QED Q's

Telephone Call/Contact Centers

Service Engineering

Queueing Science

Eurandom

September 8, 2003

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

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

1. Supporting Material (Downloadable)

M. "Call Centers: Research Bibliography with Abstracts." Version 5, July, 2003.

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." Submitted, 2003.

Erlang: "On the rational determination of the number of circuits." In "*The life and works of A.K. Erlang*," 1948.

Jagerman,, "Some properties of the Erlang loss function." The Bell System Technical Journal, 1974.

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." Accepted to QUESTA, 2003.

Whitt: "Heavy-Traffic Limits for the G/H2*/n/m Queue." working paper, 2003.

Borst, M. and Reiman: “Dimensioning Large Telephone Call Centers.” Accepted to *OR, 2003*.

Whitt: “How Multiserver Queues Scale with Growing Congestion-Dependent Demand.” Accepted to OR.

Stone: "Limit theorems for random walks, birth and death processes, and diffusion processes." 1963.

Haji and Newell: "A Relation Between Stationary Queue and Waiting Time Distributions", J. Appl. Prob. 1971.

Puhalskii: "On the invariance principle for the first passage time." MOR, 1994.

Browne and Whitt: "Piecewise-linear diffusion processes." In Advances in Queueing. Theory, Methods, and Open Problems, Ed. Dshalalow, 1995.

Whitt: “A Diffusion Approximation for the G/GI/n/m Queue.” working paper, 2003.

Contents

1. Service Engineering – Research, Teaching, Practice.
2. The World of Call Centers
3. Workforce Management (Staffing): Hierarchical View
4. Operational Regime: Quality-Driven, Efficiency-Driven

   The QED (Halfin-Whitt) Regime
5. Markovian (Birth & Death) Queues
6. Diffusion Limits/Approximations
7. M/M/N (Erlang-C); GI/D/N.

Leading to models with

7. Impatient (Abandoning) Customers
8. Predictably (Time) Varying Queues
9. Heterogeneous Customer Types and

   Partially Overlapping Server Skills

Service Engineering – a Subjective View

- Contrast with the traditional and prevalent

Service Management	(Business Schools; U.S.A.)
Industrial Engineering	(Engineering Schools; Europe)

- Goal: Develop scientifically-based design principles (rules-of-thumb) and tools (software), that support the balance of service quality and efficiency, from the (often conflicting) views of customers, servers and managers.

- Theoretical Framework: Queueing Networks
- Applications focus: Call (Contact) Centers

Example: Staffing

How many agents required for balancing service-quality and operational-efficiency.

Example: Skills-Based Routing (SBR)

VIP and Regulars, seeking Support or Purchasing, via Telephone or IVR or e.mail or Chat.

Staffing (+SBR): How Many Agents?

- Fundamental problem in service operations / call centers:

- People = 70% costs of running call centers, employing
  3% U.S. workforce; 1000's agents in a "single" Call Center.

Reality

- Workforce Management (WFM) is M/M/N-based
- Reality is complex and becoming even more so
- Solutions are urgently needed
- Technology enables smart systems
- Theory lags significantly behind needs

» Ad-hoc methods: heuristics, simulation-based

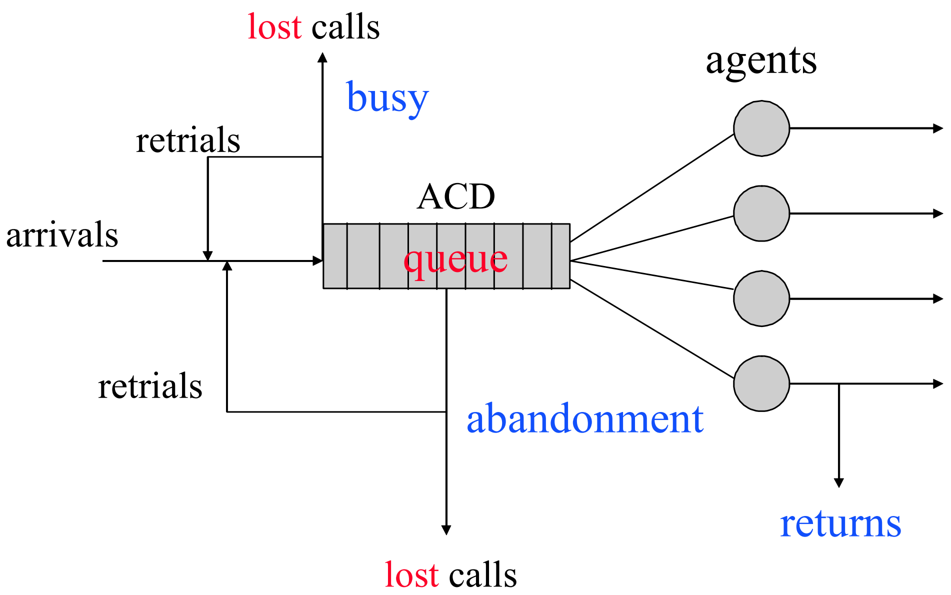
Progress is based on

- Small yet significant models for theoretical insight

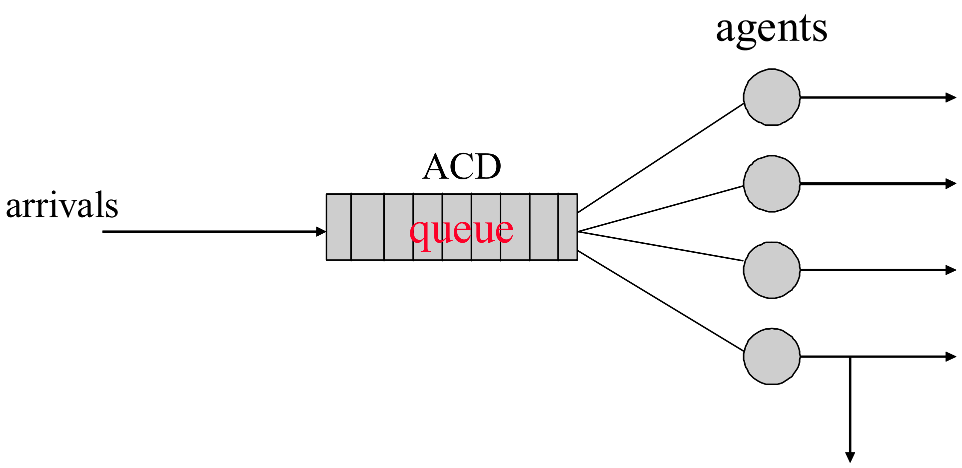
  the research of which gives rise to

- Principles, Guidelines, Tools: Service Engineering

The Basic Call Center



Erlang-C = M/M/N



"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)** = **7:30 minutes** (=**M**)

λ/hr	N Agents	ρ = OCC	L_q = Que	W_q = ASA
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

λ/hr	N	ρ = OCC	L_q = Que	W_q = ASA
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**
↓	↓	↓	↓	↓
∞	∞	**1**	∞	**7:30 !**

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

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

Indeed $\overline{W_q} \approx \overline{W_q} \mid W_q > 0 = \frac{1}{N} \cdot \frac{\rho_N}{1-\rho_N} \cdot E(S) \to E(S) = 7{:}30$ **!**

since $\rho_N = \frac{\lambda_N \times E(S)}{N} = \frac{8(N-1) \times 7.5/60}{N} = \frac{N-1}{N} = 1 - \frac{1}{N}$

$N(1-\rho_N) = 1 \quad , \quad \rho_N \to 1 \ .$

What can be achieved — At what cost

Copy of Summary Interval - Order PK

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

Time	Avg Speed Ans (W)	Avg Aban Time	ACD Calls (λ)	Avg ACD Time (1/μ)	Avg ACW Time	Aban Calls (#Aban)	% ACD Time	% Ans Calls	Avg Pos Staff (N)	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time (P)
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	98	70	149		8		
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	51	7	4	51	2	16	81
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	26	65
5:30 AM*			0			0	0		0	0		33	0	0
6:00 AM*	:00:00		12	:03:21	:00:19	0	21	100	7	2	100	9	2	19
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	29
7:00 AM*	:00:00		62	:03:34	:00:15	0	38	100	21	3	100	13	4	34
7:30 AM*	:00:00		93	:03:11	:00:34	0	36	100	30	3	100	7	4	32
8:00 AM*	:00:00		120	:03:37	:00:40	0	39	100	47	3	100	8	6	33
8:30 AM*	:00:00		193	:03:04	:00:14	0	44	100	61	3	100	10	7	37
9:00 AM*	:00:01		293	:03:25	:00:25	0	54	99	75	4	97	9	7	47
9:30 AM*	:00:02	:00:06	361	:03:45	:00:22	2	60	97	91	4	93	8	8	52
Peak → 10:00 AM*	:00:02	:00:01	418	:03:49	:00:26	1	63	97	94	4	98	5	8	55
10:30 AM*	:00:00		349	:03:35	:00:33	0	52	99	96	4	99	6	8	44
11:00 AM*	:00:00		352	:03:50	:00:27	0	51	100	102	3	100	7	8	45
11:30 AM*	:00:00		349	:03:44	:00:18	0	49	100	97	4	100	8	5	45
12:00 PM*	:00:01		354	:03:59	:00:18	0	52	95	95	4	95	8	5	47
12:30 PM*	:00:00		336	:03:38	:00:21	0	52	99	97	3	99	9	8	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	6	44
1:30 PM*	:00:00		368	: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	50	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:48	:01:37	0	54	100	98	4	100	6	7	47
4:30 PM*	:00:00		379	:03:41	:00:19	0	55	99	97	4	99	8	5	50
5:00 PM*	:00:00		411	:03:53	:00:19	0	53	100	109	4	100	9	5	46
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

Page 1 of 2

Rough Performance Analysis

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

400 calls

3:45 minutes average service time

2 seconds ASA, 1 abandonment (after 1 second)

Offered load **R** $= \lambda \times M$

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

$= 50$ Erlangs

Occupancy $\rho = R/N$

$= 50/100 = 50\%$

⇒ **Quality-Driven Operation** (Light-Traffic)

⇒ Classical Queueing Theory (M/G/N approximations)

Above: $R = 50$, $N = R + 50$, ≈ 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 - but **by how much?**

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

λ/hr	ρ	**$E(W_q)$ = ASA**	**% Wait = 0**
800	50%	0	100%
1000	62.5%	0	100%
1200	75%	0	99.7%
1400	87.5%	0:02 min.	88%
1500	93.8%	0:15 min.	60%
1550	96.9%	0:48 min.	35%
1580	98.8%	2:34 min.	15%
1585	**99.1%**	**3:34 min.**	12%

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

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

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

Changing N (Staffing) in M/M/N

			E(S) = 3:45	
λ/hr	N	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**

Heavy traffic, in the sense that OCC > 95%;

Light traffic, 50% answered **immediately**

QED Regime = Quality- and Efficiency-Driven Regime

Economies of Scale in a Frictionless Environment

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

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

Rules of Thumb: Operational Regimes

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

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

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

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

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

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

$$N = \lceil R + \beta\sqrt{R} \rceil, \quad \beta > 0$$

$\sqrt{\cdot}$ **Safety-Staffing**

Determine Regimes (Strategy), Parameters (Economics)

Strategy: Managers, Agents (Unions), Customers

Economics: Minimize agent salaries + waiting cost

QED Theorem (**Halfin-Whitt**, 1981)

Consider a sequence of M/M/N models, **N=1,2,3,…**

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

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

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

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

Extremes:

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

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

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

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

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

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

Expected Performance:

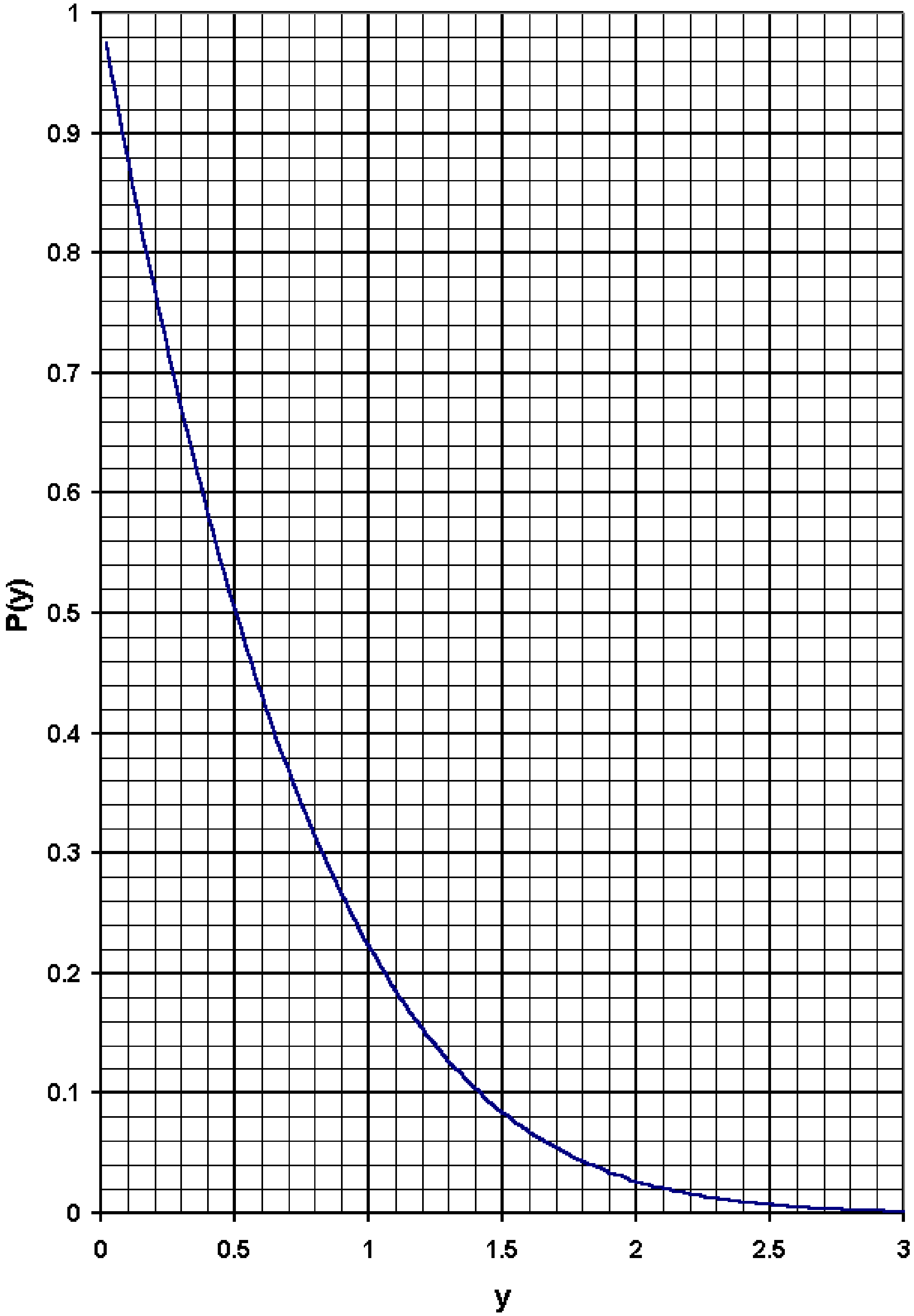
$$\%\text{ Delayed} \approx P(\beta) = \left[1 + \frac{\beta\phi(\beta)}{\varphi(\beta)}\right]^{-1}, \quad \beta > 0$$ Erlang-C

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

The Halfin-Whitt Delay Function *P(β)*



Strategy: Sustain Regime under Pooling

Economies of Scale

Base case: M/M/N with parameters λ, μ, N

Scenario: $\lambda \to m\lambda \quad (R \to mR)$

	Base Case	Efficiency-driven	Quality-driven	Rationalized
Offered load	$R = \frac{\lambda}{\mu}$	mR	mR	mR
Safety staffing	Δ	Δ	$m\Delta$	$\sqrt{m}\Delta$
Number of agents	$N = R + \Delta$	$mR + \Delta$	$mR + m\Delta$	$mR + \sqrt{m}\Delta$
Service grade	$\beta = \frac{\Delta}{\sqrt{R}}$	$\frac{\beta}{\sqrt{m}}$	$\beta\sqrt{m}$	$\boxed{\beta}$
Erlang-C = P{Wait>0}	$P(\beta)$	$P\left(\frac{\beta}{\sqrt{m}}\right) \uparrow 1$	$P(\beta\sqrt{m}) \downarrow 0$	$\boxed{P(\beta)}$
Occupancy	$\rho = \frac{R}{R+\Delta}$	$\frac{R}{R+\frac{\Delta}{m}} \uparrow 1$	$\boxed{\rho = \frac{R}{R+\Delta}}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \uparrow 1$
$\mathrm{ASA} = \mathrm{E}\left[\frac{\mathrm{Wait}}{\mathrm{E}(S)} \middle\vert \mathrm{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \mathrm{ASA}}$	$\frac{1}{m\Delta} = \frac{\mathrm{ASA}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\mathrm{ASA}}{\sqrt{m}}$
$\mathrm{TSF} = \mathrm{P}\left\{\frac{\mathrm{Wait}}{\mathrm{E}(S)} > T \middle\vert \mathrm{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \mathrm{TSF}}$	$e^{-mT\Delta} = (\mathrm{TSF})^m$	$e^{-\sqrt{m}T\Delta} = (\mathrm{TSF})^{\sqrt{m}}$

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 >> D** : Efficiency-driven
- **C << D** : Quality-driven
- **C ≈ D** : Rationalized - QED

Satisfization: N* minimal s.t. Service Constraint

Eg. %Delayed $< \alpha$.

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

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

Asymptotic-Optimality: Framework

Problem: Minimize N = Number of Servers

1. Change of Variables:

Translate the discrete optimization problem "how many agents?" into a continuous optimization problem.

2. Approximation (Asymptotically):

In each of the 3 regimes, approximate (asymptotically) the continuous optimization problem from Step 1 by an "approximating" continuous optimization problem that is easier to solve.

3. Optimality (Asymptotically):

Prove that the optimal solution to the approximating continuous problem from Step 2 provides an approximately (asymptotically) optimal solution to the original discrete optimization problem.

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

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

where **d** = delay/waiting costs

c = staffing costs

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

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

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

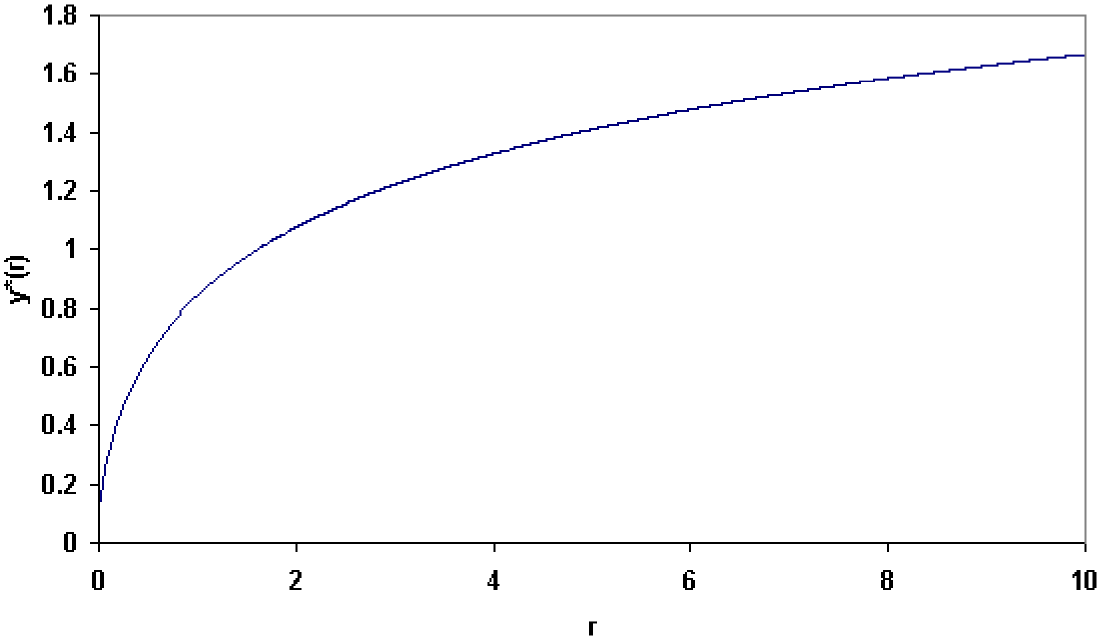
$$\mathrm{P}\{\text{Wait} > 0\} \approx \mathbf{P}(y^*) = \left[1 + \frac{y^* \phi(y^*)}{\varphi(y^*)}\right]^{-1}$$ **Erlang-C**

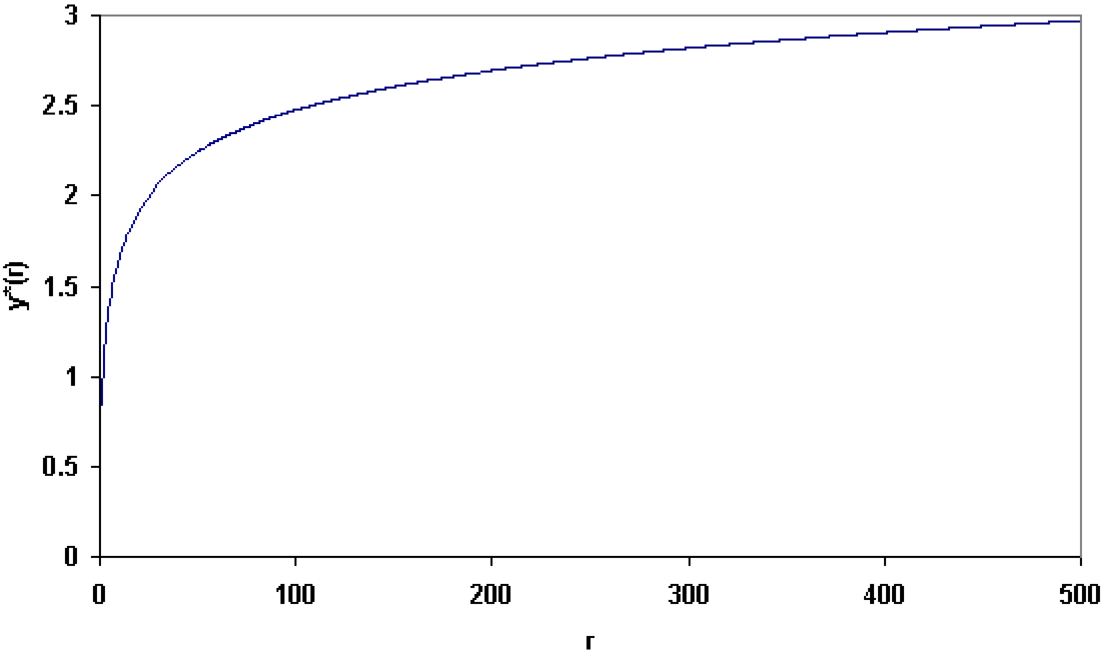
$$\text{TSF} = \mathrm{P}\left\{\frac{\text{Wait}}{\mathrm{E(S)}} > \mathrm{T} \;\middle|\; \text{Wait} > 0\right\} = e^{-\mathrm{T}\Delta}$$

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

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

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





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

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

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

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

Example: 100 calls per minute, at 4 min. per call

$\Rightarrow$ R = 400, least number of agents

$$\frac{\Delta}{R} \approx \frac{y^*(r)}{\sqrt{R}} = \frac{y^*}{20}, \quad \text{with } y^*: 0.5\text{–}1.5 \;;$$

Safety staffing: 2.5%–7.5% of R=Min ! $\Rightarrow$ **"Real" Problem?**

Performance:

N^*	% wait > 20 sec.	Utilization
400 + **11**	20%	97%
400 + **29**	1%	93%

Scenario Analysis: "Satisfization" (vs. Optimization)

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).

(Folklore: $N^* = R + \bar{\phi}^{-1}(\varepsilon)\sqrt{R}$, $\bar{\phi} = 1 - \phi$; based on "classical" normal approximations to infinite-servers models. The two essentially coincide for small ε.)

Example: λ = 1,800 calls at peak hour (avg)

M = 4 min. service time (avg)

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

Service level constraint: less than 15% delayed, equivalently at least 85% answered immediately.

$\Rightarrow N^* = R + P^{-1}(0.15)\sqrt{R} = 120 + 1.22\sqrt{120} = 133$ agents

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

ASA = E[Wait] = **2.7** sec. average wait

ASA | Wait > 0 = **18** sec. average wait of delayed

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: λ = 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.

QED Staffing: State of Art (8/2003)

1. GI/**M**/N $N \approx R + \beta\sqrt{R}, \quad \beta > 0$

   - Conceptual: Erlang; **Halfin-Whitt**
   - Dimensioning: Borst, Reiman

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

   - Conceptual: Garnett, Reiman; Zeltyn; **Whitt**
   - Dimensioning: (Borst, Reiman, Zeltyn) in progress

3. Time-Varying (Non-homogenous Poisson arrivals)

   - Infinite-server heuristics: Jennings, Massey, Whitt
   - Conceptual: (Massey, Rider) in progress
   - Dimensioning: ?

4. Skills-Based Routing:

   - Conceptual: Atar, Reiman; Gurvich (V-Model)
   - Dimensioning: **Borst, Seri** (General); Gurvich (V);

     Armony (Reversed-V)**;**

5. Service Time Duration:

   - Conceptual: Whitt H2*/G; Jelenkovic, Momcilovic D

QED M/G/N: ???

E(Wq|Wq>0) vs. β