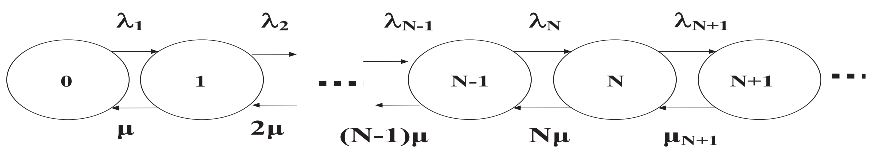
Markovian N-Server Queues

(Birth & Death)



Arrivals   Poisson (λ) ;

Services   exp(μ) $(E(S) = 1/\mu)$;

Servers N **statistically identical**, serving FCFS ;

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

$\rho = R/N$ Erlangs per server.

$Q(t)$ = number in system (served + queued) at time t.

$W(k)$ = queueing time of k-th arrival.

Steady State: $Q(\infty)$, $W(\infty)$, when exists.

Non-idling $\Rightarrow$ $[Q(t) - N]^+$ = queue-length, and

$[Q)(t) - N]^-$ = number of idle servers.

Examples: M/M/N, Blocking, Abandonment

M/M/∞: $\lambda_k \equiv \lambda, \quad \mu_k = k \cdot \mu \quad , \quad k \geq 1.$

Steady state $\pi_k = e^{-R} R^k / k!$ Poisson (R)

M/M/N/N: $\lambda_k \equiv 0, \quad k \geq N.$

Steady state $\pi_k = \dfrac{R^k}{k!} \Bigg/ \sum_{n=0}^{N} \dfrac{R^n}{n!}$, $0 \leq k \leq N$;

Erlang-B $P\{\text{Blocked}\} \doteq \boldsymbol{E_{1,N}} = \pi_N$, by PASTA

M/M/N: $\lambda_k \equiv \lambda, \quad \mu_k = (k \wedge N) \cdot \mu \quad , \quad k \geq 1.$

Steady state $\Leftrightarrow$ $\rho = \dfrac{\lambda}{N\mu} < 1$, servers' utilization.

Erlang-C $P\{\text{Wait} > 0\} \doteq \boldsymbol{E_{2,N}} = \sum_{k \geq N} \pi_k$

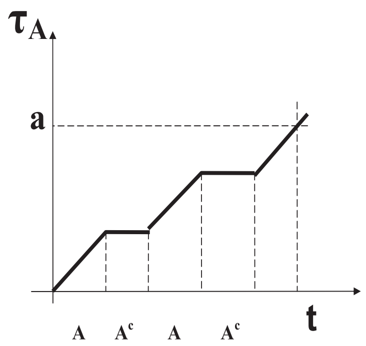
$$W(\infty) \mid W(\infty) > 0 \stackrel{d}{=} \exp\left(\text{mean} = \frac{1}{N\mu(1-\rho)}\right)$$

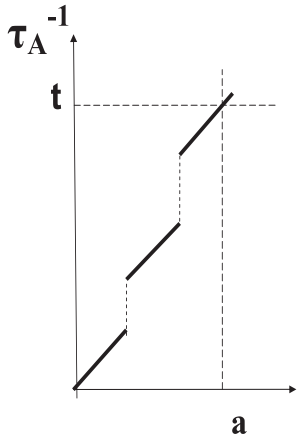
M/M/N+M: $\lambda_k \equiv \lambda, \quad \mu_k = (k \wedge N) \cdot \mu + (n-k)^+ \theta \quad , \quad k \geq 0.$

Restriction to a Set via Time-Change

X Markov , $X(\infty) \stackrel{d}{=} \pi$

$\tau_A(t) = \int_0^t 1_{\{X(u)\in A\}} du$ time in A





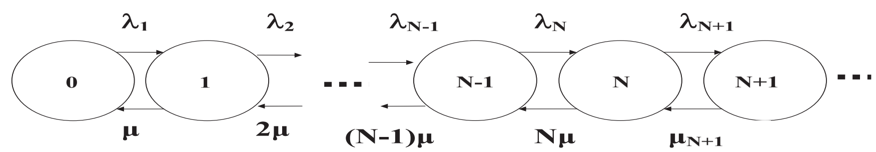
$X_A(t) = X\left(\tau_A^{-1}(t)\right)$ X restricted to A: Markov

$X_A(\infty) \stackrel{d}{=} X(\infty) | X(\infty) \in A$

Example: M/M/N/N $\stackrel{d}{=}$ M/M/∞ restricted to $\{0, 1, \ldots, N\}$.

$$E_{1,N} = Pr\{X_R = N\}/Pr\{X_R \le N\},$$
$$X_R \stackrel{d}{=} \text{Poisson } (R)$$

M/M/N/N (Erlang-B): Restricted M/M/∞



$M/M/\infty$:

$Q(t)$ = number in system at time $t \ge 0$, $Q(\infty) \stackrel{d}{=} \pi$

Then

$Q_- = Q$ restricted to $\{0, 1, \ldots, N\}$: M/M/N/N ; (λ, μ).

$$P(Q_-(\infty) = N) = \frac{\pi(N)}{\sum_{i=0}^N \pi(i)} = E_{1,N}$$

Proof: Q_- is a Markov Process:

Sojourn time in state $i \in \{0, 1, \ldots, N-1\}$ is $\exp(\lambda + i\mu)$.

State N: Each visit lasts $\exp(r = \lambda + N\mu)$

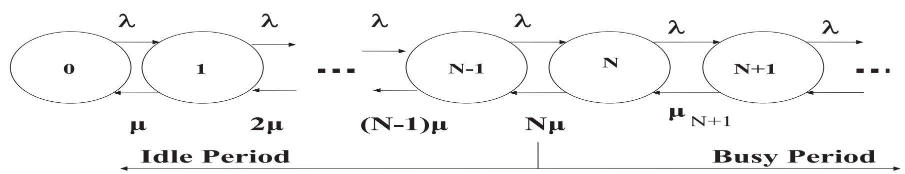
Number of visits is $Geo(p = \frac{N\mu}{\lambda + N\mu})$

Fact: Geometric(p) sum of i.i.d. exponentials $(r) \stackrel{d}{=} \exp(p \cdot r)$

$\Rightarrow$ Sojourn time in state $N \stackrel{d}{=} \exp(N\mu)$.

Conclude: $Q_- \stackrel{d}{=} M/M/N/N$.

Up/Down Crossings



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

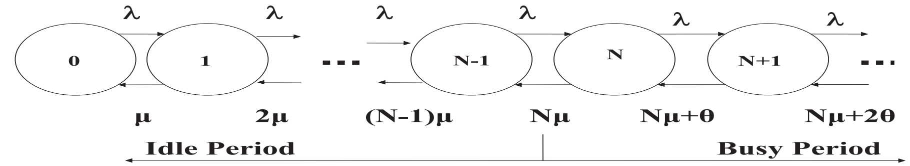
$T_{N,N-1} =$ E $\left[\, 1^{st} \text{ hitting time of } N-1 | Q(0) = N \right]$.

Proof:

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

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

M/M/N+M (Erlang-A)



$Q(t)$ = number in system at time $t \geq 0$

$Q_- = Q$ restricted to $\{0, 1, \ldots, N-1\}$: M/M/N-1/N-1 ; λ, μ.

$Q_+ = Q$ restricted to $\{N, N+1, \ldots\}$: M/M/1+M ; $\lambda, N\mu + \theta, \theta$.

Evolution of Q: Alternates between M/M/1+M (Q_+) and M/M/N-1/N-1 (Q_-).

$$P(\text{Wait} > 0) = \frac{T_{N,N-1}}{T_{N,N-1} + T_{N-1,N}} = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1}, \text{ by PASTA,}$$

where

$$T_{N,N-1} = \frac{1}{\mu_N \pi_+(0)},$$

$$T_{N-1,N} = \frac{1}{\lambda_{N-1}\pi_-(N-1)} = \frac{1}{\lambda\, E_{1,N-1}}.$$

M/M/N+M (Erlang-A): Continued

Hence,

$$P(Wait > 0) = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1} = \left[1 + \frac{\pi_+(0)}{\rho\, E_{1,N-1}}\right]^{-1},$$

in which

$$\boldsymbol{E}_{1,N-1} = \frac{R^{N-1}/(N-1)!}{\sum_{k=0}^{N-1} R^k/k!} \quad ,$$

$$\pi_+(0) = \frac{(\lambda/\theta)^{N\mu/\theta}/(N\mu/\theta)!}{\sum_{i=0}^{\infty} (\lambda/\theta)^{N\mu/\theta+i}/(N\mu/\theta+i)!} \quad .$$

Recall: $E_{1,N-1} = P(Y = N-1 | Y \le N-1)$,

where $Y \stackrel{d}{=} Pois(R)$, $\quad R = \lambda/\mu$.

"Similarly"

$\pi_+(0) \stackrel{d}{=} P(X = N\mu/\theta | X \ge N\mu/\theta)$,

where $X \stackrel{d}{=} Pois(\lambda/\theta)$

Hence

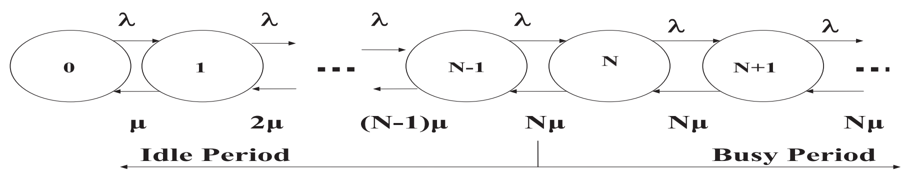
$$P(Wait > 0) = \left[1 + \frac{P(X = N\mu/\theta | X \ge N\mu/\theta)}{\rho\, P(Y = N-1 | Y \le N-1)}\right]^{-1},$$

Note: Only $\pi_+(0)$ changes by the model.

Special Cases: Erlang-C, Erlang-B, $\mu = \theta$

Erlang-C

Infinitely patient customers ($\theta = 0$)



$$P(Wait > 0) = E_{2,N} = \lim_{\theta \to 0} \left[1 + \frac{P(X = N\mu/\theta | X \ge N\mu/\theta)}{\rho\, P(Y = N-1 | Y \le N-1)}\right]^{-1}$$

Lemma: Let $X \stackrel{d}{=} Pois(\lambda/\theta)$. Then

$\lim_{\theta \to 0} P(X = N\mu/\theta | X \ge N\mu/\theta) = 1 - \rho$.

Proof:

$$\lim_{\theta \to 0} \frac{(\lambda/\theta)^{N\mu/\theta}/(N\mu/\theta)!}{\sum_{i=0}^{\infty} (\lambda/\theta)^{N\mu/\theta+i}/(N\mu/\theta+i)!} =$$

$$\lim_{\theta \to 0} \left[\sum_{i=0}^{\infty} (\tfrac{\lambda}{\theta})^i \cdot (\tfrac{\theta}{N\mu})^i / \prod_{j=1}^{i} (1 + \tfrac{j\theta}{N\mu})\right]^{-1} =$$

$$\left[\sum_{i=0}^{\infty} (\lambda/N\mu)^i\right]^{-1} = 1 - \rho.$$

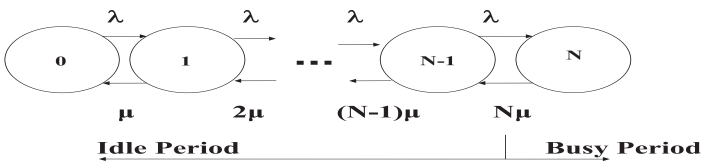
Conclude: Erlang-C Formula

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

Special Cases, Continued

Erlang-B

Infinitely impatient customers ($\theta = \infty$)



$$P(Q_t \geq N) = \lim_{\theta \to \infty} \left[1 + \frac{\pi_+(0)}{\rho\, E_{1,N-1}} \right]^{-1}$$

Lemma: Let

$$\pi_+(0) = \frac{(\lambda/\theta)^{N\mu/\theta} \exp(-\lambda/\theta)}{\frac{N\mu}{\theta}\gamma(\frac{N\mu}{\theta}, \frac{\lambda}{\theta})}.$$

Then

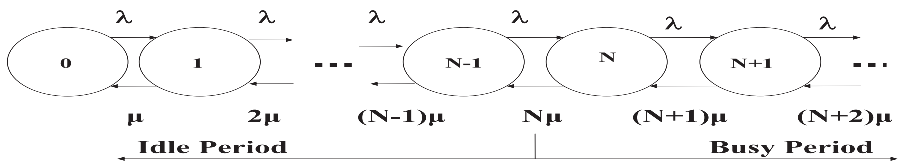
$$\lim_{\theta \to \infty} \pi_+(0) = 1.$$

Conclude: Erlang-B Formula

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

Special Cases, Continued

$\mu = \theta \qquad (M/M/\infty)$



$$P(Wait > 0) = \left[1 + \frac{P(Y = N | Y \geq N)}{\rho\, P(Y = N-1 | Y \leq N-1)} \right]^{-1},$$

$$\approx \left[1 + \frac{P(Y = N | Y \geq N)}{P(Y = N | Y \leq N)} \right]^{-1},$$

for large N and $\rho \approx 1$.

Recall: $\quad Y \stackrel{d}{=} Pois(R = \lambda/\mu)$.

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

Assume: μ fixed, while $\lambda_N \uparrow \infty$ as $N \uparrow \infty$.

Recall: $R = R_N = \lambda_N/\mu, \quad \rho = \rho_N = R_N/N.$

Erlang-B: $E_{1,N} = Pr\{Y_R = N\}/Pr\{Y_R \le N\},$

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

R large $\quad Y_R \stackrel{d}{\approx} R + Z\sqrt{R}$, where $Z \stackrel{d}{=} N(0,1)$,

suggesting $\quad N \sim R + \beta\sqrt{R}, \ -\infty < \beta < \infty.$

Then

$$\begin{aligned} E_{1,N} &= Pr\{N-1 < Y_R \le N\}/Pr\{Y_R \le N\} \\ &\approx Pr\left\{\beta - \tfrac{1}{\sqrt{R}} < Z \le \beta\right\}/Pr\{Z \le \beta\} \\ &\approx \tfrac{1}{\sqrt{R}}\,\phi(\beta)/\Phi(\beta) \ \approx\ \tfrac{1}{\sqrt{N}}\,\tfrac{\phi(\beta)}{\Phi(\beta)} = \\ &= \tfrac{1}{\sqrt{N}}\,\tfrac{\phi(-\beta)}{1-\Phi(-\beta)} \ =\ \tfrac{1}{\sqrt{N}}\,h(-\beta), \end{aligned}$$

h = hazard rate of $N(0,1)$.

Algebra: μ fixed, $N \uparrow \infty$.

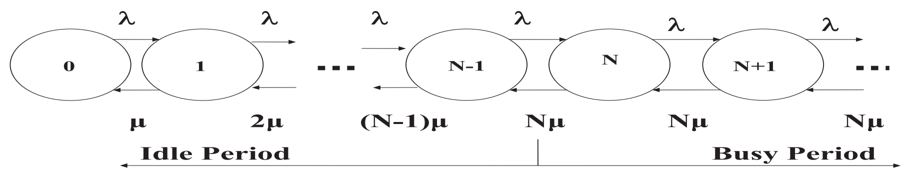
QED: $\quad N \sim R + \beta\sqrt{R}$, for some $\beta, -\infty < \beta < \infty$

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

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

Theorem: (Jagerman, 1974) $\quad$ QED $\Leftrightarrow \lim_{N\to\infty} \sqrt{N}\,E_{1,N} = h(-\beta).$

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



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

Recall $\quad E_{2,N} = \left[1 + \dfrac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1} = \left[1 + \dfrac{1-\rho_N}{\rho_N E_{1,N-1}}\right]^{-1}.$

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

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

Also $\quad T_{N,N-1} = \dfrac{1}{N\mu(1-\rho_N)} \sim \dfrac{1/\mu}{\beta\sqrt{N}}$

which applies as above, but for $0 < \beta < \infty$.

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

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

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

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

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

QED Theorem (Halfin-Whitt, 1981)

Consider a sequence of $M/M/N$ models, $N = 1, 2, 3, \ldots$

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

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

Here

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

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

Extremes:

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

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

M/M/N+M (Erlang-A) with Many Servers: $\pi_+(0)$

Assume: $\mu,\ \theta$ fixed, while $\lambda_N \uparrow \infty$ as $N \uparrow \infty$.

Recall: $R = R_N = \lambda_N/\mu, \qquad \rho = \rho_N = R_N/N.$

$$\pi_+(0) = Pr\{X_\lambda = N\mu/\theta\}/Pr\{X_\lambda \geq N\mu/\theta\},$$

where $\quad X_\lambda \stackrel{d}{=}$ Poisson (λ/θ).

N large $\quad X_\lambda \stackrel{d}{\approx} \lambda/\theta + Z\sqrt{\lambda/\theta}$, where $Z \stackrel{d}{=} N(0,1)$,

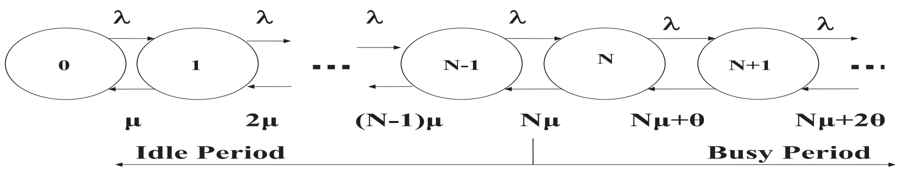
suggesting $\quad N \sim R + \beta\sqrt{R}, \ -\infty < \beta < \infty.$

Then
$$\begin{aligned}
\pi_+(0) &= Pr\{N\mu/\theta - 1 < X_\lambda \leq N\mu/\theta\}/Pr\{X_\lambda \geq N\mu/\theta\} \\
&\approx Pr\left\{\delta - \tfrac{1}{\sqrt{\lambda/\theta}} < Z \leq \delta\right\}/Pr\{Z \geq \delta\},\ \delta = \beta\sqrt{\mu/\theta} \\
&\approx \tfrac{1}{\sqrt{\lambda/\theta}}\,\phi(\delta)/(1 - \Phi(\delta)) \approx \tfrac{1}{\sqrt{N}}\sqrt{\tfrac{\theta}{\mu}}\,\tfrac{\phi(\delta)}{1-\Phi(\delta)} = \\
&= \tfrac{1}{\sqrt{N}}\sqrt{\tfrac{\theta}{\mu}}\,h(\delta)\ ,\ h = \text{hazard rate of } N(0,1).
\end{aligned}$$

Algebra: μ, θ fixed, $N \uparrow \infty$.

Proposition: QED $\Leftrightarrow \lim_{N\to\infty} \sqrt{N}\,\pi_+(0) = \sqrt{\frac{\theta}{\mu}}h(\delta)$.

M/M/N+M (Erlang-A) with Many Servers: QED



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

Recall $\quad P(Wait > 0) = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1} = \left[1 + \frac{\pi_+(0)}{\rho_N E_{1,N-1}}\right]^{-1}.$

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

Also $\quad T_{N,N-1} = \frac{1}{N\mu\pi_+(0)} \sim \frac{1}{\sqrt{N}} \frac{1/\mu}{h(\delta)/\delta}, \quad \delta = \beta\sqrt{\mu/\theta}$

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

Hence, $\quad P(Wait > 0) \sim \left[1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta}\right]^{-1}$, Garnett Function.

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

$\Leftrightarrow \quad \lambda \sim \mu N - \beta\mu\sqrt{N}$

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

Theorem (Garnett, M., Reiman 2002)

QED $\Leftrightarrow \lim_{N\to\infty} P(Wait > 0) = \left[1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta}\right]^{-1}.$

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

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

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

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

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

(the square-root staffing level)

Covers also the Extremes:

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

$\alpha = 0: \quad N = R + \gamma R$ **Quality-Driven**