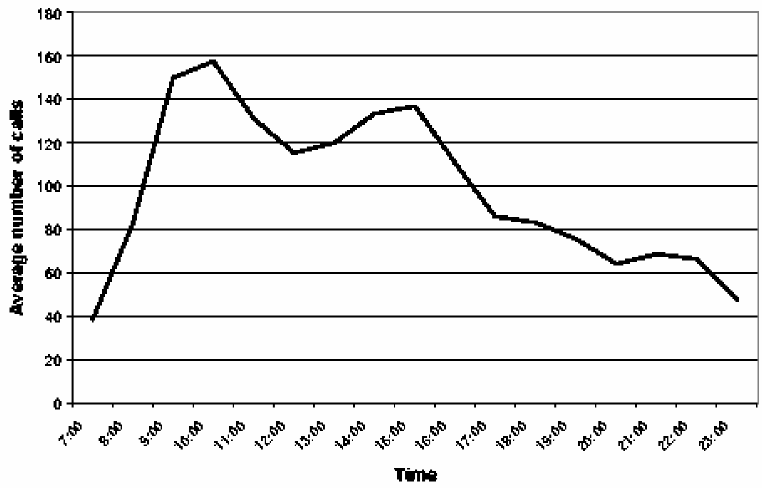
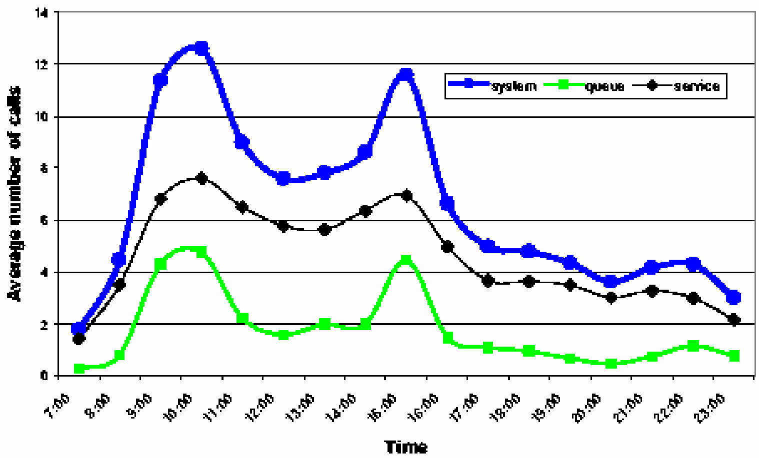
Predictable Variability = "Piecewise Stationary" load

Goal: Predictably Stable Performance, but HOW?

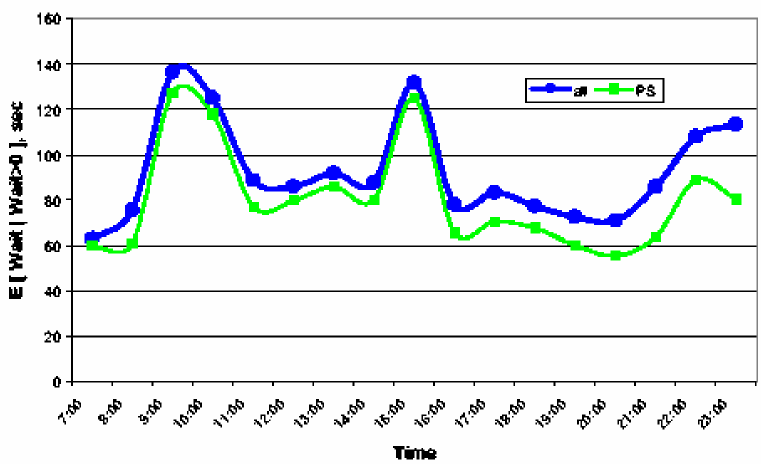
Arrivals



Queues



Waiting



8

Staffing Time-Varying Queues:

Two Common Approaches:

SSA – Simple Stationary Approximation.

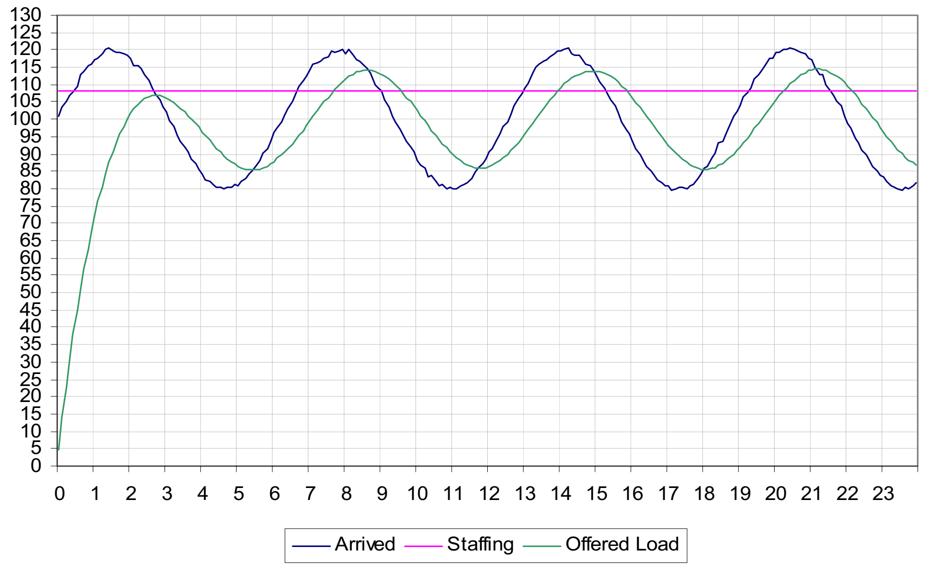
Constant staffing levels, based on steady-state M/M/N, with λ=long-run average number of arrivals.

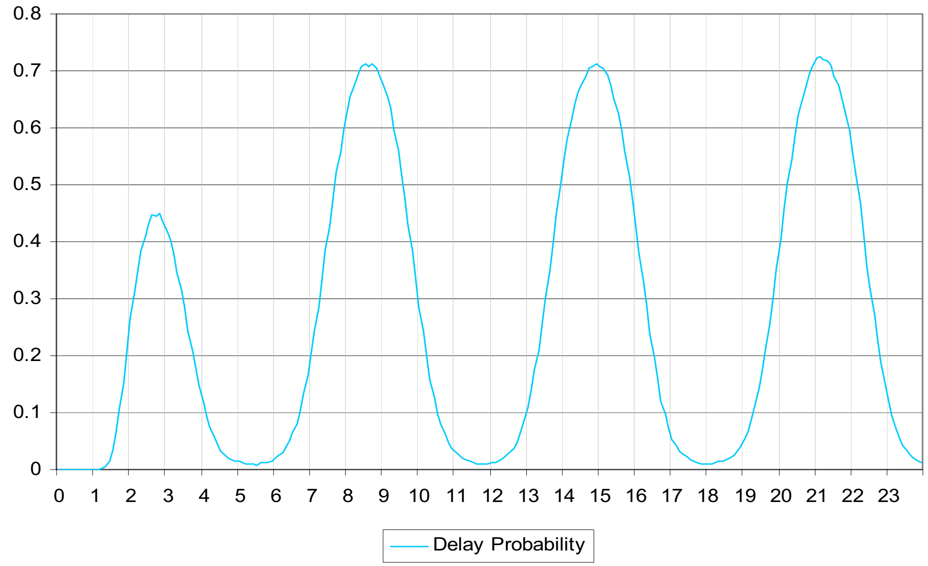
PSA – Point-wise Stationary Approximation.

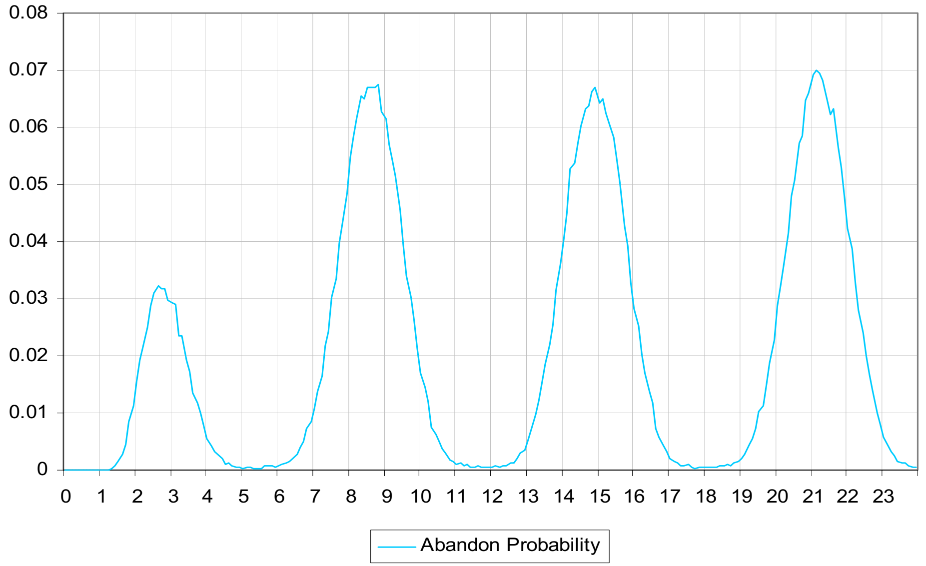
Time-varying staffing levels, based on steady-state M/M/N, with $\lambda = \lambda(t)$ at each time t.

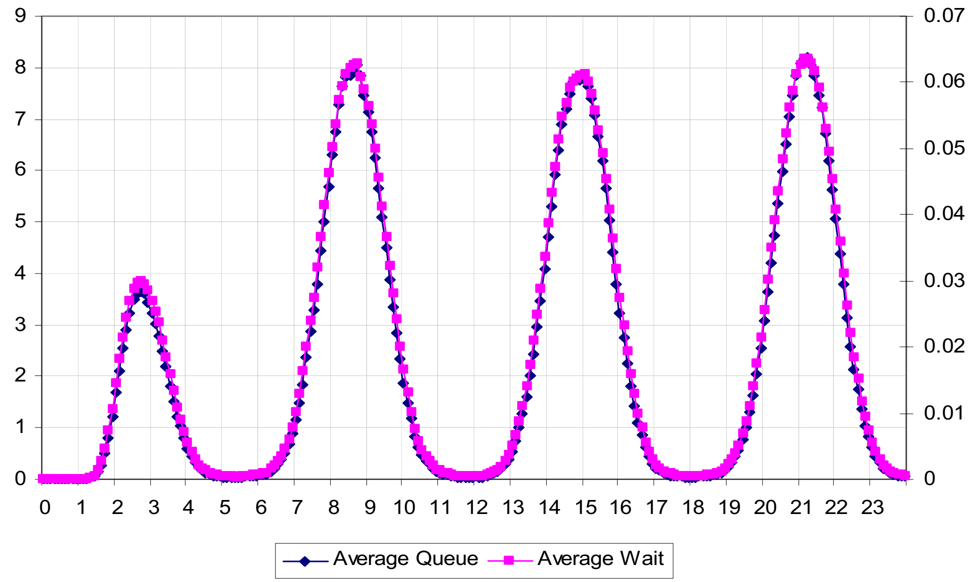
Could result in time-varying (highly oscillating) performance (utilization, service), which is undesirable.

Simple Stationary Approximation (SSA, α=0.2)

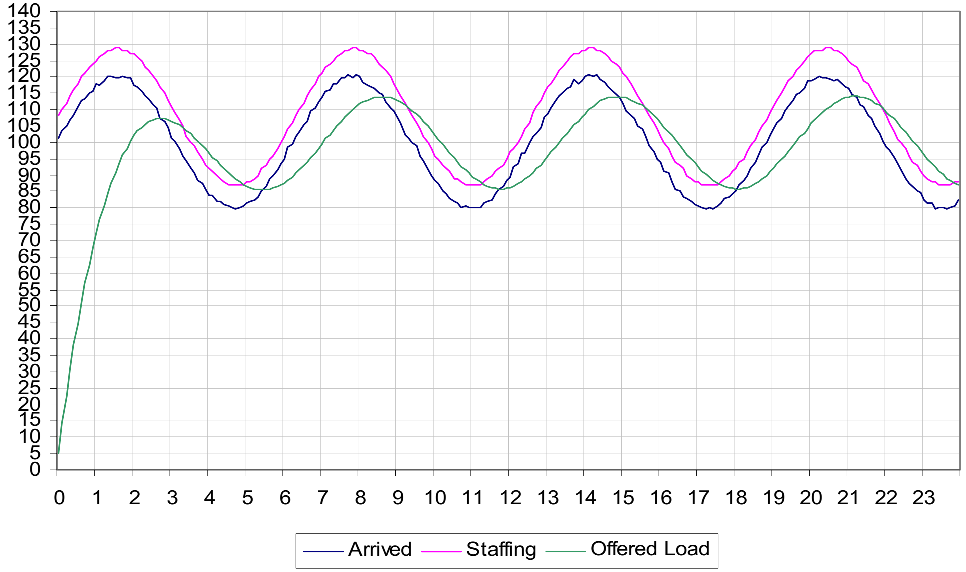


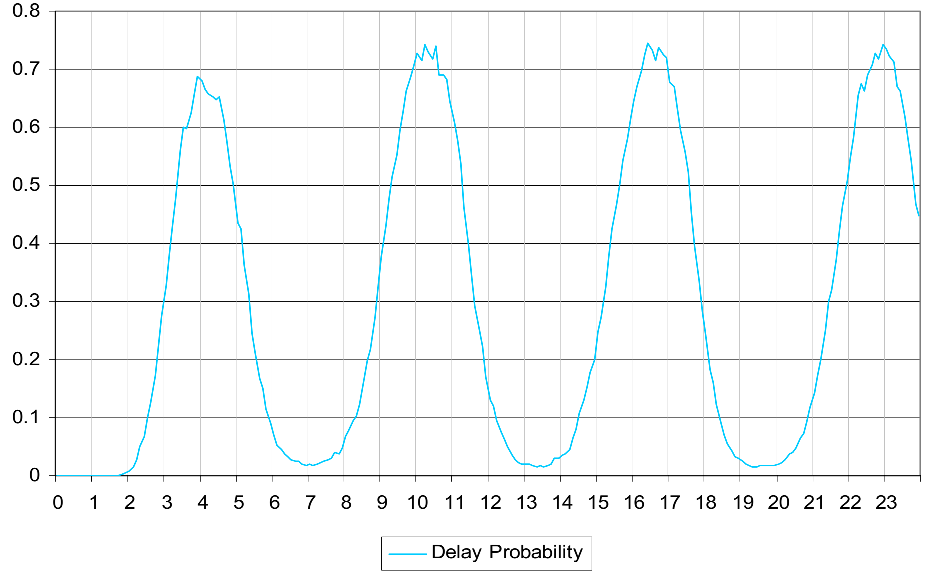


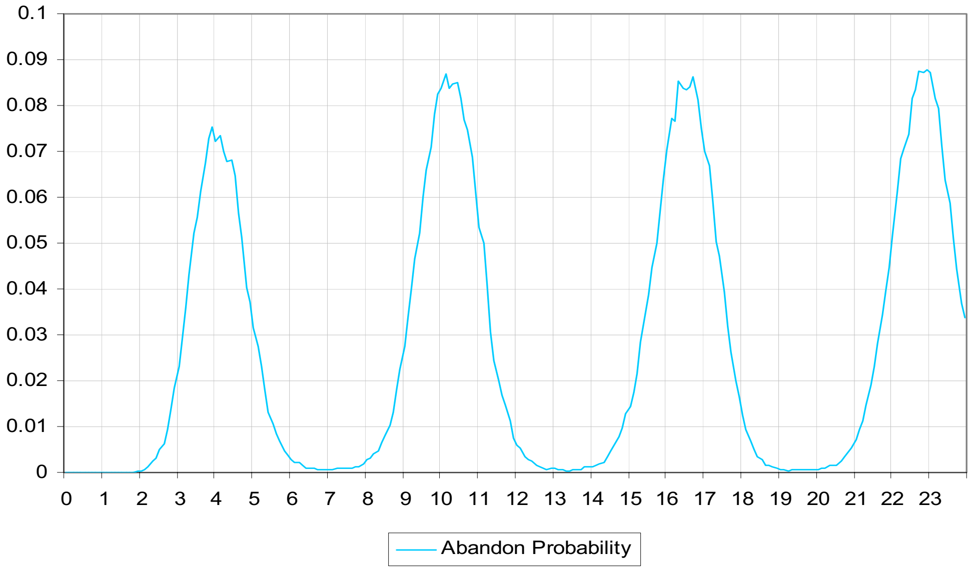


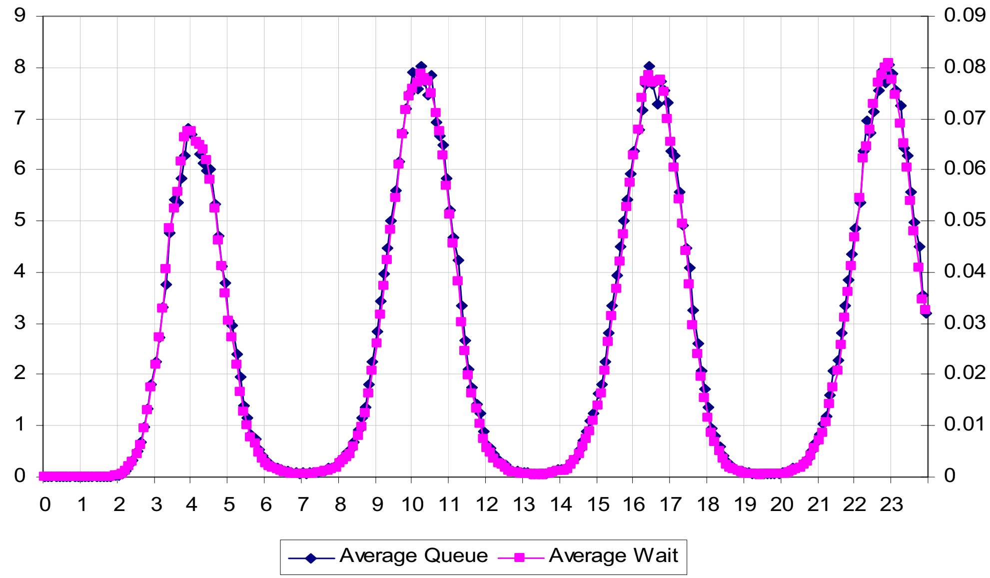


Point-wise Stationary Approximation (PSA, α=0.2)





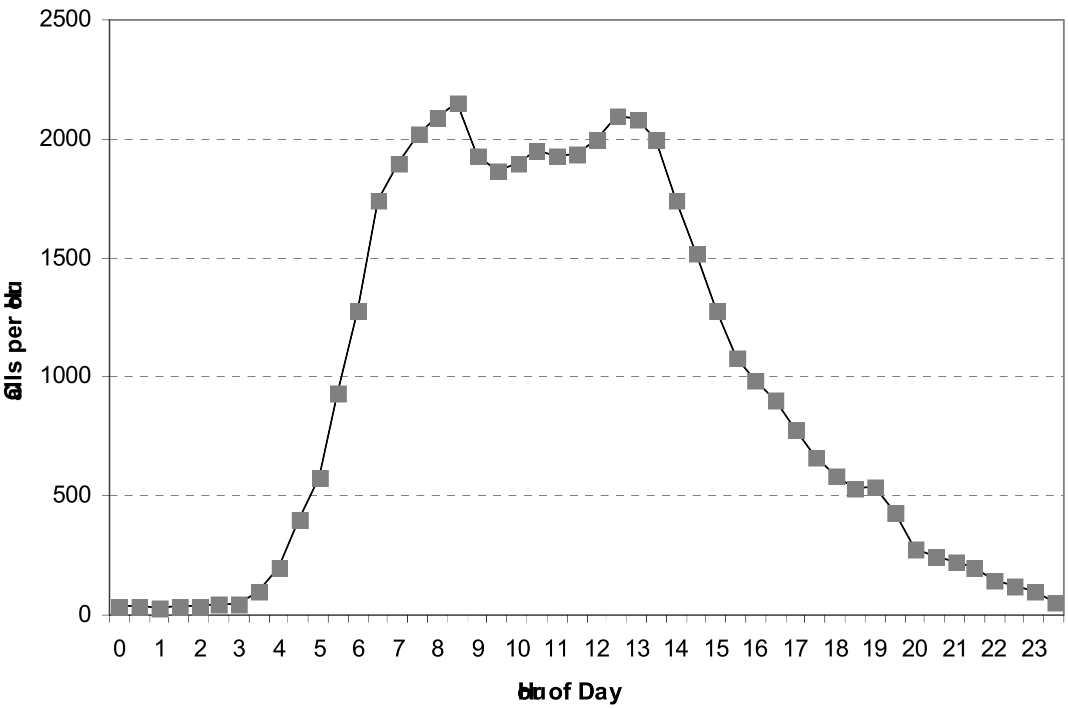




Example: "Real" Call Center

Two-hump arrival functions are common

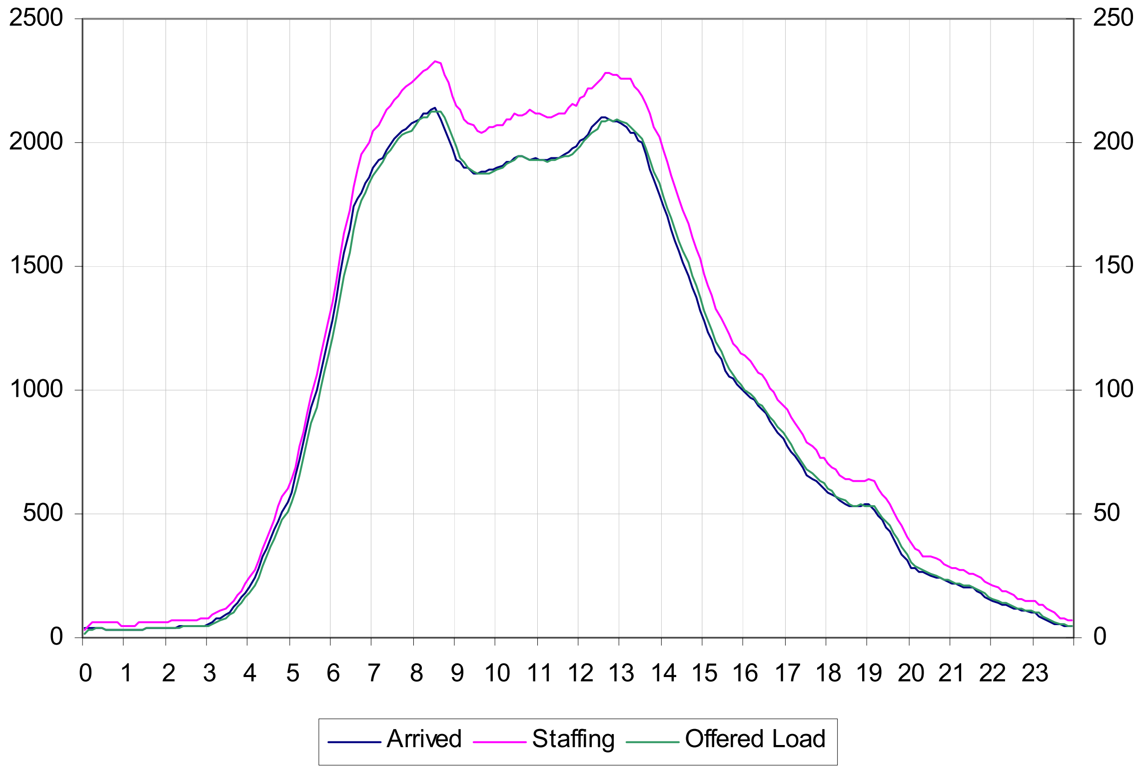
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)



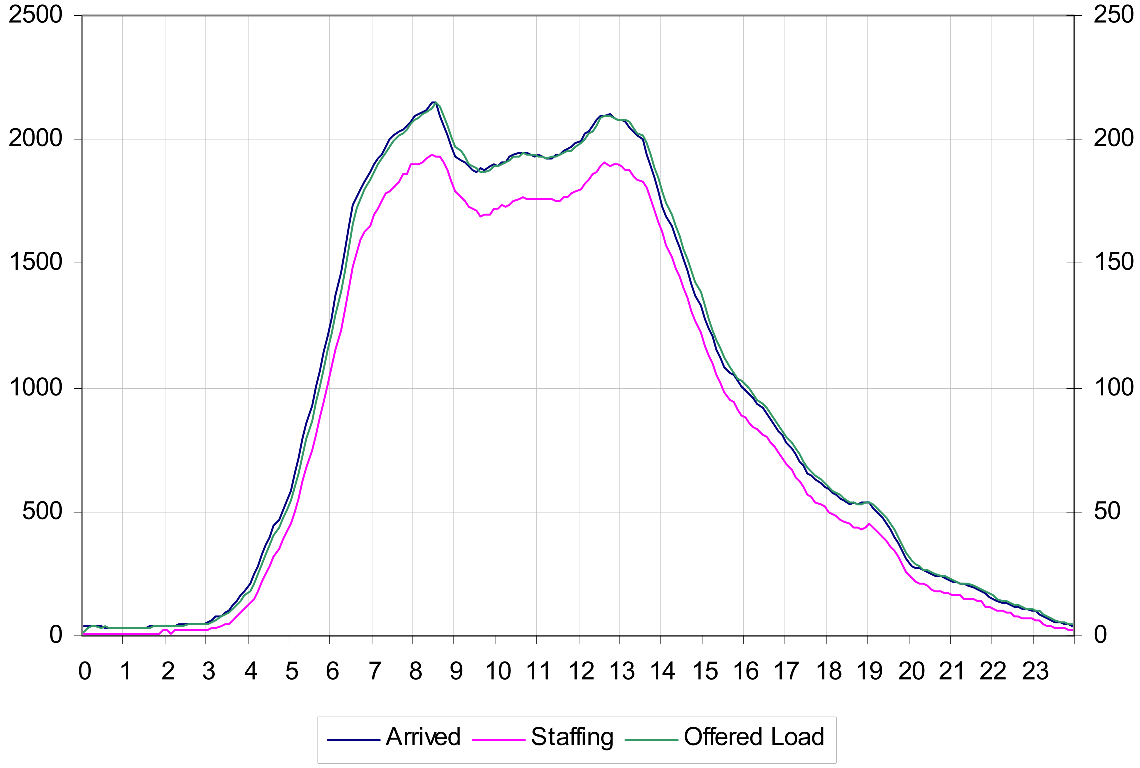
Assume: Service and abandonment rates are both

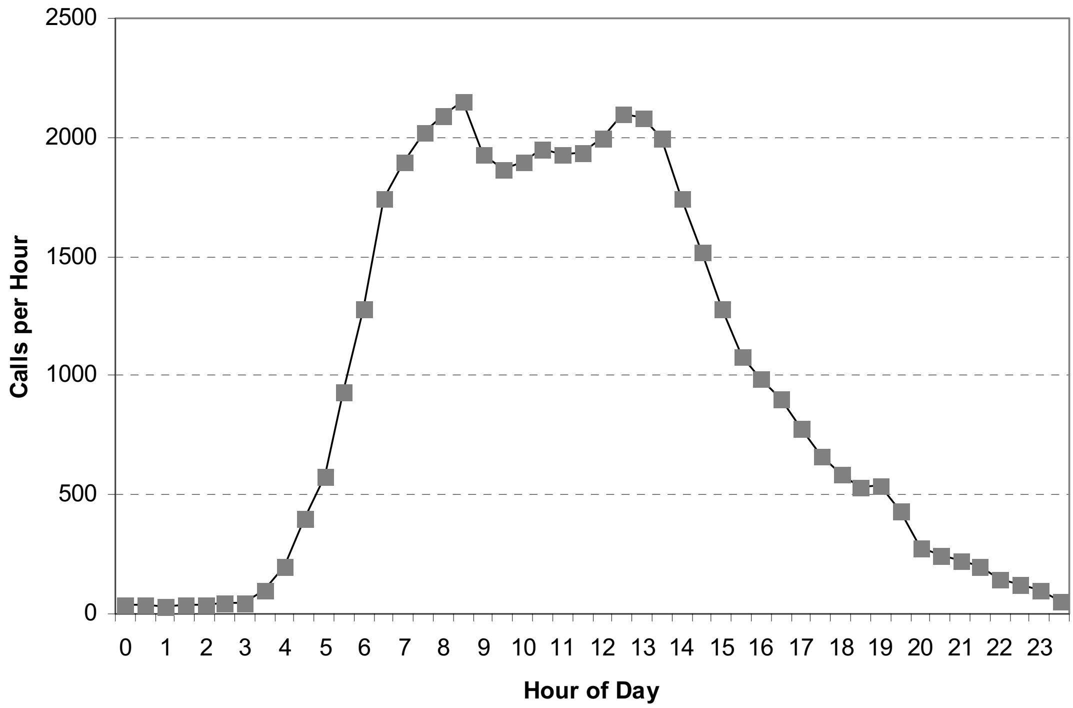
exponential having mean 0.1 (6 min.)

QD Staffing (α=0.1)

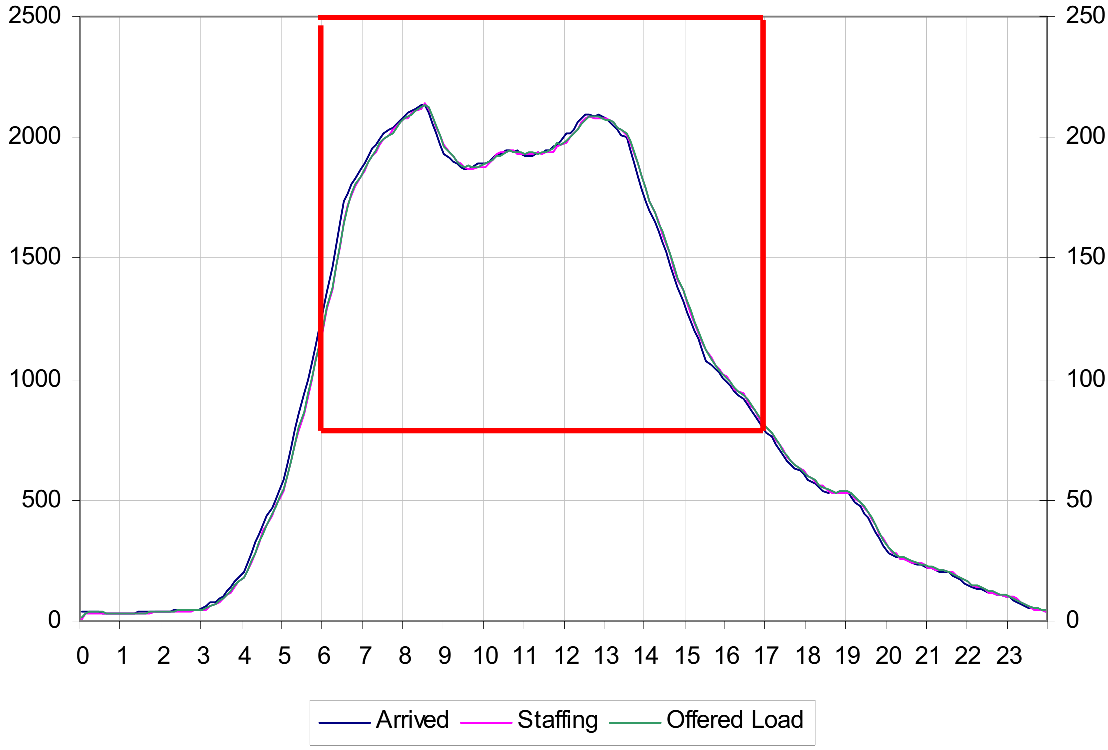


ED Staffing (α=0.9)

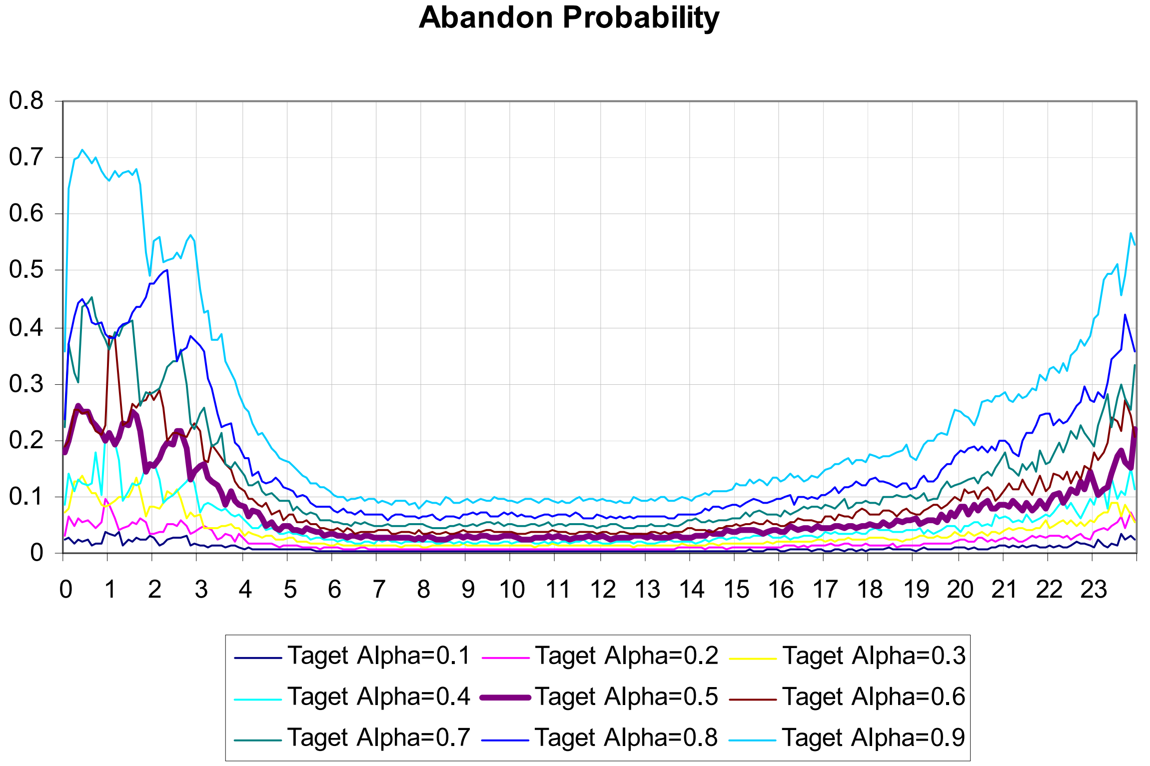


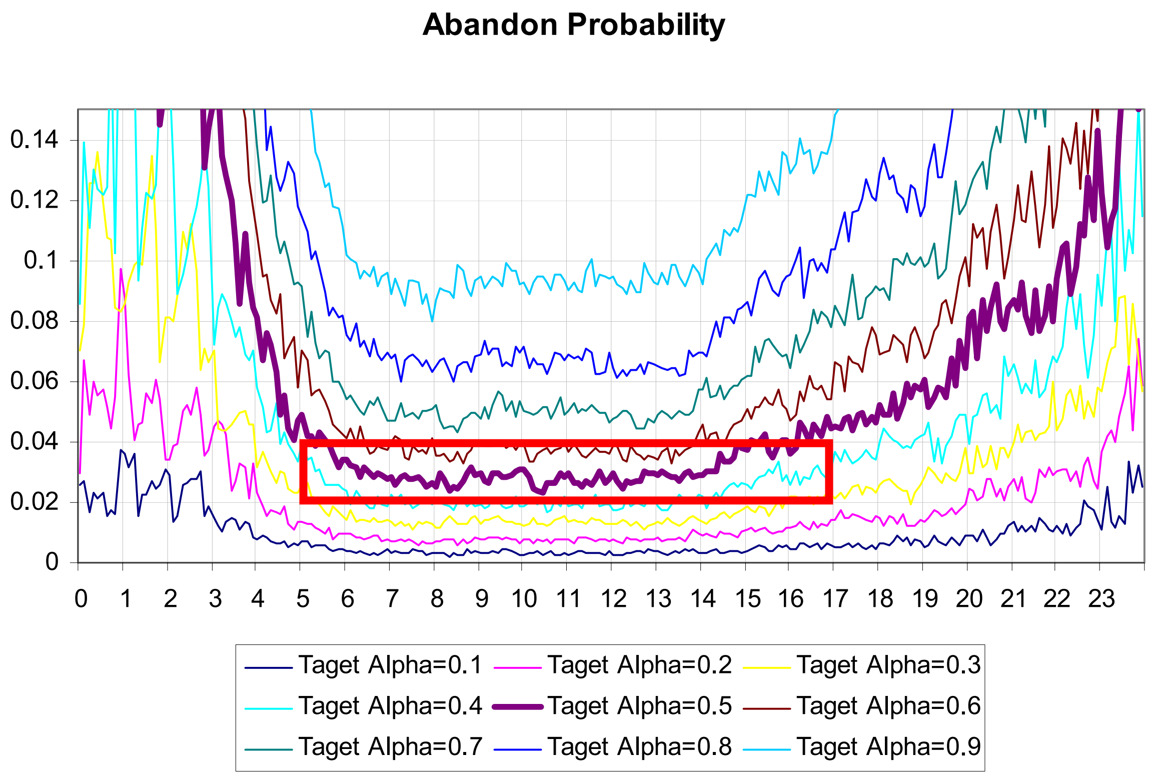


QED Staffing (α=0.5)

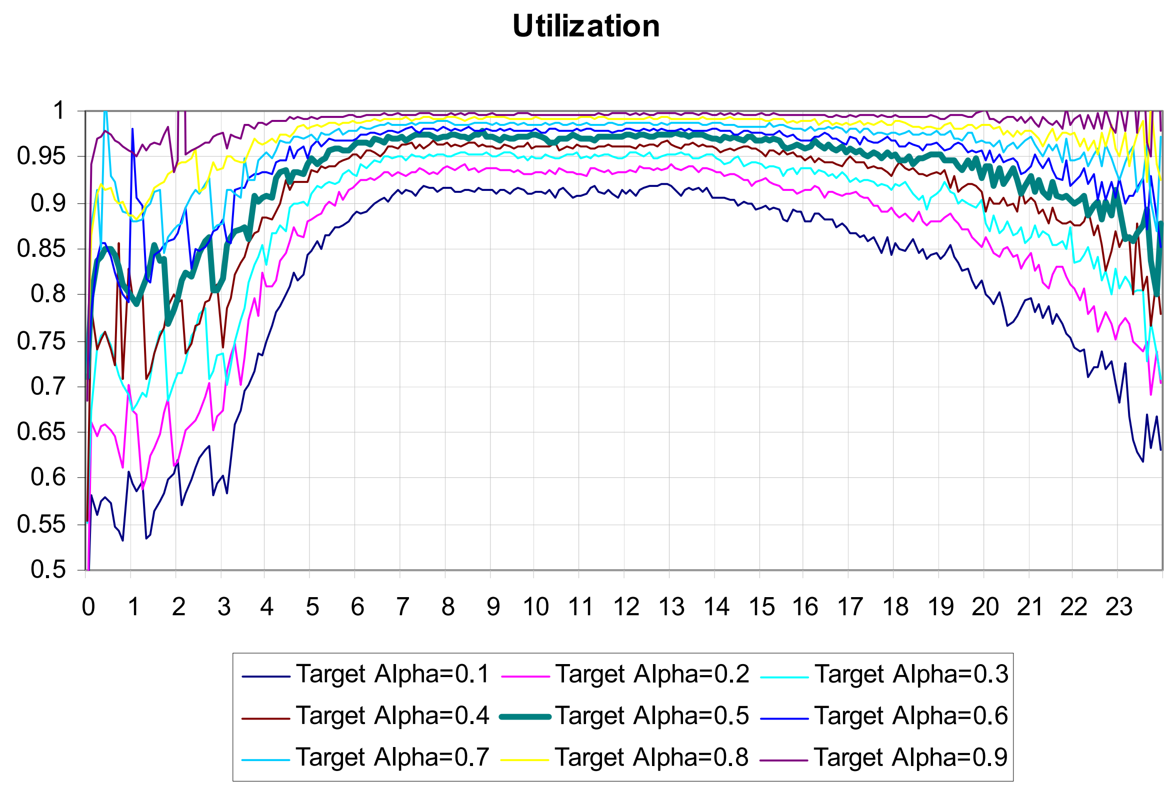


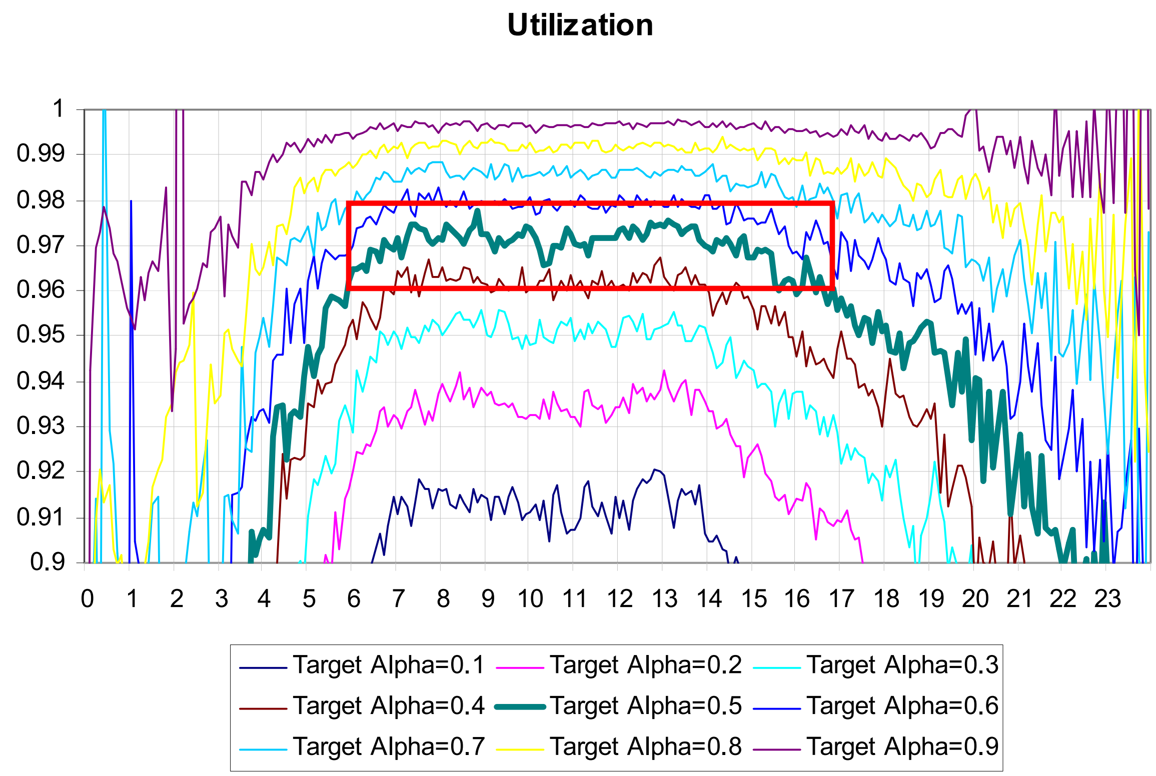
Abandon Probability





Utilization

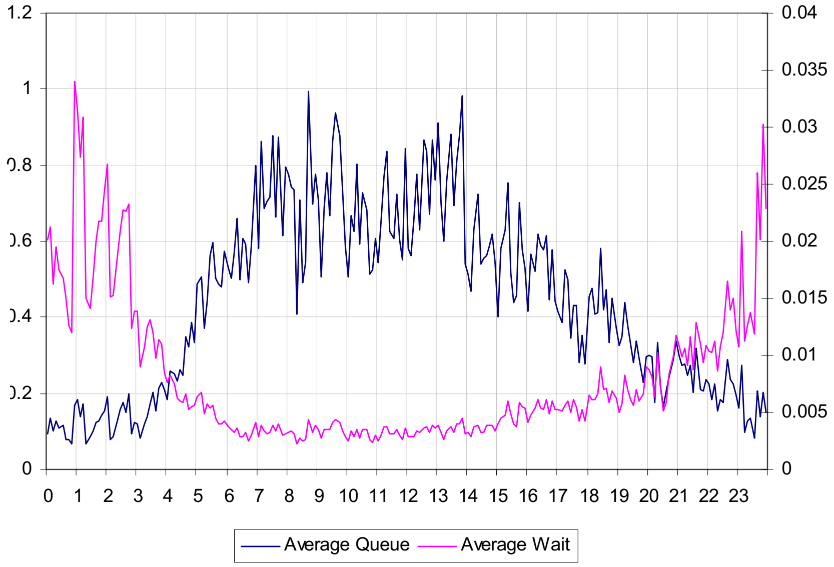




Congestion (Queue, Wait)

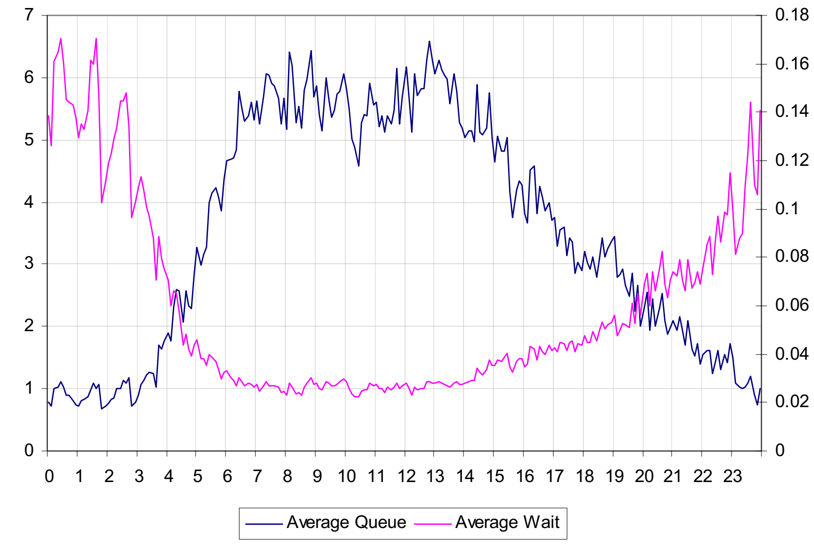
QD
α=0.1

Negligible



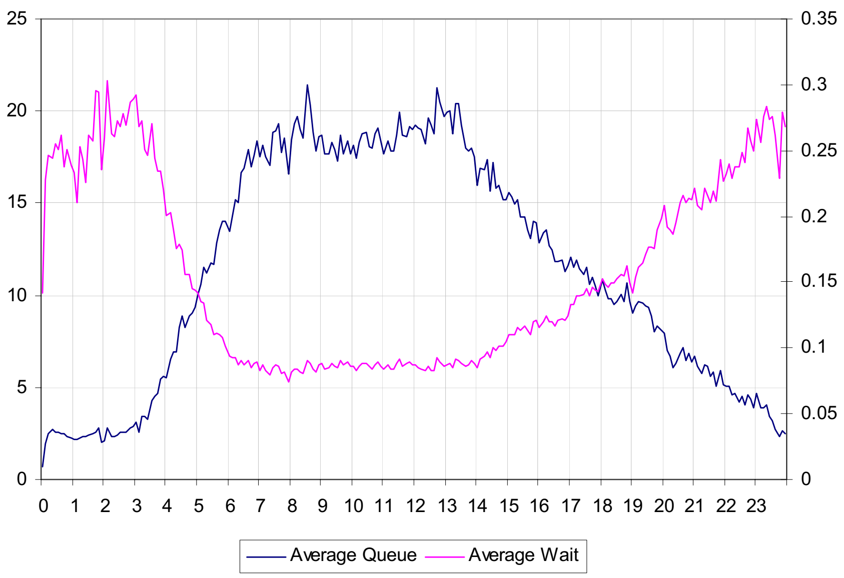
QED
α=0.5

Seconds



ED
α=0.9

Minutes

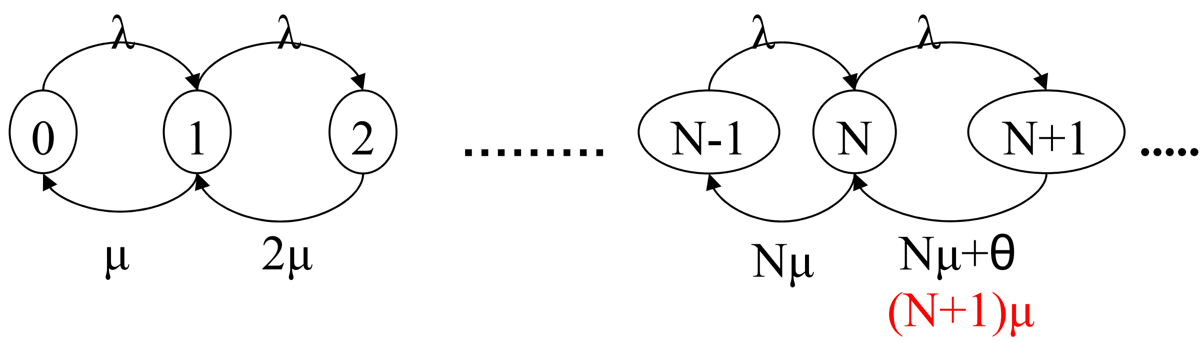


Erlang-A: Moderate (Im)patience

- M/M/N + M queue, with
  service rate μ equals θ abandonment rate

- L_t: number-in-system at time t (Birth & Death)

- For **any** N, transition-rates for $\{L_t, t \geq 0\}$:



Note: The same transition rates as **M/M/∞**

Square-Root Staffing: Motivation

$$P\{W_q(M/M/N+M)>0\} \underset{PASTA}{=} P\{L(M/M/N+M)\ge N\} \underset{\theta=\mu}{=} P\{L(M/M/\infty)\ge N\}$$

Fact: $L(M/M/\infty) \sim \text{Poisson}(R)$; $R=\lambda/\mu$ offered load

For R not too small:

$$\mathrm{L(M/M/\infty)} \overset{\mathrm{d}}{\approx} \mathrm{Normal(R,R)} \overset{\mathrm{d}}{=} \mathrm{R} + \mathrm{Z}\sqrt{\mathrm{R}}$$

$$\Rightarrow \quad P\{W_q>0\} \approx P\left\{Z \ge \frac{N-R}{\sqrt{R}}\right\} = 1-\phi\left(\frac{N-R}{\sqrt{R}}\right)$$

Given target delay-probability $\alpha = 1-\phi\left(\frac{\mathrm{N-R}}{\sqrt{\mathrm{R}}}\right)$

$$\Rightarrow \quad N = R+\beta\cdot\sqrt{R}, \quad \text{with} \quad \beta=\phi^{-1}(1-\alpha)$$

N is the "least integer for which" $P\{W_q>0\}\le\alpha$

Time-Varying Arrivals

Extension: $M_t / M / N_t + M \quad (\mu=\theta)$

$$N_t = R_t + \beta \cdot \sqrt{R_t} \quad \textbf{?}$$

Fact: $L_t \sim \text{Poisson}(R_t)$

R_t – the offered load at time t, namely:

$$R_t = E\lambda(t - S_e) \cdot E(S) = E \int_{t-S}^{t} \lambda(u) du$$

S_e – excess service $\left(E(S_e) = E(S) \dfrac{1 + c_s^2}{2} \right)$

Time-Varying Arrivals

Extension: $M_t / M / N_t + M \quad (\mu=\theta)$

$$N_t = R_t + \beta \cdot \sqrt{R_t} \quad \textbf{?}$$

Fact: $L_t \sim \text{Poisson}(R_t)$

R_t – the offered load at time t, namely:

$$R_t = E\lambda(t - S_e) \cdot E(S) = E \int_{t-S}^{t} \lambda(u)du$$

Se – excess service $\left(E(S_e) = E(S) \frac{1 + c_s^2}{2} \right)$

$L_t \overset{d}{\approx} N(R_t, R_t)$ hence, as before:

$$\Rightarrow \quad N_t = \left\lceil R_t + \beta \cdot \sqrt{R_t} \right\rceil, \quad \beta = \phi^{-1}(1-\alpha)$$

hopefully yields time-stable delay probability α:

Indeed, but in fact **TIME-STABLE PERFORMANCE !**

What if $\mu \neq \theta$?

Use an *Iterative Algorithm* that is *Simulation-Based*

Performance Measures

- ***Delay probability** in interval t*, calculated by the fraction of customers who are not served immediately upon arrival, out of all arriving customers during the t time-interval

- ***Average waiting time** in interval t*, calculated by the average waiting time of all customers arriving during the t time-interval.

- ***Average queue length** in interval t,* taken constant over the time-interval. The queue length is averaged over all replications

- ***Tail probability** in interval t*, calculated as the probability that queue size equals or exceeds some threshold (e.g. 3 times average queue)

- ***Servers' Utilization** in interval t*, calculated as the fraction of busy-servers during a time-interval (accounting for servers who are busy only a fraction of the interval)

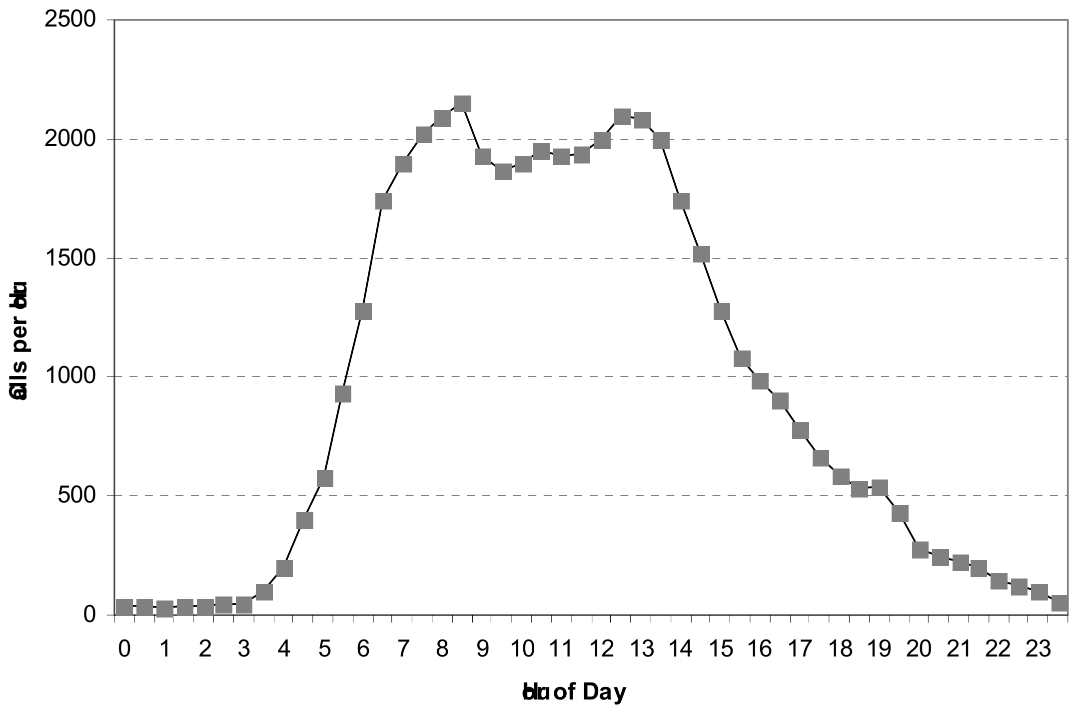
- ***Service grade** β_t in interval t*, which arises from the following "Square-Root Staffing" rule:

$$N_t = R_t + \beta_t \sqrt{R_t}$$

Example: "Real" Call Center

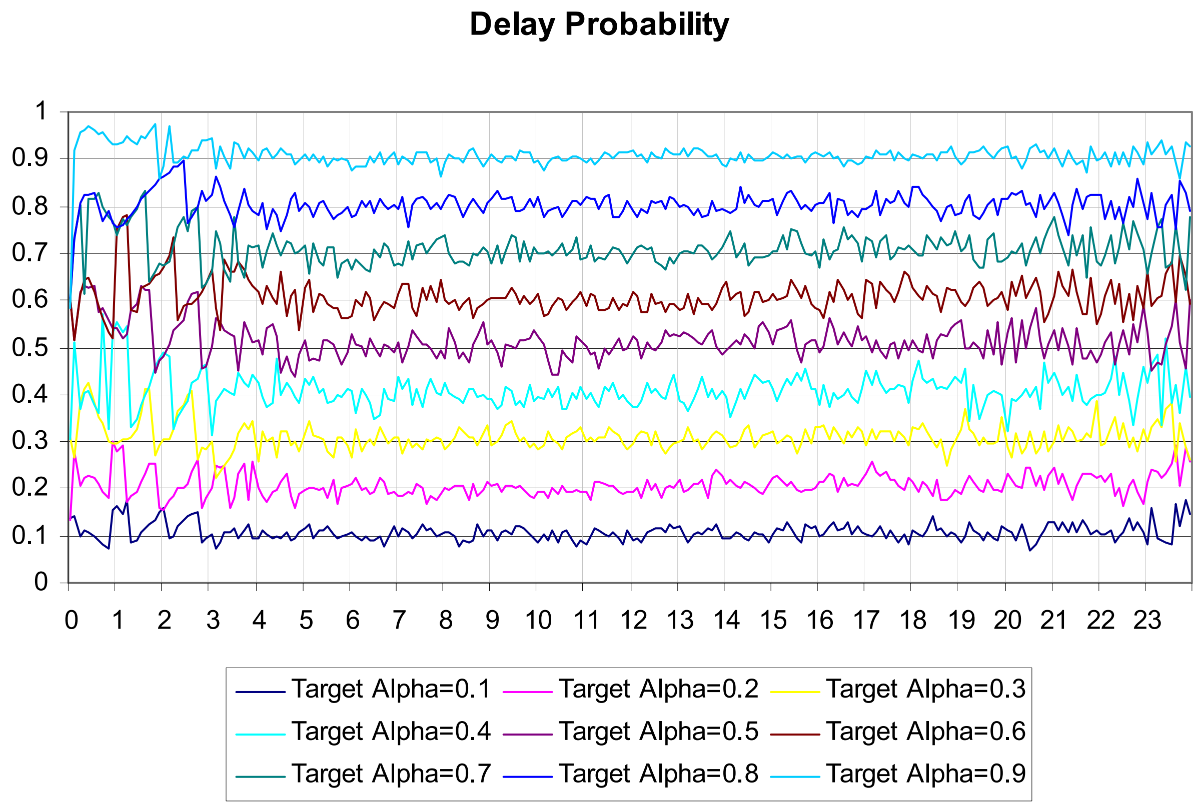
Two-hump arrival functiona are typical

(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)

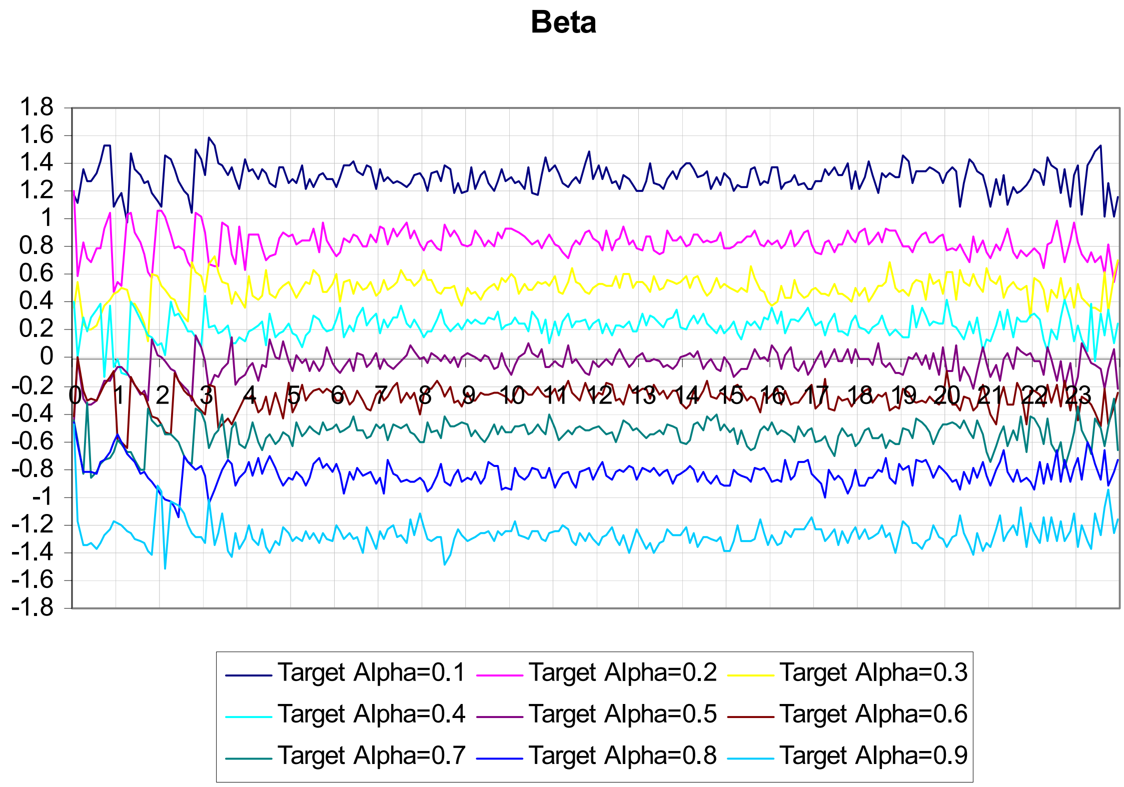


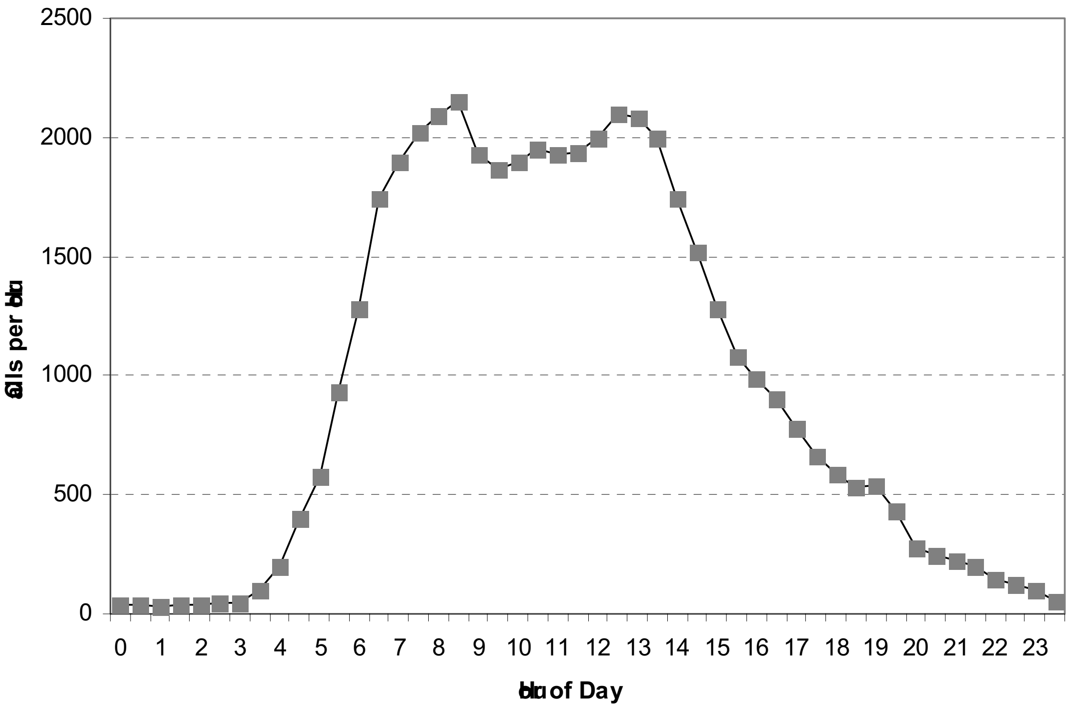
- Service and abandonment rates are both exponential having mean 0.1 (6 min.)

Delay Probability α

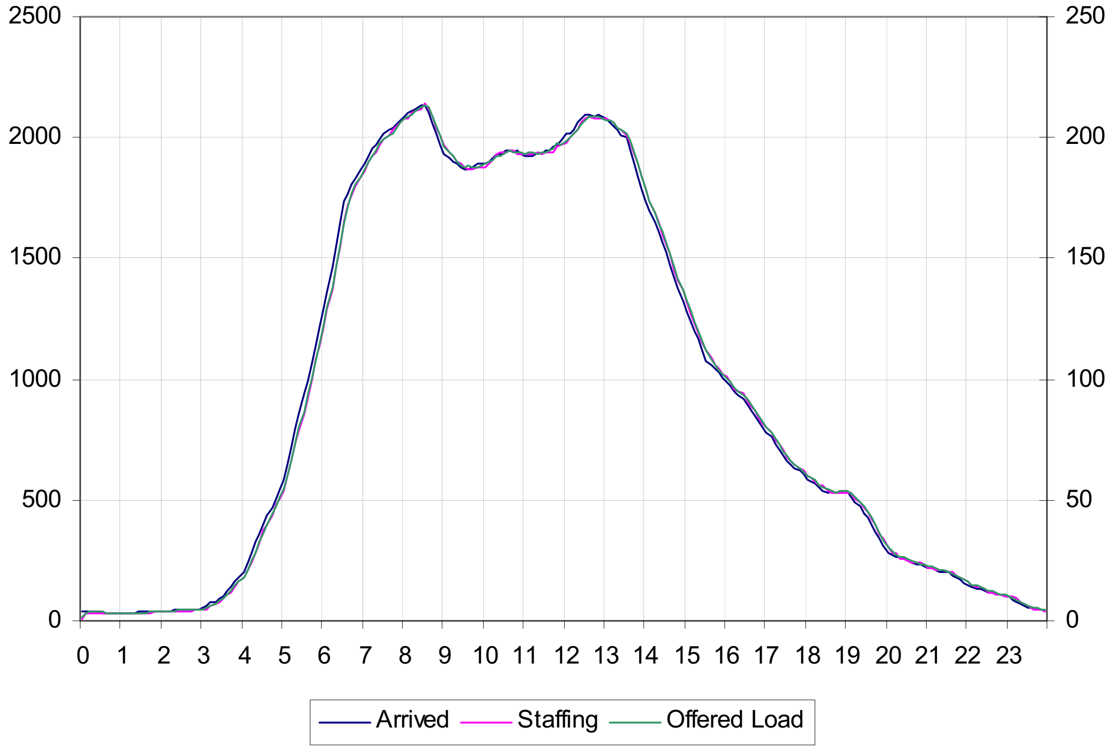


Service Grade β



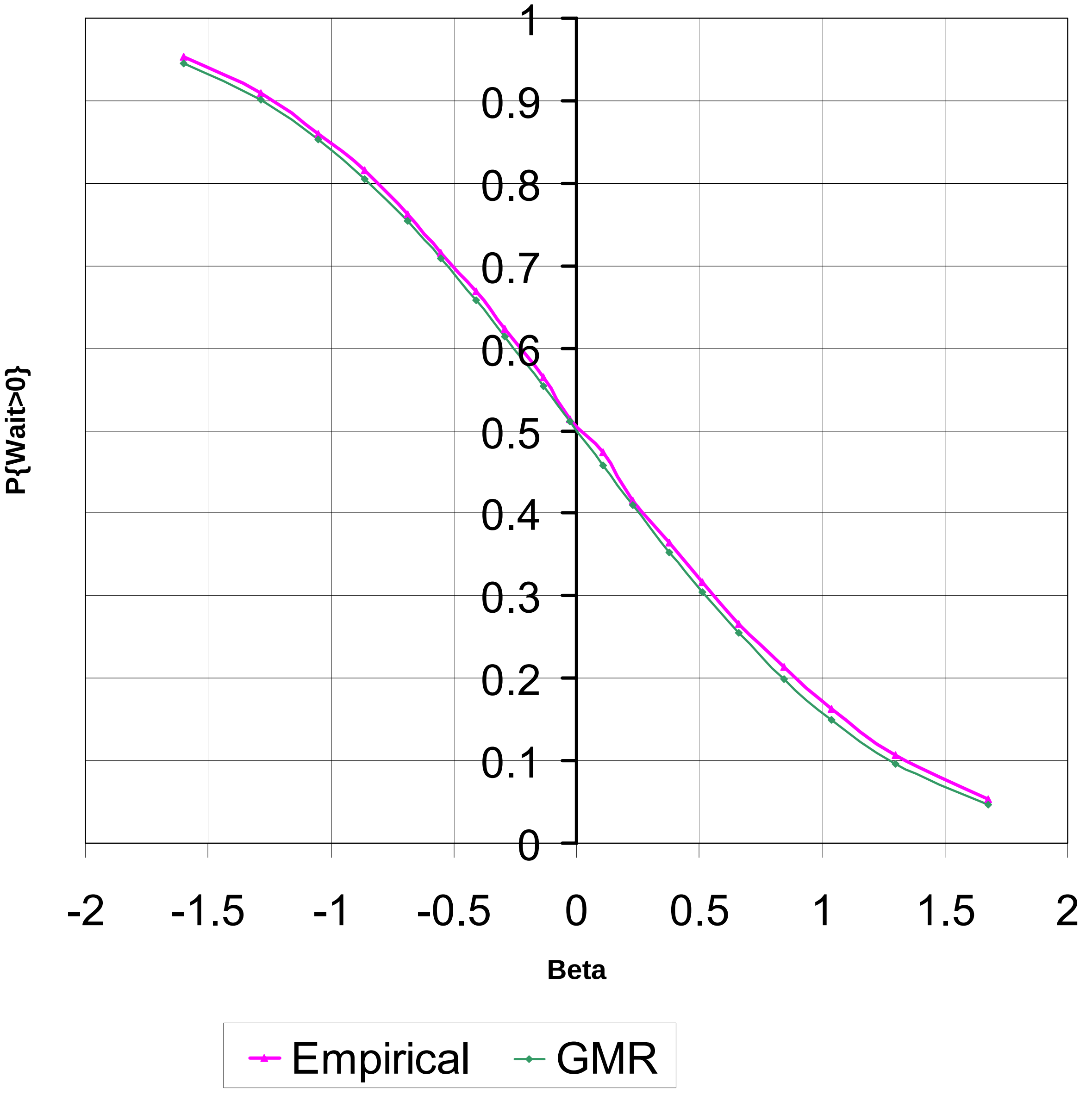


QED Staffing (α=0.5)

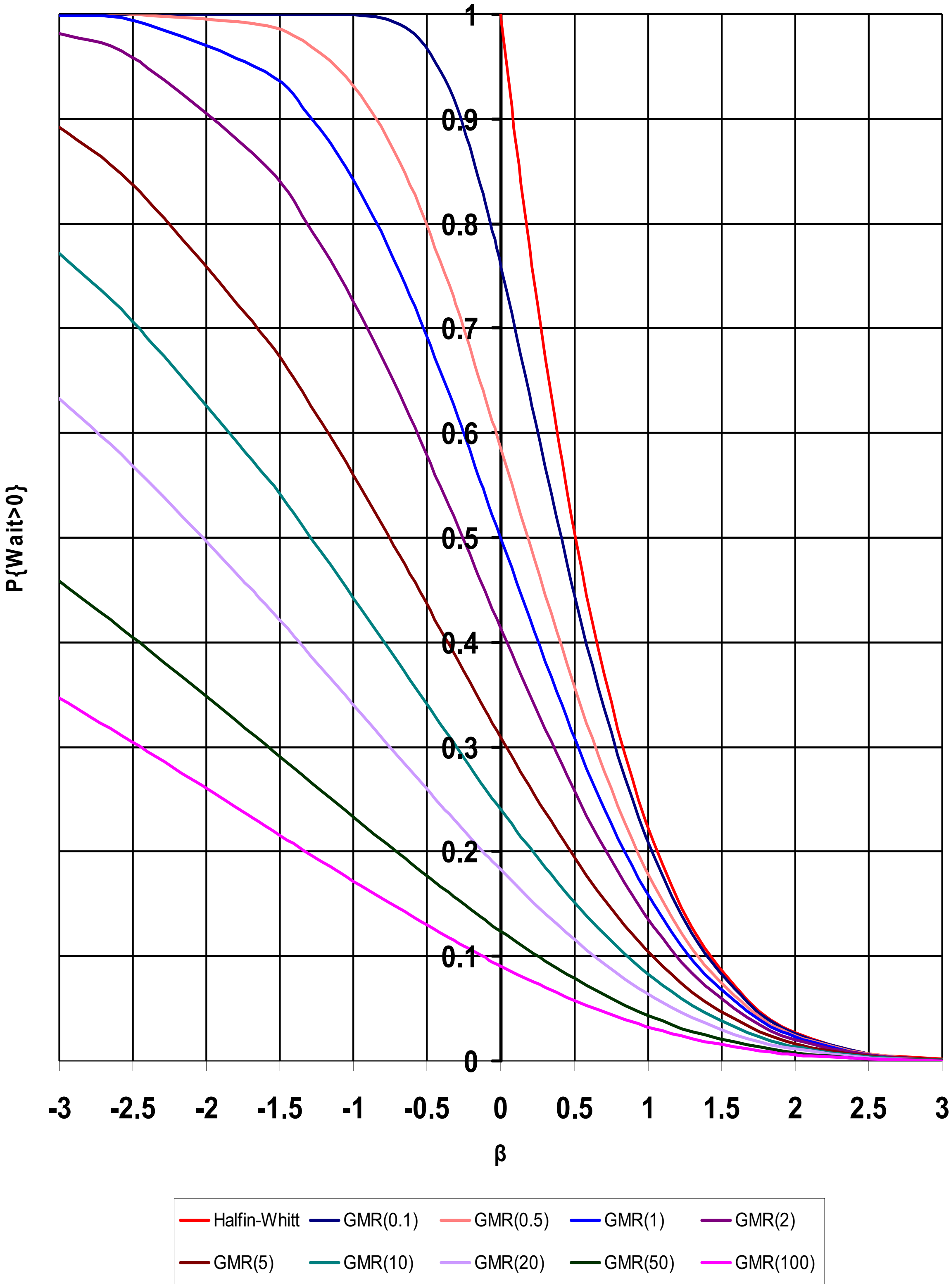


Erlang-A: Theoretical vs. Empirical
P{Wait>0}=α vs. β (N=R+β√R)

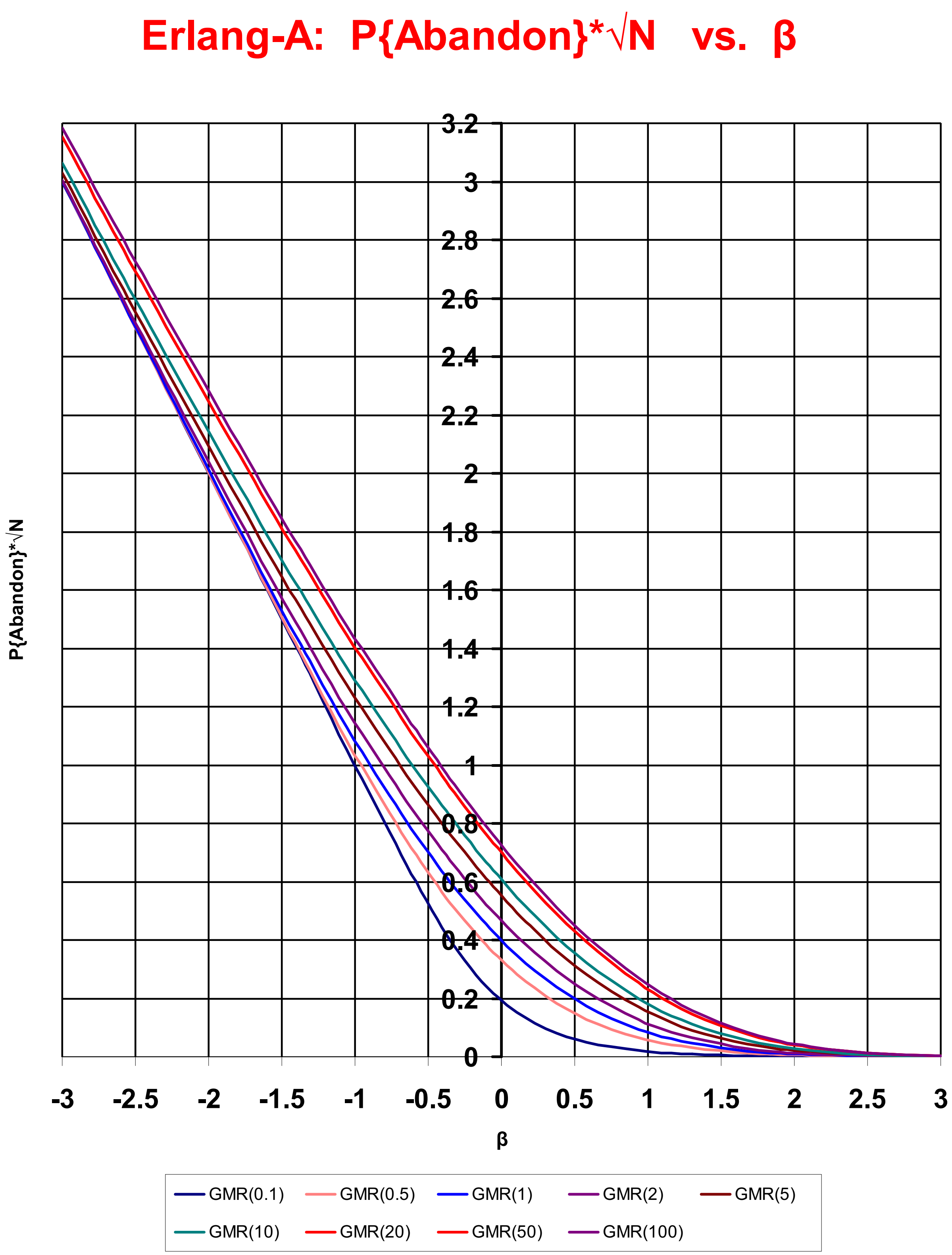
Moderate Patience



Erlang-A: P{Wait>0}=α vs. β (N=R+β√R)



GMR(x) describes the asymptotic probability of delay as a function of β when $\theta / \mu = x$. Here, θ and μ are the abandonment and service rate, respectively.


Erlang-A: P{Abandon}*√N vs. β
P{Abandon}*√N
β
3.2
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
GMR(0.1)
GMR(0.5)
GMR(1)
GMR(2)
GMR(5)
GMR(10)
GMR(20)
GMR(50)
GMR(100)

Iterative Algorithm

Inputs

- ➢ System primitives:

  arrival function, service-time distribution,

  patience distribution (when relevant) ;
- ➢ Target delay probability α ;
- ➢ Time horizon [0,T] .

Outputs

- ✓ Staffing function, aiming at

  a delay probability α is over [0,T] .

Starting point: The *infinite-server heuristics* by

Jennings, M., Massey, Whitt (1996)

Algorithm (cont.)

Notation: $\forall t \in [0,T]$ (*practically* t=0, Δ, $2\cdot\Delta,\ldots$)

$N_i(t)$ – staffing level at time t,

determined in iteration i=1,2,…

$L_i(t)$ – number in the system at t,

under staffing function $s_i(t)$.

Algorithm:

(1) i=0; $N_0(t)\equiv\infty$ (delay probability =0)

(2) Evaluate the distribution of $L_i(t)$, using simulation.

(3) Determine $N_{i+1}(t)$ as follows:

$$N_{i+1}(t) = \arg\min\{c : P\{L_i(t) \geq c\} < \alpha\}, \quad 0 \leq t \leq T.$$

(4) Check stopping condition:

if $\|N_{i+1}(\cdot) - N_i(\cdot)\|_\infty \leq 1$, then $N_{i+1}(\cdot)$ is our staffing level;

else i := i+1, and go back to (2) .

(∞) Last iteration. The algorithm converges to a

Staffing Function $N_\infty(\cdot)$ least for which

$$P\{L_\infty(t) \geq N_\infty(t)\} \leq \alpha, \ \mathbf{0 \leq t \leq T}.$$

BONUS SUPPLEMENT: E-TAILING'S FUTURE

e.b GEN

www.businessweek.com

BusinessWeek

OCTOBER 23, 2000

A PUBLICATION OF THE McGRAW-HILL COMPANIES

Mutual Funds
How to avoid a big tax bill

Wall Street
Will tech's slide keep spreading?

Dot-coms
The search for new business models

Managed Care
Employers seek a new solution

WHY SERVICE STINKS

Companies know just how good a customer you are – and unless you're a high roller, they would rather lose you than fix your problem

PAGE 118

AOL Keyword: BW