*Service Engineering Seminar*

# Nurse-to-Patient Ratios in Hospital staffing: A Queuing Perspective

Véricout and Jennings

# Objective

Determine Nurse-to-Patient staffing rules in order to obtain uniform quality of care across all hospitals

# *The solution should be...*

- Simple to use
- Simple data requirements (Parameters)
- Economical

# Quality of Care Measurements

- Probability of excessive delay
  - The likelihood that a needy patient's waiting period before getting access to a nurse is longer than a time threshold $T \geq 0$
- Probability of Delay ($T=0$)
- Probability of Timely Service ($T>0$)

# Fixed nurse-to-patient ratios

| Hospital Unit | Minimum ratio |
|---|---|
| Critical care | 1:2 |
| Pediatric care | 1:4 |
| Emergency room | 1:6 |
| Critical trauma | 1:1 |

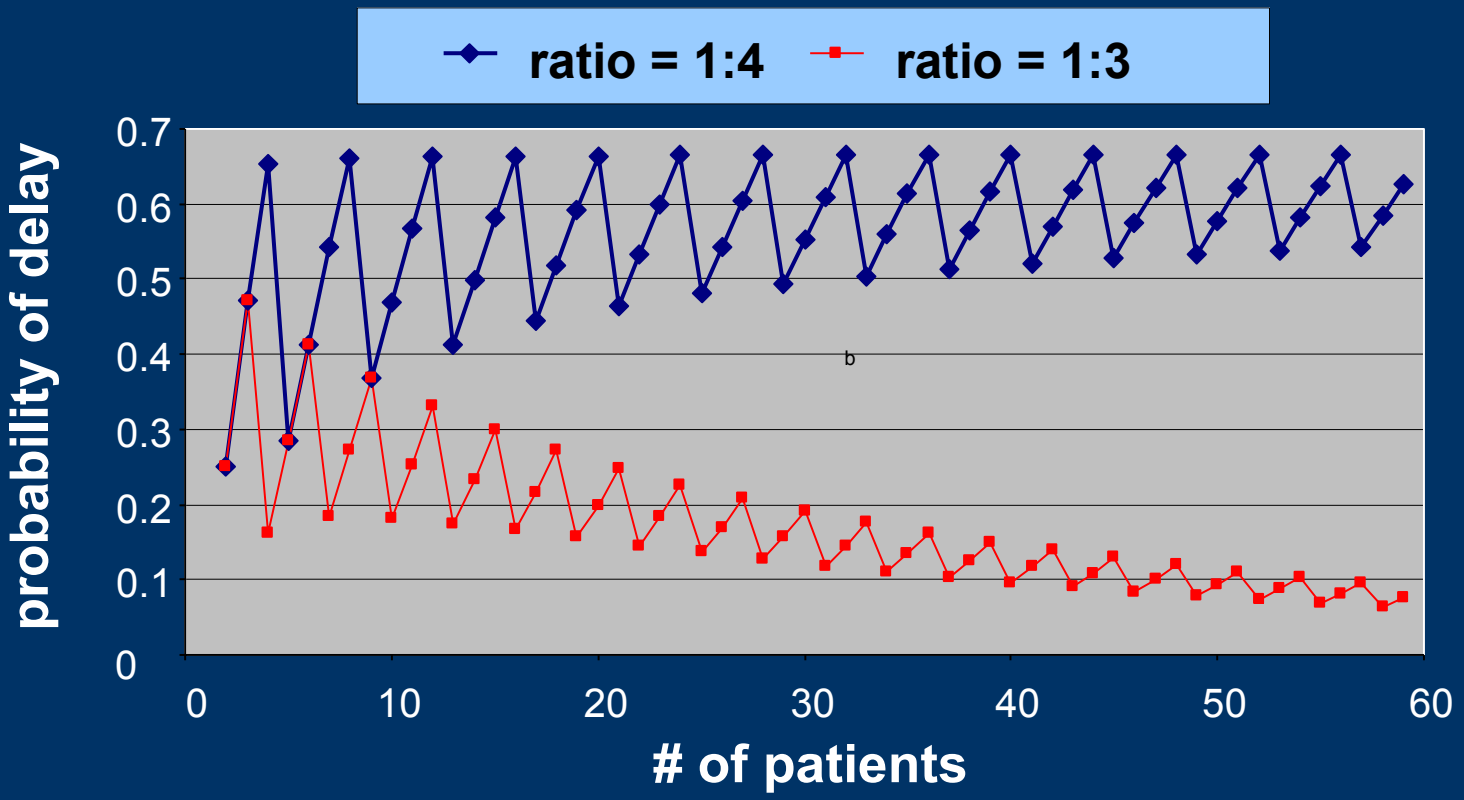
Nominal ratio policy: $S_n = \lceil rn \rceil$

# Ratio setting

- Offered load = [Average number of nursing hours required per patient] X [ number of patients]
- Categorization of patients by their direct nursing care needs.
  - Physical condition / Activity studies

  Average mix of patient per unit => staffing for each unit.
- Categorization of nursing skills => Staffing for each personnel category.

# Theoretical Performance of Pediatric Units




Fraction of nursing time required (r) = 1/4

# 1st Conclusion

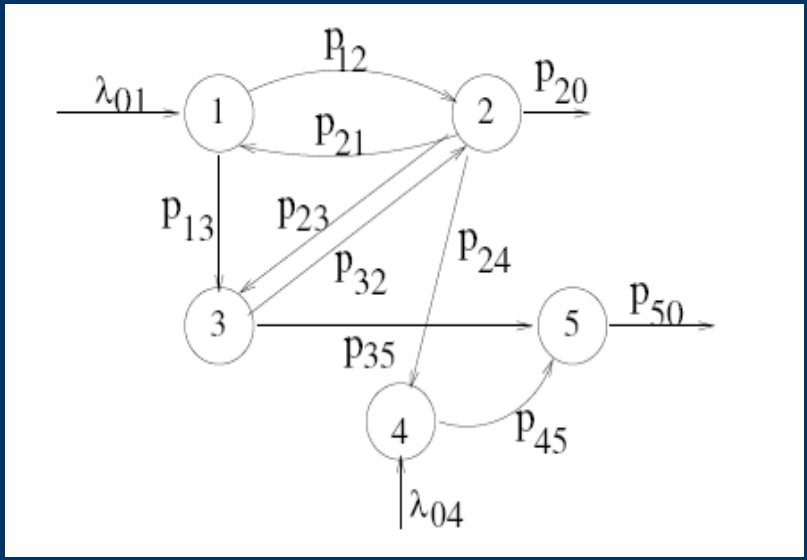
- The fixed nurse-to-patient ratio policy does not provide uniform quality-of-service across hospital

SIZE MATTERS…

# Modeling Assumptions

- Patients:
  - Homogeneous patient population
  - Two states: needy or dormant
  - n (the number of patients) is constant
- Nurses:
  - Homogeneous nursing staff
  - Pooling of nursing staff
- Closed Queuing Network
  - Service time ~ exp(μ)
  - Activation time ~ exp(λ)
  - FCFS

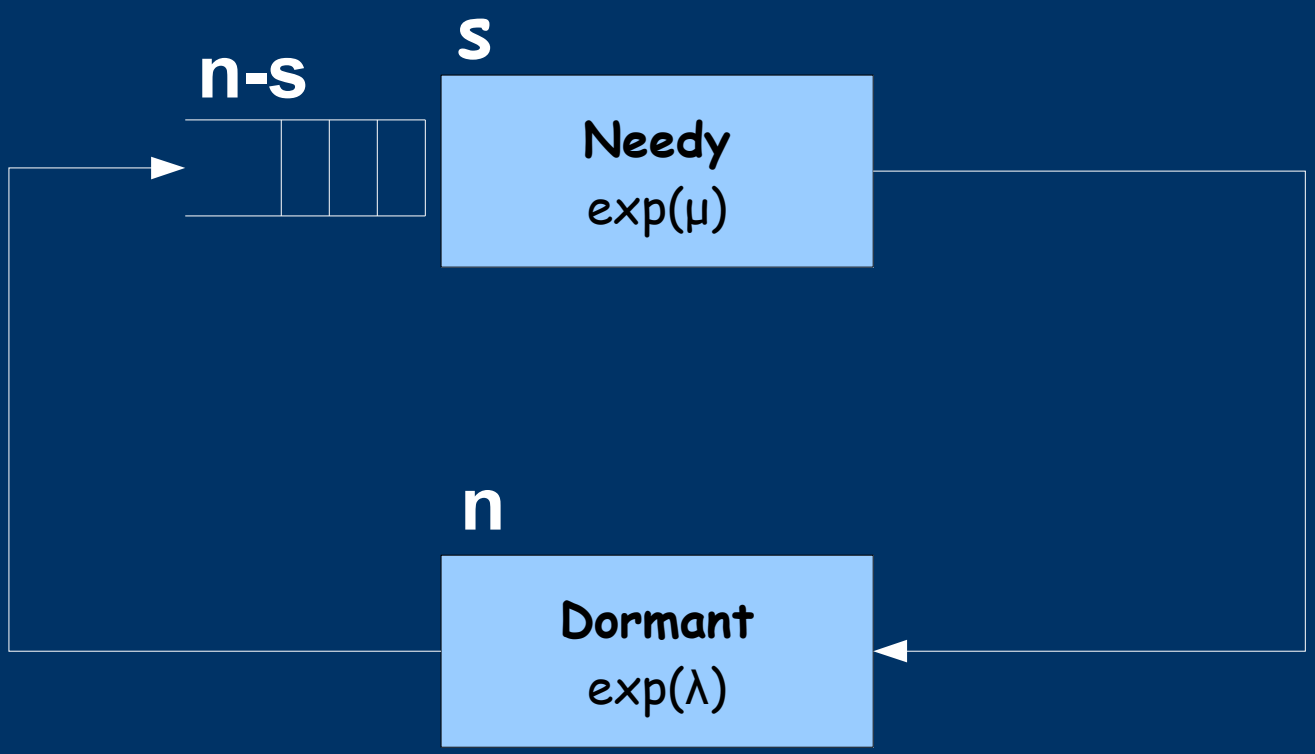
# Closed Queuing Network
## (Closed Jackson Network)




1. Arrivals from 'outside' to node i follow Poisson Process (PP) with an arrival rate $\lambda_{oi}$;
2. Service times for each server at node i are independent, exponentially distributed with parameter $\lambda_i$;
3. The probability of moving from node i to node j (after the service at node i is completed) is $P_{ij}$ and it is state-independent;
4. $P_{i0}$ - probability the customer will leave the system from node i.

If $\lambda_{0i} = 0$; for all i, and $P_{i0} = 0$ - the network is called a closed Jackson Network.

# Our Closed Queuing Network (M/M/s/∞/n queue)

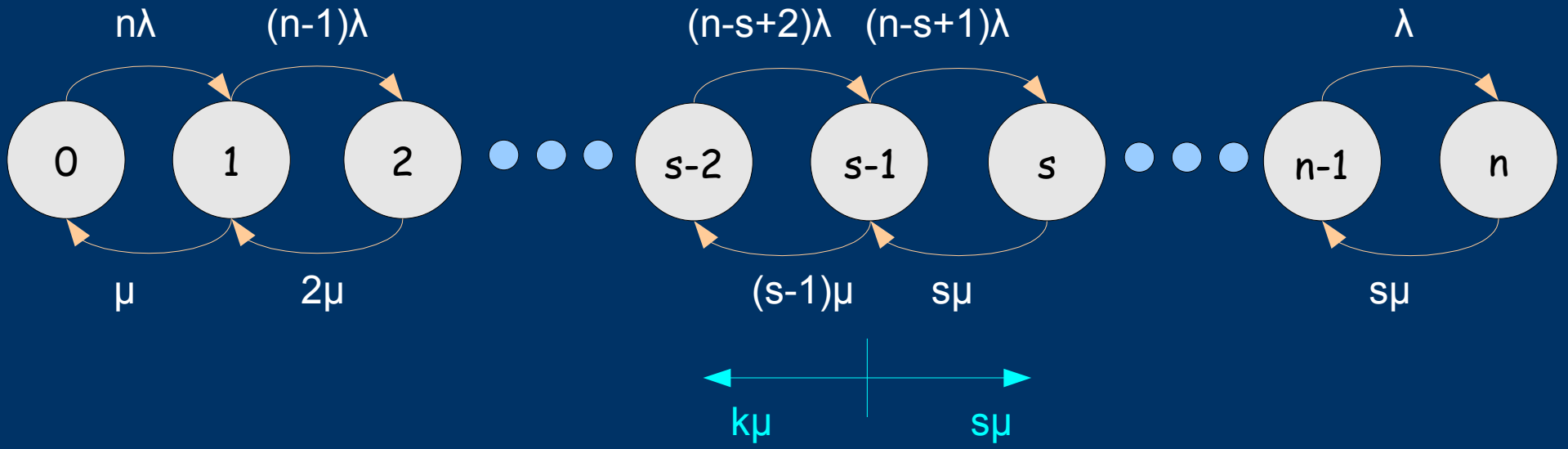



n - Number of beds

s - Number of nurses

# The model as a birth-death process

$X_t$ - Number of needy patients




PASTA does not hold for small n!

Virtual hitting time ≠ Needy patient waiting time

(The Arrival Theorem: In closed network system with *m* customers, the system as seen by arrivals to server *j* is distributed as the steady-state distribution in the same network with only *m -1* customers)

# Product form solution of the stationary distribution

$$\pi^0(\alpha,\beta)=\begin{cases}\dfrac{\pi^1(\alpha)\pi^2(\beta)}{\sum_{i+j=N}\pi^1(i)\pi^2(j)}, & \alpha+\beta=n;\\ 0 & otherwise.\end{cases}$$

$\pi^l(.)$ is the steady state probability for node $l=1,2$; $l$ is an M/M/s or M/M/n system.

# Steady state probability distribution of the number of needy patients in the system

$$\pi_1 \mu = \pi_0 n\lambda \qquad \pi_1 = \frac{n\lambda}{\mu}\pi_0 = N\rho\pi_0$$

$$\pi_2 2\mu = \pi_1 (n-1)\lambda \qquad \pi_2 = \frac{(n-1)\lambda}{2\mu}\pi_1 = \frac{n(n-1)\lambda^2}{2\mu^2}\pi_0 = \binom{n}{2}\rho^2\pi_0$$

$$\vdots$$

$$\pi_s s\mu = \pi_{s-1}(n-s+1)\lambda \qquad \pi_s = \frac{(n-s+1)\lambda}{s\mu}\pi_1 = \binom{n}{s}\rho^s\pi_0$$

$$\pi_{s+1} s\mu = \pi_s (n-s)\lambda \qquad \pi_{s+1} = \frac{(n-s)\lambda}{(s+1)\mu}\pi_1 = \binom{n}{s+1}\frac{(s+1)!}{s!\,s}\rho^{s+1}\pi_0$$

# Steady state probability distribution of the number of needy patients in the system

$$\pi(k)=\begin{cases}\pi_0\binom{n}{k}\rho^k & \text{if } k<s;\\ \pi_0\binom{n}{k}\frac{k!}{s!}s^{s-k}\rho^k & \text{if } k\geq s.\end{cases}$$

$$\rho:=\lambda/\mu=r/\bar{r}$$

$$r:=\lambda/(\lambda+\mu); \qquad \bar{r}:=1-r$$

# The probability of excessive delay

- Virtual hitting time (V)
  - V is the time required for the number of busy servers to fall below *s*, provided no new jobs arrive in the interim. Thus, this is an objective measure of the system.
  - P(V>t|N=k) ~ Erlang (k-s+1, sμ)  (k≥s)
  - The tail of steady state distribution is:

$$P(V>t)=e^{-s\mu t}\sum_{k=s}^{n}\pi_k\sum_{j=0}^{k-s}\frac{(s\mu t)^j}{j!}$$

# The probability of excessive delay

- Virtual hitting time (W)
  - *W* denote the steady state, in-queue waiting time for a hypothetical newly needy patient. Thus, this is a subjective measure, from the point of view of the patient
  - Activation rate: $\lambda_k := \lambda \cdot (n-k)$
  - The tail of steady state distribution is:

$$p_n(s,t) := P(W>t) = \sum_{k=s}^{n} \frac{\lambda_k \pi_k}{\sum_{i=0}^{n} \lambda_i \pi_i} P(V>t \mid N=k)$$

$$= e^{-s\mu t} \sum_{k=s}^{n} \frac{(n-k)\pi_k}{\sum_{i=0}^{n} (n-i)\pi_i} \sum_{j=0}^{k-s} \frac{(s\mu T)^j}{j!}$$

- Data required: r and μT

# The optimization Problem

- Given a target probability of delay, ε, for each potential unit size n, select the staffing level $s_n$, such that the probability of delay (or the probability of timely service) is less than ε.

$$\begin{aligned} &Minimize && s_n, && for\ s_n \in [1,\ldots,n] \\ &s.t. && p_n(s_n, T) < \epsilon \end{aligned}$$

- What's the problem?

# Optimal Vs. Nominal ration staffing

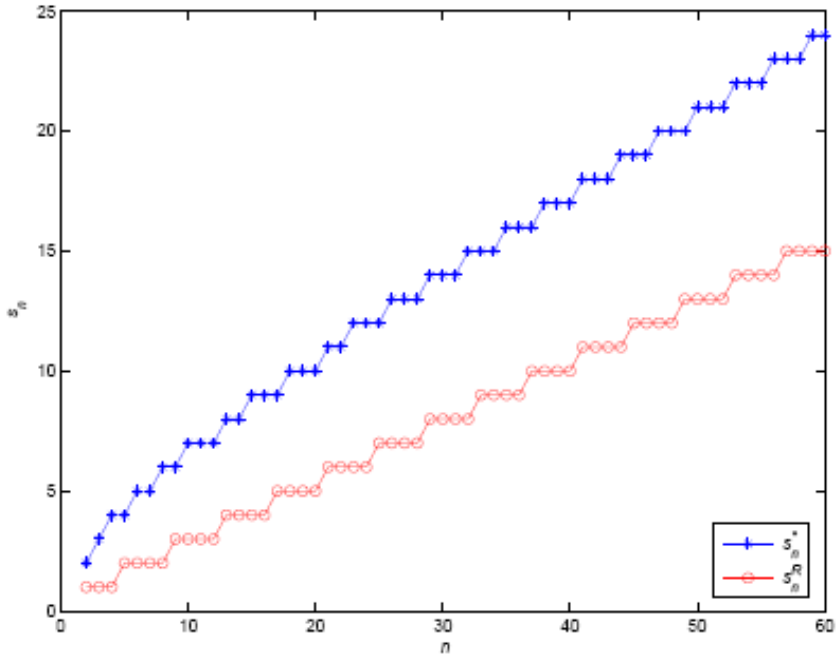



Figure 6: Optimal versus nominal ratio staffing, $r = \frac{1}{4}$, $\epsilon = 1\%$, $T = 0$

T=0, Epsilon = 1%, r=1/4

# Optimal Vs. Nominal ration staffing

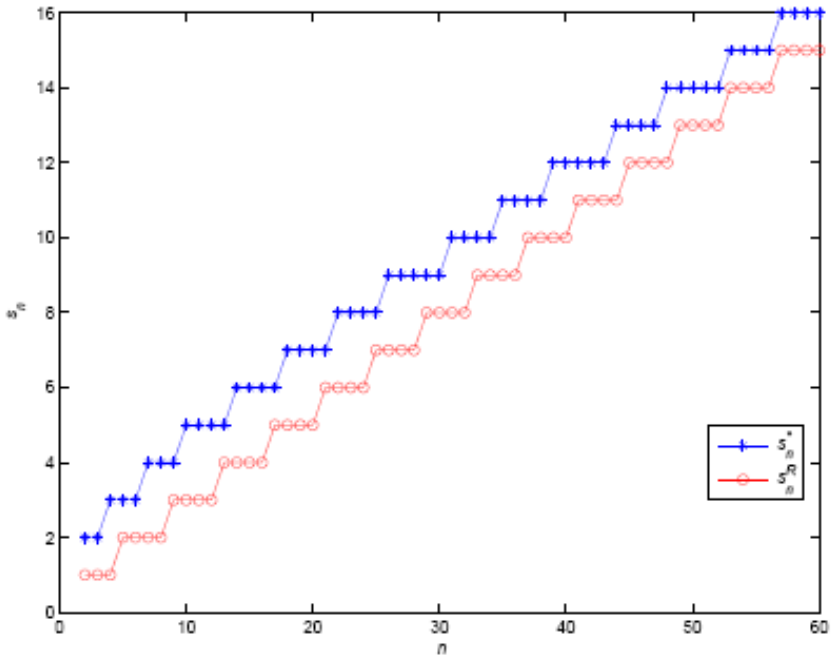



Figure 7: Optimal versus nominal ratio staffing, $r = \frac{1}{4}$, $\epsilon = 1\%$, $T = \frac{1}{\mu}$

T=1/μ, Epsilon = 1%, r=1/4

# Optimal Vs. Nominal ration staffing

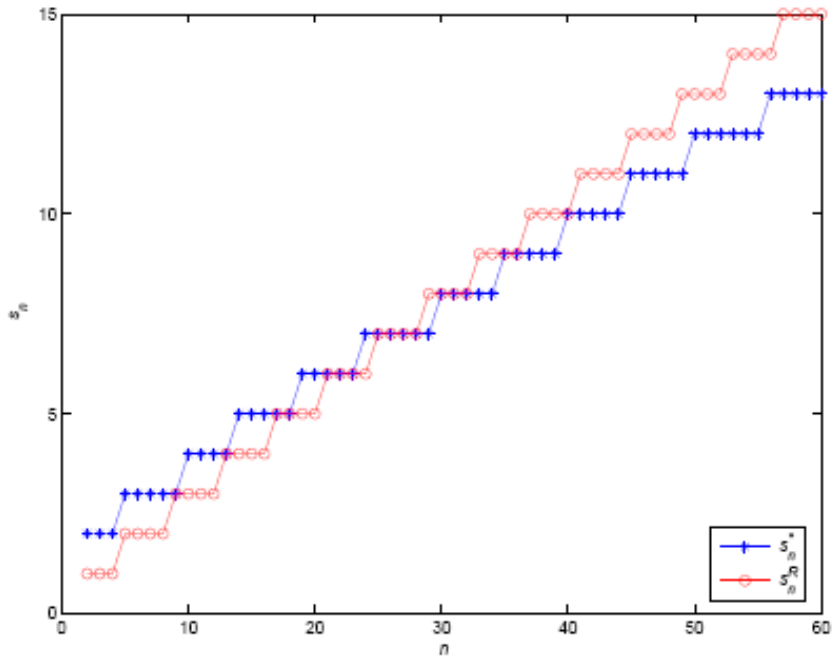



Figure 8: Optimal versus nominal ratio staffing, $r = \frac{1}{4}$, $\epsilon = 1\%$, $T = \frac{2}{\mu}$

T=2/μ, Epsilon = 1%, r=1/4

# More conclusion (2nd set of c.)

1. Optimal staffing is not a ratio policy
2. Optimal Staffing depends on T
3. At least for Large T - Small units are understaffed and large units are overstaffed by ratio policy

# Heuristics

- Based on many-server asymptotics (n->∞)
- Define: $\bar{s} = \lim_{n \to \infty} s_n / n$

- If $\bar{s}$<r: ED staffing regime (T>0)
- If $\bar{s}$=r: QED staffing regime (T≥0 and small)
- If $\bar{s}$>r: QD staffing regime (T=0)

# QED Staffing regime

- Optimal staffing levels are slight deviations from the nominal ratio policy $s_n^{QED} = \lceil rn + \beta\sqrt{n} \rceil$

$$\epsilon = \alpha(\beta)\Phi\left(\frac{-\beta(1+r\mu T)}{r\sqrt{\bar{r}}}\mu r T\sqrt{\frac{n}{\bar{r}}}\right)/\Phi\left(\frac{-\beta}{r\sqrt{\bar{r}}}\right)$$

$$\alpha(\beta) = \left(1 + e^{\frac{-\beta^2}{2r^2}}\sqrt{r}\frac{\Phi\left(\frac{\beta}{\sqrt{r\bar{r}}}\right)}{\Phi\left(\frac{-\beta}{r\sqrt{\bar{r}}}\right)}\right)^{-1} = \left(1 + \sqrt{r}\frac{h\left(\frac{\beta}{r\sqrt{\bar{r}}}\right)}{h\left(\frac{-\beta}{\sqrt{r\bar{r}}}\right)}\right)^{-1}$$

(See proof at the end)

# ED Staffing regime

- Optimal staffing level is an order √n deviation from $\hat{r}_T n$: $s_n^{ED} = \lceil \hat{r}_T n + \beta_T \sqrt{n} \rceil$

$$\hat{r}_T = \frac{r}{1 + r\,\mu\,T} \qquad \beta_T = -r\,\Phi^{-1}(\epsilon)\sqrt{\frac{\bar{r} + r\,\mu\,T}{(1 + r\,\mu\,T)^3}}$$

$$\Rightarrow \quad \epsilon = \Phi\left(\frac{-\beta}{r}\sqrt{\frac{(1 + r\,\mu\,T)^3}{\bar{r} + r\,\mu\,T}}\right)$$

# Open systems staffing rules (reminder)

- Erlang-A: M/M/N + M queue, as N->∞

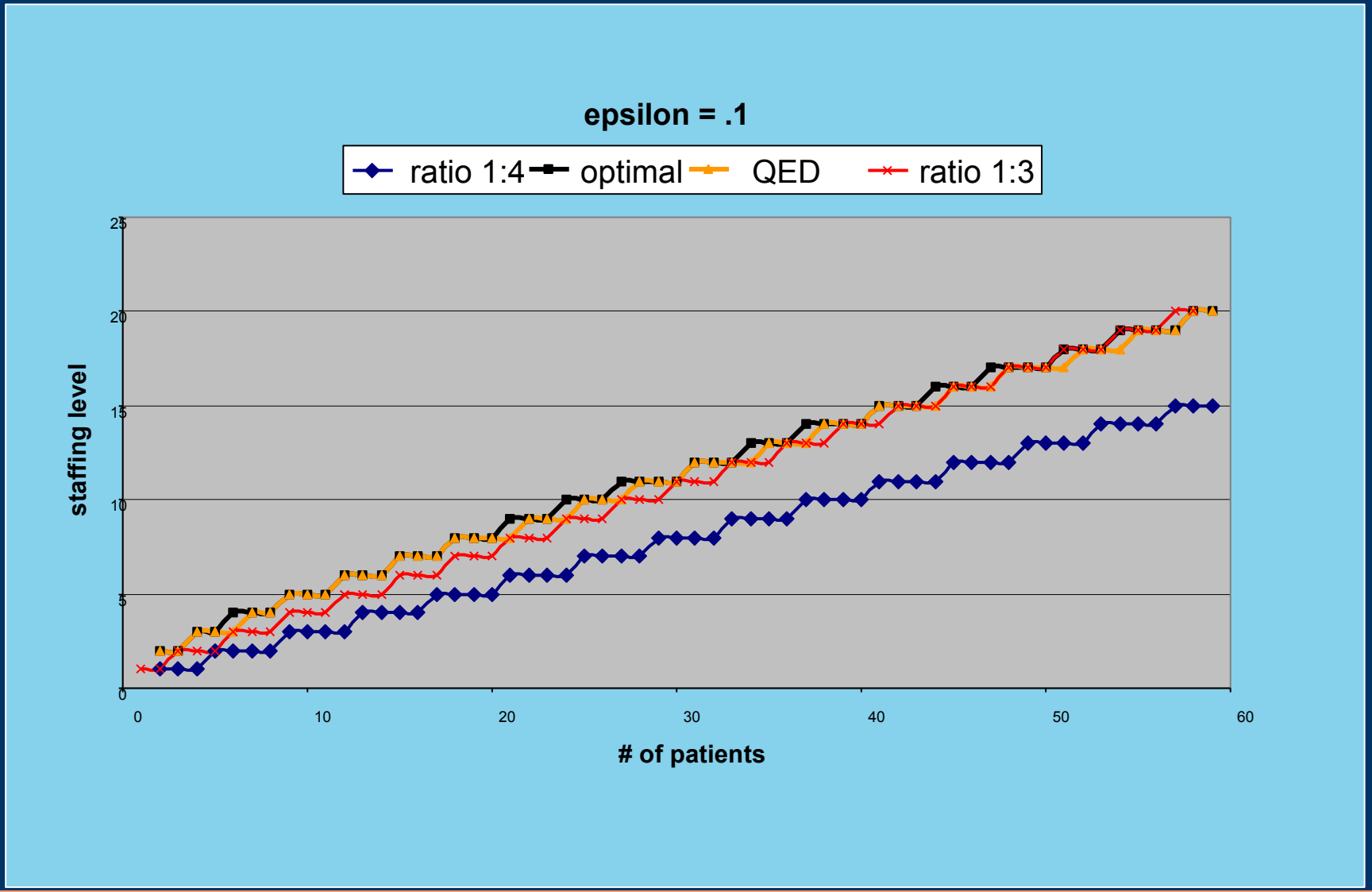
$$N^{QED} = R + \beta\sqrt{R}, \quad for\ some\ \beta, \quad -\infty < \beta < \infty$$

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

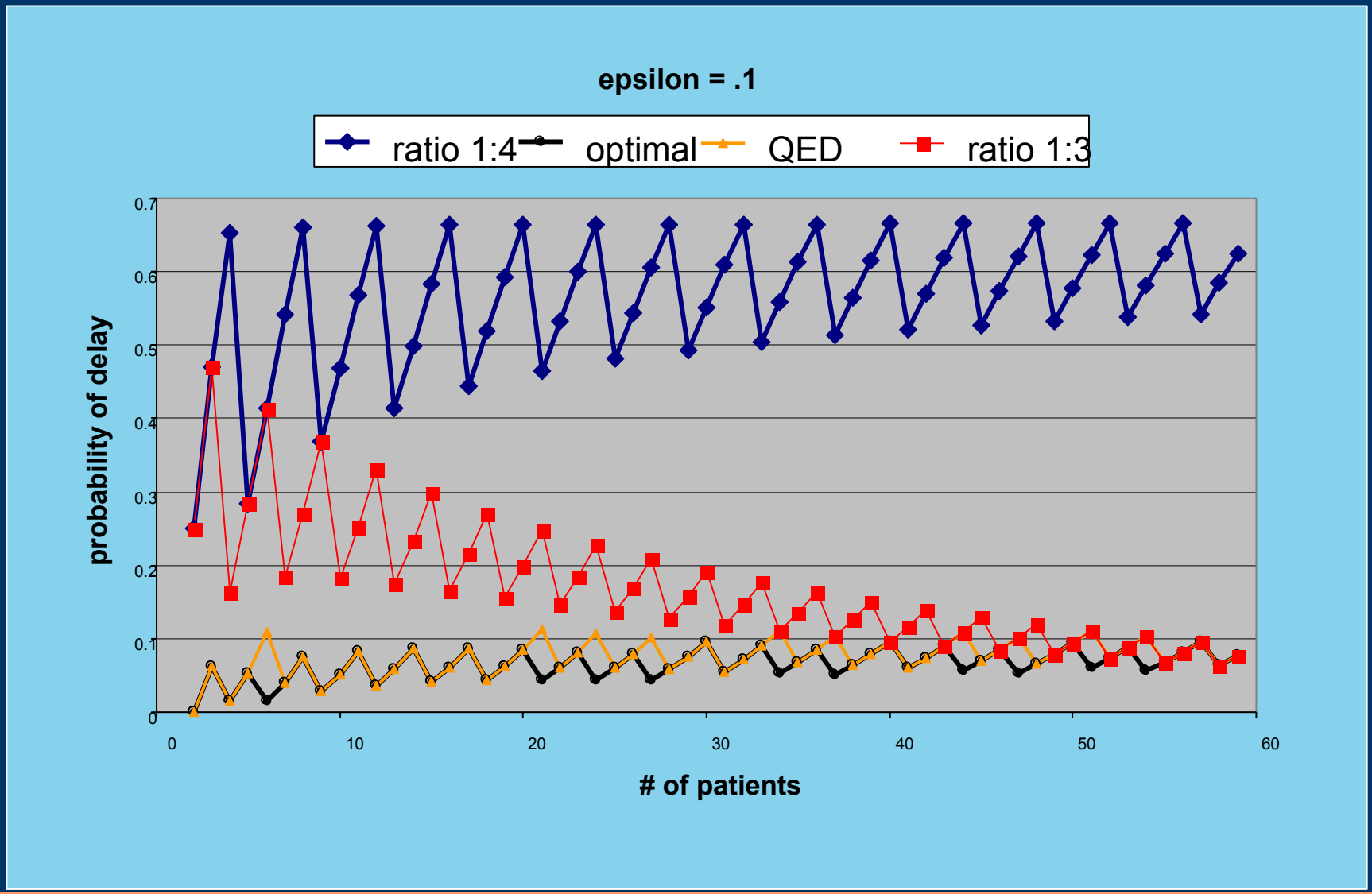
$where \quad \delta = \beta\sqrt{\mu/\theta},$

$\beta = \lim_{N\to\infty} \sqrt{N}(1-\rho_N).$

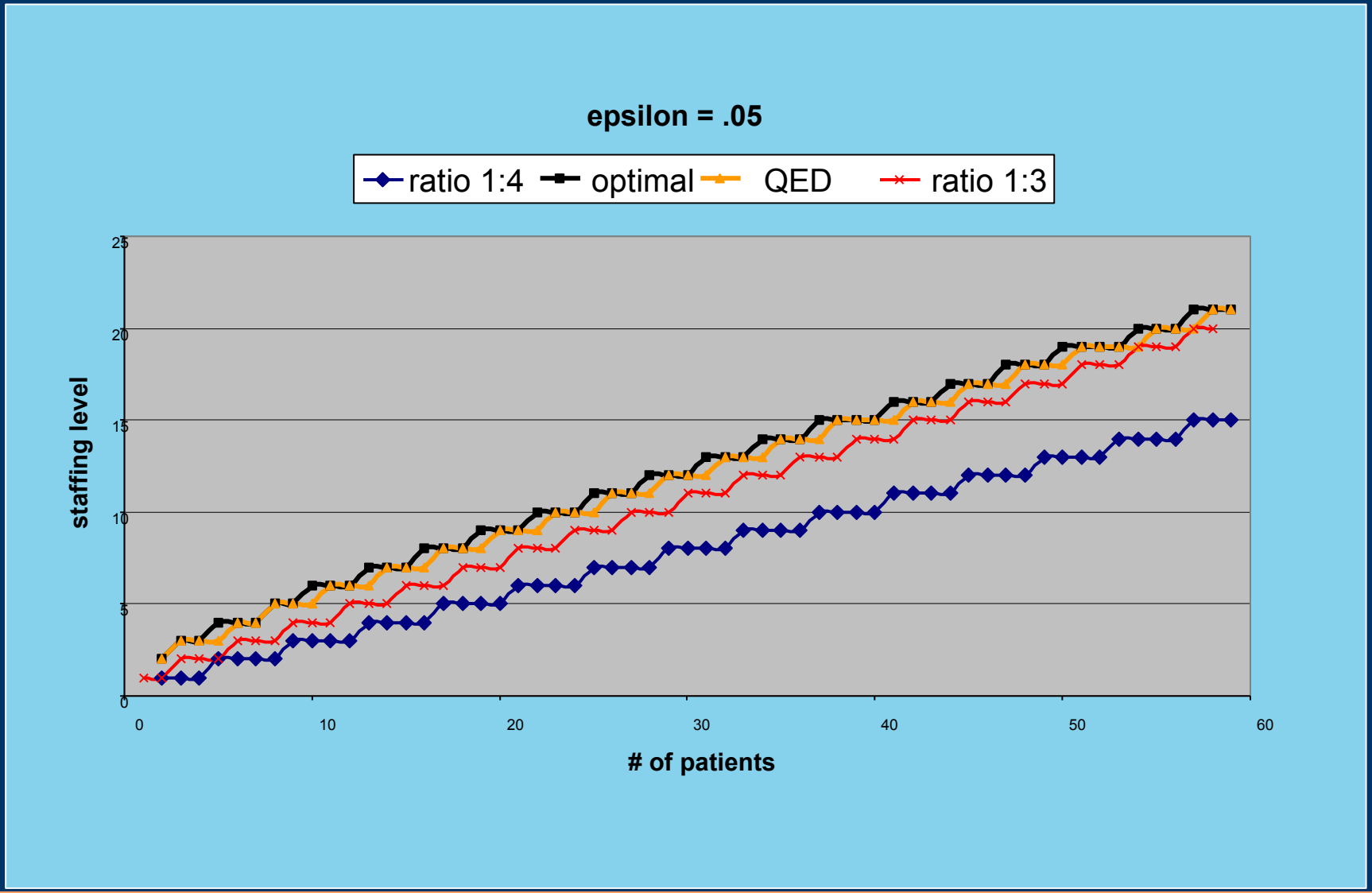
# Staffing:
# Ratio vs. Optimal vs. QED (T=0)



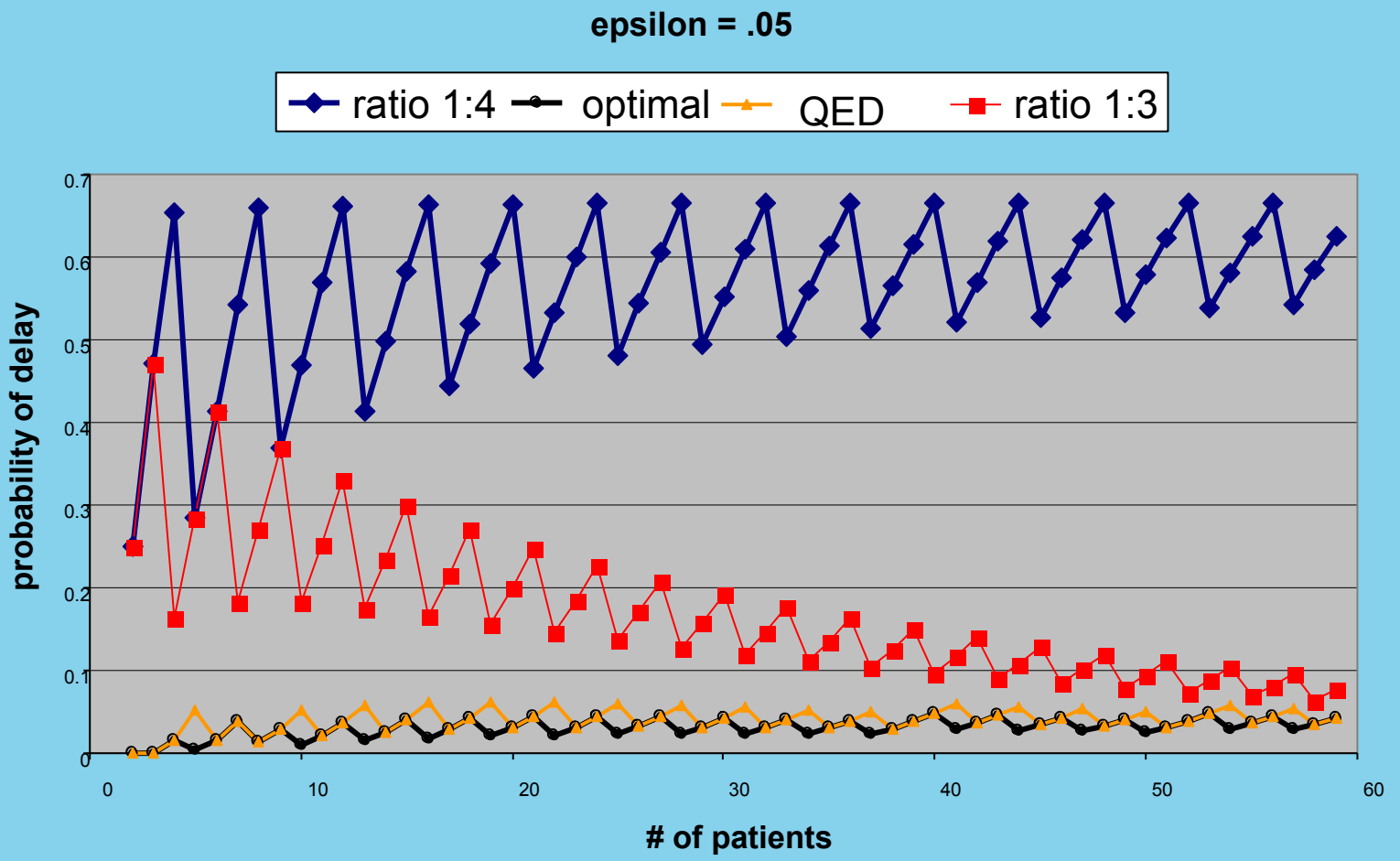
# Performance:
# Ratio vs. Optimal vs. QED (T=0)



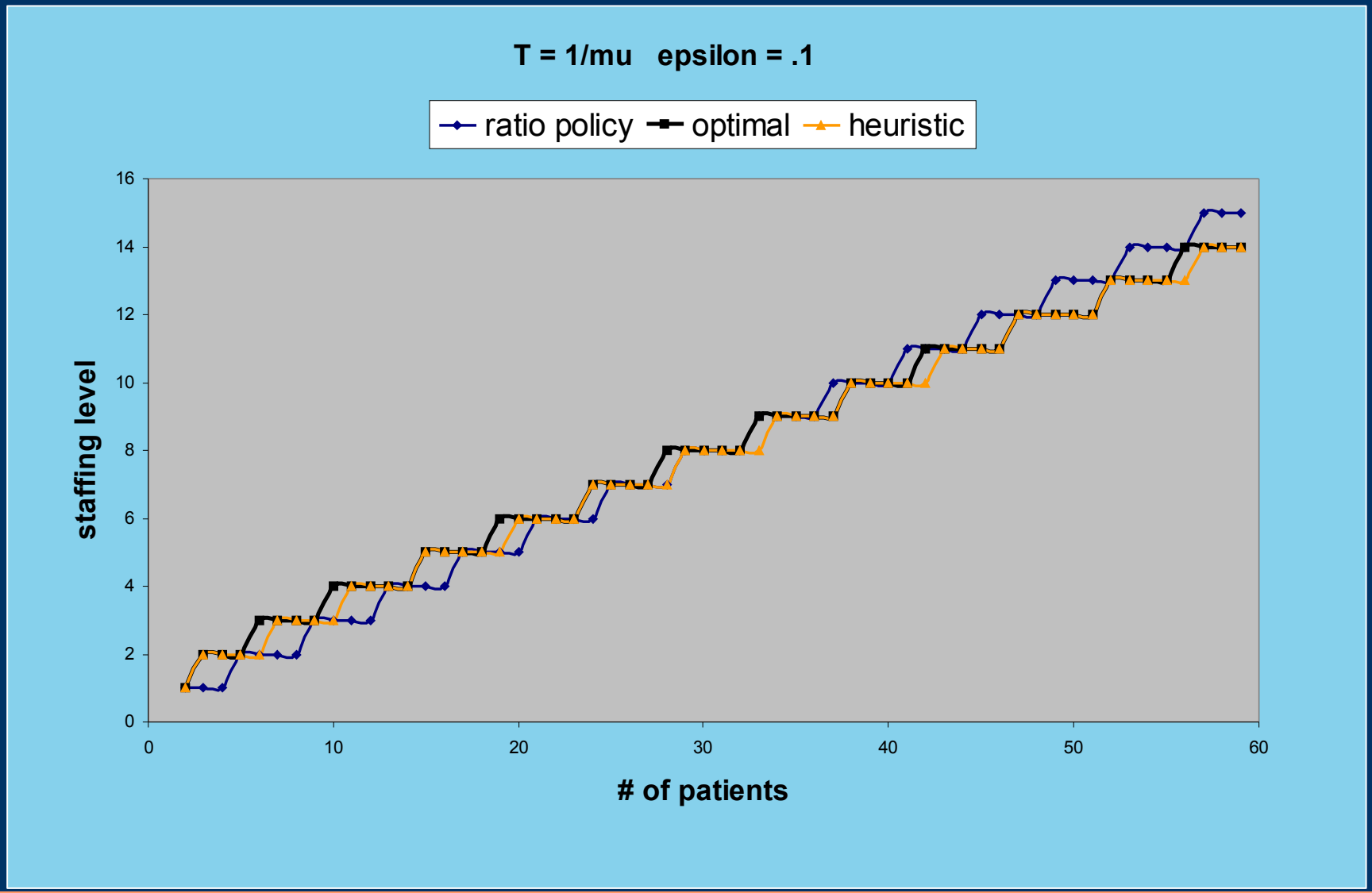
# Staffing:
# Ratio vs. Optimal vs. QED (T=0)



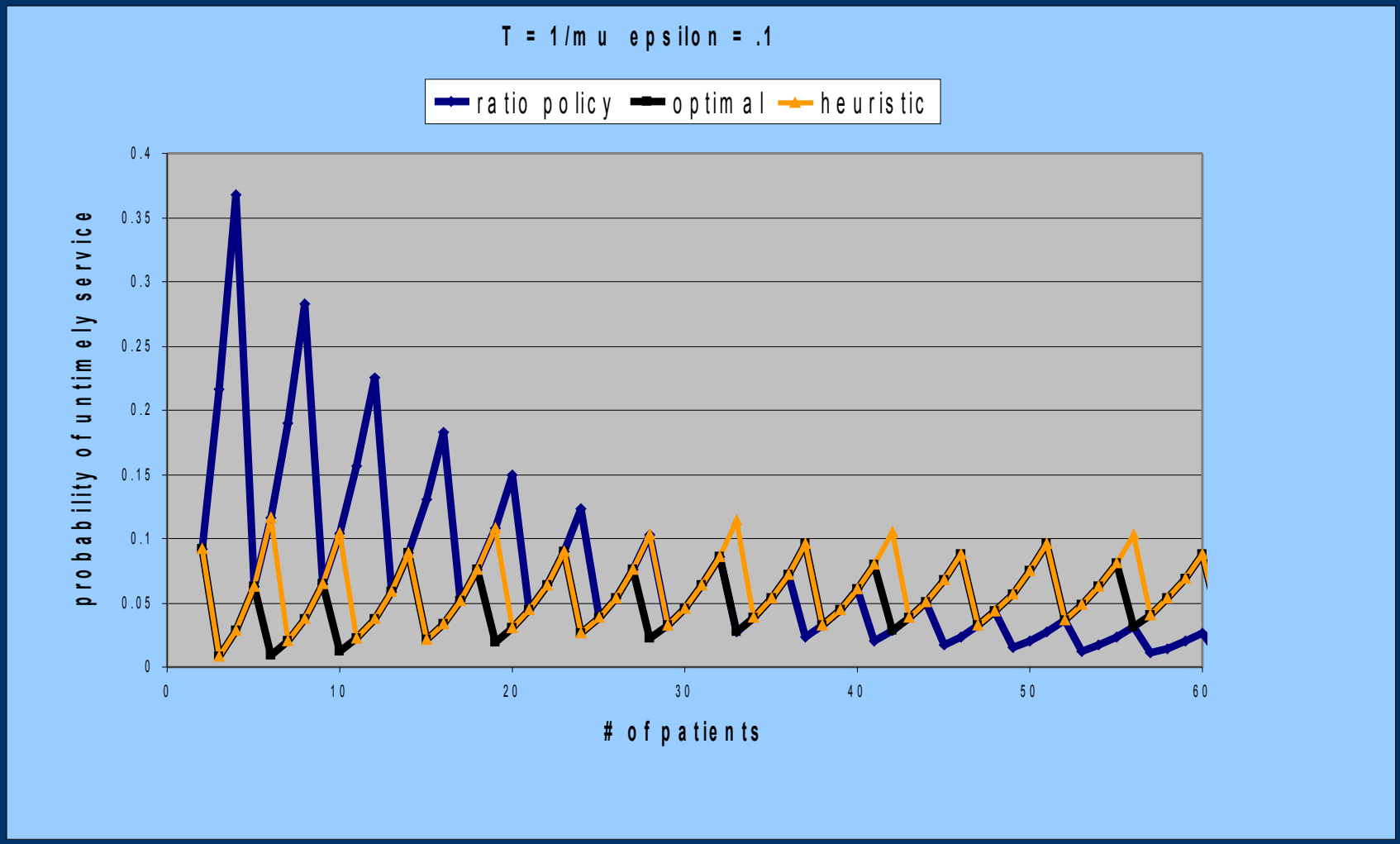
# Performance:
# Ratio vs. Optimal vs. QED (T=0)



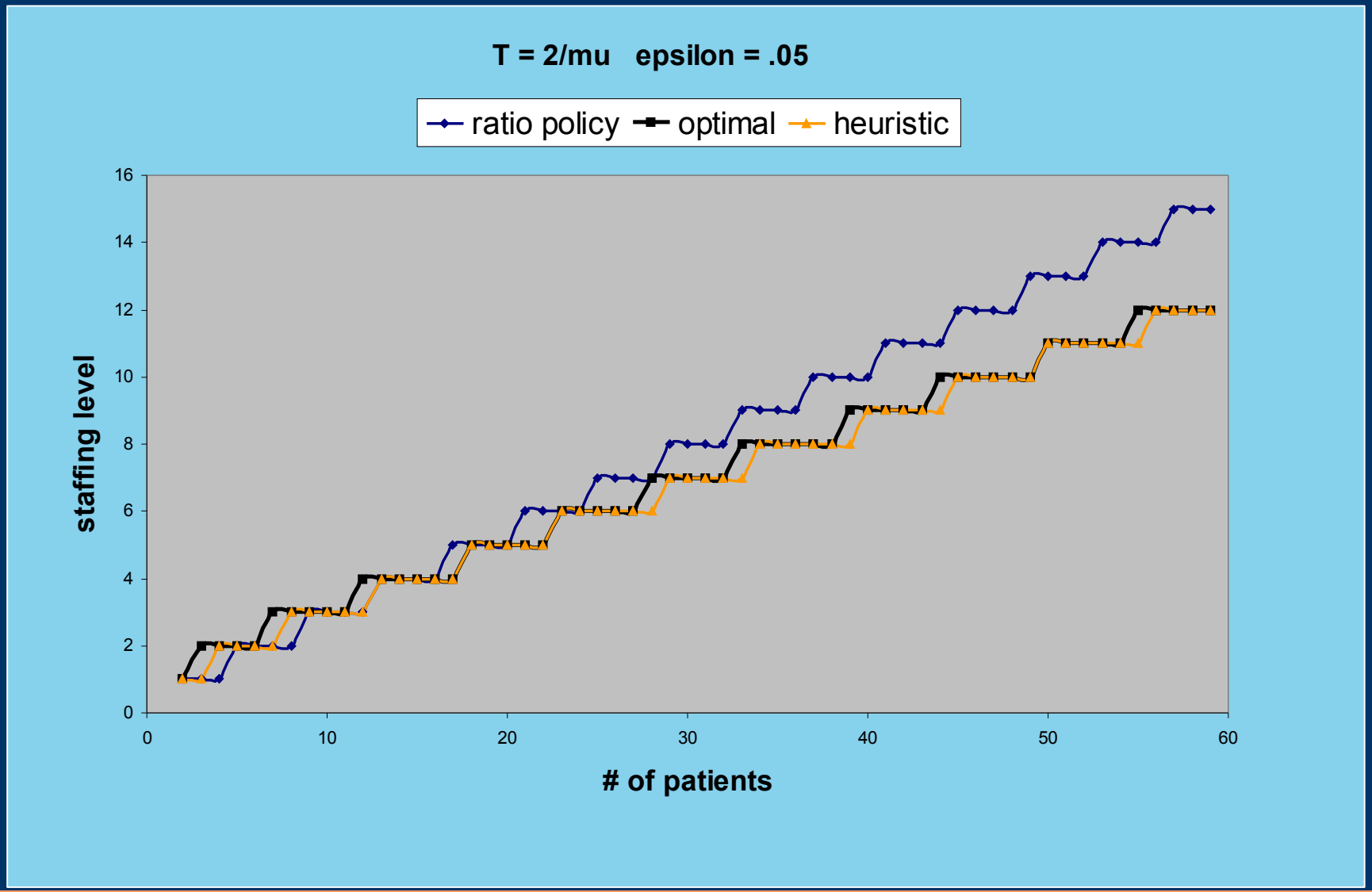
# Timely Service Staffing: Ratio vs. Optimal vs. Heuristic ED



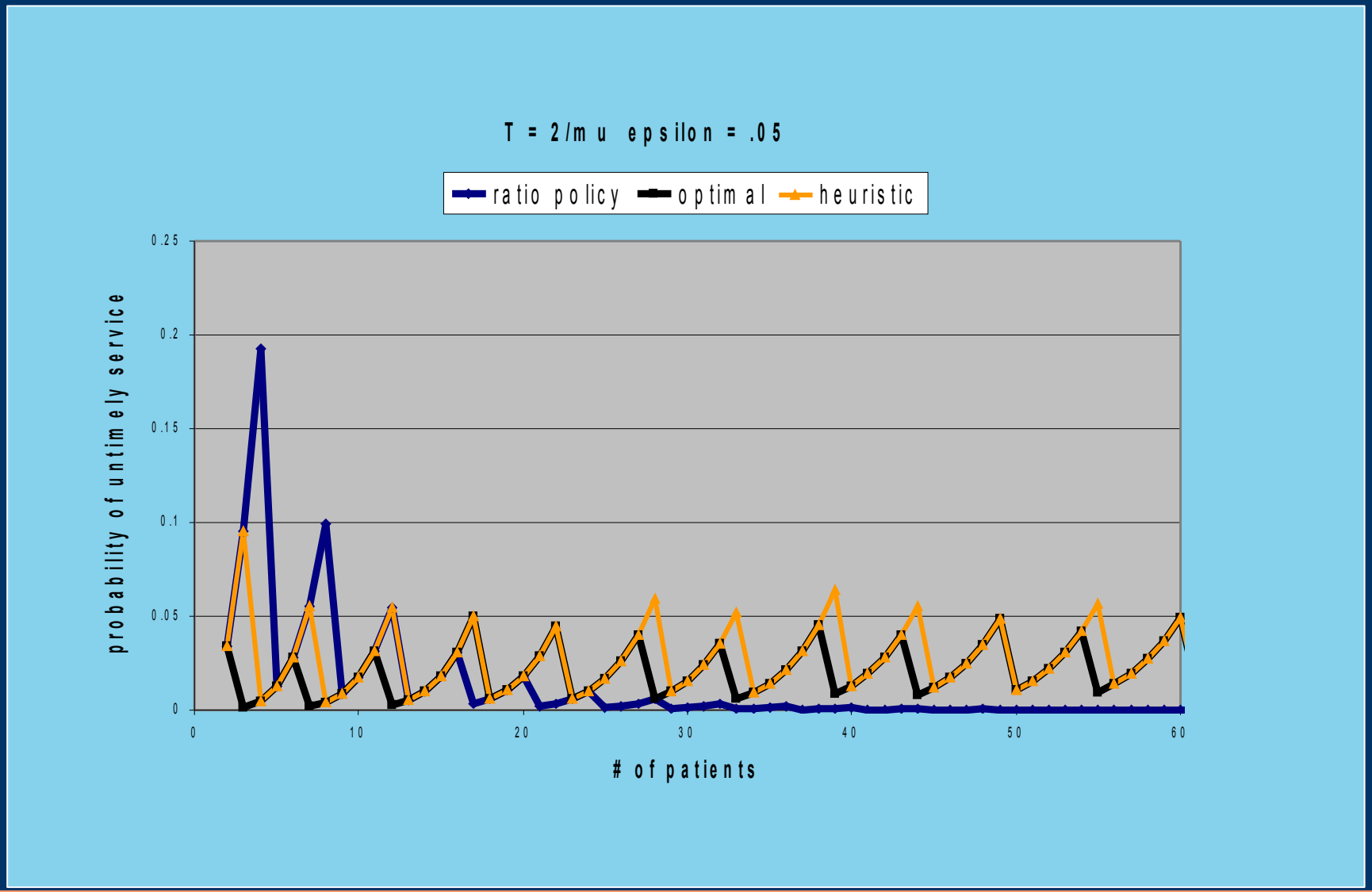
# Timely Service Performance: Ratio vs. Optimal vs. Heuristic ED



# Timely Service Staffing: Ratio vs. Optimal vs. Heuristic ED



# Timely Service Performance: Ratio vs. Optimal vs. Heuristic ED



# Conclusions / Managerial Insights

- Size Matters
  - Under mild nurse pooling assumptions, the ratios will not provide uniform QOS across units of varying sizes.
- Typically, small units are understaffed and large units are overstaffed.
- Inconsistencies under ratio policies worsen as T increases and ε decreases.
- QED and ED Staffing Rules
  - Nurse-patient ratios provide a (first order) starting point. QED and ED staffing rules provide the necessary adjustments so that uniform QOS is achieved.

# Future research

- Heterogeneous workforce
- Heterogeneous patients - Acuity levels
- Different probability distribution for service and activation times
- Varying N and S simultaneously

# QED Staffing Policy

**Proposition**: The approximate probability of delay has a nondegenerate limit $\alpha \in (0,1)$ if and only if

$$\beta_n = \left( \frac{s_n}{n} - r \right) \sqrt{n} \to \beta, \qquad as \quad n \to \infty,$$

for some $\beta \in (-\infty, \infty)$, with

$$\alpha = \left( 1 + e^{\frac{-\beta^2}{r^2}} \sqrt{r} \frac{\Phi\left( \frac{\beta}{\sqrt{r\bar{r}}} \right)}{\Phi\left( \frac{-\beta}{r\sqrt{\bar{r}}} \right)} \right)^{-1}$$

# QED Staffing Policy

**Proof**: A. Represent the probability of delay as combination of two random variable (Xn and Yn)

B. Apply CLT for each one.

C. Determine the limit of each variable as n tend to infinity.

D. Determine the limit of the probability of delay.

## Appendix A: Proof of Proposition 1

**Proof:** Assume first that Condition (2) holds. Note then that

$$P(Q_n \geq s_n) = \left(1 + \frac{A_n}{B_n}\right)^{-1} \tag{29}$$

with

$$A_n = \sum_{k=0}^{s_n-1} \binom{n}{k} \rho^k$$
$$B_n = \sum_{k=s_n}^{n} \binom{n}{k} \frac{k!}{s_n!} s_n^{s_n-k} \rho^k$$

Using the definitions of $r$ and $\overline{r}$, note that $A_n$ can be rewritten as follows,

$$\begin{aligned} A_n &= (1+\rho)^n \sum_{k=0}^{s_n-1} \binom{n}{k} r^k \overline{r}^{n-k} \\ &= (1+\rho)^n P(X \leq s_n - 1) \end{aligned}$$

where $X$ is a binomial random variable with parameters $n$ and $r$, i.e. $X \sim Bi(n, r)$. Similarly, $B_n$ is equal to

$$\begin{aligned} B_n &= \frac{n!}{s_n!} \frac{\rho^n}{s_n^{n-s_n}} e^{s_n/\rho} \sum_{k=s_n}^{n} \frac{1}{(n-k)!} \left(\frac{s_n}{\rho}\right)^{n-k} e^{-s_n/\rho} \\ &= \frac{n!}{s_n!} \frac{\rho^n}{s_n^{n-s_n}} e^{s_n/\rho} P(Y \leq n - s_n) \end{aligned}$$

where $Y$ is a Poisson random variable with rate $s_n/\rho$, i.e. $Y \sim P(s_n/\rho)$. As a result the probability of delay can be written as

$$P(Q_n \geq s_n) = \left(1 + C_n \frac{P(X \leq s_n - 1)}{P(Y \leq n - s_n)}\right)^{-1} \tag{30}$$

where

$$C_n = \frac{s_n!}{n!} \left(\frac{s_n}{\rho}\right)^n \frac{(1+\rho)^n}{s_n^{s_n}} e^{-s_n/\rho} \tag{31}$$

Note first that

$$P(X \leq s_n - 1) = P\left(\frac{X - nr}{\sqrt{nr\overline{r}}} \leq \frac{s_n - 1 - nr}{\sqrt{nr\overline{r}}}\right) \tag{32}$$

where the convergence of $(s_n - 1 - nr)/\sqrt{n}$ to the limit $\beta$ is equivalent to Condition (2). It follows that, using the central limit theorem,

$$P(X \leq s_n - 1) \to \Phi\left(\beta/\sqrt{r\overline{r}}\right). \tag{33}$$

Similarly,

$$\begin{aligned} P(Y \leq n - s_n) &= P\left(\frac{Y - s_n/\rho}{\sqrt{s_n/\rho}} \leq \frac{n - s_n - s_n/\rho}{\sqrt{s_n/\rho}}\right) \\ &= P\left(\frac{Y - s_n/\rho}{\sqrt{s_n/\rho}} \leq \frac{rn - s_n}{\sqrt{r\overline{r}}\sqrt{s_n}}\right) \end{aligned}$$

where the second equality is obtained using $\rho = r/\overline{r}$. From Condition (2), $(rn - s_n)/\sqrt{s_n} \to -\beta/\sqrt{r}$ when $n \to +\infty$. Using again the central limit theorem, we have then

$$P(Y \leq n - s_n) \to \Phi\left(\frac{-\beta}{r\sqrt{\overline{r}}}\right). \tag{34}$$

It remains to study the limit of $C_n$ as $n \to +\infty$. To that end, we apply the Stirling's formula $n! \sim (2\pi n)^{1/2} n^n e^{-n}$ twice, to $n$ and to $s_n$. It follows then from the definition of $C_n$ (Equation 31),

$$C_n \sim \sqrt{r}\left(1 + \frac{\beta}{r\sqrt{n}}\right)^{n+1/2} e^{n - s_n/r}. \tag{35}$$

Since $(1 + \beta/(r\sqrt{n}))^{n+1/2} \sim e^{\beta\sqrt{n}/r - \beta^2/r^2}$ and $n - s_n/r = -\beta\sqrt{n}/r$ from Condition (2), we deduce that $C_n \to e^{-\beta^2/r^2}\sqrt{r}$ and

$$P(Q_n \geq s_n) \to f(\beta) \tag{36}$$

where $f : \mathbb{R} \mapsto (0,1)$ is a strictly decreasing function such that

$$f(\beta) = \left(1 + e^{-\beta^2/r^2}\sqrt{r}\frac{\Phi\left(\frac{\beta}{\sqrt{r\overline{r}}}\right)}{\Phi\left(\frac{-\beta}{r\sqrt{\overline{r}}}\right)}\right)^{-1} \tag{37}$$

Assume now that $a_n$ has a limit $\alpha$, $0 < \alpha < 1$ and that $\beta \neq f^{-1}(\alpha)$ is a (possibly infinite) limit point of $\{(s_n/n - r)\sqrt{n}\}$. Assume for now that $\beta > f^{-1}(\alpha)$. Construct a sequence $\{s'_n\}$ such that $s'_n \leq s_n$ and $(s'_n/n - r)\sqrt{n} \to \beta' = \min((\beta + f^{-1}(\alpha))/2, f^{-1}(\alpha) + 1)$, as $n \to \infty$. Notice that $f^{-1}(\alpha) < \beta' < \infty$, which implies $\alpha > f(\beta')$. Let $Q'_n$ denote the number of users in the $n$th system with $s'_n$ servers. Since $s'_n \leq s_n$, $P(Q'_n \geq s'_n) \geq P(Q_n \geq s_n)$. However, taking the limit of both sides yields $f(\beta') \geq \alpha$, a contradiction. A similar argument shows that $\beta < f^{-1}(\alpha)$ is also impossible. Hence, the convergence $a_n \to \alpha \in (0,1)$ implies $\{(s_n/n - r)\sqrt{n}\}$ has a unique limit as well.□