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.

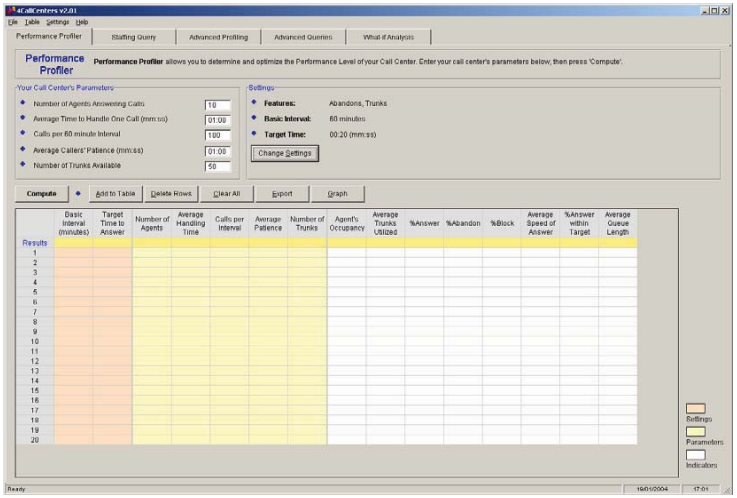
Halfin and Whitt: "Heavy Traffic Limits for Queues with Many Exponential Servers." OR, 1981.

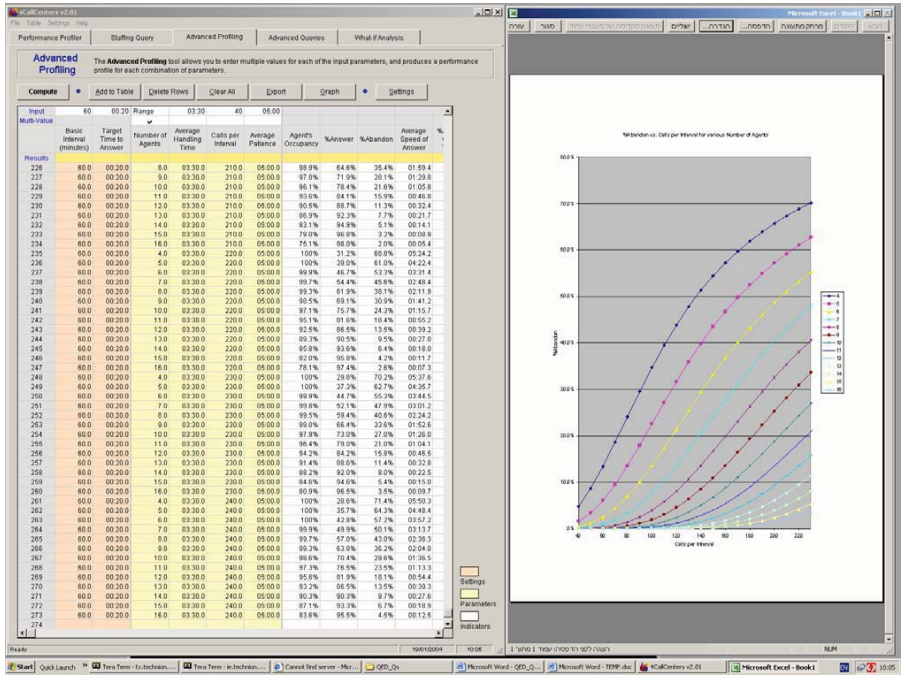
Jelelnkovic, M. and Momcilovic: "Heavy Traffic Limits for Queues with Many Deterministic Servers." QUESTA, 2004.

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

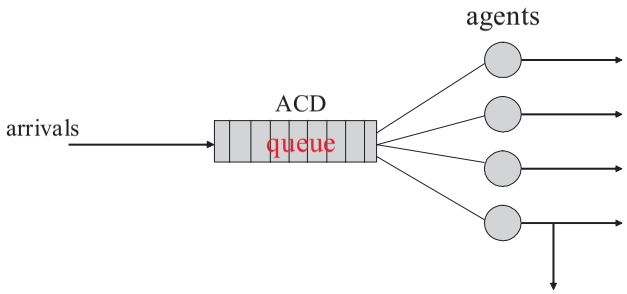
Whitt's website: both the ED and QED regimes.

4CallCenters: Personal Tool for WFM

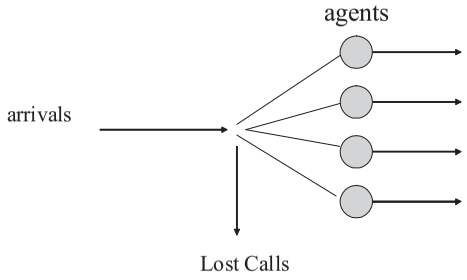




Erlang-C (M/M/N): # Agents



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 ⇒ **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 !**

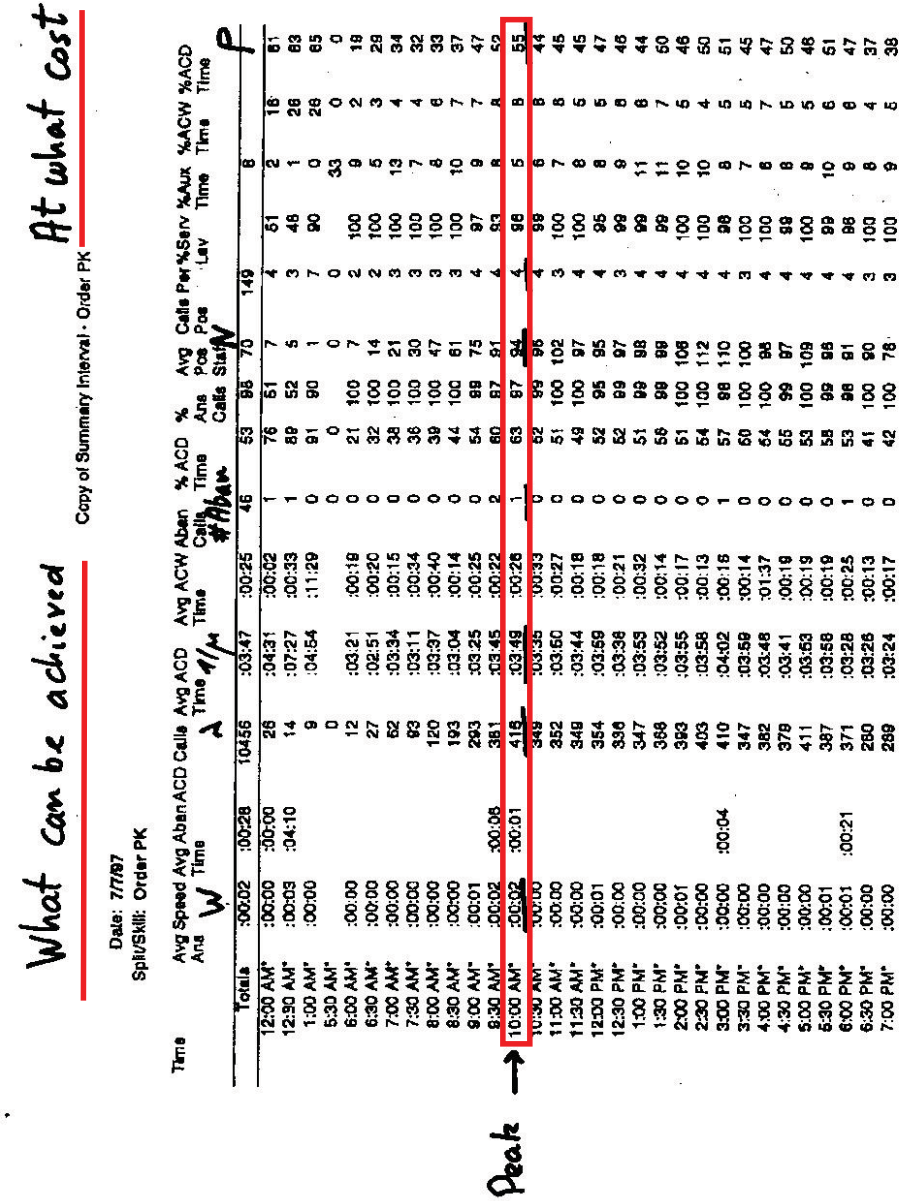
⇒ **Efficiency-Driven Operation** **(Heavy-Traffic)**

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

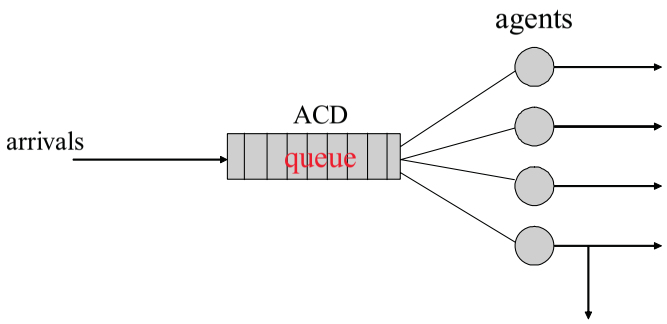
Indeed $\overline{W}_q \approx \overline{W}_q \mid W_q > 0 = \frac{1}{N} \cdot \frac{\rho_N}{1-\rho_N} \cdot E(S) \to E(S) = 7{:}30$!

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

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



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

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

$= 50$ 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 $R = \lambda \times E(S)$

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

$= 50$ 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	$E(W_q)$ = 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	$E(W_q)$ = 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) = 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)**

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

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

Above: R = 99, N = R + 1, ≈ **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	N	OCC	ASA	% Wait = 0
1585	100	99.1%	3:34	12%

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

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%

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

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%**

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

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**

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,…**

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

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

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

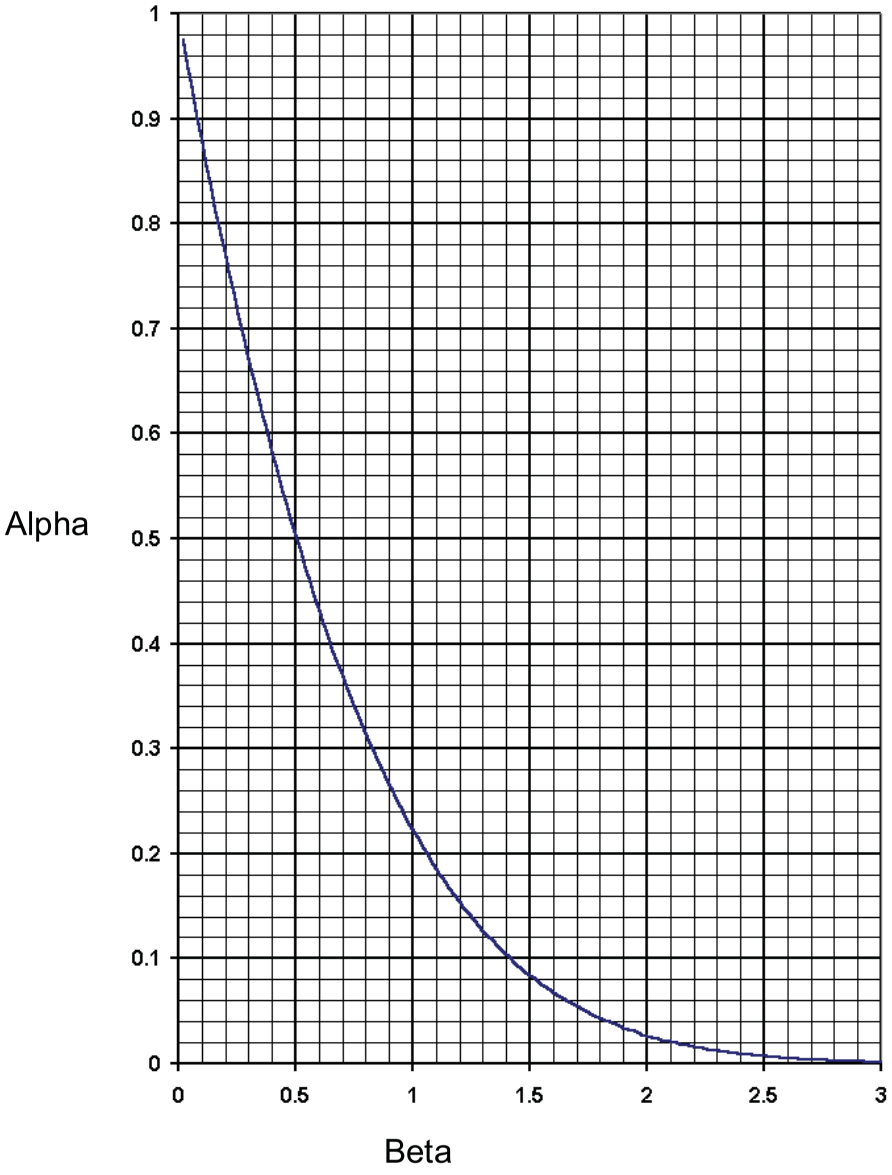
where $\varphi(\cdot)/\phi(\cdot)$ is the standard normal density/distribution.

Extremes:

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

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

The Halfin-Whitt Delay Function



√· Safety-Staffing: Performance

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

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

$= R + \Delta$ √· 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

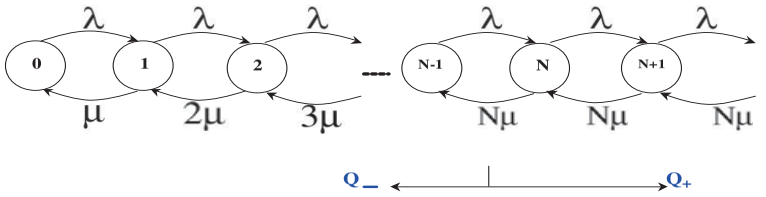
QED : Some Intuition (Assume $\mu = 1$)

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

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

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

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 $\sqrt{N}\,(1-\rho_N) \to \beta, \ -\infty < \beta < \infty.$

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

Rules of Thumb: Operational Regimes

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

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

$N = \lceil R + \gamma \rceil$, $\gamma > 0$ service grade

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

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

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

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

Determine Regimes (Strategy), Parameters (Economics)

Strategy: Managers, Agents (Unions), Customers

Economics: Minimize agent salaries + waiting cost

Strategy: Sustain Regime under Pooling

Base: $\lambda = 300$/hr, AHT = 5 min, N = 30 agents

$R = 300 \times \frac{5}{60} = 25$, OCC = 83.3% ASA = 15 sec

$y = (N - R)/\sqrt{R} = (30 - 25)/\sqrt{25} = 1$, P(1) = 22%

4 CC: $\lambda = 1200$, AHT = 5, R = 100; N=?

Quality-Driven: maintain OCC at 83.3%.

N = 120, ASA = .5 sec, $y = (120 - 100)/10 = 2$

Efficiency-Driven: maintain ASA at 15 sec.

N = 107, OCC = 95%, y = 0.8

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

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

9 CC: $\lambda = 2700$, AHT = 5, R = 225

Q: N = 270

E: N = 233

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

Strategy: Sustain Regime under Pooling

Economies of Scale

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

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

	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	$\rho = \frac{R}{R+\Delta}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \uparrow 1$	$\boxed{\rho = \frac{R}{R+\Delta}}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \uparrow 1$
ASA = $\mathrm{E}\left[\frac{\text{Wait}}{\mathrm{E}(S)} \middle\vert \text{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \text{ASA}}$	$\frac{1}{m\Delta} = \frac{\text{ASA}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\text{ASA}}{\sqrt{m}}$
TSF = $\mathrm{P}\left\{\frac{\text{Wait}}{\mathrm{E}(S)} > T \middle\vert \text{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \text{TSF}}$	$e^{-mT\Delta} = (\text{TSF})^m$	$e^{-\sqrt{m}T\Delta} = (\text{TSF})^{\sqrt{m}}$

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

Satisfization: N* minimal s.t. Service Constraint

Eg. %Delayed < α .

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

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

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

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

where **d** = delay/waiting costs

c = staffing costs

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

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

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

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

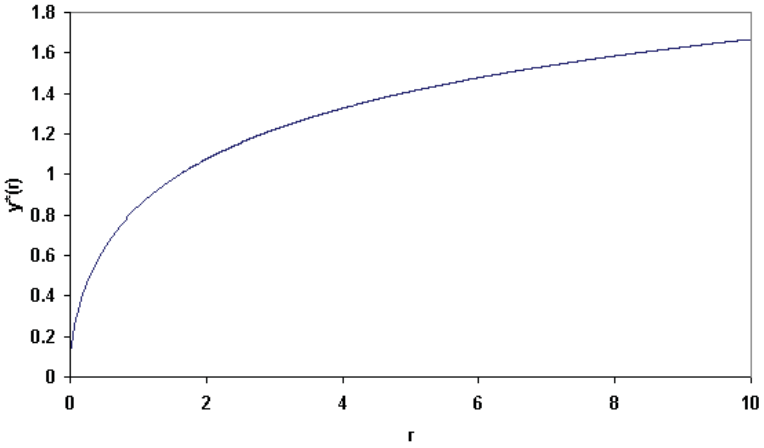
$\text{TSF} = P\left\{\frac{\text{Wait}}{E(S)} > T \,\middle|\, \text{Wait} > 0\right\} = e^{-T\Delta}$

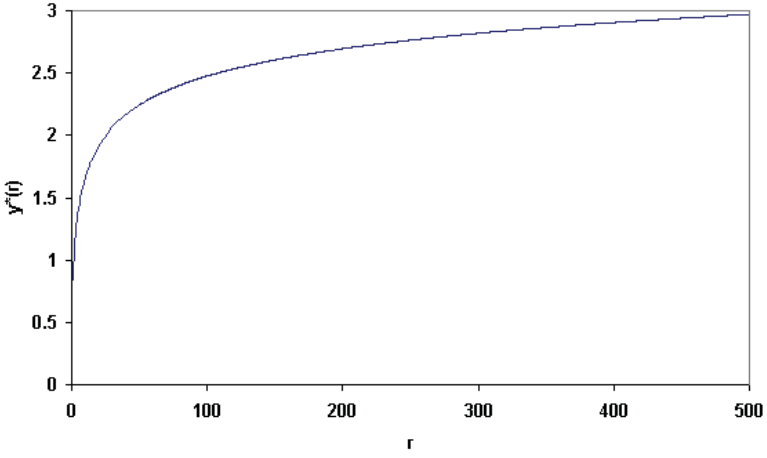
$\text{ASA} = E\left[\frac{\text{Wait}}{E(S)}\middle|\text{Wait} > 0\right] = \frac{1}{\Delta}$

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

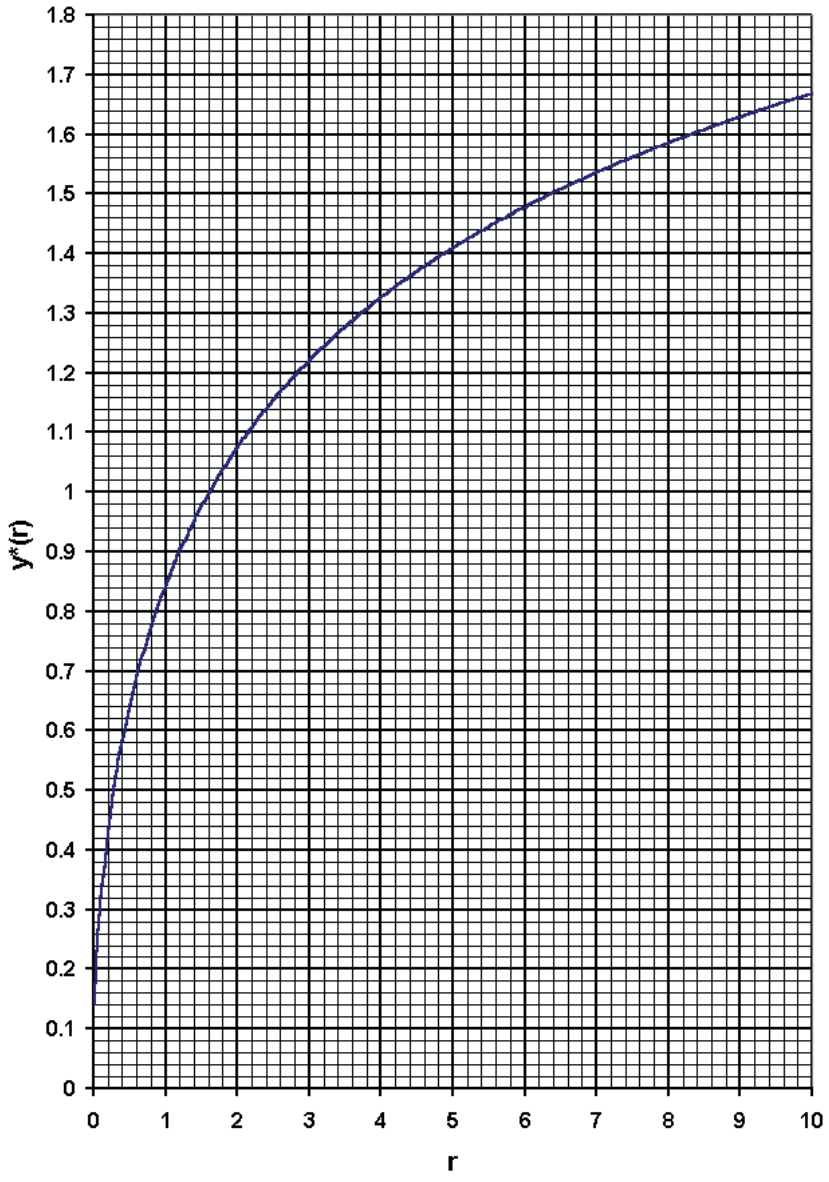
Square-Root Safety Staffing: $N = R + y^*(r)\sqrt{R}$

r = cost of delay / cost of staffing





$y^*(r)$, ***r* = cost of delay / 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: $\mathbf{N = R + 0.5\sqrt{R}}$

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)

$y^*(\frac{d}{c}) = 0.53$, 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 (idelness) = 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: 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: 1 out of 100 delayed (avg), namely 99% answered immediately.

$\Rightarrow N^* = R + P^{-1}(0.01)\sqrt{R} = 120 + 2.38\sqrt{120} = 146$ agents

$\Rightarrow \frac{d}{c} = (y^*)^{-1}(2.38) = 75$: 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.**