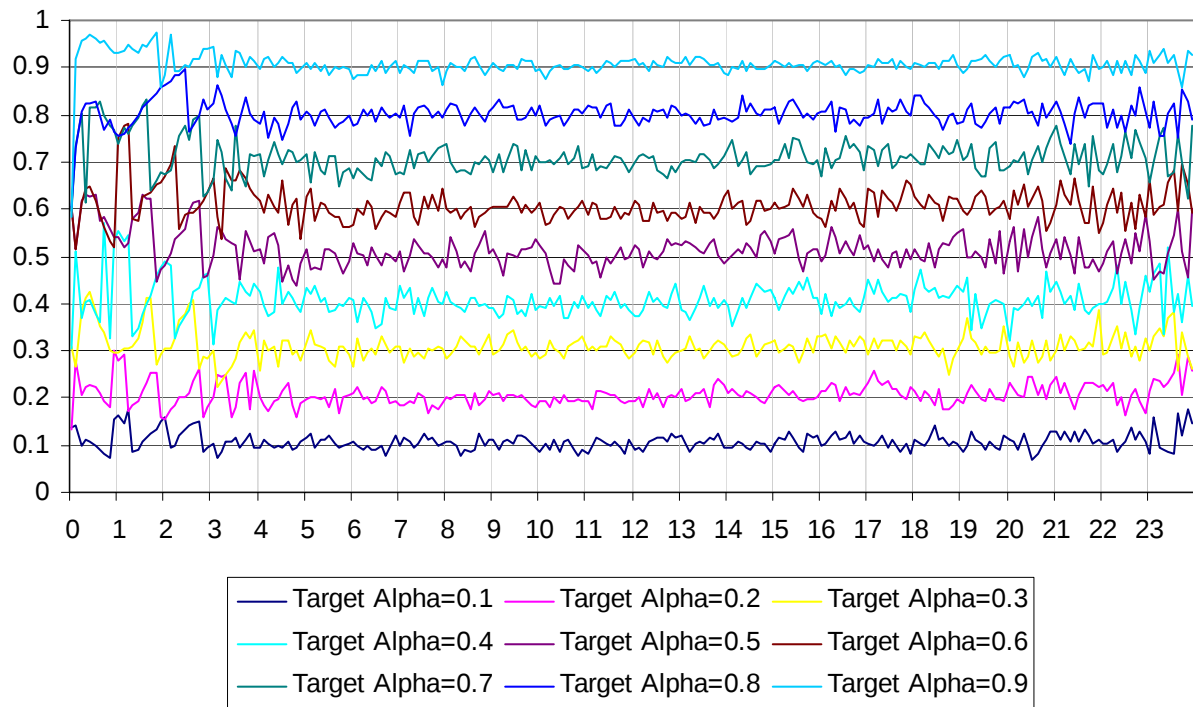
use the Iterative **Simulation-Based Staffing Algorithm**
in Feldman, M., Massey and Whitt, 2005.)

HW/GMR Delay Functions



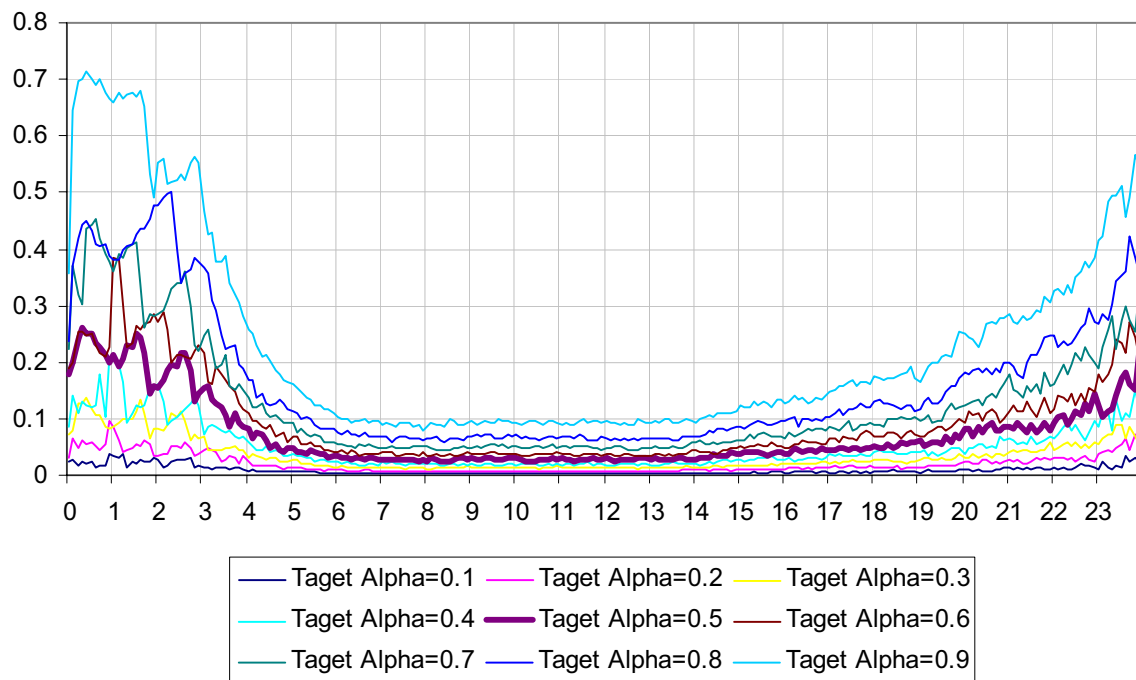
Delay Probability α

Delay Probability

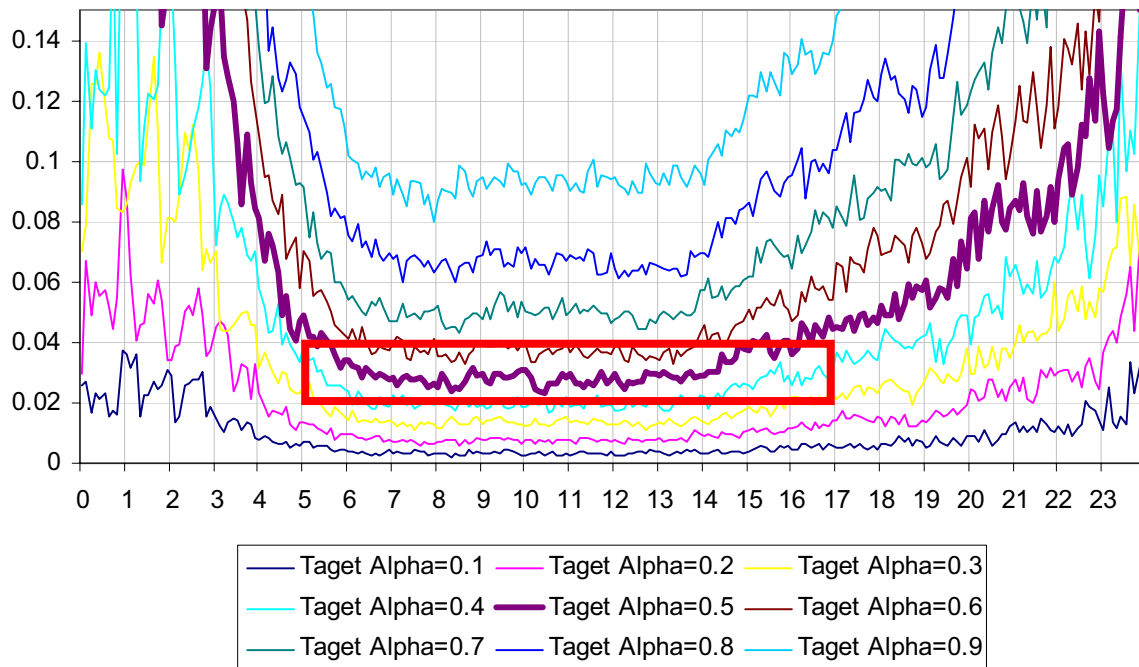


Abandon Probability

Abandon Probability



Abandon Probability

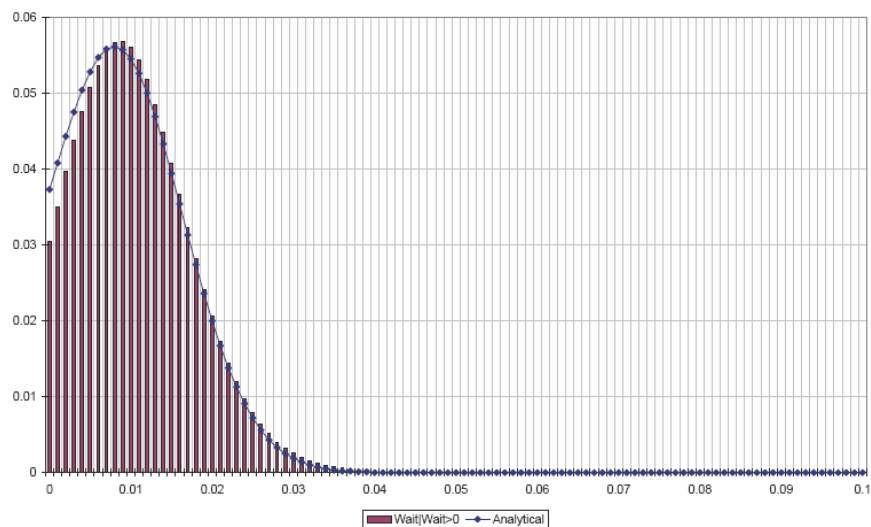
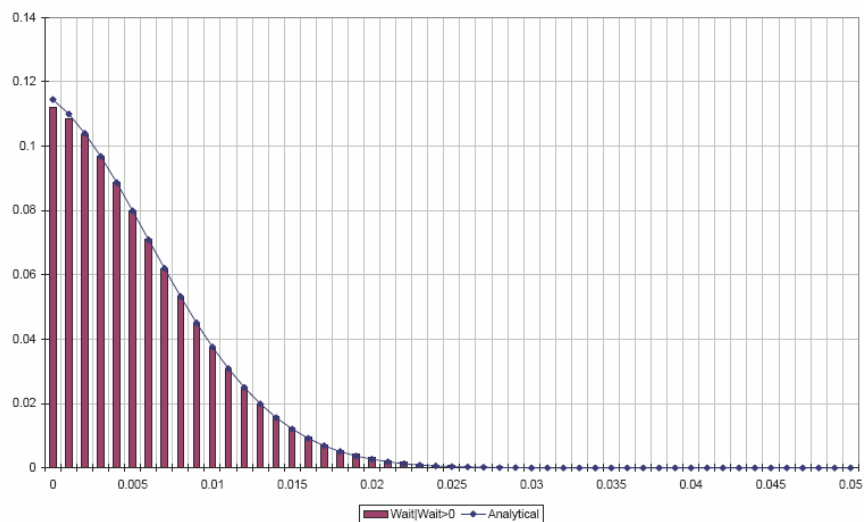
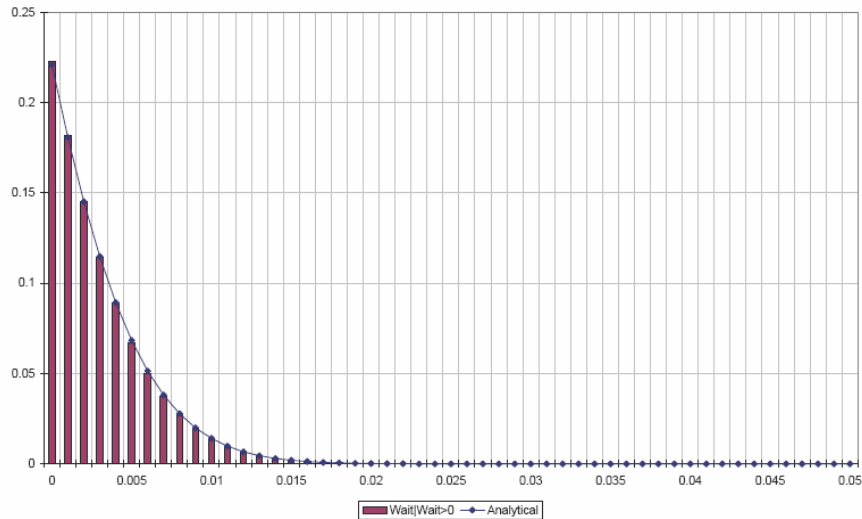


Real Call Center: Empirical waiting time, given positive wait

(1) $\alpha=0.1$ (QD)

(2) $\alpha=0.5$ (QED)

(3) $\alpha=0.9$ (ED)



The "Right Answer" (for the "Wrong Reasons")

Prevalent Practice

$$N_t = \lceil \lambda(t) \cdot E(S) \rceil \quad (\text{PSA})$$

"Right Answer"

$$N_t \approx R_t + \beta \cdot \sqrt{R_t} \quad (\text{MOL})$$

$$R_t = E\lambda(t - S) \cdot E(S)$$

The "Right Answer" (for the "Wrong Reasons")

Prevalent Practice $N_t = \lceil \lambda(t) \cdot E(S) \rceil$ (PSA)

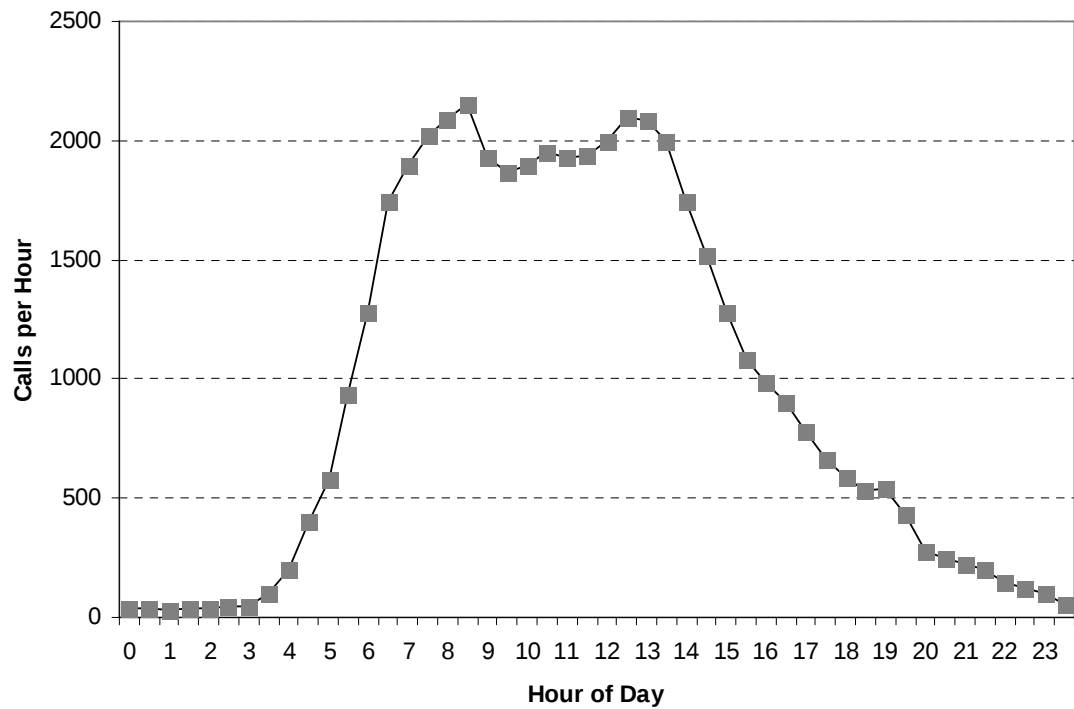
"Right Answer" $N_t \approx R_t + \beta \cdot \sqrt{R_t}$ (MOL)

$$R_t = E\lambda(t - S) \cdot E(S)$$

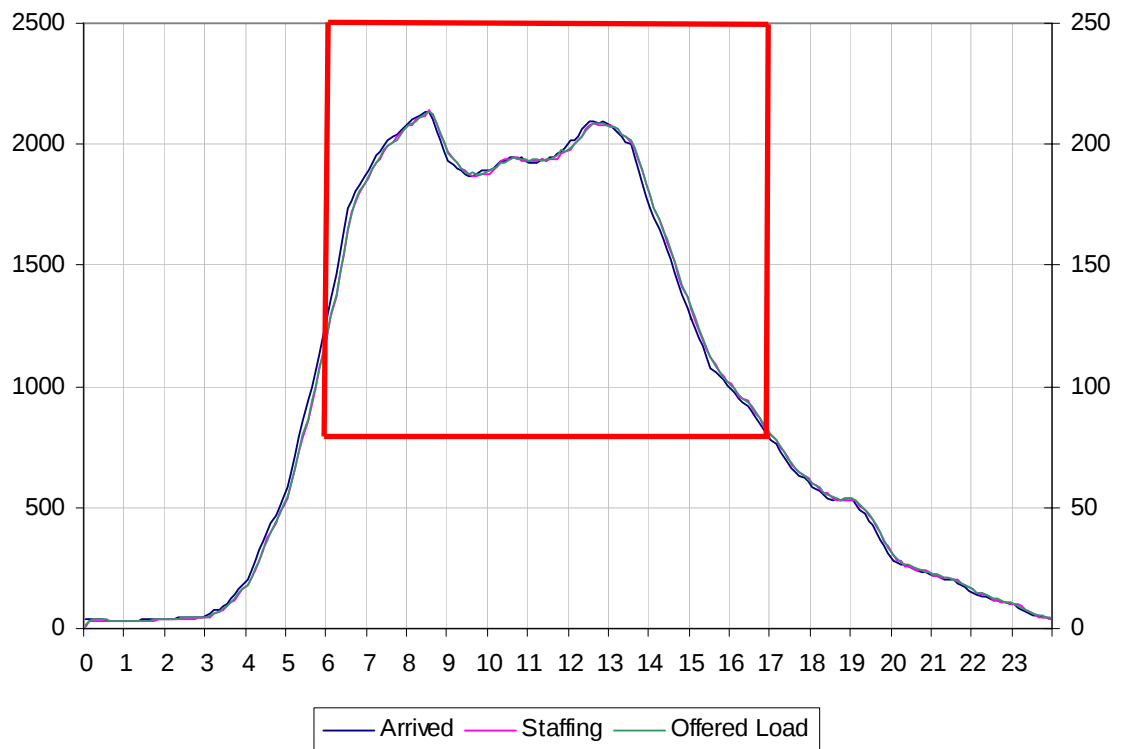
Practice $\approx$ "Right" $\beta \approx 0$ (QED)

and $\lambda(t) \approx$ stable over service-durations

Practice Improved $N_t = \lceil \lambda[t - E(S)] \cdot E(S) \rceil$



QED Staffing ($\beta=0$ iff $\alpha=0.5$)



The "Right Answer" (for the "Wrong Reasons")

Prevalent Practice $N_t = \lceil \lambda(t) \cdot E(S) \rceil$ (PSA)

"Right Answer" $N_t \approx R_t + \beta \cdot \sqrt{R_t}$ (MOL)

$$R_t = E\lambda(t - S) \cdot E(S)$$

Practice $\approx$ "Right" $\beta \approx 0$ (QED)

and $\lambda(t) \approx$ stable over service-durations

Practice Improved $N_t = \lceil \lambda[t - E(S)] \cdot E(S) \rceil$

When **Optimal**? for moderately-patient customers:

1. **Satisfization** $\Leftrightarrow$ At least 50% to be serve immediately
2. **Optimization** $\Leftrightarrow$ Customer-Time = 2 x Agent-Salary

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ **?** θ

?

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

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ **?** θ

?

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

$\mu = \theta$: $L_t \stackrel{d}{=} \text{Poisson}(R_t) \stackrel{d}{\approx} N(R_t, R_t)$, since $M_t / M \rightarrow \infty$

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

offered load

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ **?** θ

?

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

$\mu = \theta$: $L_t \stackrel{d}{=} \text{Poisson}(R_t) \stackrel{d}{\approx} N(R_t, R_t)$, since $M_t / M \rightarrow \infty$

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

Given $L_t \approx R_t + Z\sqrt{R_t}$, $Z \stackrel{d}{=} N(0,1)$

choose $N_t = R_t + \beta \sqrt{R_t}$

$$\Rightarrow \alpha = \underset{\text{PASTA}}{P(W_t > 0)} \approx P(L_t \geq N_t) = P(Z \geq \beta) = 1 - \phi(\beta)$$

$$\Rightarrow \beta = \phi^{-1}(1 - \alpha) \quad \text{time-stable } \alpha \equiv P(W_t > 0) \text{ ?}$$

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ **?** θ

?

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

$\mu = \theta$: $L_t \stackrel{d}{=} \text{Poisson}(R_t) \stackrel{d}{\approx} N(R_t, R_t)$, since $M_t / M \rightarrow \infty$

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

Given $L_t \approx R_t + Z\sqrt{R_t}$, $Z \stackrel{d}{=} N(0,1)$

choose $N_t = R_t + \beta \sqrt{R_t}$

$$\Rightarrow \alpha = \underset{\text{PASTA}}{P(W_t > 0)} \approx P(L_t \geq N_t) = P(Z \geq \beta) = 1 - \phi(\beta)$$

$$\Rightarrow \beta = \phi^{-1}(1 - \alpha) \quad \text{time-stable } \alpha \equiv P(W_t > 0) ?$$

Indeed, but in fact

TIME-STABLE PERFORMANCE

Time-Varying Arrivals: $\sqrt{\cdot}$ Safety-Staffing

Model $M_t / M / N_t + M$

Parameters $\lambda(t)$ μ **?** θ

?

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

$\mu = \theta$: $L_t \stackrel{d}{=} \text{Poisson}(R_t) \stackrel{d}{\approx} N(R_t, R_t)$, since $M_t / M / \infty$

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

Given $L_t \approx R_t + Z\sqrt{R_t}$, $Z \stackrel{d}{=} N(0,1)$

choose $N_t = R_t + \beta \sqrt{R_t}$

$$\Rightarrow \alpha = \underset{\text{PASTA}}{P(W_t > 0)} \approx P(L_t \geq N_t) = P(Z \geq \beta) = 1 - \phi(\beta)$$

$$\Rightarrow \beta = \phi^{-1}(1 - \alpha) \quad \text{time-stable } \alpha \equiv P(W_t > 0) ?$$

Indeed, but in fact

TIME-STABLE PERFORMANCE

($\mu \neq \theta$, or generally : Iterative **Simulation-Based Algorithm**)