Service Engineering

Class 12

QED (QD, ED) Queues
Erlang-B/C: Some Proofs, Facts and Analysis

- Erlang-B in the QED-Regime (Jagerman);
- Erlang-C in the QED-Regime (Halfin & Whitt);
- QED Erlang-C: Some Intuition;
- Erlang-C in the ED-Regime;
- Conceptual Framework;
- Pooling;
- Cost Optimization for Erlang-C (with Borst & Reiman);
- Constraint-Satisfaction; The 80-20 Rule.

The Erlang-B Queue in the QED-Regime

Recall the **Erlang-B Formula:**

$$E_{1,n} \triangleq \mathrm{P}\{\text{Blocked}\} = \frac{R^n}{n!} \Big/ \sum_{j=0}^{n} \frac{R^j}{j!}$$

Consider a sequence of **M/M/n/n** queues,
indexed by the number of servers $n = 1, 2, \ldots$.

- λ_n = arrival-rate, varies with n;
- μ = service-rate, fixed (independent of n).
- $R_n = \lambda_n/\mu$ (Offered Load) ; $\rho_n = R_n/n$ (Load per Server);
  We shall use R and ρ, without the subscript n, for simplicity.

<u>**Theorem**</u> (**QED Erlang-B**; Jagerman, 1974)

As $n \to \infty$, the following 3 statements are equivalent:

1. Customers: $E_{1,n} \approx \frac{\gamma}{\sqrt{n}}$, for some $\gamma > 0$;

2. Servers: $\rho \approx 1 - \frac{\beta}{\sqrt{n}}$, for some $-\infty < \beta < \infty$;

3. Manager: $n \approx R + \beta\sqrt{R}$ (square-root "staffing");

in which case

$$\gamma = h(-\beta) = \frac{\phi(-\beta)}{\bar{\Phi}(-\beta)} = \frac{\phi(\beta)}{\Phi(\beta)},$$

where $\phi, \Phi, \bar{\Phi}$ and h are the density, cdf, survival function and hazard rate of $N(0,1)$ (standard-normal), respectively.
Note: **Servers' Occupancy** $\approx 1 - \frac{\beta+\gamma}{\sqrt{n}}$, accounting for blocking.

QED Erlang-B: Proof

Proof:

2 $\iff$ **3** is straightforward algebra from the definitions.

3 $\Rightarrow$ **1**. Assume $n = R + \beta\sqrt{R}$. The key observation is a **Poisson-Representation** of the Erlang-B Formula:

$$E_{1,n} = \frac{\mathrm{P}\{X_R = n\}}{\mathrm{P}\{X_R \le n\}},$$

where $X_R \stackrel{d}{=} \text{Poisson}(R)$.

$$\begin{aligned} \mathrm{P}\{X_R \le n\} &= \mathrm{P}\left\{\frac{X_R - R}{\sqrt{R}} \le \frac{n-R}{\sqrt{R}}\right\} \\ &\stackrel{CLT,3}{\approx} \mathrm{P}\{N(0,1) \le \beta\} = \Phi(\beta)\,. \end{aligned}$$

$$\begin{aligned} \mathrm{P}\{X_R = n\} &= \mathrm{P}\{n-1 < X_R \le n\} \\ &= \mathrm{P}\left\{\frac{n-R-1}{\sqrt{R}} < \frac{X_R - R}{\sqrt{R}} \le \frac{n-R}{\sqrt{R}}\right\} \\ &\approx \mathrm{P}\{\beta - \frac{1}{\sqrt{R}} \le N(0,1) \le \beta\} \\ &\approx \frac{1}{\sqrt{R}} \cdot \phi(\beta) \approx \frac{1}{\sqrt{n}} \cdot \phi(\beta)\,. \end{aligned}$$

Finally, $\frac{\phi(\beta)}{\Phi(\beta)} = \frac{\phi(-\beta)}{1-\Phi(-\beta)} = h(-\beta)$, by the symmetry of $N(0,1)$.

QED Erlang-B: Proof (Continued)

1 ⇒ 3. $n = R + \beta\sqrt{R} + o(\sqrt{R})$ iff
$\forall \epsilon > 0, \quad R + (\beta - \epsilon)\sqrt{R} \le n \le R + (\beta + \epsilon)\sqrt{R}$ for large enough n.

Assume **3** does not hold. This implies that along some subsequence:

$$n > R + (\beta + \epsilon)\sqrt{R}\,.$$

$E_{1,n}$ decreasing in $n \quad \Rightarrow \quad \limsup \sqrt{n}E_{1,n} < h(-\beta - \epsilon)$.

$h(\cdot)$ increasing function $\quad \Rightarrow \quad h(-\beta - \epsilon) < h(-\beta)$

$\Rightarrow$ Contradicting **1**. **q.e.d.**

Erlang-C: Previously Known Facts

Recall:

1. The **Erlang-C Formula**:

$$E_{2,n} \triangleq \mathrm{P}\{W_q > 0\} = \sum_{i \ge n} \pi_i = \frac{R^n}{n!} \frac{1}{1-\rho} \cdot \pi_0 \,,$$

where

$$\pi_0 = \left[\sum_{j=0}^{n-1} \frac{R^j}{j!} + \frac{R^n}{n!(1-\rho)} \right]^{-1} .$$

2. **Palm's Relation** between Erlang-C and Erlang-B:

$$E_{2,n} = \frac{E_{1,n}}{(1-\rho) + \rho E_{1,n}} \,.$$

3. The **Waiting-Time** distribution:

$$\frac{W_q}{1/\mu} = \begin{cases} 0 & \text{wp} \;\; 1 - E_{2,n} \\ \exp\left(\text{mean} = \frac{1}{n} \cdot \frac{1}{1-\rho}\right) & \text{wp} \;\; E_{2,n} \end{cases}$$

The Erlang-C Queue in the QED-Regime

Theorem (**QED Erlang-C**; Halfin & Whitt, 1981)

As $n \to \infty$, the following 4 statements are equivalent:

0. QED: $E_{2,n} \approx \alpha$, for some $0 < \alpha < 1$;

1. Manager: $n \approx R + \beta\sqrt{R}$, for some $0 < \beta < \infty$;

2. Servers: $\rho \approx 1 - \frac{\beta}{\sqrt{n}}$;

3. Customers: $\mathrm{E}[W_q | W_q > 0] \approx \frac{1}{\sqrt{n}} \cdot \frac{1}{\mu\beta}$;

in which case

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

which we call the **Halfin-Whitt Delay-Function**.

Note: $\beta\sqrt{R} \;=\;$ **Safety-Staffing**, in analogy to Safety-Stock.

Proof:

1 $\Longleftrightarrow$ **2** as in Erlang-B.

0 $\Longleftrightarrow$ **2** is a consequence of Palm's relation and QED Erlang-B:

$$E_{2,n} \;=\; \frac{E_{1,n}}{(1-\rho) + \rho E_{1,n}}$$

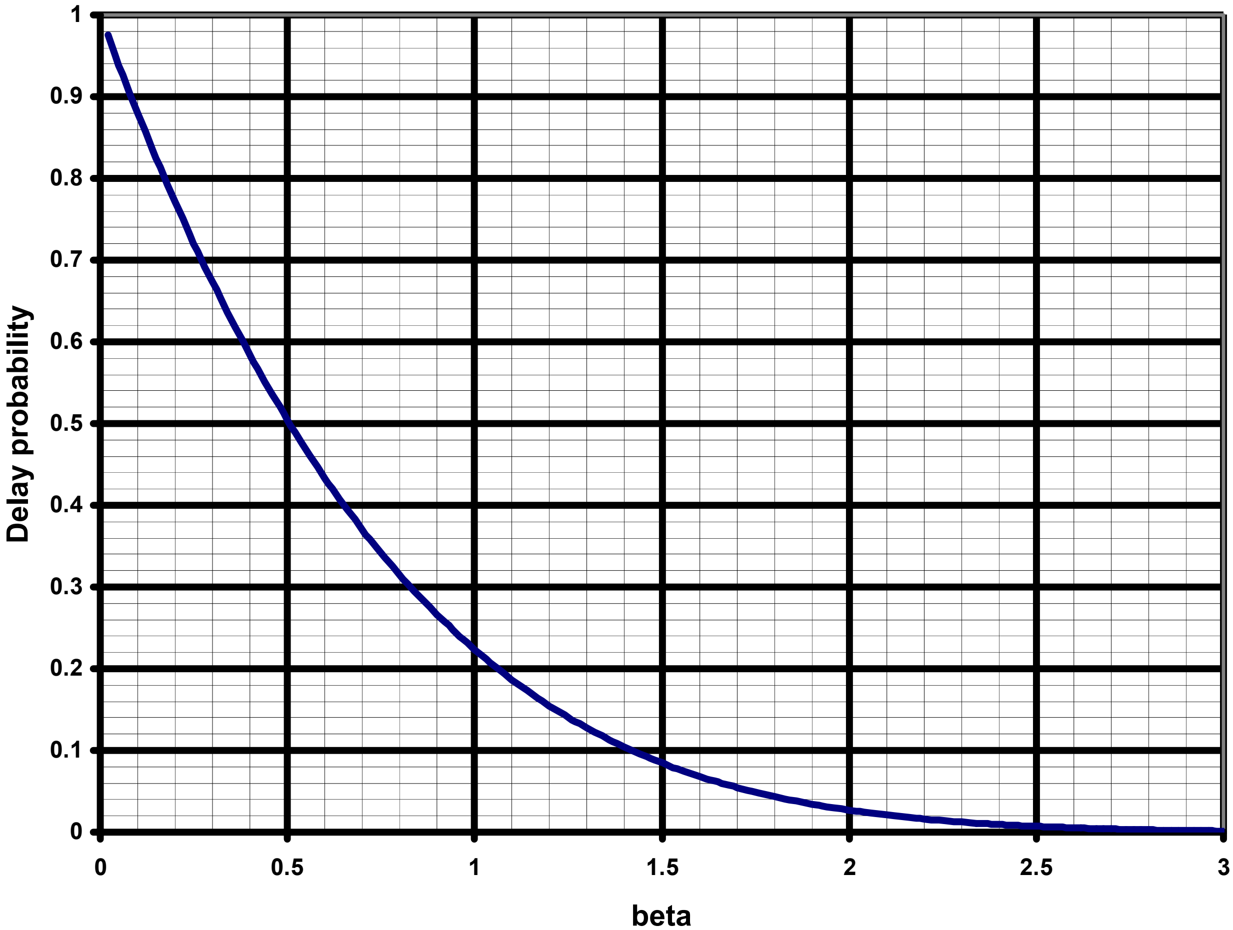
$$\approx\; \frac{h(-\beta)/\sqrt{n}}{\beta/\sqrt{n} + h(-\beta)/\sqrt{n}} \;=\; \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}.$$

Finally, **3** $\Longleftrightarrow$ **2** by the Waiting-Time distribution of Erlang-C.

q.e.d.

The Halfin-Whitt Delay-Function

$$E_{2,n} \triangleq \mathrm{P}\{W_q > 0\} \approx \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}$$



- $\beta = 0.5$ (safety-staffing $= 0.5 \cdot \sqrt{R}$) $\Rightarrow$ $\mathrm{P}\{W_q > 0\} \approx 0.5$;
- $\beta = 2$ (safety-staffing $= 2 \cdot \sqrt{R}$) $\Rightarrow$ $\mathrm{P}\{W_q > 0\} \approx 0.02$;
- $\beta = 3$ $\Rightarrow$ $\mathrm{P}\{W_q > 0\} \approx 0$, QD Regime;

For example, with offered-loads

- $R = 100$: 100+5=105 and 100+20=120;
- $R = 1000$: 1000+16=1016, and 1000+63=1063.

QED Erlang-C: Exact 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

QED Erlang-C: Intuition via Waiting-Time

- Recall: The **Waiting-Time** distribution is given by

$$\frac{W_q}{E(S)} = \begin{cases} 0 & \text{wp } \ 1 - E_{2,n} \ ; \\ \exp\left(\text{mean} = \frac{1}{n} \cdot \frac{1}{1-\rho}\right) & \text{wp } \ E_{2,n} \ . \end{cases}$$

- **Given** $\{W_q > 0\}$, the distribution of W_q is thus Exponential, with mean

$$\mathrm{E}(S)\frac{1}{n}\frac{1}{1-\rho}$$

.

- In the **QED-Regime**: $\sqrt{n} \cdot (1-\rho) \approx \beta$.

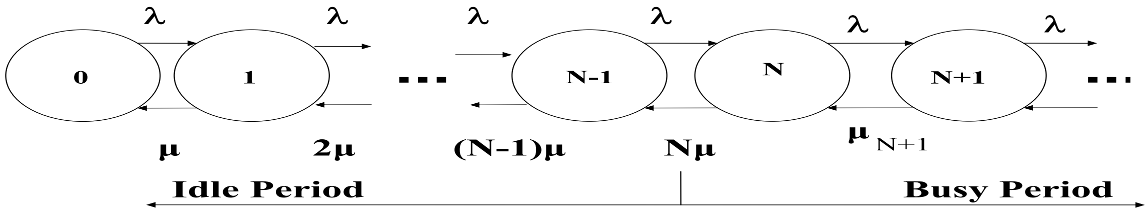
- Hence, given $\{W_q > 0\}$, the distribution of W_q is approximately Exponential, with mean

$$\mathrm{E}(S)\frac{1}{\sqrt{n}}\frac{1}{\beta}.$$

- In particular, with say n=100's, average waiting time is one order of magnitude less than average service time.

Still unclear:
In the QED-Regime, why is the delay probability α **strictly** between 0 and 1? Answer via Busy- and Idle-Period analysis.

Excursions: Busy- & Idle-Periods



Define: Idle Period

$T_{N-1,N} =$ E $\left[1^{st} \text{ hitting time of } N | Q(0) = N-1 \right]$.

Then
$$T_{N-1,N} = \frac{\sum_{i=0}^{N-1} \pi_i}{\lambda_{N-1}\pi_{N-1}} = \frac{1}{\lambda \pi_-(N-1)},$$

where π_- is the distribution of the restricted Q_-.

Similarly: Busy Period

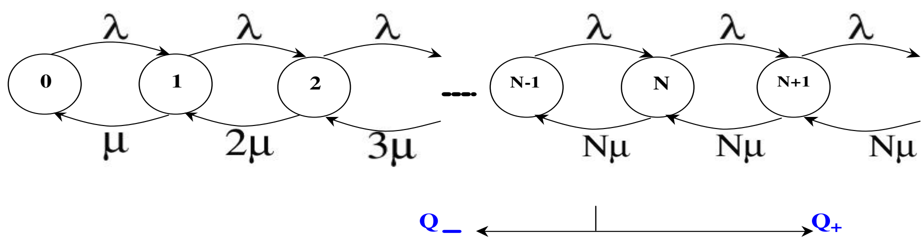
$T_{N,N-1} =$ E $\left[1^{st} \text{ hitting time of } N-1 | Q(0) = N \right]$.

Proof:

Number of Idle Excursions $\stackrel{d}{=} Geometric_{\geq 0}\left(\frac{\lambda_{N-1}}{\lambda_{N-1}+\mu_{N-1}}\right)$

$$T_{N-1,N} = \underbrace{\frac{1}{\pi_-(N-1)\mu_{N-1}}}_{E(Idle\ Excursion)} \quad \times \quad \underbrace{\frac{\mu_{N-1}}{\lambda_{N-1}}}_{E(\#\ of\ Excursions)}$$

QED Erlang-C: Why $0 < \alpha < 1$? Intuition via Busy-Idle Periods



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

QED: $N \sim R + \beta\sqrt{R}$ for some β, $0 < \beta < \infty$

$\Leftrightarrow$ $\lambda_N \sim \mu N - \beta\mu\sqrt{N}$

$\Leftrightarrow$ $\rho_N \sim 1 - \frac{\beta}{\sqrt{N}}$, namely $\lim_{N\to\infty} \sqrt{N}\,(1-\rho_N) = \beta$.

Theorem (Halfin-Whitt, 1981) QED $\Leftrightarrow$ $\lim_{N\to\infty} E_{2,N} = \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}$.

Erlang-C in the ED-Regime

Assume "stingy" safety-staffing: $\boldsymbol{n = R + \gamma, \quad \gamma > 0}$.

Then

1. $n \cdot (1 - \rho) = \gamma$,

2. $\mathrm{P}\{W_q > 0\} \approx 1$,

3. $W_q \overset{d}{\approx} \exp(\gamma\mu) \quad \left(\Rightarrow \frac{E[W_q]}{E[S]} = \frac{1}{\gamma} \right.$: think $\gamma = 1, 2, \ldots, 10, \ldots)$

Example (via 4CallCenters)

$\mathrm{E}[S] = 6$ min $(\mu = 10)$, $\gamma = 1$.

λ/hr	n	ρ	$\mathrm{P}\{W_q > 0\}$	$\mathrm{E}[W_q]$
10	2	50%	33.3%	2:00
50	6	83.3%	58.8%	3:32
250	26	96.2%	78.2%	4:42
1000	101	99%	88.3%	5:18
9000	901	99.9%	95.9%	5:45
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
∞	∞	1	1	6:00

Note:

- $\mathrm{E}[W_q | W_q > 0]$ remains **constant** (6:00).
- Very **sensitive**: decrease n by merely **1** $\Rightarrow$ queue "explodes".

A Conceptual Framework

How to determine the Regime?

Strategy, accounting for tradeoff between efficiency and service quality; or for union-constraints; or for managerial constraints; or,...

How to determine the parameters?

Analysis, via **Constraint Satisfaction** or **Cost/Profit Optimization**.

In principle, can do an analysis with **4CallCenters**.

One then gets the answers but typically these **lack insight**.

Ideally, combine 4CallCenters with ED/QD/QED guidelines.

We shall now demonstrate all this through **examples**.

- Strategy: via Pooling
- Constraint Satisfaction (easy, prevalent)

$$\begin{aligned} \min n \ s.t. \ & P_n\{W_q > T\} \le a \\ & E_n[W_q] \le b \\ & P_n\{Ab\} \le c \end{aligned}$$

- Cost / Profit Optimization

QED Erlang-C: Pooling ($y \leftrightarrow \beta$)

Base: λ = 300/hr, AHT = 5 min, N = 30 agents

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

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

4 CC: λ = 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).

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

9 CC: λ = 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

QED Erlang-C: Pooling Theoretical Support

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

Erlang-C: Cost- or Profit-Optimization

Suppose that revenues depend only on the number of served customers (eg. linearly, or fixed per call). Now observe that, for Erlang-C in steady-state, all customers are eventually served. It follows that staffing levels do not effect revenues. Hence,
profit-maximization is equivalent to cost-minimization.

Conceptual Framework:

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

Mathematical Framework:
Asymptotic Analysis, as the number-of-servers $n \uparrow \infty$.
(Reference: with Borst & Reiman, 2004)

Erlang-C: Cost Minimization

(Reference: Borst, M., Reiman, 2004)

$\mathbf{Cost} = \boldsymbol{c} \cdot \boldsymbol{n} + \boldsymbol{d} \cdot \boldsymbol{\lambda}\mathbf{E}[\boldsymbol{W_q}]$ **,**

c = Staffing cost;
d = Delay cost.

Optimal staffing level:

$$n^* \approx R + \beta^*(r)\sqrt{R}, \qquad r = \text{delay-cost / staffing-cost} .$$

$\beta^*(r)$ = optimal service-grade, independent of λ:

$$\beta^*(r) = \arg \min_{0<y<\infty} \left\{ y + \frac{r \cdot P_w(y)}{y} \right\} ,$$

where

$$P_w(y) = \left[1 + \frac{y}{h(-y)} \right]^{-1} .$$

Very good approximation:

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

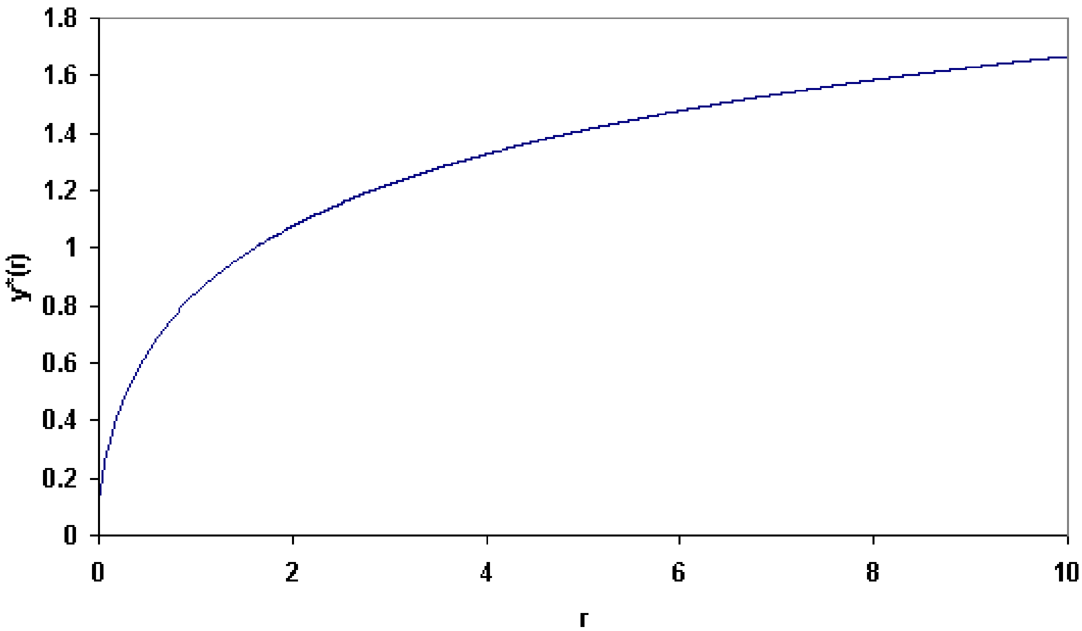
$$\approx \left(2 \ln \frac{r}{\sqrt{2\pi}} \right)^{1/2} , \qquad r \geq 10 .$$

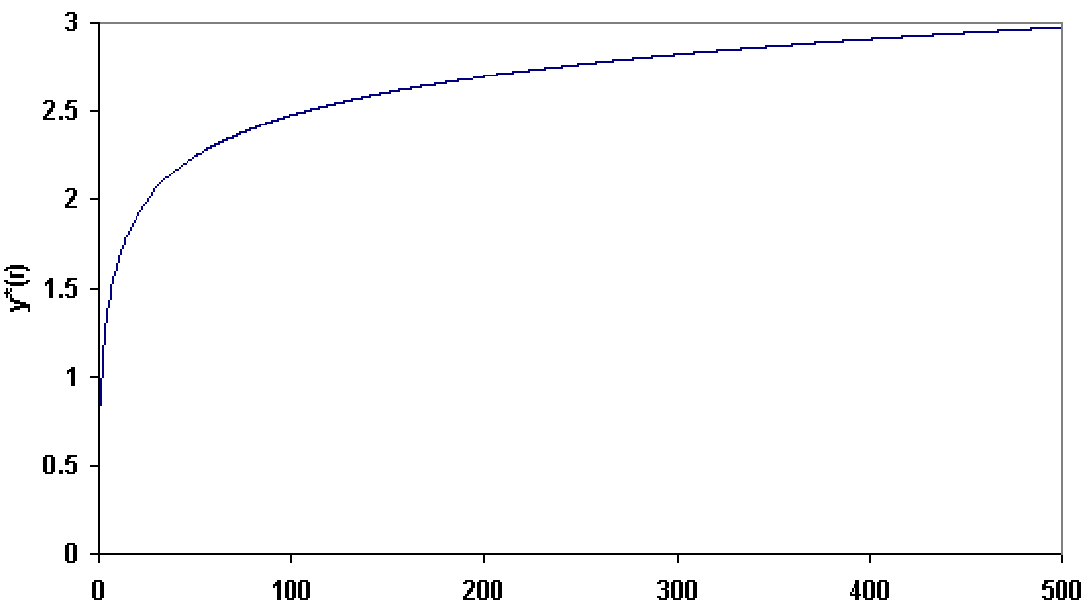
Final comment: r small (large) $\Rightarrow$ ED (QD).

Erlang-C: Optimal Square-Root Staffing

$$n = R + \beta^*(r)\sqrt{R} \qquad (\beta^* \leftrightarrow y^*)$$

r = cost-of-delay / cost-of-staffing

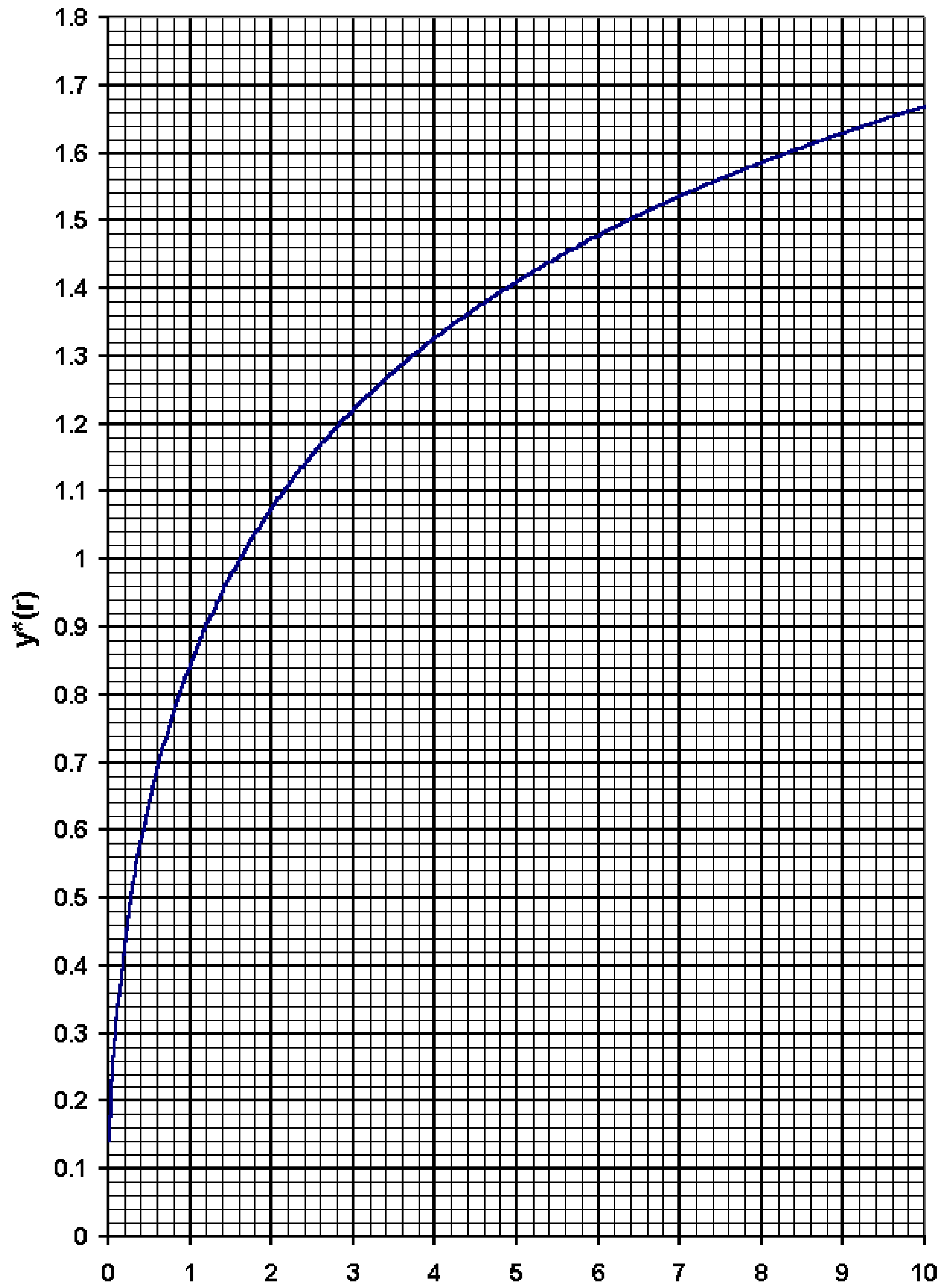




Erlang-C: Optimal Square-Root Staffing

$$n = R + \beta^*(r)\sqrt{R}$$

r = cost-of-delay / cost-of-staffing



Erlang-C: "The 80-20 Rule"

Prevalent: At least 80% customers served within 20 seconds;

Formally, %({Wait $\leq$ 20 sec.} $\geq$ 80%.

Call center: $\lambda = 6000/\text{hr}$, E[S]=4 min $\Rightarrow$ R=400 Erlangs.

4CallCenters: $n = 411$ agents needed.

The above is a solution to the staffing-problem via **Constraint Satisfaction**.
But how does one "understand" (internalize) the 80-20 rule?

$n = 411 \quad \Rightarrow \quad \beta^* = (411 - 400)/20 = 0.55.$

According to cost-graph (or formula), $\boldsymbol{r = d/c \approx 0.32}$. Yet:
Congestion-Index = E[Wait/E[S]] $\approx \frac{P\{Wait>0\}}{411-400} \approx \frac{1}{33}$. We observe:
The 80-20 Rule: Low valuation of customers' time, at 1/3 agents' time, yet very-good performance? enabled by scale!

What if $\boldsymbol{d/c = 5}$? $\beta^* = 1.4$:

- $n^* = 428$ (vs. 411 before);
- Agents' accessibility (idleness) = 7% (vs. 3% before);
- 1 out of 100 wait over 20 seconds (vs. 1 out of 5).

Conclude: **Constraint-Satisfaction is easier to formulate but Optimization is easier to internalize**.