Dimensioning Call Centers with Abandonment: Constraint Satisfaction and Cost Minimization

INFORMS Annual Meeting, San Francisco, November 16, 2005

Presenting Author: Sergey Zeltyn, Technion, Israel

Faculty of Industrial Engineering and Management

zeltyn@ie.technion.ac.il

Co-Author: Avi Mandelbaum, Technion, Israel

Faculty of Industrial Engineering and Management

avim@tx.technion.ac.il

Related Literature

Borst, Mandelbaum, Reiman. Dimensioning large call centers, *OR*, 2004.

M/M/n+G queue:

- Baccelli and Hebuterne, 1981;
- Brandt and Brandt, 1999, 2002;
- Zeltyn and Mandelbaum, 2005.

QED operational regime:

- Halfin and Whitt, 1981;
- Garnett, Mandelbaum and Reiman, 2002.

Global service level constraint:

- Koole and van der Sluis, 2003.

Call Center Industry

Overall expenditure. More than \$300 billion per year.

US. Several million employees (4% of workforce); 1000's agents in a "single" call center.

Call Center Industry

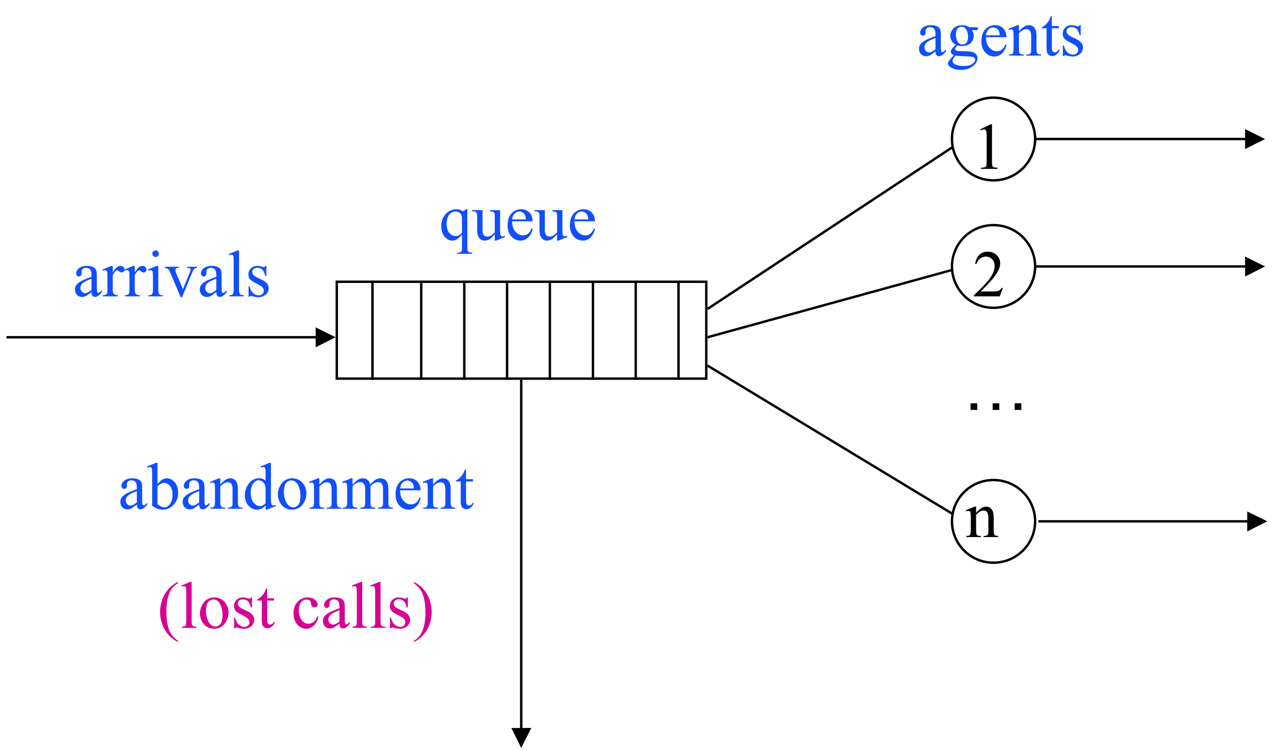
Overall expenditure. More than $300 billion per year.

US. Several million employees (4% of workforce); 1000's agents in a "single" call center.

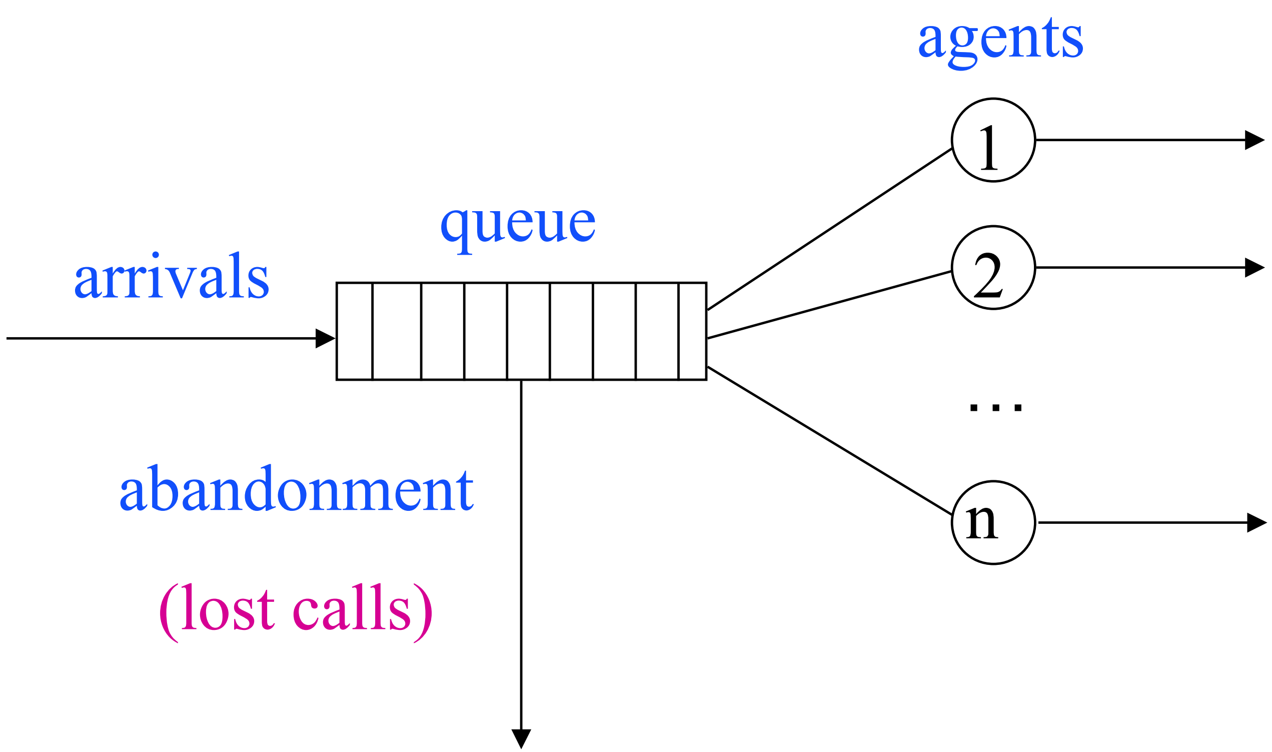
Quality/Efficiency Tradeoff.

- Personnel costs: **65-80%** of expenditure on a call center;
- More than **90%** of US consumers form company's image via call center experience;
- More than **60%** stop using company's products based on negative call center experience.

Modelling a Basic Call Center: M/M/n+G Queue



Modelling a Basic Call Center: M/M/n+G Queue



- λ – Poisson **arrival** rate;
- n service agents;
- Infinite queue;
- μ – exponential **service** rate;
- G – **patience** distribution;
- First Come First Served.

Modelling Abandonment

- **Patience time** $\tau \sim G$: time a customer **is willing** to wait for service;
- **Offered wait** V: time a customer **must** wait;
- If $\tau \leq V$, customer abandons; otherwise, gets service;
- **Actual wait** $W = \min(\tau, V)$;
- Patience times are not observed directly (censoring).

Modelling Abandonment

- **Patience time** $\tau \sim G$: time a customer **is willing** to wait for service;
- **Offered wait** V: time a customer **must** wait;
- If $\tau \le V$, customer abandons; otherwise, gets service;
- **Actual wait** $W = \min(\tau, V)$;
- Patience times are not observed directly (censoring).

M/M/n+M (Erlang-A) Queue

Patience time are **exponential**(θ).

Widely used in modern call centers.

Operational Performance Measures

- $\mathbf{P}\{\mathbf{Ab}\}$ – probability to abandon;
- $\mathbf{E}[\boldsymbol{W}]$ – average wait;
- $\mathbf{P}\{\boldsymbol{W} > \mathbf{0}\}$ – delay probability;
- $\mathbf{P}\{\boldsymbol{W} > \boldsymbol{T}\}$ – probability to exceed deadline.

Constraint satisfaction

Fix λ, μ, G.

$$n^* = \min\{n : \mathrm{P}_n\{\mathrm{Ab}\} \leq \alpha\} ,$$

$$n^* = \min\{n : \mathrm{E}_n[W] \leq T\} ,$$

$$n^* = \min\{n : \mathrm{P}_n\{W > T\} \leq \alpha\} ,$$

where α, T – **constraint** values.

Constraint satisfaction

Fix λ, μ, G.

$$n^* = \min\{n : \mathrm{P}_n\{\mathrm{Ab}\} \le \alpha\} ,$$

$$n^* = \min\{n : \mathrm{E}_n[W] \le T\} ,$$

$$n^* = \min\{n : \mathrm{P}_n\{W > T\} \le \alpha\} ,$$

where α, T – **constraint** values.

Cost Minimization

n^* should minimize

$$C_s \cdot n + (C_a \cdot \mathrm{P}_n\{\mathrm{Ab}\} + C_w \cdot \mathrm{E}_n[W]) \cdot \lambda ,$$

where C_s, C_a and C_w are **costs** of staffing, abandonment and waiting.

Exact Calculations in M/M/n+G

For example,

$$P\{Ab\} \;=\; \frac{1 + (\lambda - n\mu)J}{\mathcal{E} + \lambda J},$$

where

$$J \;=\; \int_0^\infty \exp\left\{\lambda \int_0^x \bar{G}(u)du - n\mu x\right\} dx\,, \qquad \mathcal{E} \;=\; \frac{\sum_{j=0}^{n-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}}\,.$$

No intuition and can be hard to compute.

Exact Calculations in M/M/n+G

For example,

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{1 + (\lambda - \mathrm{n}\mu)\mathrm{J}}{\mathcal{E} + \lambda \mathrm{J}}\,,$$

where

$$J \;=\; \int_0^\infty \exp\left\{\lambda \int_0^x \bar{G}(u)du - n\mu x\right\} dx\,, \qquad \mathcal{E} \;=\; \frac{\sum_{j=0}^{n-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}}\,.$$

No intuition and can be hard to compute.

Research Goals

Derive approximations for constraint satisfaction and cost minimization.

Compare with exact solution.

Asymptotic Operational Regimes

Offered load: $R = \lambda \times \mathrm{E}[S]$,

minutes of work (=service) that arrive per minute.

Asymptotic Operational Regimes

Offered load: $R = \lambda \times \mathrm{E}[S]$,

minutes of work (= service) that arrive per minute.

Efficiency-Driven (ED):

$$n \;\approx\; R - \gamma R \,, \qquad\qquad \gamma > 0\,.$$

Understaffing with respect to offered load.

Asymptotic Operational Regimes

Offered load: $R = \lambda \times \mathrm{E}[S]$,

minutes of work (=service) that arrive per minute.

Efficiency-Driven (ED):

$$n \approx R - \gamma R\,, \qquad \gamma > 0\,.$$

Understaffing with respect to offered load.

Quality and Efficiency-Driven (QED):

$$n \approx R + \beta\sqrt{R}, \qquad -\infty < \beta < \infty\,.$$

Square-Root Staffing Rule: Described by Erlang in 1924!

Asymptotic formulae for $\lambda, n \to \infty$ are available for both regimes.

Probability to Abandon: Approximations

ED: $\mathbf{P\{Ab\}} \;\approx\; \gamma$, **QED:** $\mathbf{P\{Ab\}} \;\approx\; \dfrac{1}{\sqrt{\lambda}} \cdot P_a(\beta) P_w(\beta)$,

where

$$P_w(\beta) \;:=\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \qquad P_a(\beta) \;:=\; \sqrt{g_0} \cdot \left(h(\hat{\beta}) - \hat{\beta}\right),$$

$g_0 :=$ patience density at the origin, $\quad \hat{\beta} \;:=\; \beta\sqrt{\dfrac{\mu}{g_0}}$,

$$h(x) \;=\; \frac{\phi(x)}{1 - \Phi(x)} \;=\; \frac{\phi(x)}{\bar{\Phi}(x)} \; : \text{ hazard rate of standard normal distribution.}$$

Compute **asymptotic optimal staffing level** via approximations.

Constraint Satisfaction: Probability to Abandon

Let $\mu = 1$ (time-units are minutes) everywhere.

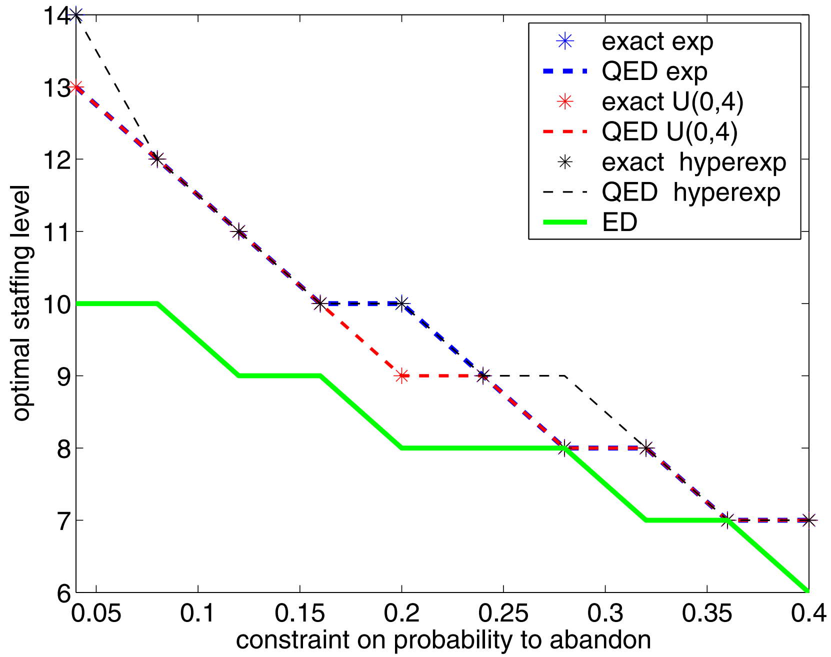
Optimal staffing	n^*	n^*_{QED}	n^*_{ED}
$\mathbf{P\{Ab\} \leq 4\%}$, $R = 50$, $\tau \sim \exp(\text{mean=30 sec})$	53	53	48 (8.8%)
$\mathbf{P\{Ab\} \leq 40\%}$, $R = 1000$, $\tau \sim U(0,1)$	601	600	600

Constraint Satisfaction: Probability to Abandon

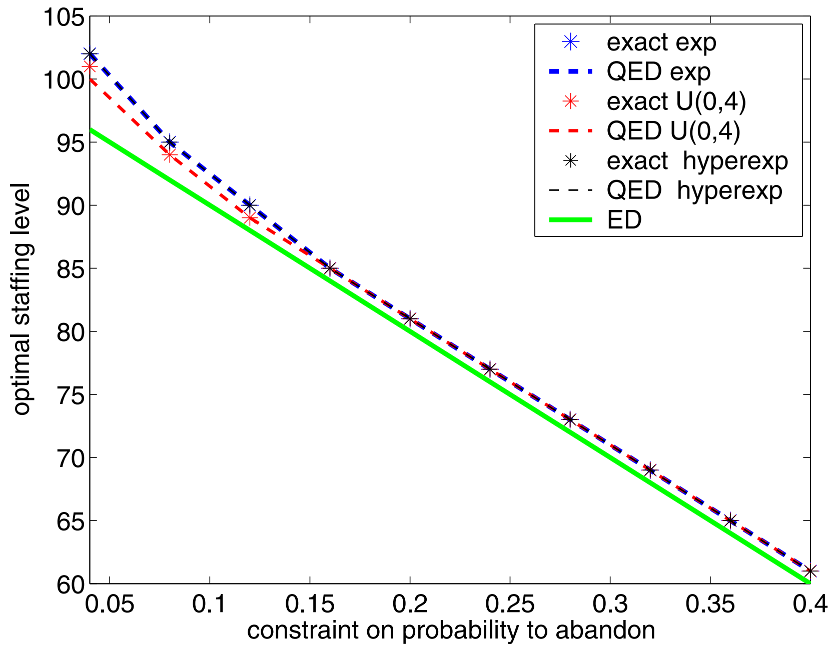
Let $\mu = 1$ (time-units are minutes) everywhere.

Optimal staffing	n^*	n^*_{QED}	n^*_{ED}
$\mathbf{P\{Ab\} \le 4\%}$, $R = 50$, $\tau \sim \exp(\text{mean=30 sec})$	53	53	48 (8.8%)
$\mathbf{P\{Ab\} \le 40\%}$, $R = 1000$, $\tau \sim U(0,1)$	601	600	600

$R = 10$



$R = 100$



Constraint Satisfaction: Average Wait

ED: $\mathbf{E}[\boldsymbol{W}] \;\approx\; \int_0^{G^{-1}(\gamma)} \bar{\boldsymbol{G}}(\boldsymbol{u})d\boldsymbol{u}$, **QED:** $\mathbf{E}[\boldsymbol{W}] \;\approx\; \frac{\mathbf{1}}{\sqrt{\boldsymbol{\lambda}}} \cdot \frac{\mathbf{1}}{\boldsymbol{g_0}} \cdot \boldsymbol{P_a}(\boldsymbol{\beta})\boldsymbol{P_w}(\boldsymbol{\beta})$.

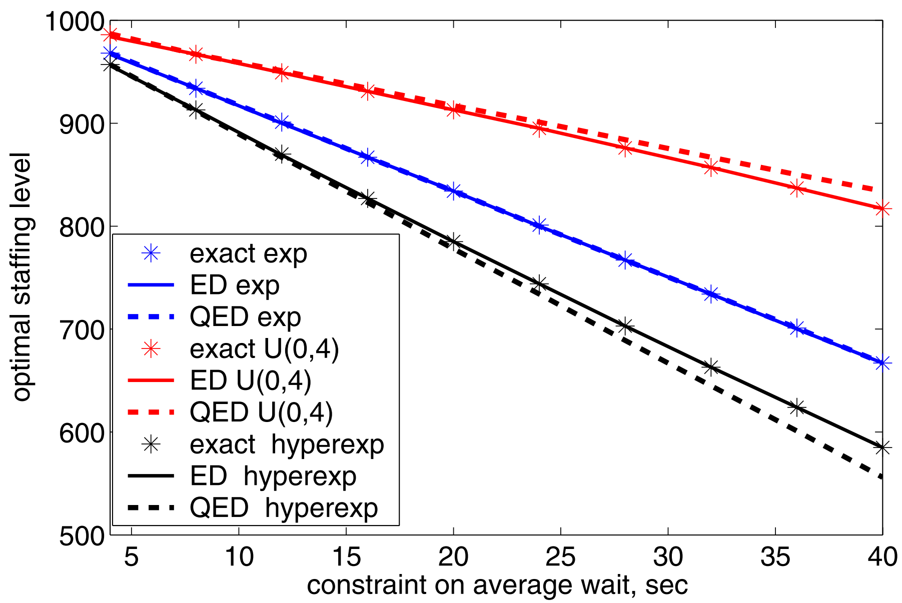
Optimal staffing	n^*	n^*_{QED}	n^*_{ED}
$\mathbf{E}[\boldsymbol{W}] \le$ **4 sec** , $R = 50,\ \tau \sim U(0,4)$	54	54	50 (8.7 sec)
$\mathbf{E}[\boldsymbol{W}] \le$ **40 sec** , $R = 1000,\ \tau \sim U(0,4)$	817	834	817

Constraint Satisfaction: Average Wait

ED: $\mathbf{E}[W] \approx \int_0^{G^{-1}(\gamma)} \bar{G}(u)du$, **QED:** $\mathbf{E}[W] \approx \frac{1}{\sqrt{\lambda}} \cdot \frac{1}{g_0} \cdot P_a(\beta) P_w(\beta)$.

Optimal staffing	n^*	n^*_{QED}	n^*_{ED}
$\mathbf{E}[W] \le$ **4 sec** , $R = 50$, $\tau \sim U(0,4)$	54	54	50 (8.7 sec)
$\mathbf{E}[W] \le$ **40 sec** , $R = 1000$, $\tau \sim U(0,4)$	817	834	817

$$R = 1000$$



Constraint Satisfaction: $\mathbf{P}\{W > T\}$

ED+QED: $n \approx R - \gamma R + \delta\sqrt{R}$, $\quad \gamma > 0, \ -\infty < \delta < \infty.$

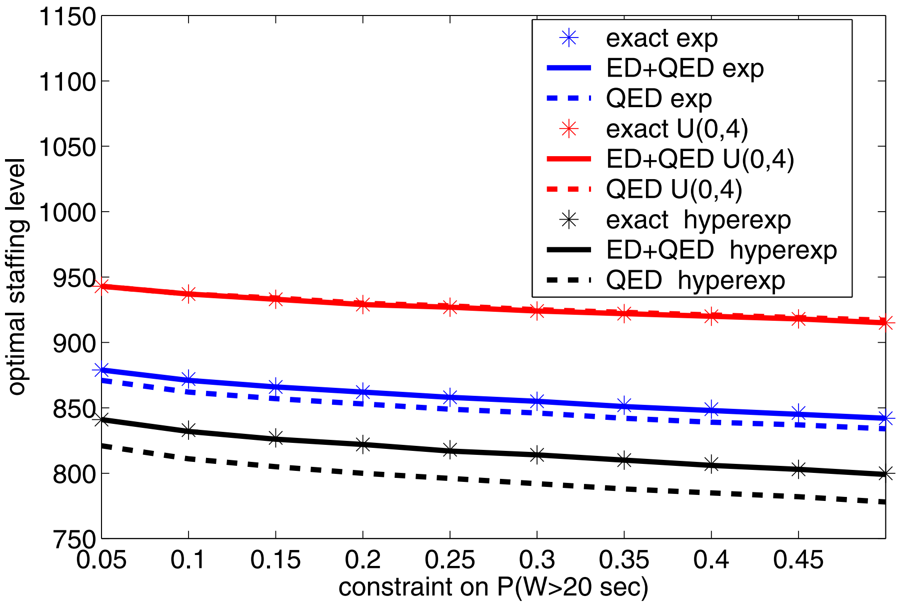
Optimal staffing	n^*	n^*_{QED}	n^*_{ED+QED}
$\mathbf{P}\{W < 20 \text{ sec}\} \geq 80\%$, $R = 100$, $\tau \sim \exp(\text{mean}=2)$	90	90	90
$\mathbf{P}\{W < 20 \text{ sec}\} \geq 80\%$, $R = 1000$, $\tau \sim \exp(\text{mean}=2)$	862	853 (68%)	862

Constraint Satisfaction: P$\{W > T\}$

ED+QED: $n \approx R - \gamma R + \delta\sqrt{R}$, $\quad \gamma > 0, \; -\infty < \delta < \infty.$

Optimal staffing	n^*	n^*_{QED}	n^*_{ED+QED}
P$\{W < 20$ sec$\} \geq 80\%$, $R = 100$, $\tau \sim \exp(\text{mean}=2)$	90	90	90
P$\{W < 20$ sec$\} \geq 80\%$, $R = 1000$, $\tau \sim \exp(\text{mean}=2)$	862	853 (68%)	862

$R = 1000$



Constraint Satisfaction: Conclusions

Staffing around offered load ("tight" constraint) – QED should be used.

Significant understaffing ("loose" constraint, large offered load)

- **Probability to abandon:** both QED and ED are good.
- **Average wait:** use ED for non-exponential patience, QED for Erlang-A.
- $\mathbf{P}\{\boldsymbol{W > T}\}$**:** use ED+QED.

Constraint Satisfaction: Global Constraint

Day of work consists of K time intervals.

Fractions of daily arrival rate $r_i,\ 1 \leq i \leq K$.

Staffing costs $c_i,\ 1 \leq i \leq K$.

Minimize $\sum_{i=1}^{K} c_i n_i$ given constraint on daily performance.

Constraint Satisfaction: Global Constraint

Day of work consists of K time intervals.

Fractions of daily arrival rate $r_i, \ 1 \le i \le K$.

Staffing costs $c_i, \ 1 \le i \le K$.

Minimize $\sum\limits_{i=1}^{K} c_i n_i$ given constraint on daily performance.

$\mathbf{P\{Ab\}} \le \boldsymbol{\alpha}$: use QED staffing, $\lceil n_i \ = \ R_i + \beta_i \sqrt{R_i} \rceil$,

where $\sum\limits_{i=1}^{K} \beta_i c_i \sqrt{r_i} \ \longrightarrow \ \min$

given $\sum\limits_{i=1}^{K} \sqrt{r_i} P_w(\beta_i) P_a(\beta_i) \ = \ \alpha \sqrt{\lambda}\,.$

$\mathbf{P\{}\boldsymbol{W} > \mathbf{0\}} \le \boldsymbol{\alpha}$: either use QED staffing or “close gate” $(n_i = 0)$.

Cost Minimization

$\text{Cost} = C_s \cdot n + (C_a \cdot \mathrm{P}\{\mathrm{Ab}\} + C_w \cdot \mathrm{E}[W]) \cdot \lambda$.

Use QED staffing, where β depends on C_a/C_s and C_w/C_s.

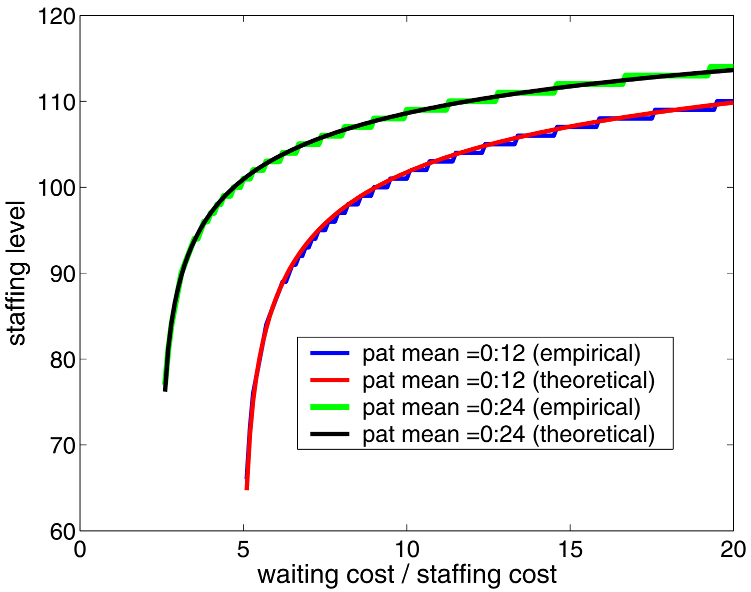
Cost Minimization

$\text{Cost} = C_s \cdot n + (C_a \cdot \text{P}\{\text{Ab}\} + C_w \cdot \text{E}[W]) \cdot \lambda$.

Use QED staffing, where β depends on C_a/C_s and C_w/C_s.

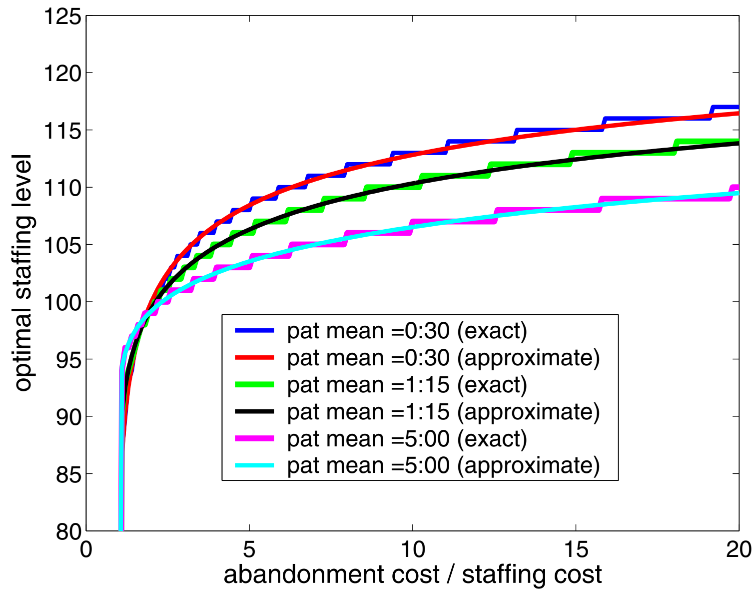
$\mathbf{Cost} = \boldsymbol{C_s \cdot n + C_w \cdot \lambda \cdot \mathrm{E}[W]}$

exponential patience



$\mathbf{Cost} = \boldsymbol{C_s \cdot n + C_a \cdot \lambda \cdot \mathrm{P}\{\mathrm{Ab}\}}$

uniform patience



Excellent fit, except some cases with waiting costs and non-exponential patience.

Possible Future Research

- Cost minimization: theoretical validation;
- Random arrival rate;
- Time-inhomogeneous arrival rate;
- Generally distributed service times.