CASE STUDY: Ambulance Diversion in Boston, USA.

Root Cause Analysis of Emergency Department Crowding and Ambulance Diversion in Massachusetts

A report submitted by the Boston University
Program for the Management of Variability in Health Care Delivery
under a grant from the
Massachusetts Department of Public Health

October, 2002

Emergency Room Diversion Study: Analysis and Findings

Phase I

Phase I of these investigations involved formulation of a conceptual model that would permit data collection and analysis germane to the problem of ambulance diversion. As preparation for this study, a wide range of relevant medical publications, policy statements and commissioned studies were reviewed. This was followed by personal interviews with representatives in government, hospital administration, public health and the Emergency Medicine community. Information was gathered from throughout Massachusetts and from other key states. Particular attention was given to experience in areas where crowding is particularly severe including metropolitan Boston, San Francisco, Los Angeles and the states of Arizona and Florida. Overall, numerous potential root causes of diversion had been articulated both in the medical literature and lay press, but empirical data to support them were lacking. Available research tended to be descriptive, documenting the extent of crowding without clear delineation of its sources. Various solutions had been proposed and implemented, all without consistent benefit. A partial summary of this analysis has been previously released by the Massachusetts Health Policy Forum of Brandeis University.

An operations management perspective suggested straightforward input-throughput-output analysis. Hospital utilization data provided by the Division of Health Care Finance and Policy was therefore reviewed alongside diversion data provided by regional EMS providers. Analysis of this information revealed the likely operation of mechanisms both internal and external to emergency departments. In addition to simple supply/demand imbalances for emergency care, diversion and utilization patterns suggested bottlenecks and backlogs related to the competition of emergency and non-emergency patients for similar resources. The interrelationships of hospital services then became the focus of attention and patient care pathways were explored with administrators from the two study hospitals.

Two paradigms for the quantitative study of interrelationships among hospital departments were considered. The first involved an analytical approach wherein each relationship was identified, its stochastic character estimated, and appropriate mathematical models applied. The second involved a simulation approach, wherein stochastic relationships were embedded into computer software that translated real patient flow inputs into utilization and capacity information. Computer simulation was ultimately selected as the route of choice because of its scalability and adaptability.

Phase II

Data Collection/Analysis Effort:

The study was performed at two hospitals in Massachusetts: Hospital A, a large tertiary academic hospital, and Hospital B, a medium-sized acute care community hospital. The following data were collected:

- 42 days of information covering:
- 6000+ admissions
- 8000+ ED visits
- 2000+ staffing/capacity data points
- 300,000+ patient movement/status data points

In order to analyze the relationship between diversion status and other factors within the hospital environment all measures were split into observations at one hour increments. The study period of 42 days, with 24 hours per day, yielded a total of 1008 full sets of observations. The analysis required collection of patient flow data well beyond the usual capabilities of contemporary hospital information systems.

Point-biserial coefficients of correlation, with diversion status as the binary variable, were examined against a variety of factors. Comparisons when using full hours of diversion versus partial hours as the “true” condition did not reveal significant differences, so partial diversion hours were evaluated as the “true” binary throughout the analysis for the sake of consistency.

It is important to note that in the real world the decisions to commence or cease diversion status are, but their nature, highly subjective. Because the purpose of the study was to examine the root causes of diversion, we did not approach the task from the standpoint of critiquing or attempting to influence this inherent operational subjectivity. As a result, any such analysis is itself subjective to a certain degree.

Because both hospitals straddled EMS regional borders and diversion rules vary by region, each hospital’s data was used for the sake of determining diversion status rather than using centralized EMS data. Also, all diversions were considered equally rather than separately analyzing the factors related to each individual diversion type.

Patterns of diversion were also examined as averages across the hours of the day and the days of the week in order to ascertain relevant hour of the day and day of the week patterns. This data analysis was performed separately for each of the hospitals.

Hospital A:

Diversion Pattern "Hospital A - Diversion Minutes by Hour"

- There were a total of 22 episodes of diversion which started and ended within the study, with an average length of 814 minutes. There was one episode that began prior to the study and ended after the study began and so is not included in this calculation, nor in any calculations which involve the beginning of diversion episodes.
- The hourly diversion pattern shows diversion is highest in the evening hours, settles back down during the early morning hours, and then stays steady until the mid to late afternoon (see Fig. 1).
- The goal of the project was to determine the drivers which create this pattern.

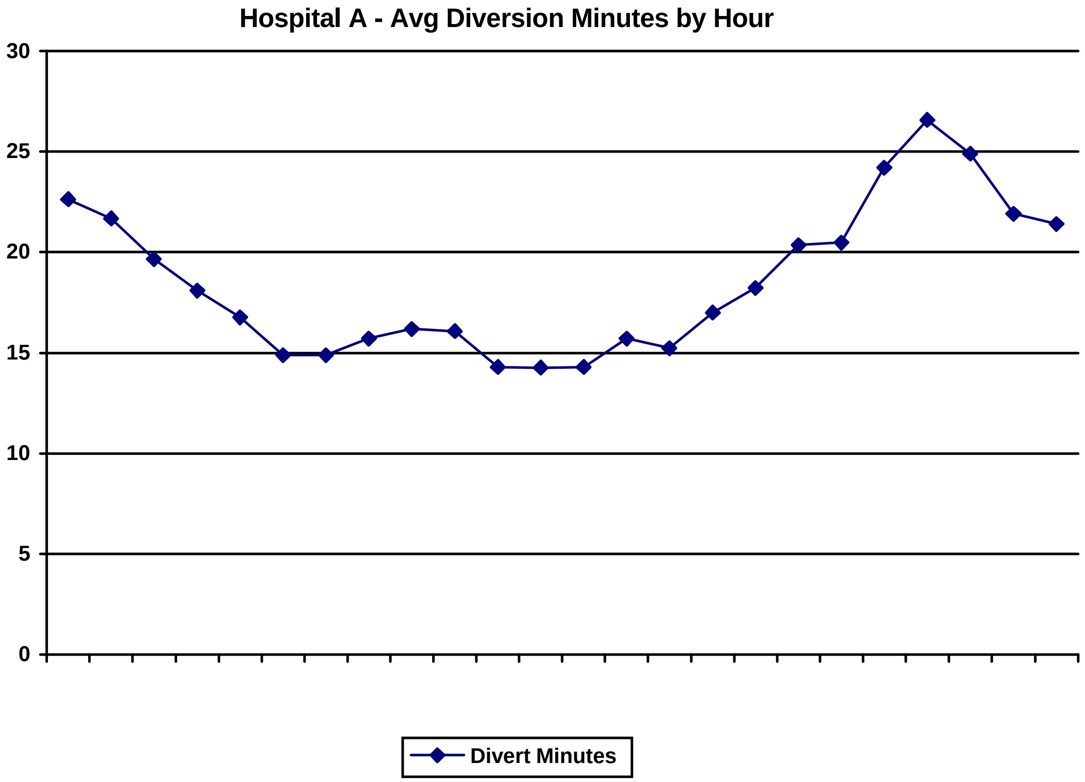


Fig. 1

The following 3 hypotheses were tested to determine the drivers of diversions:

1. ED arrival rate is too high, leading to diversion when the ED becomes full.
2. ED processing of patients is too slow, causing backups that lead to diversion
3. ED arrival and processing rates are fine, but there are not enough beds in the hospital to accommodate the admissions.

There are seven sets of data (see Fig. 2), each representing a different view of arrivals into the ED. The "Arrivals_0" category only includes new arrivals from the hour in question. Each subsequent category, from "Arrivals_1" to "Arrivals_6" includes one more hour's worth added to the total. In other words, "Arrivals_1" includes arrivals from the current hour added to the arrivals from the previous hour, "Arrivals_2" includes all of "Arrivals_1" plus the arrivals from two hours ago, and so on. This is what accounts for the stacked shape as each additional hour is layered on top. Because average length of stay was 340 minutes, 6 hours is used as the maximum lag. Correlation coefficients from each of these cumulatives to Avg Diversion Minutes by hour are as follows:

Arrivals_0 = -0.073
Arrivals_1 = 0.001
Arrivals_2 = 0.078
Arrivals_3 = 0.165
Arrivals_4 = 0.259
Arrivals_5 = 0.359
Arrivals_6 = 0.460

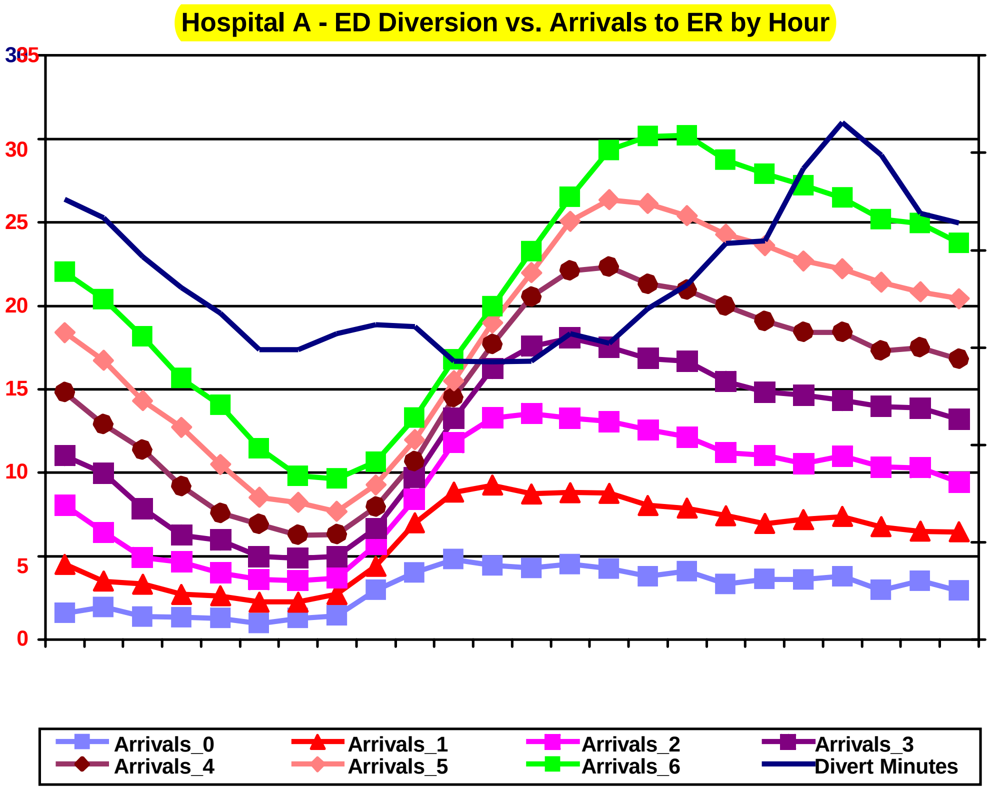


Fig. 2

There is also a possible corollary to hypothesis #1, that overall ED census is a driver of diversion. When counting the non-boarding census and comparing it to diversion status, however, the resulting point-biserial coefficient (r = -0.051) makes clear that this potential explanation should be rejected as well.

again points towards examining hospital capacity as the primary determinate of diversion.

Census/Admissions/Discharges: Hospital B

The overall relationship between inpatient census and ED boarders in Hospital B was similar to that of Hospital A. However, detailed analysis of admission sources in Hospital B is not presented because scheduled demand played a far smaller role than that observed in Hospital A.

During the study period, there were 1,158 weekday unscheduled admissions (average: 38.6/day) and 208 weekday scheduled admissions (average: 6.9/day). This suggests very little operational flexibility in controlling the variability or timing of scheduled arrivals. This likely reflects a fundamental difference between most community hospitals and larger academic centers.

Hospital B Conclusions:

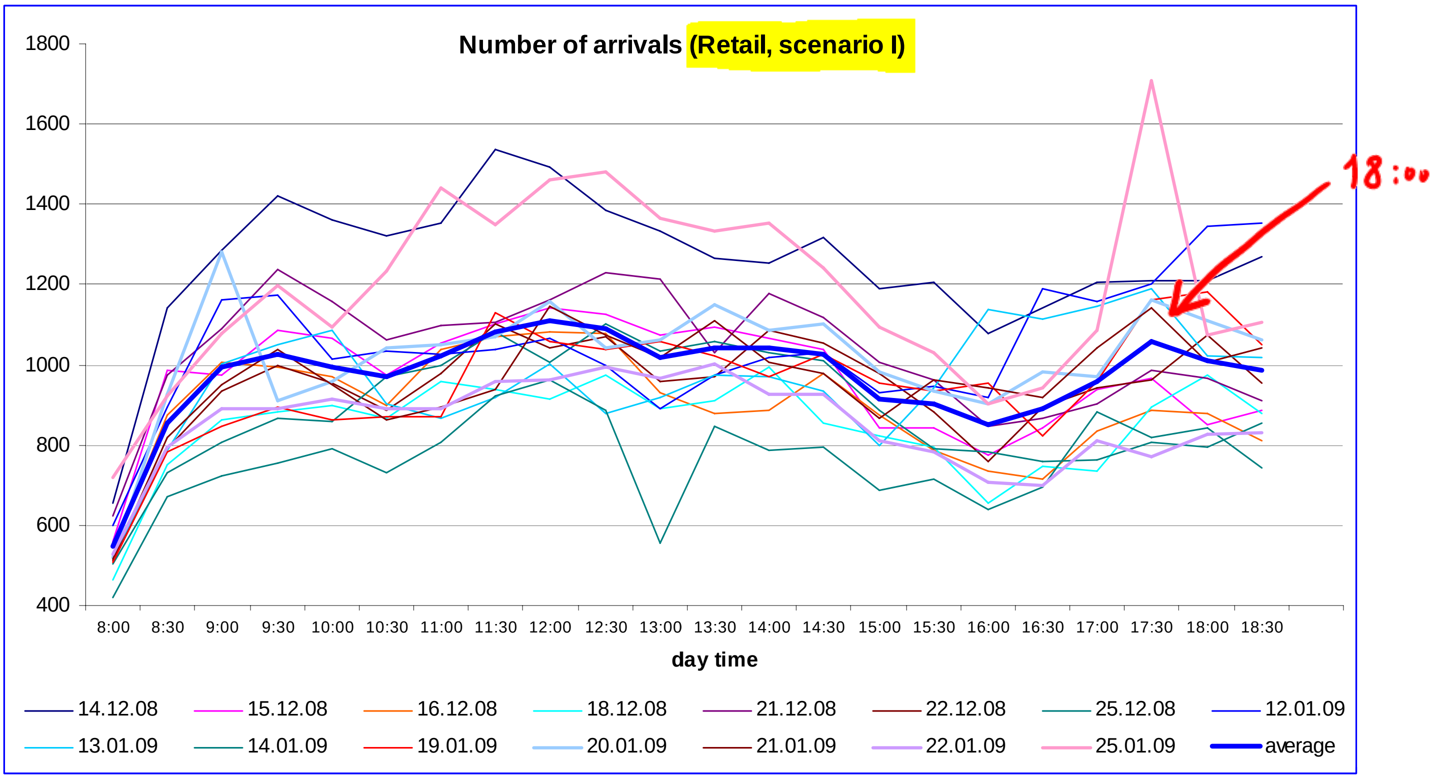
The findings at Hospital B are consistent with and reinforce those at Hospital A. Specifically, there was no evidence that ED process times were temporally or mechanistically related to ED diversion while the relationship between ED arrival rate and diversion was weak. Instead, the data suggest that factors outside of the ED that combine to increase boarders and limit ED capacity are more important.

Phase II Summary:

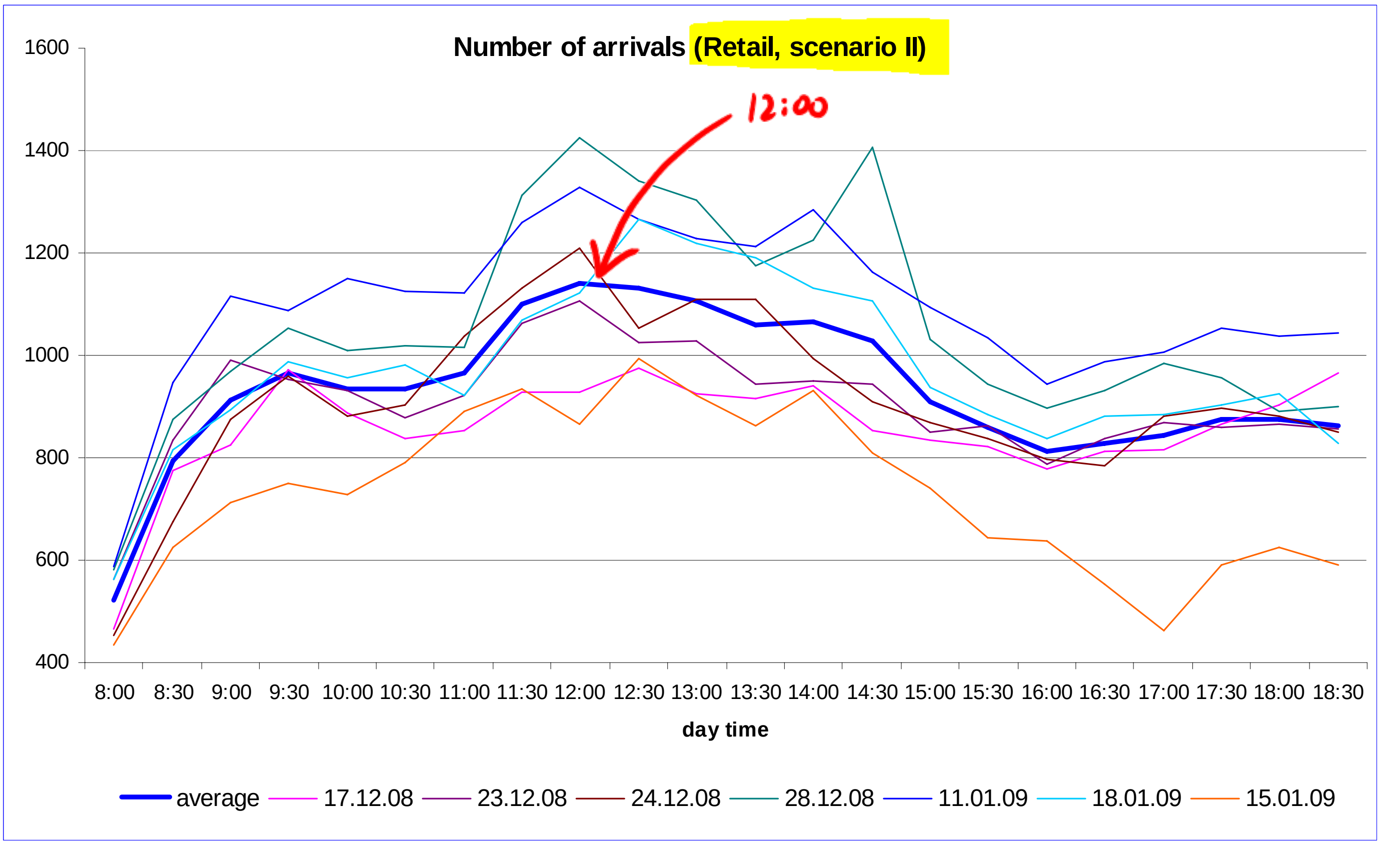
Detailed flow analysis in two very different types of hospitals yielded similar findings with respect to the root cause of emergency department crowding and ambulance diversion. Neither increased patient inflow nor increased process time could be strongly related to diversion status. Instead, diversion was seen as an outflow problem, with busy emergency departments crowding as patients await transfer to crowded inpatient services. This problem is exacerbated in hospitals with large volumes of scheduled admissions, since these necessarily compete for the same resources. The “collision” of scheduled and unscheduled patient flows results in diversion patterns that are specific and reproducible. Because scheduled patient flows are theoretically controllable, better understanding of this phenomenon may suggest means of decreasing diversion. If the experience here may be generalized, we conclude that institutions with small (or uncontrollable) scheduled patient flows will require addition of resources *on the inpatient side* if diversion is to be substantially reduced.

CASE STUDY : Staffing (must be driven by the Offered-Load)

Calls arrive by different scenarios: sometimes arrivals during a day have a bell-form with peak around 12:00 (scenario II) and in some days we ~~can~~ see peaks in evenings around 18:00 (scenario I).

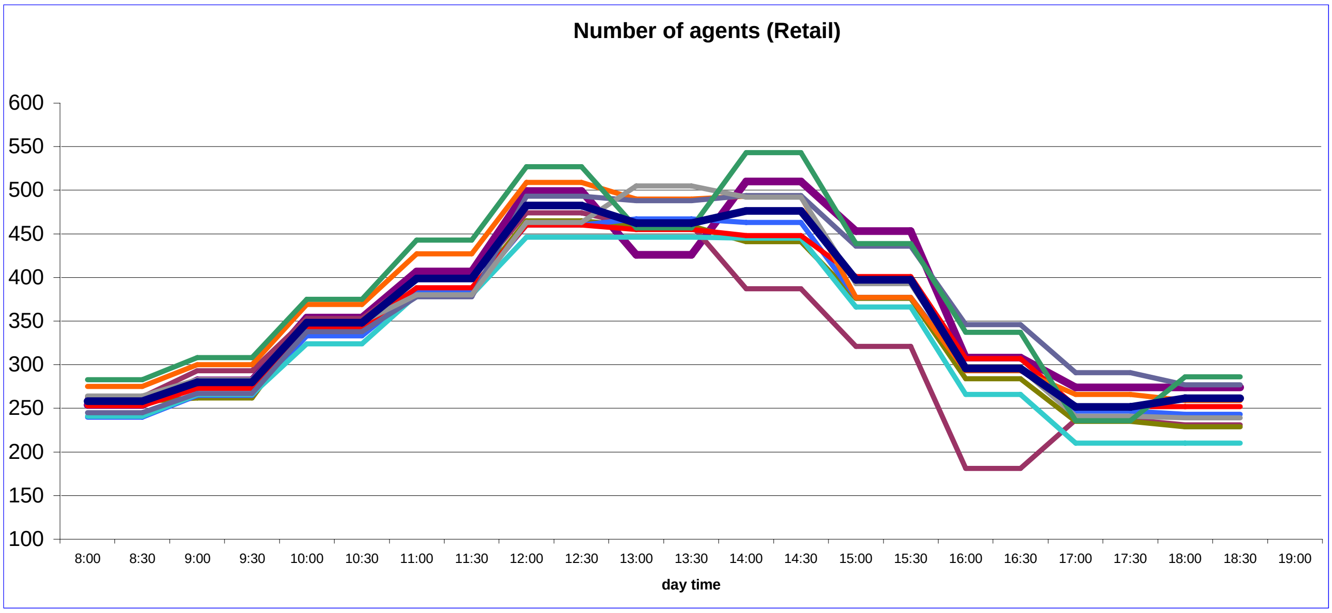


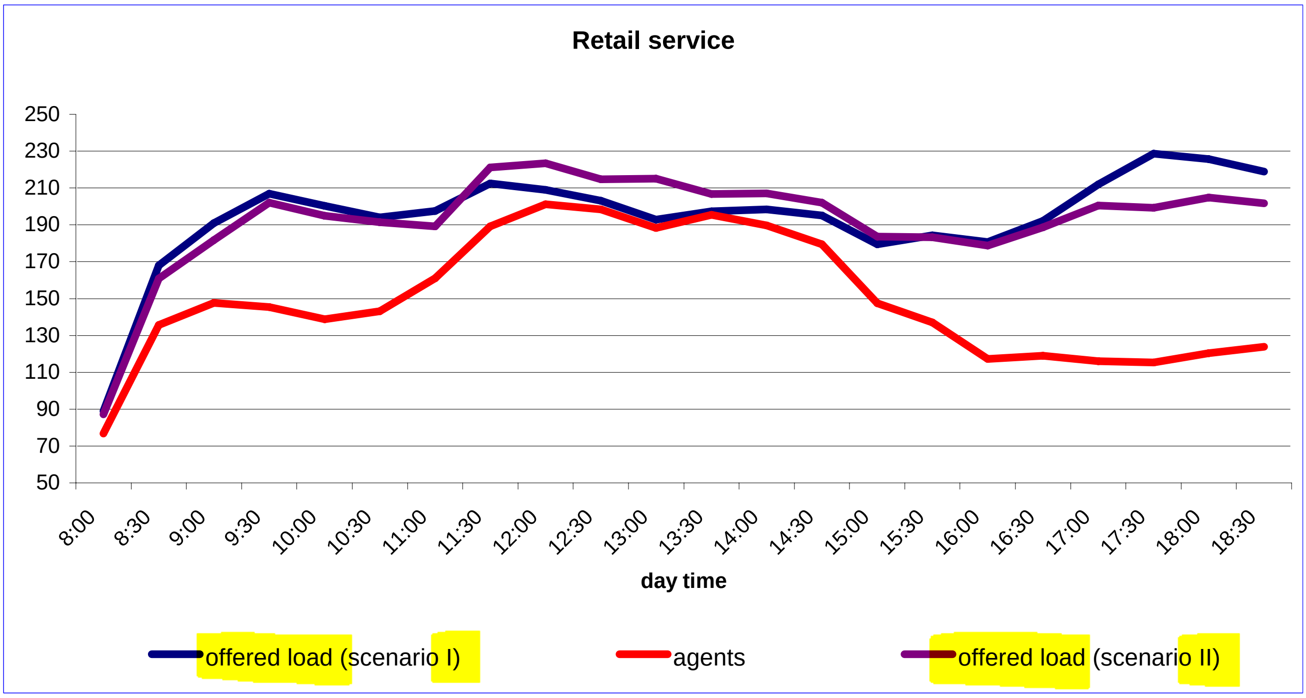
18:00



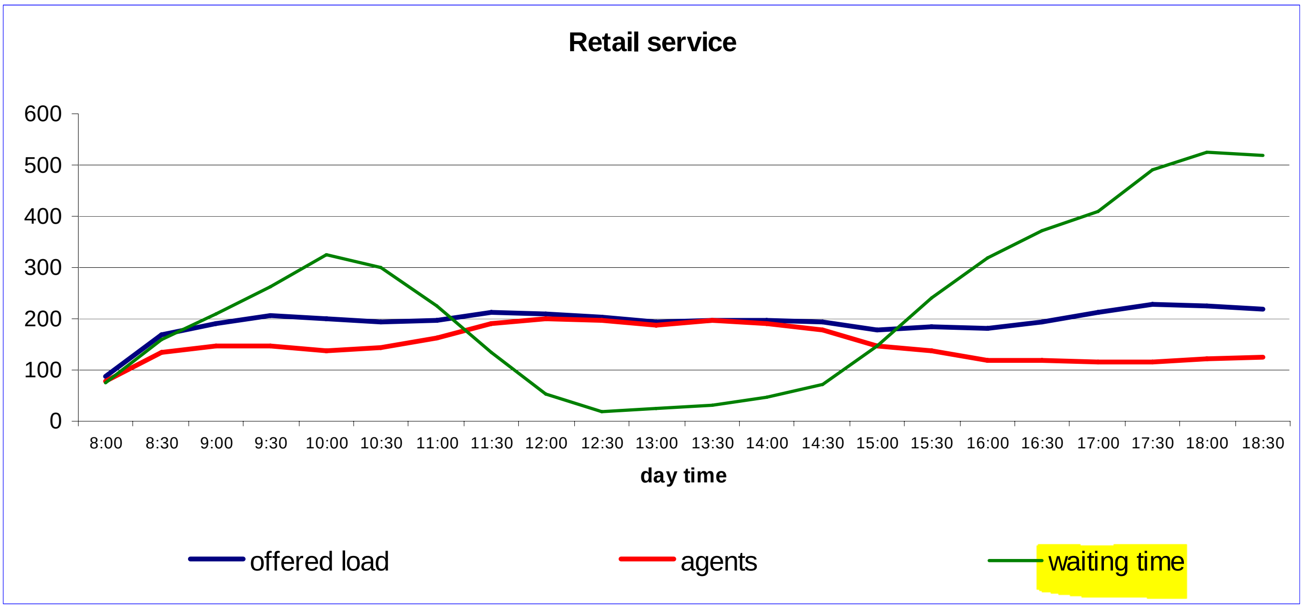
12:00

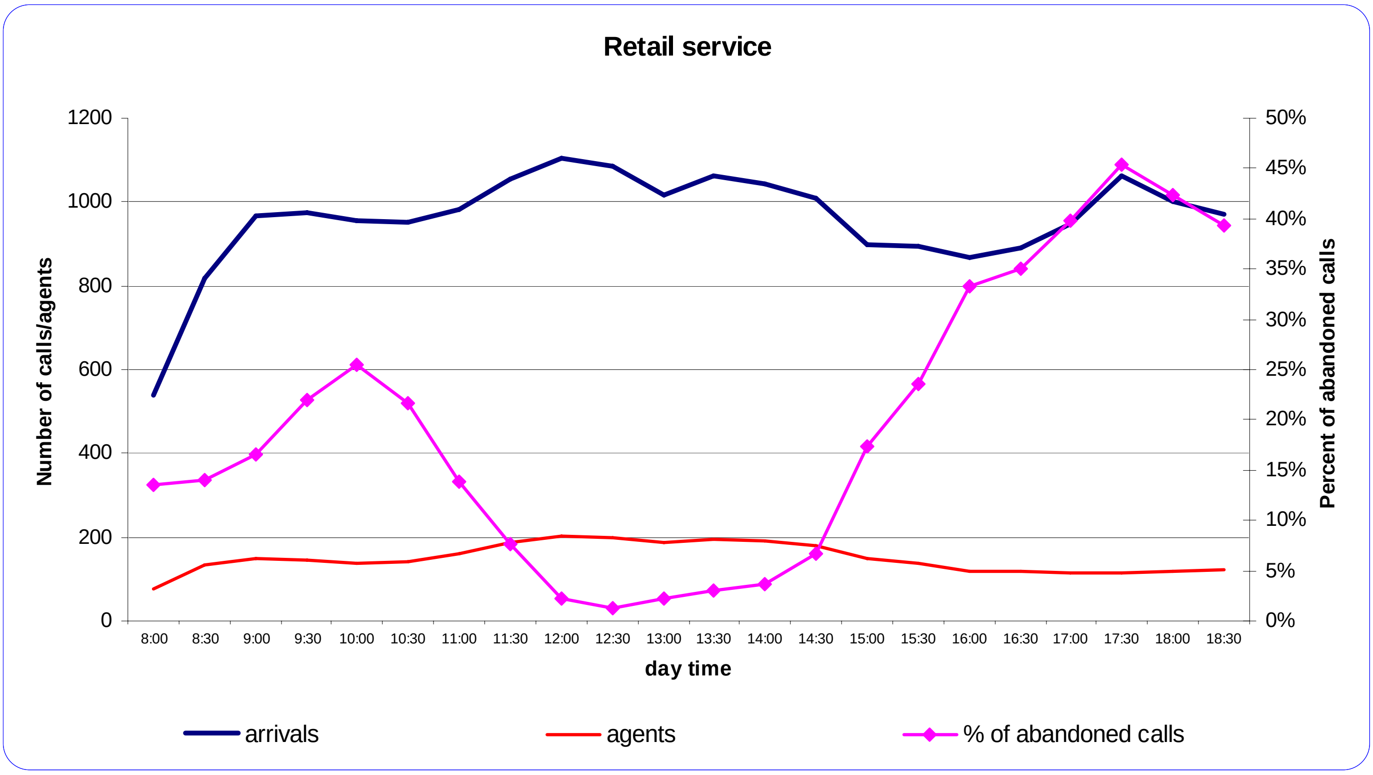
At the same time, the form of the agents' staffing does not changed for the days with different scenario for arrivals.



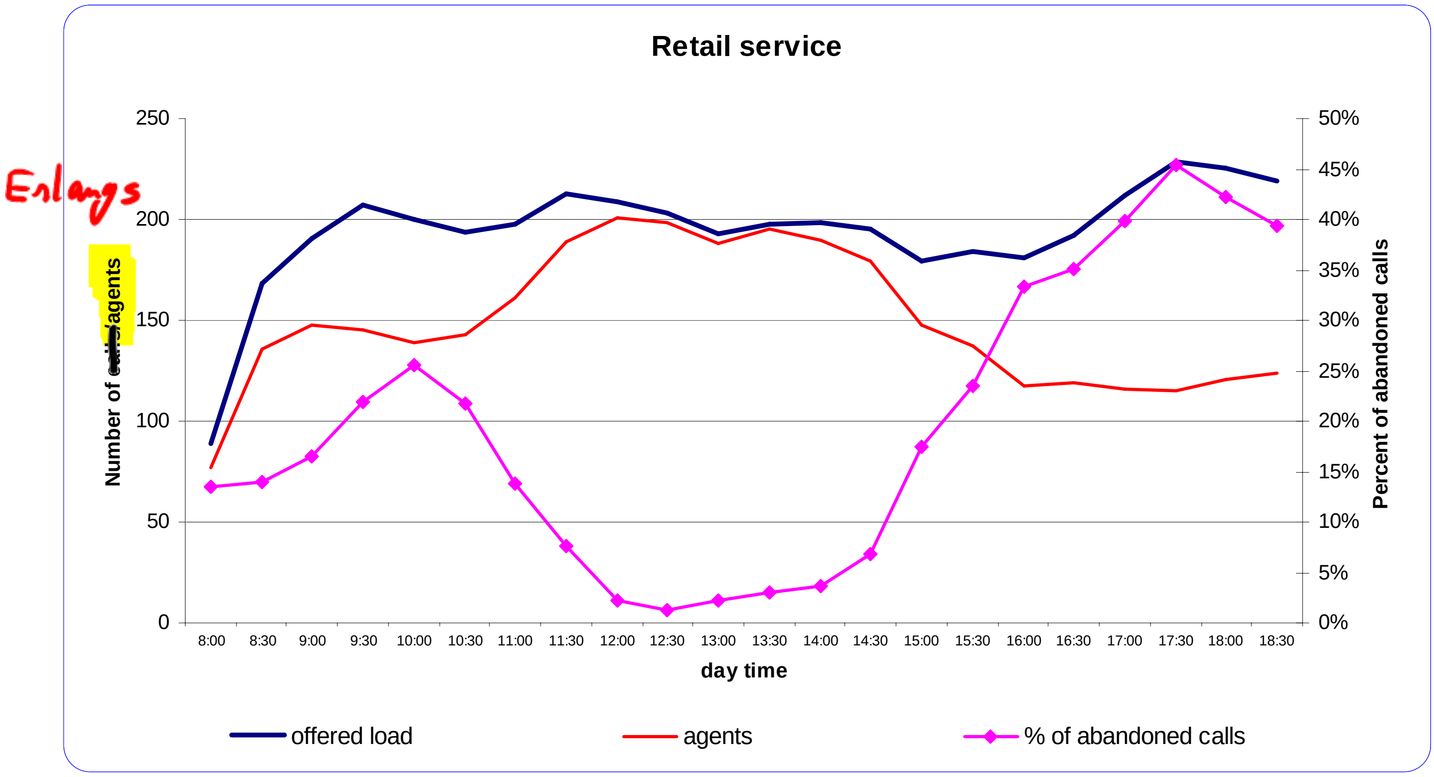


The figure above shows inconsistency in forms of the offered load and agents' staffing. All values presented in this figure are averages, of the offered load in two different scenarios and the number of agents, calculated in each 30 minutes interval.





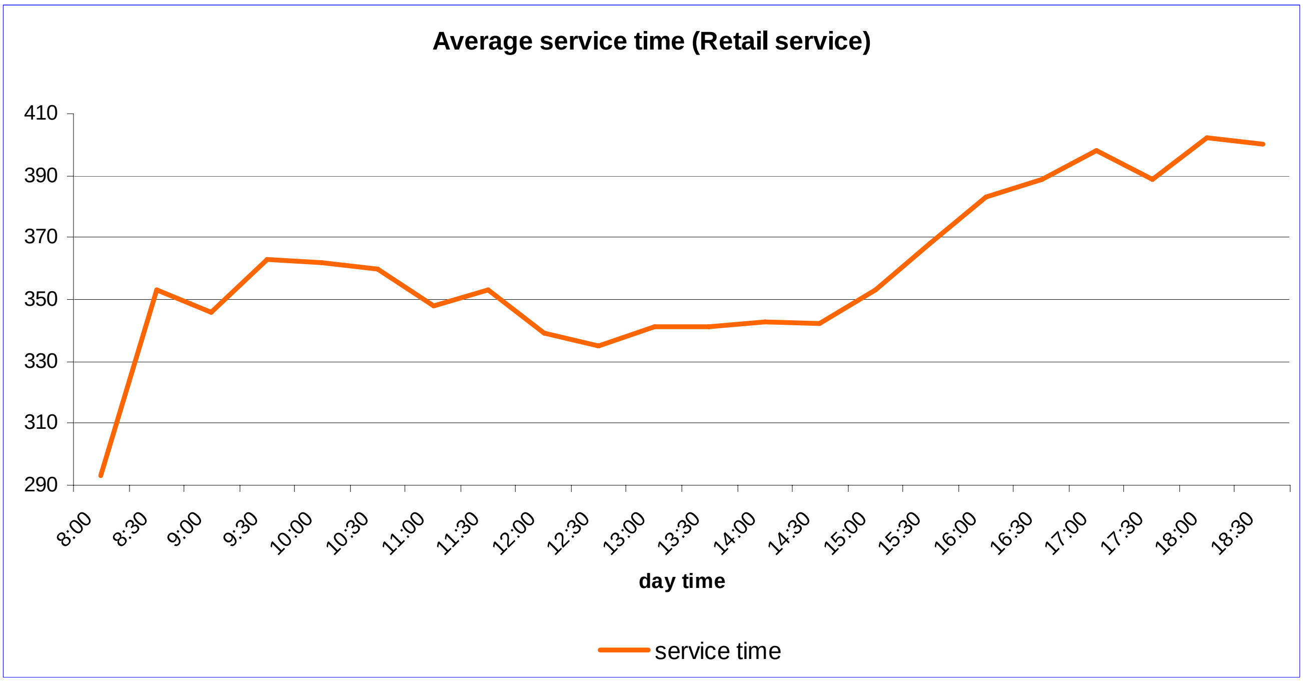
Hard to deduce from the above graph
(which is the one most frequently used)
the source of the problem?



<u>An example of the best case scenario</u>

We choose Sunday 18.01.09 as a day with the best performance characteristics between weekdays 18-15.01.09. Even in this day we see problems in the service in morning and evening hours.

Average service time changes dramatically during the day. We ~~can~~ see that in the evening average service time is much bigger than in the midday. This means that the offered load in the evening is bigger than in the midday.



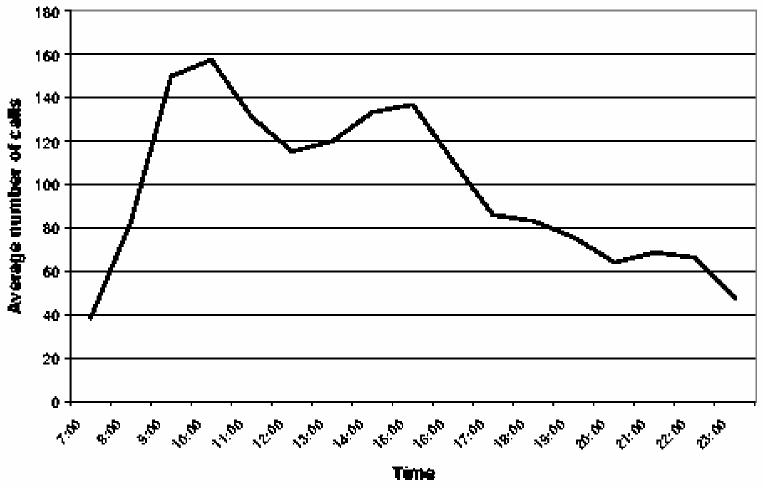
Nightmare: longest services during most congested times!

Why? Preventable?

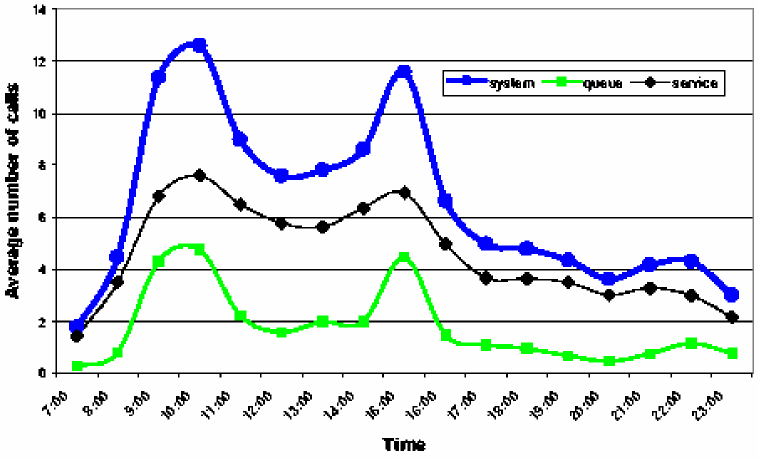
Predictable Variability

Goal: Predictably Stable Performance, but HOW?

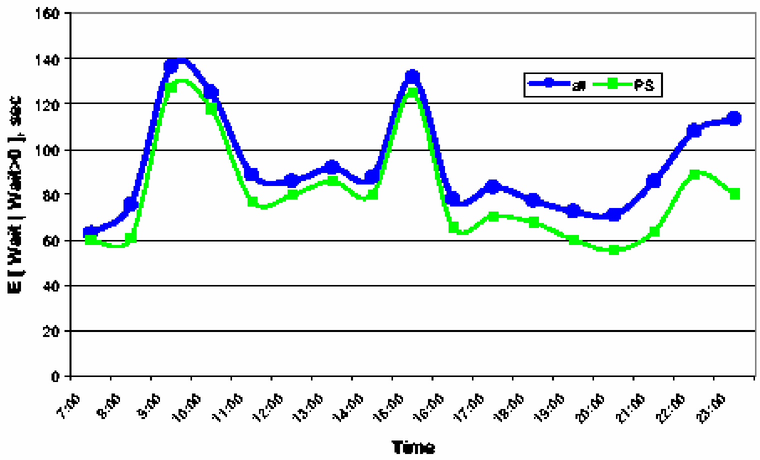
Arrivals



Queues



Waiting



Staffing Time-Varying Queues:

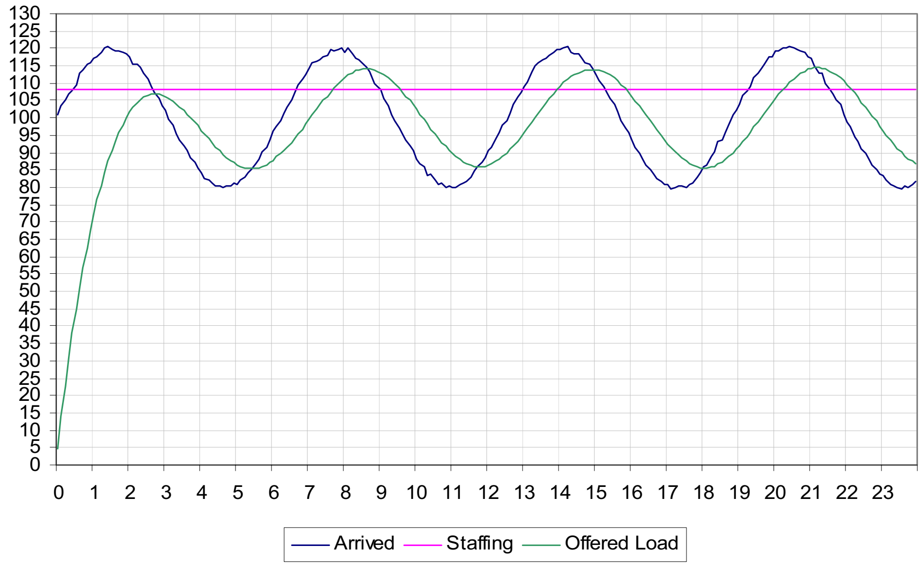
Two Common Approaches:

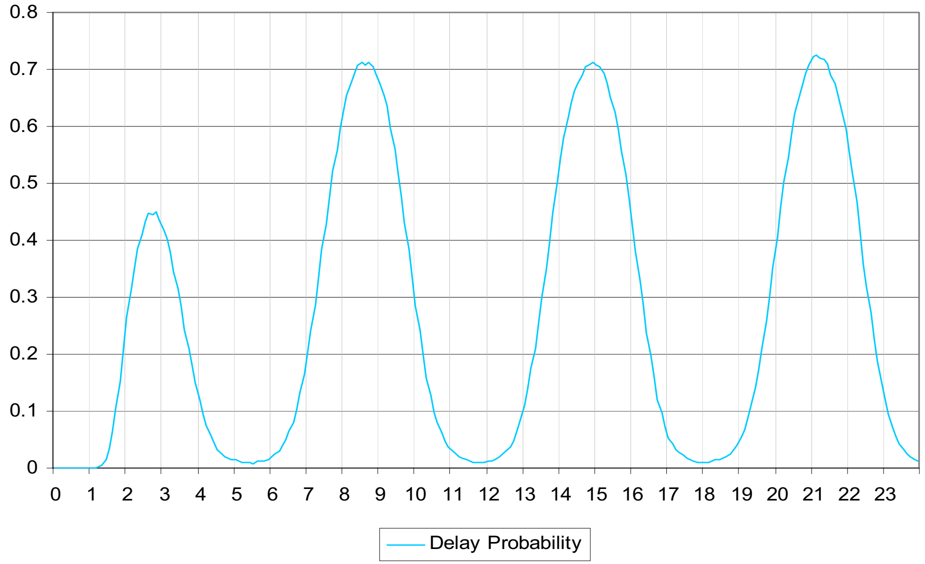
SSA – Simple Stationary Approximation.
Constant staffing levels, based on steady-state M/M/N,
with λ=long-run average number of arrivals.

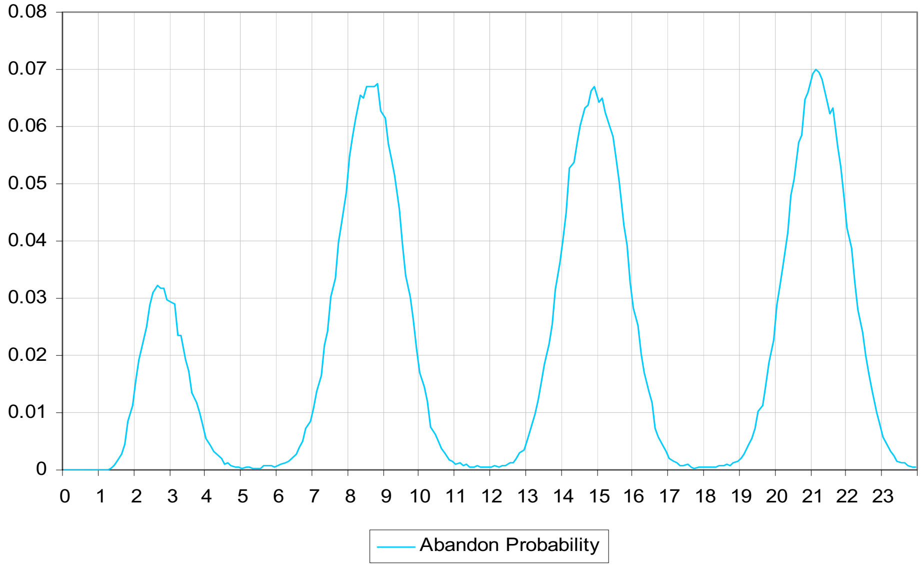
PSA – Point-wise Stationary Approximation.
Time-varying staffing levels, based on steady-state
M/M/N, with λ= λ(t) at each time t.

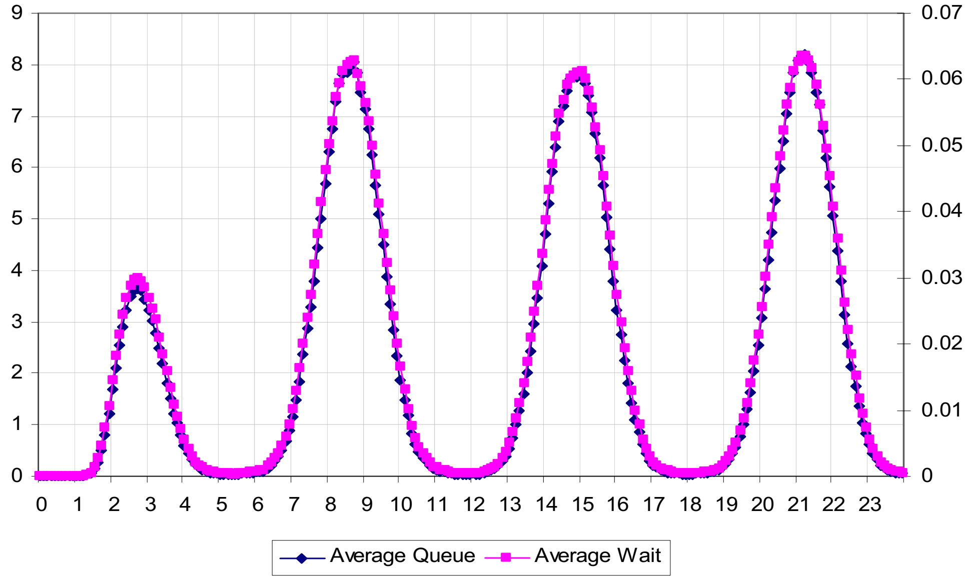
Could result in time-varying (highly oscillating) performance (utilization, service), which is undesirable.

Simple Stationary Approximation (SSA, α=0.2)

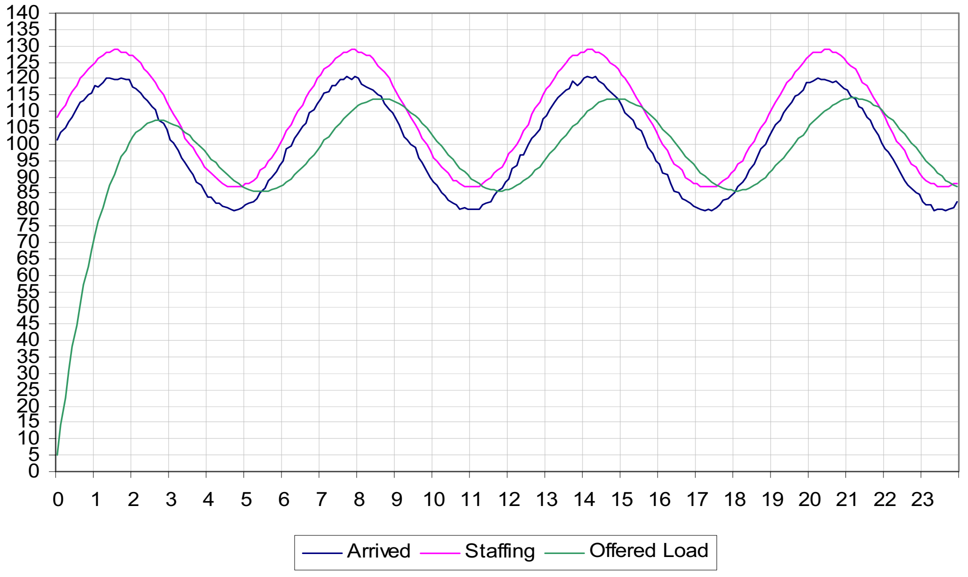


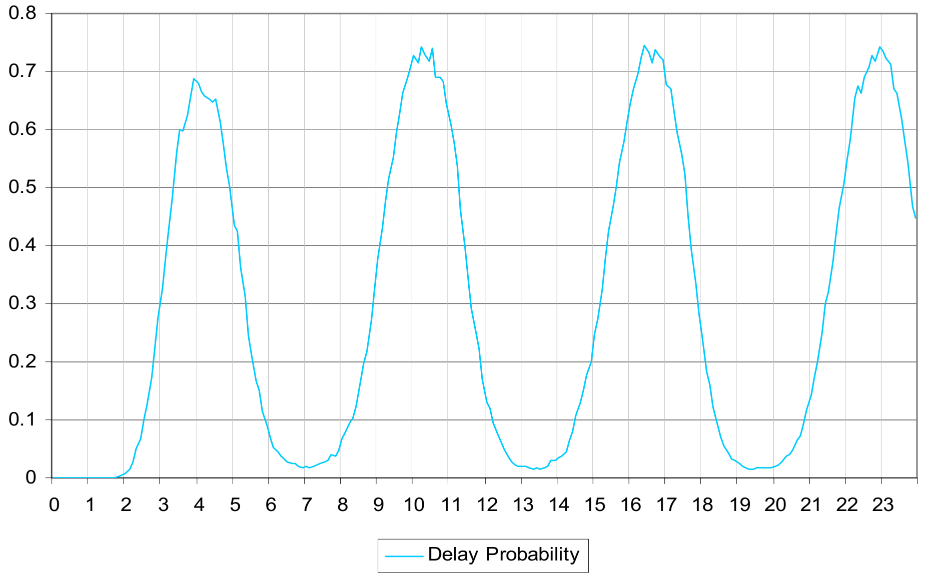


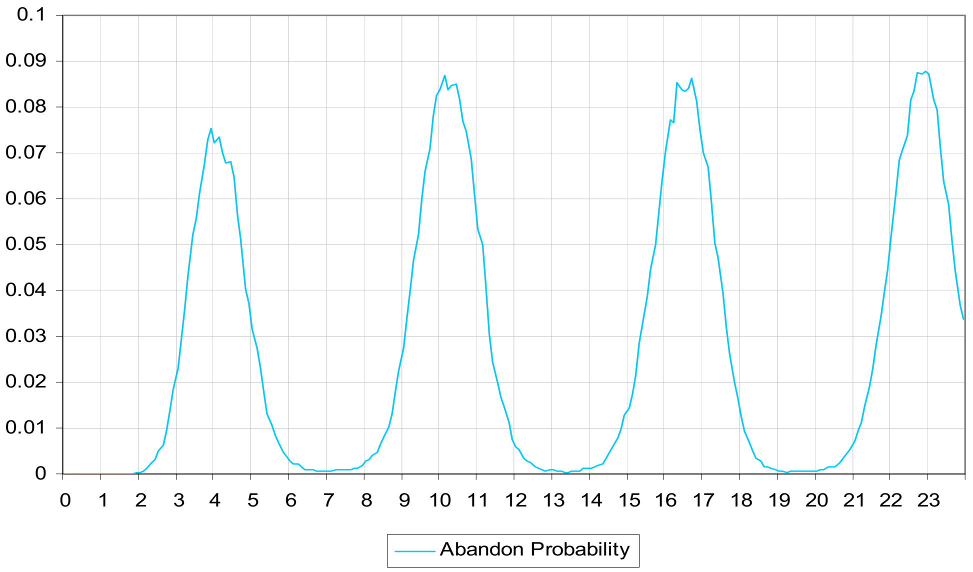


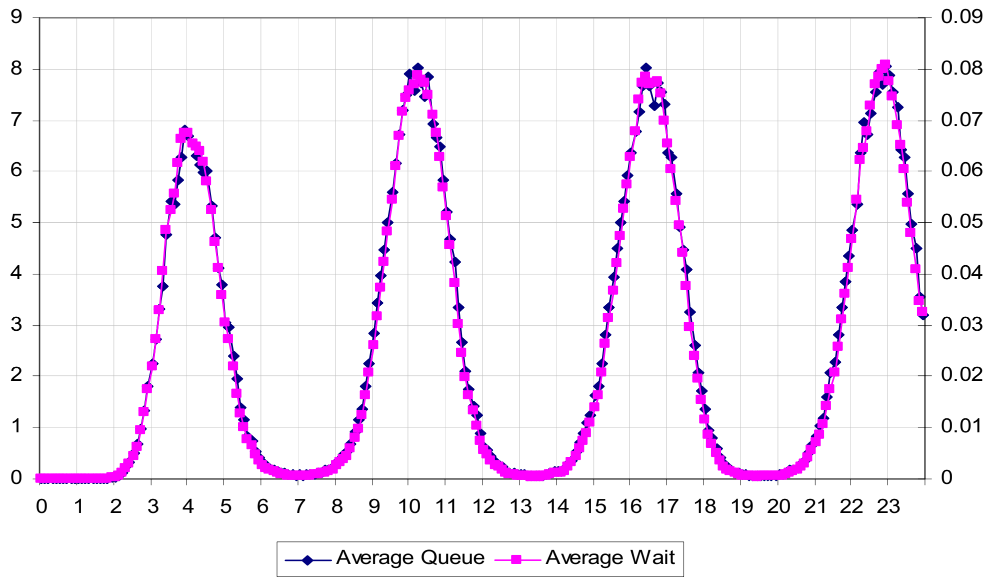


Point-wise **S**tationary **A**pproximation (PSA, **α=0.2**)





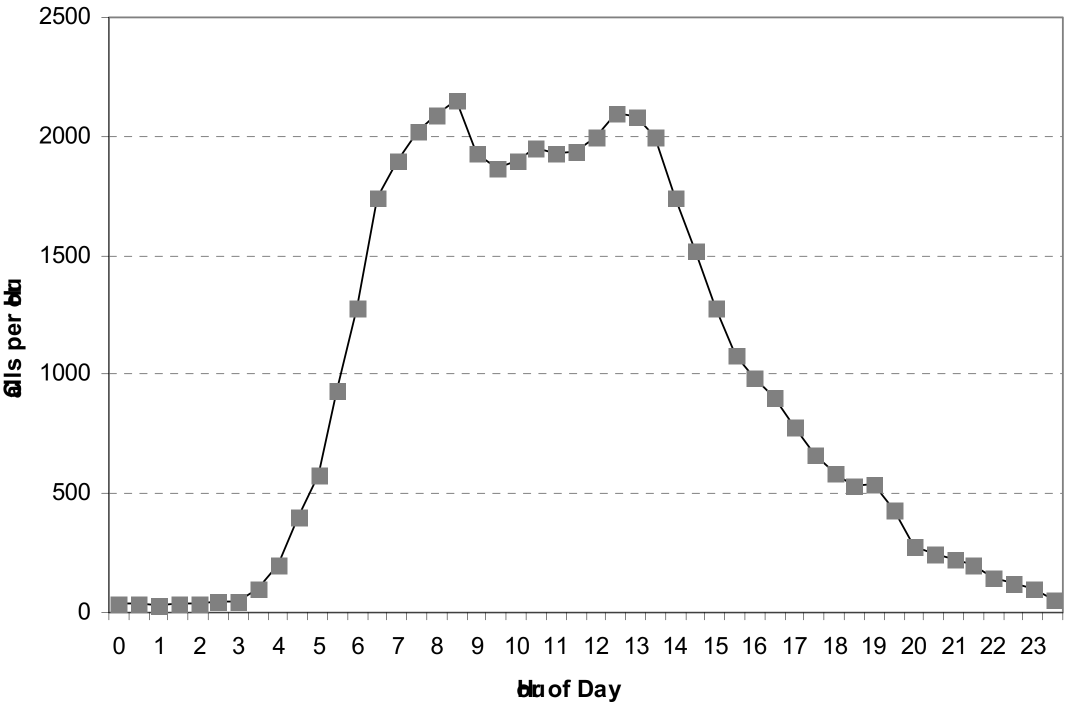




Example: "Real" Call Center

Two-hump arrival functions are common

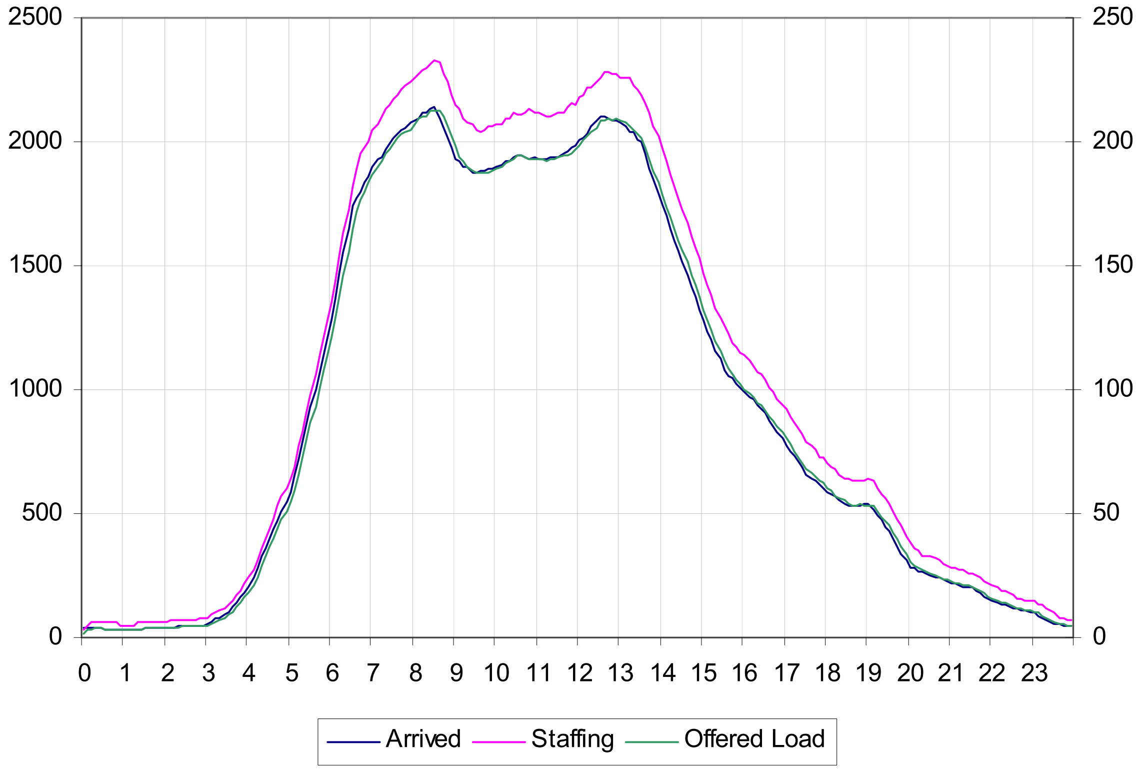
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)



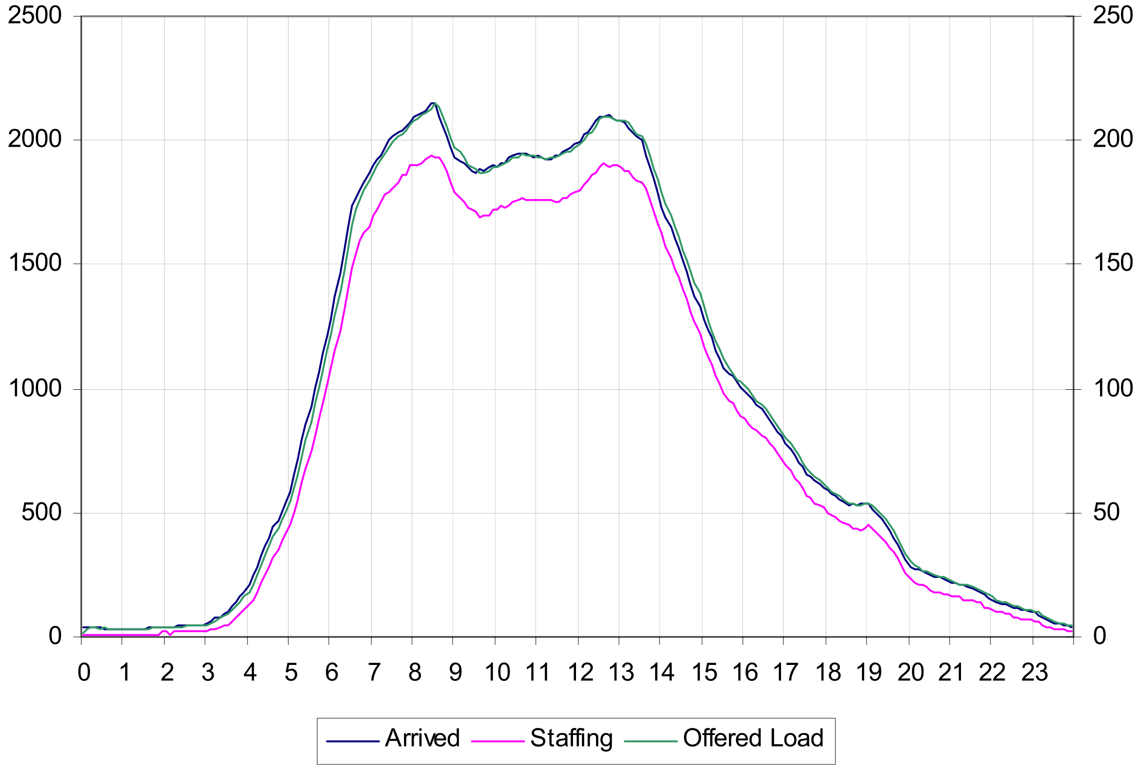
Assume: Service and abandonment rates are both

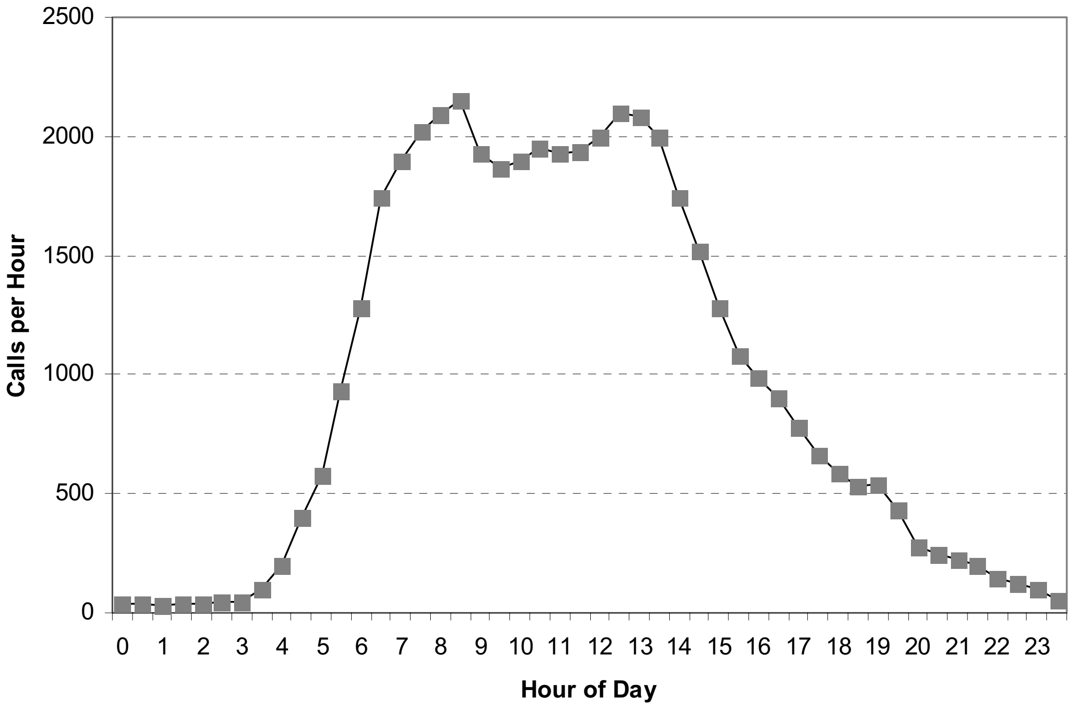
exponential having mean 0.1 (6 min.)

QD Staffing (α=0.1)

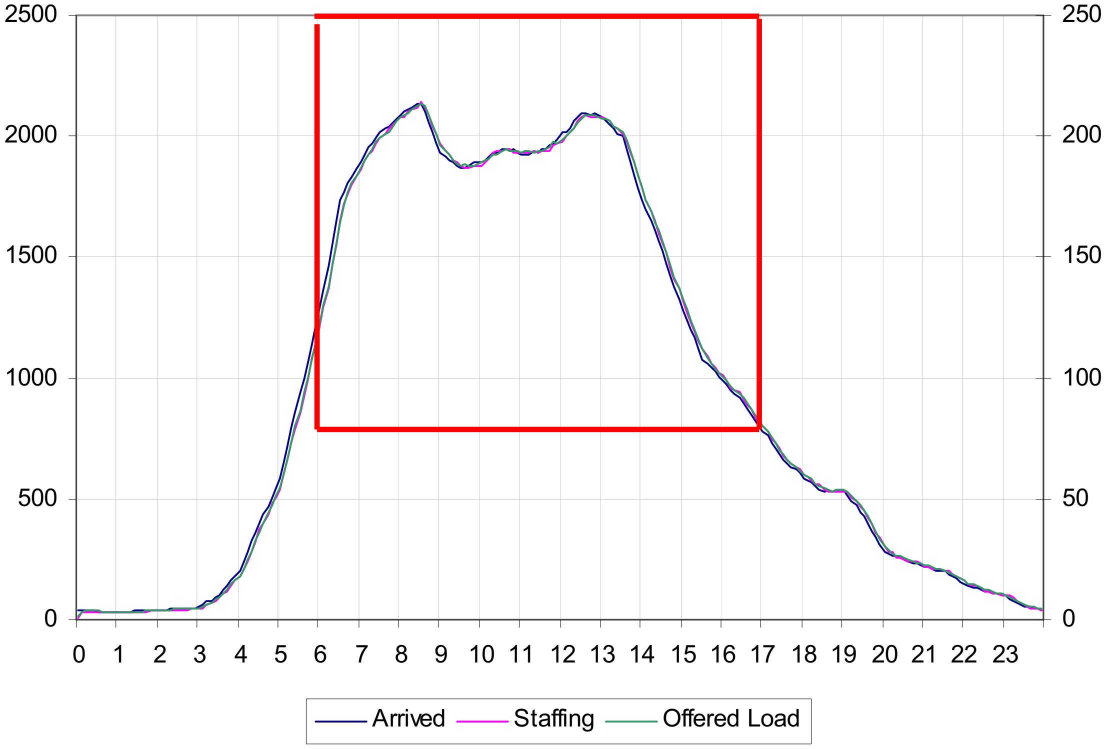


ED Staffing (α=0.9)

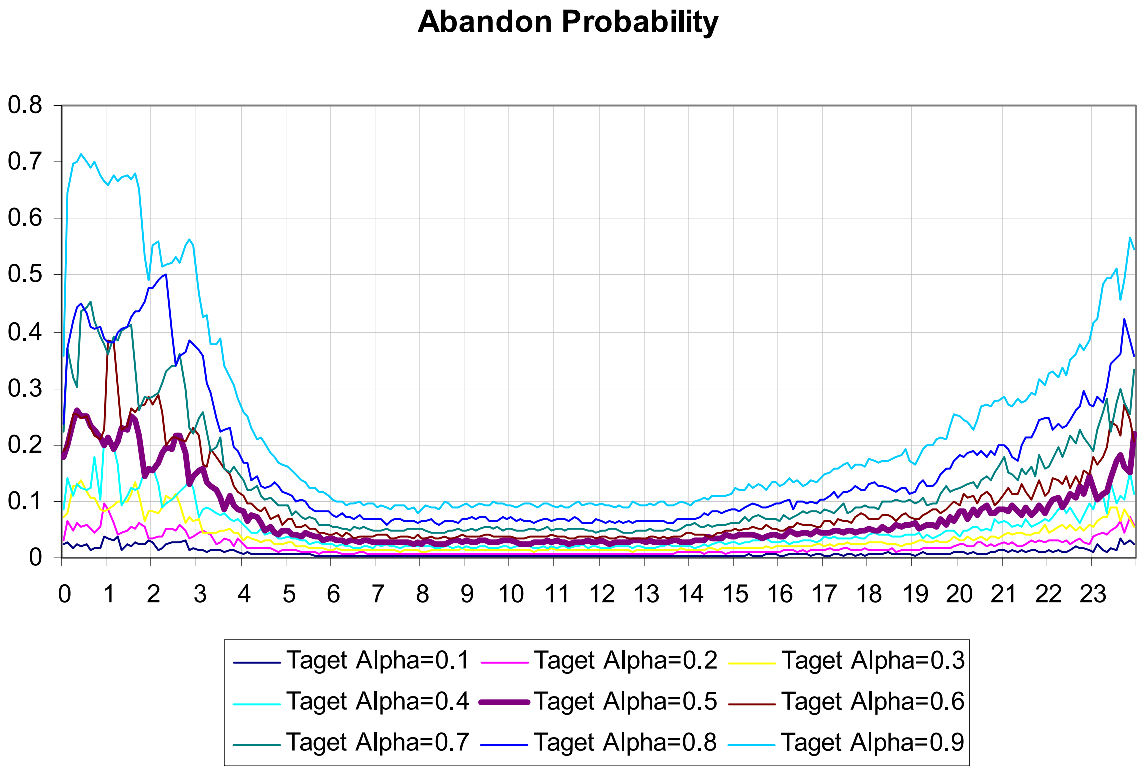


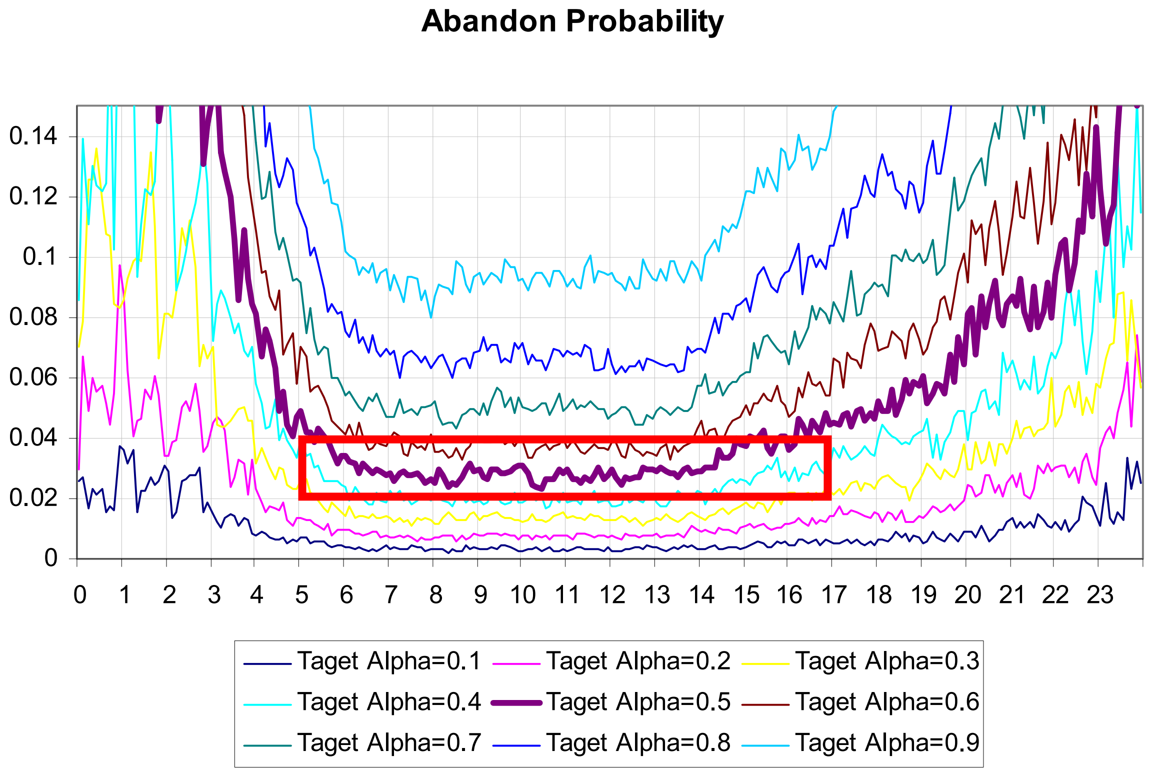


QED Staffing (α=0.5)

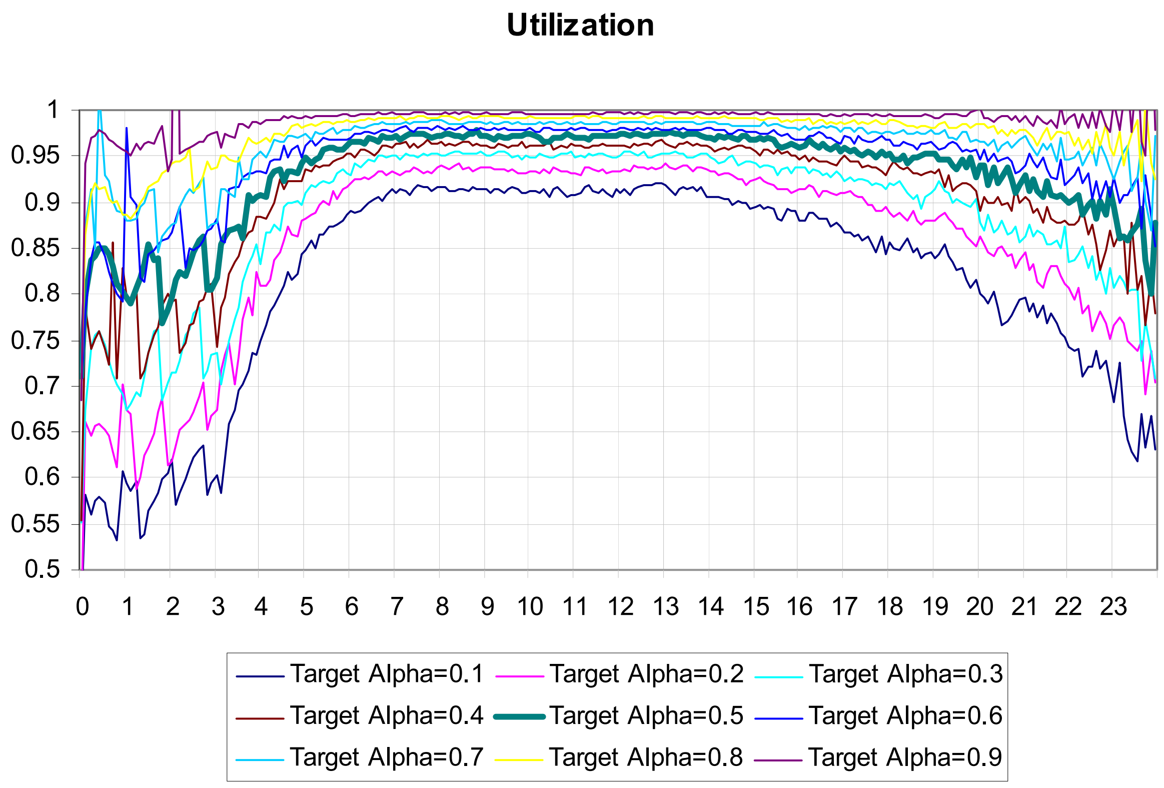


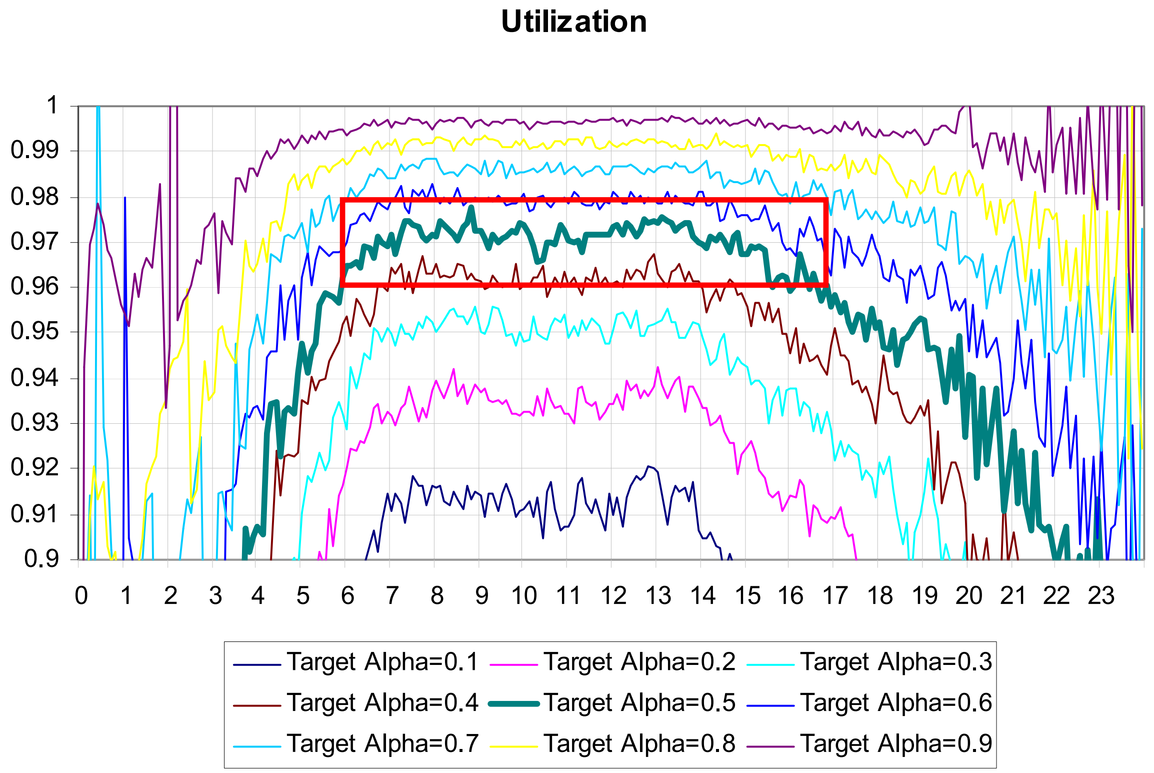
Abandon Probability





Utilization

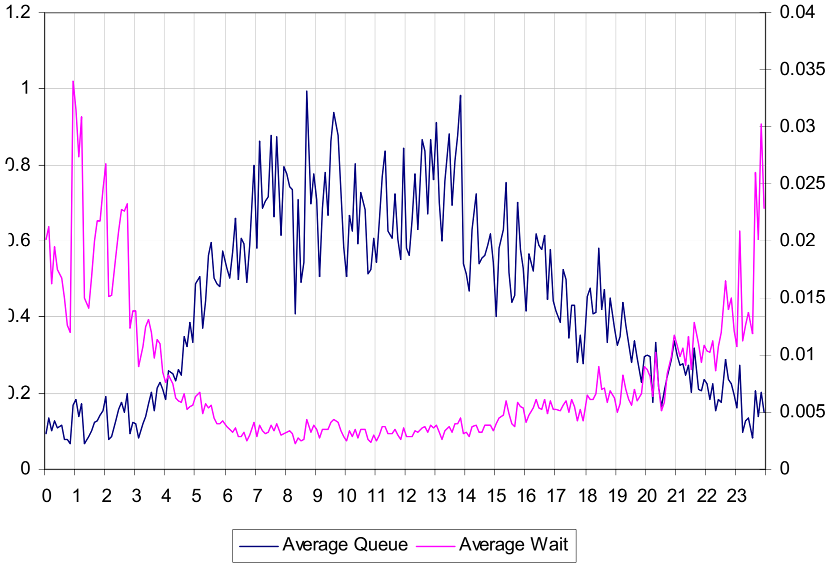




Congestion (Queue, Wait)

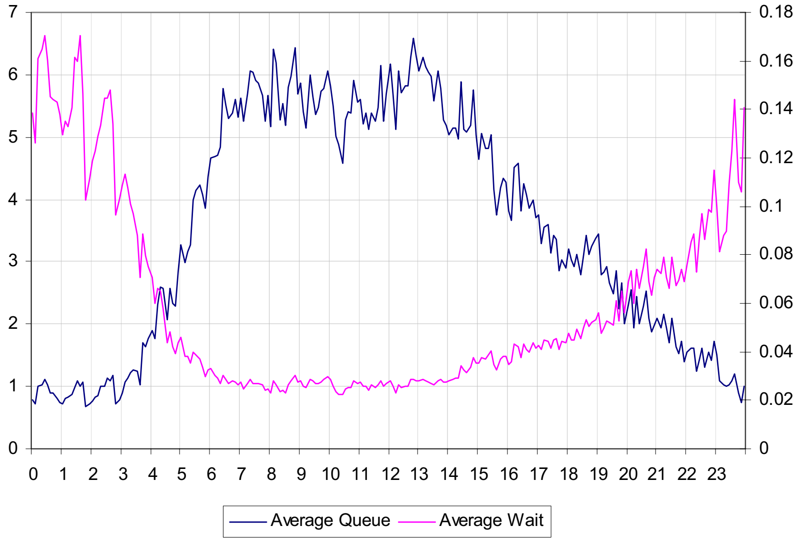
QD
α=0.1

Negligible



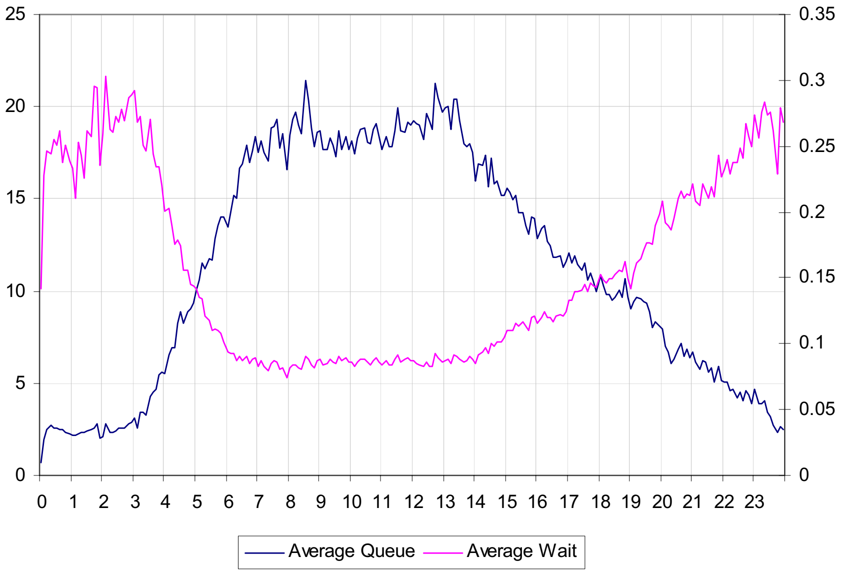
QED
α=0.5

Seconds



ED
α=0.9

Minutes



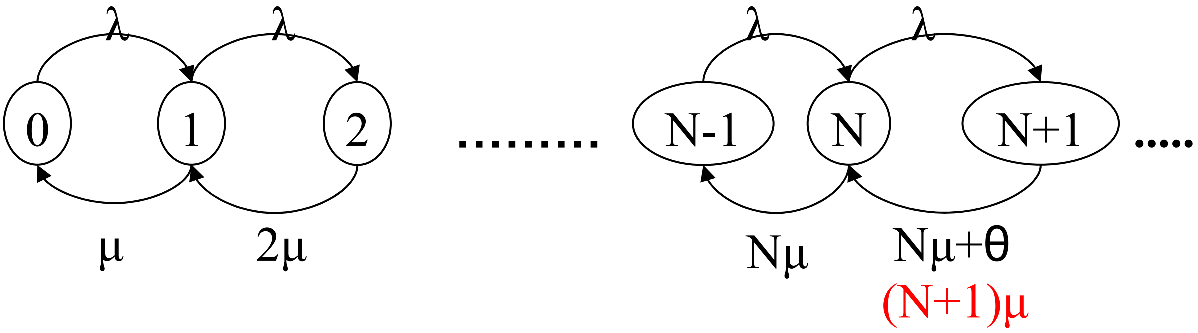
Erlang-A: Moderate (Im)patience

- 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

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

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

For R not too small:

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

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

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

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

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

Time-Varying Arrivals

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

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

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

Time-Varying Arrivals

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

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

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

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

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

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)

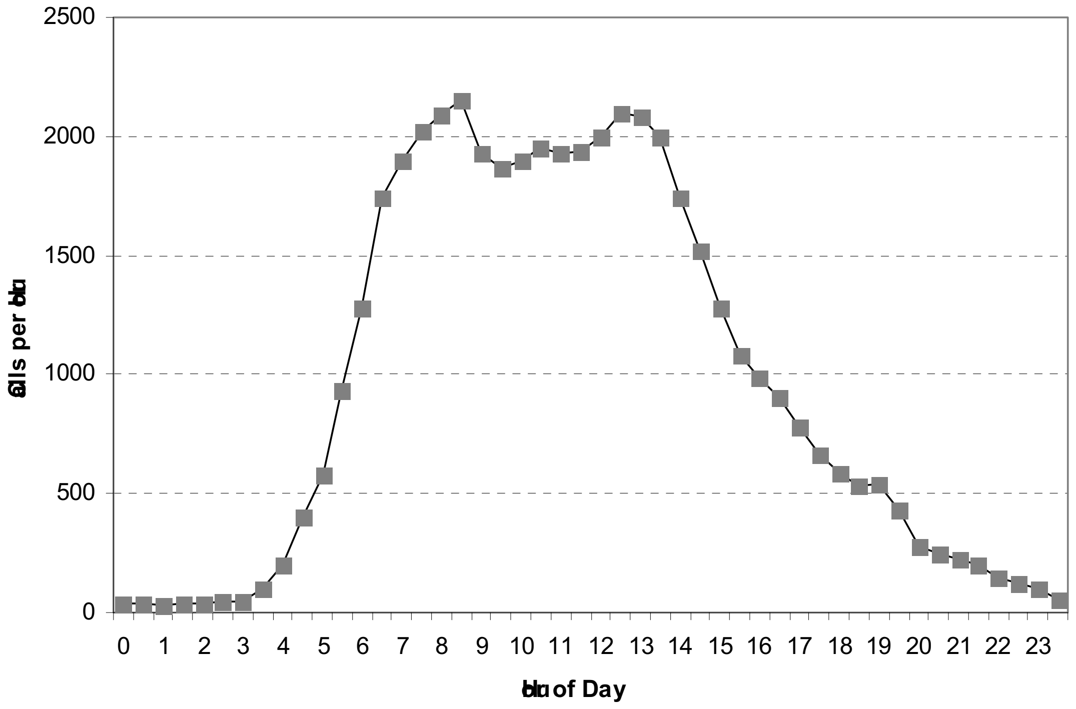
- ***Service grade*** β_t *in interval t*, which arises from the following "Square-Root Staffing" rule:

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

Example: "Real" Call Center

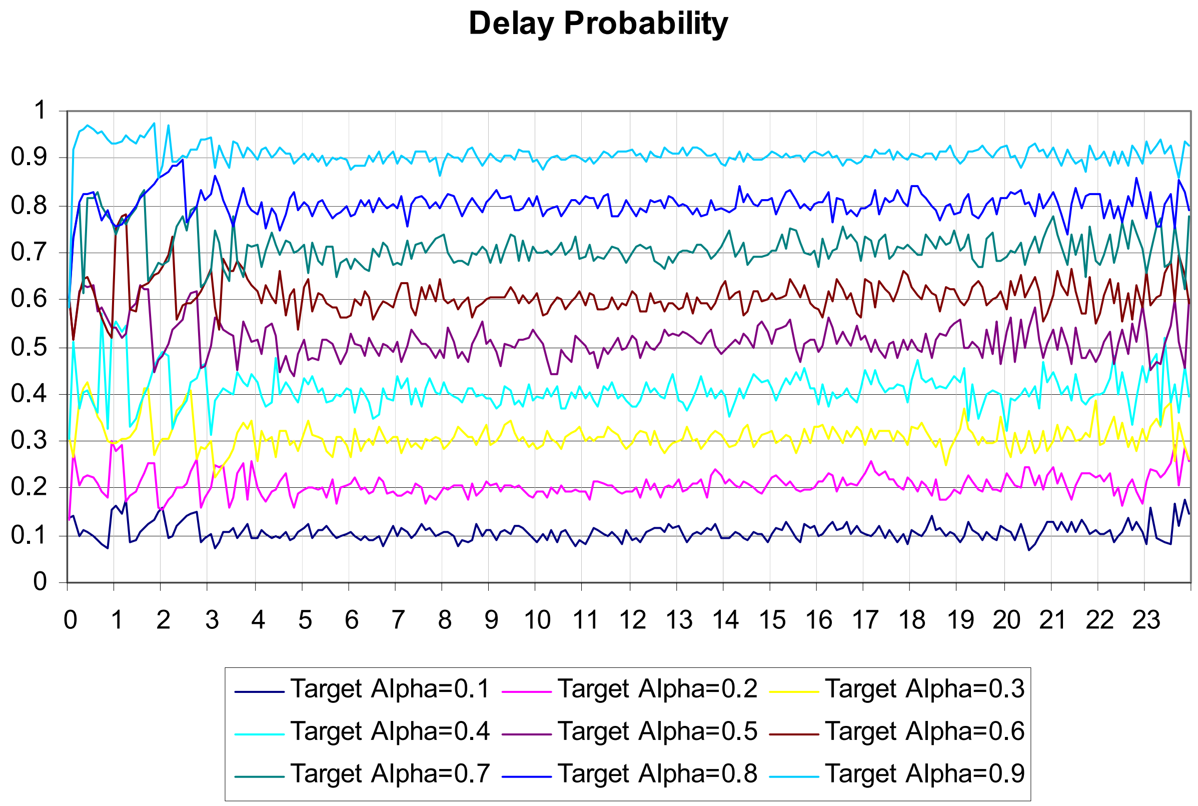
Two-hump arrival functiona are typical

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

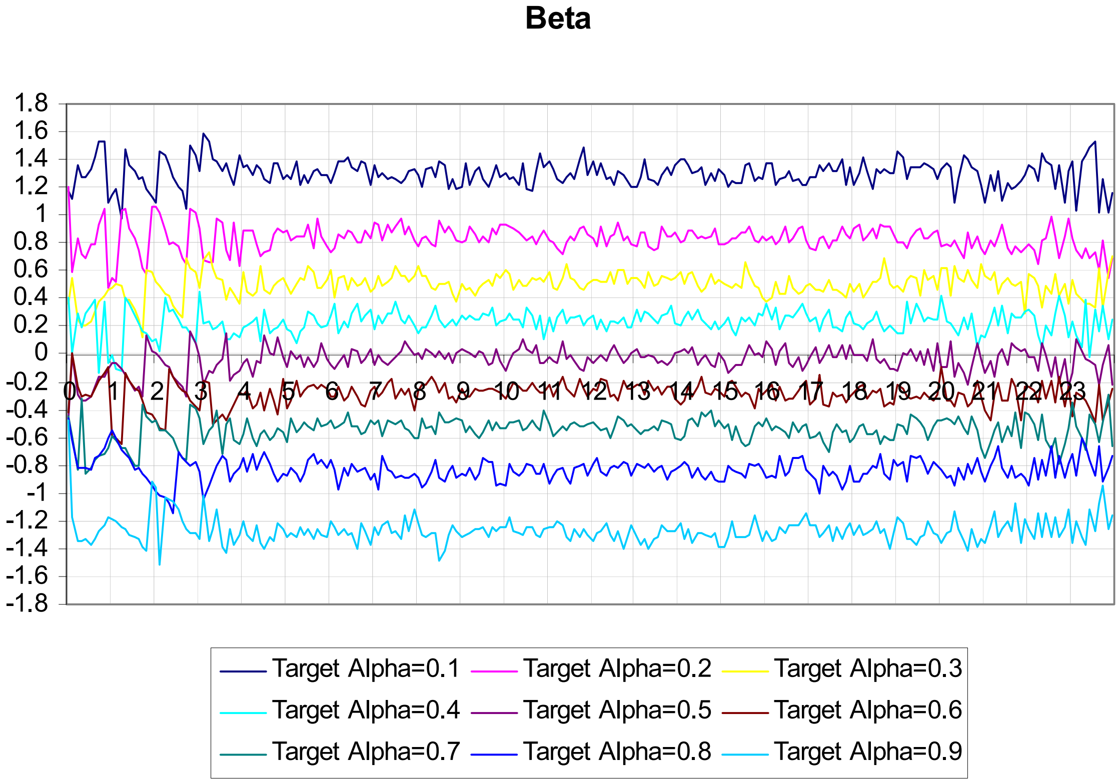


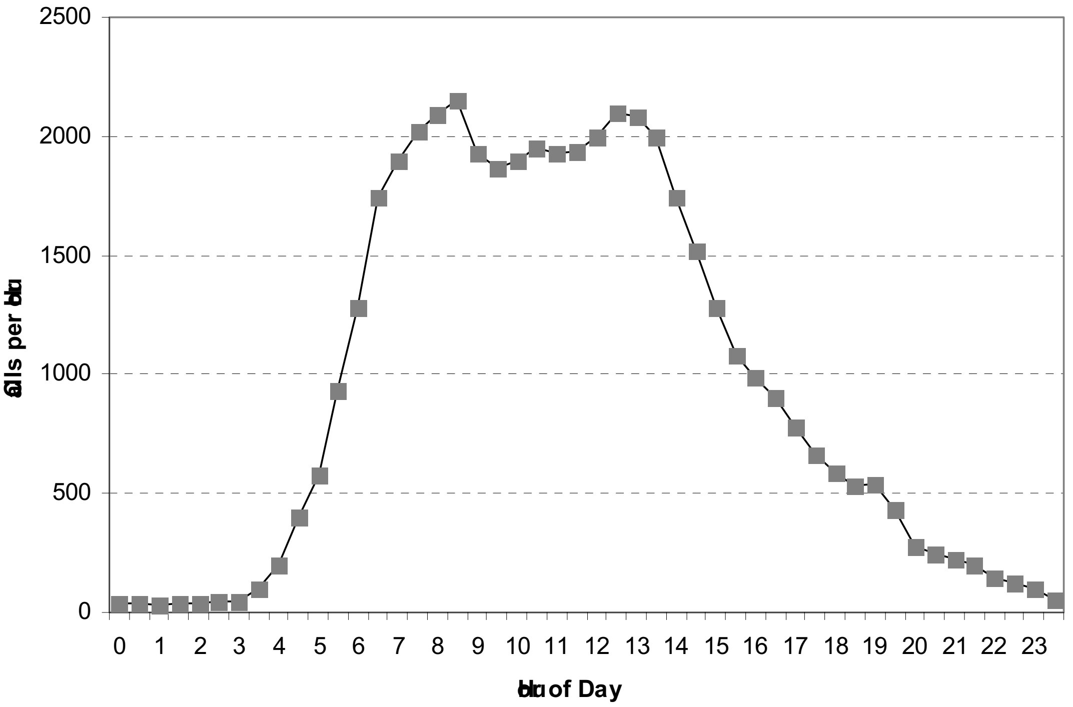
- Service and abandonment rates are both exponential having mean 0.1 (6 min.)

Delay Probability α

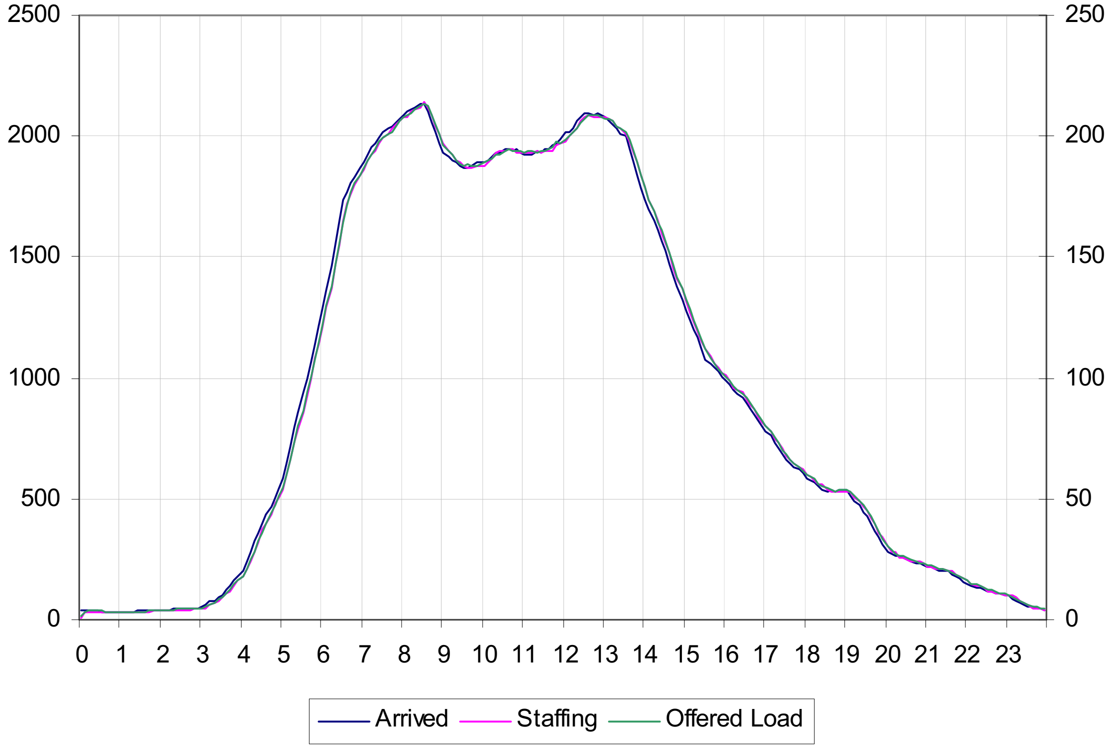


Service Grade β



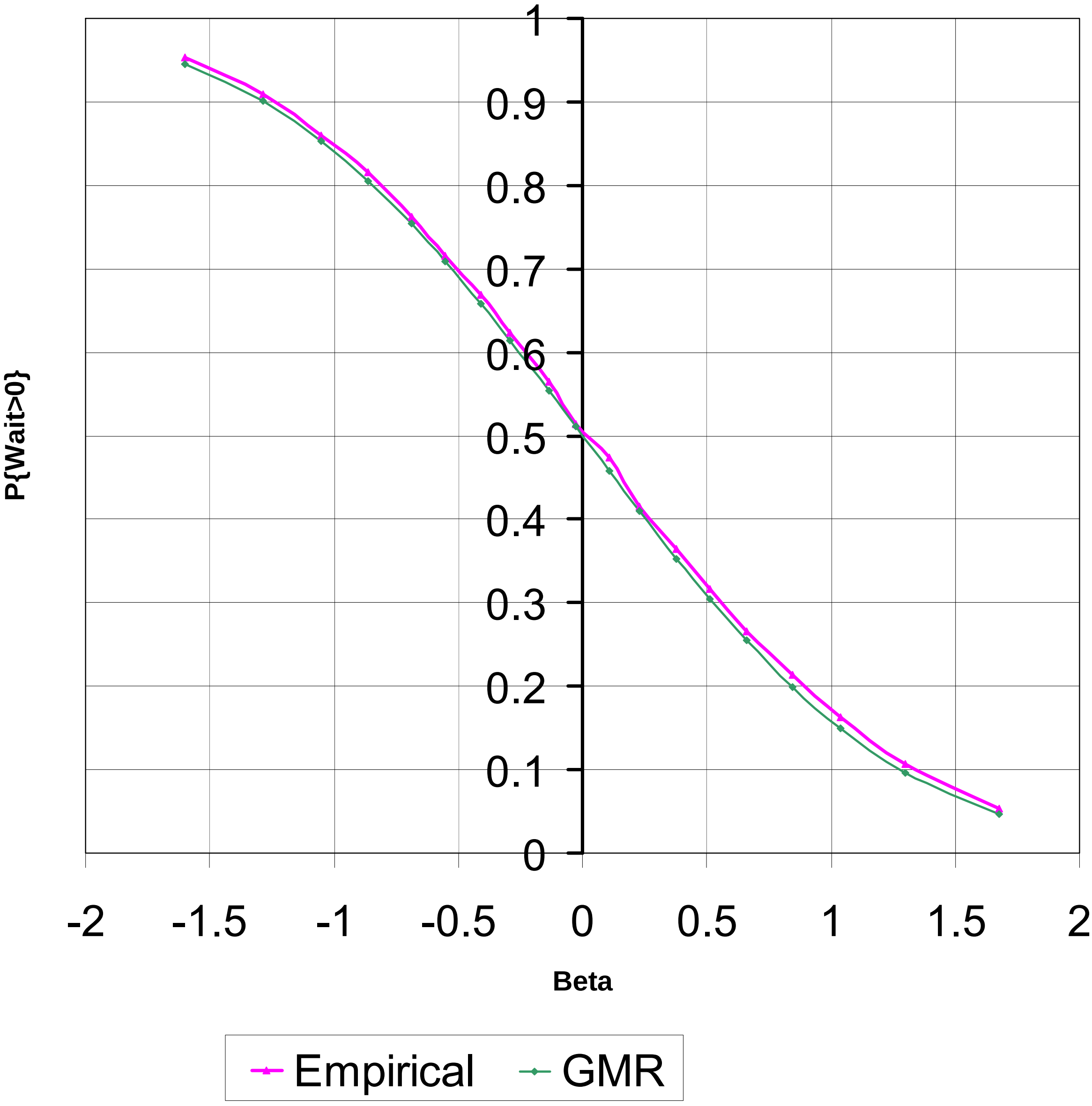


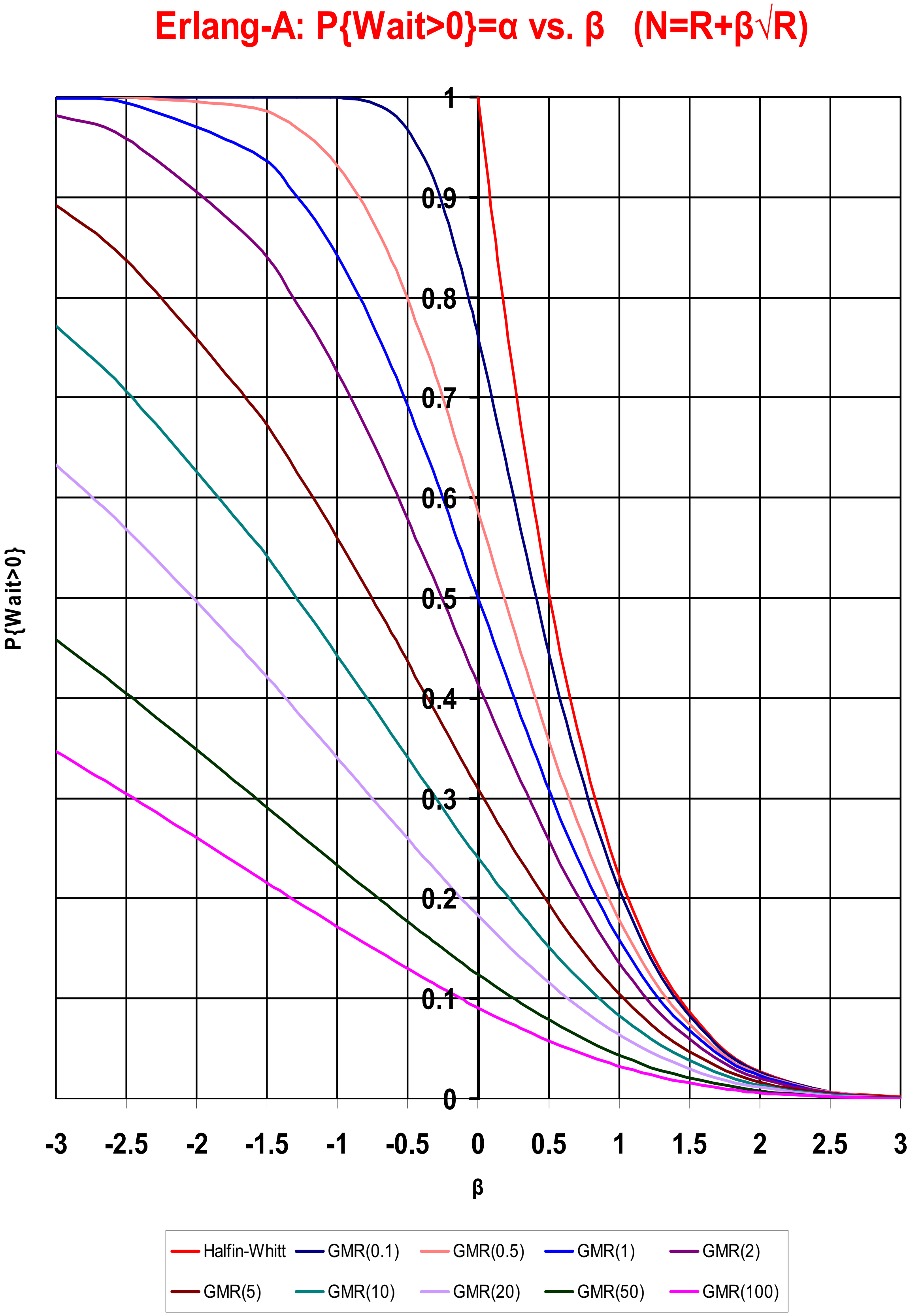
QED Staffing (α=0.5)



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

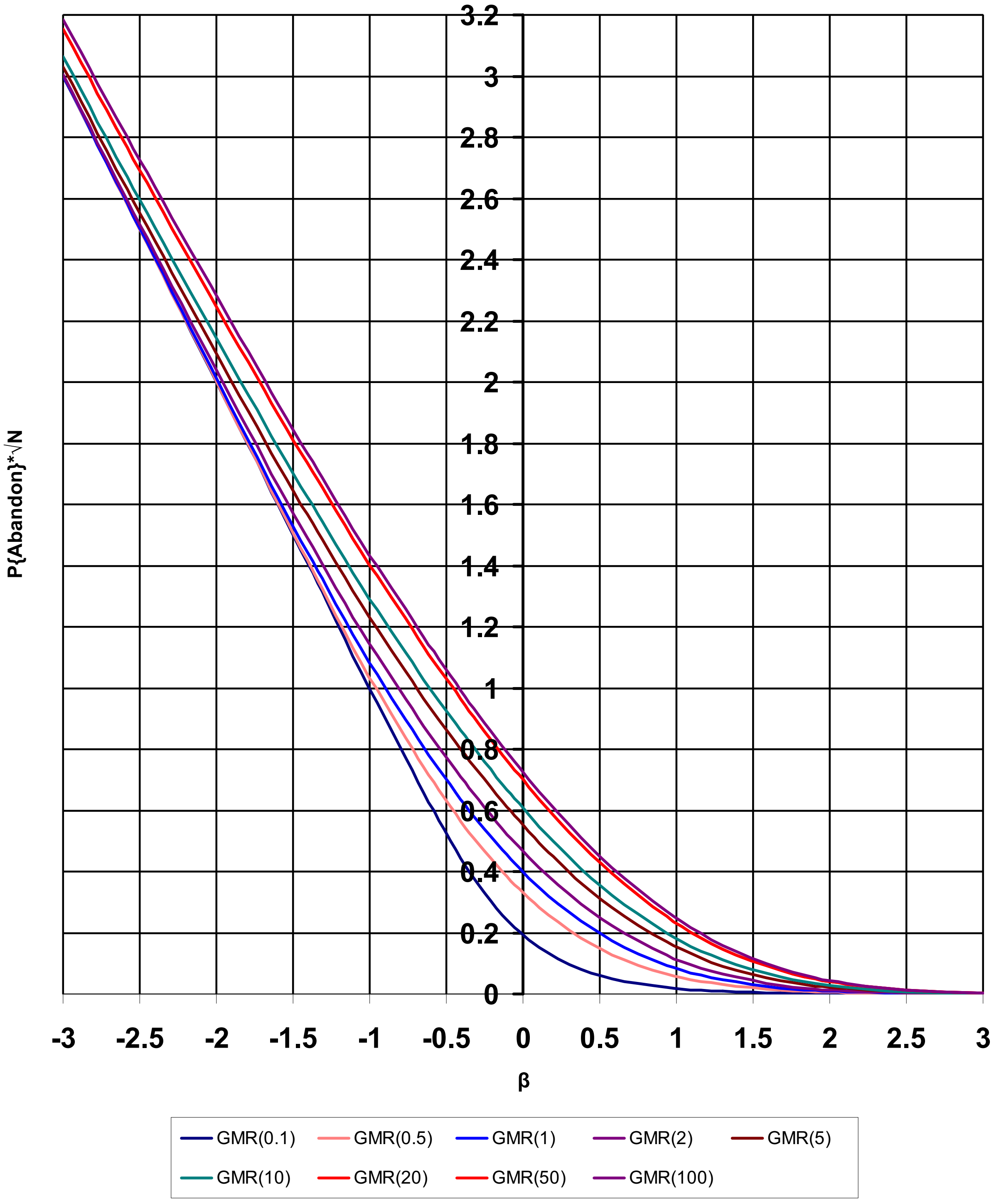
Moderate Patience





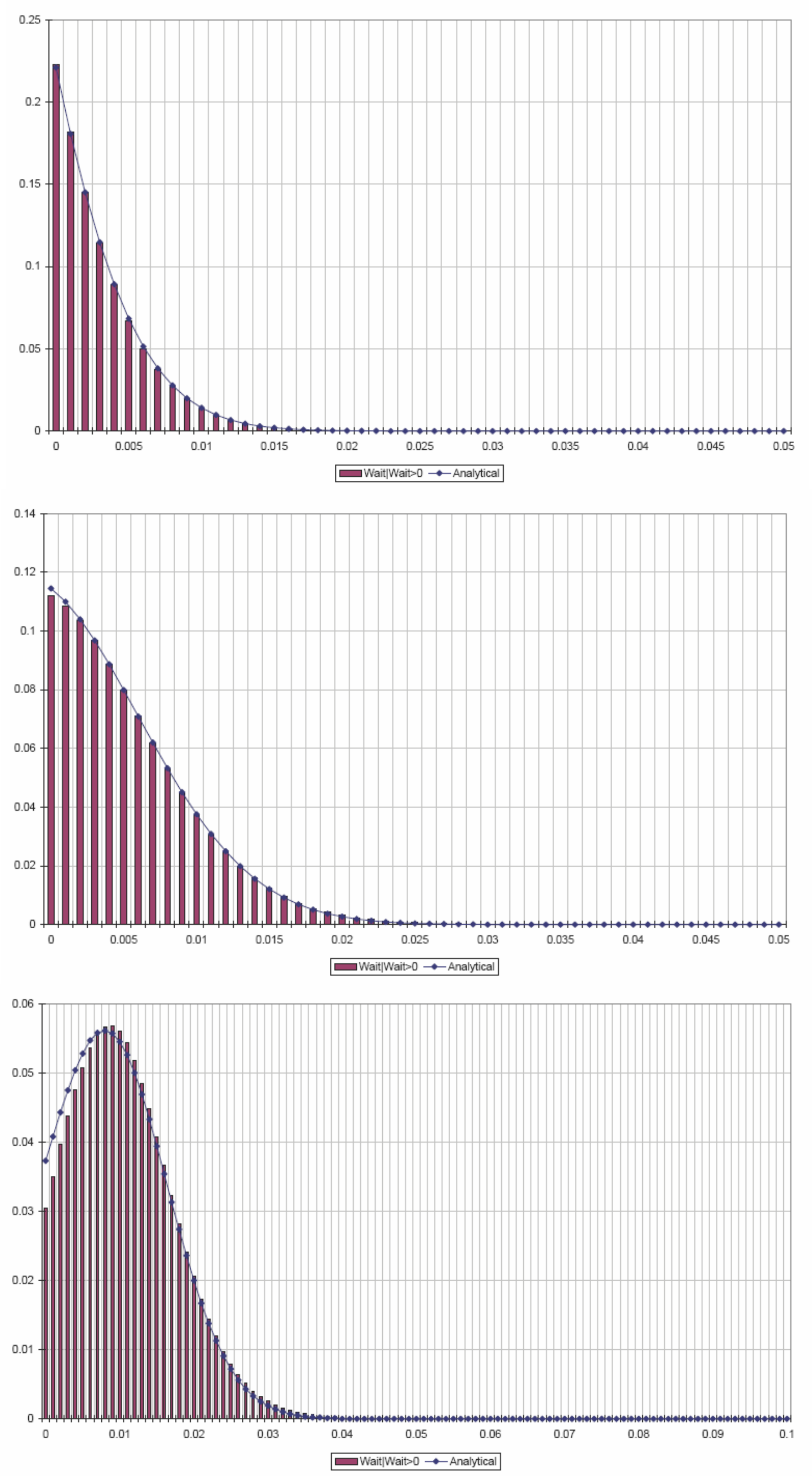
GMR(x) describes the asymptotic probability of delay as a function of β when $\theta/\mu = x$. Here, θ and μ are the abandonment and service rate, respectively.

Erlang-A: P{Abandon}*√N vs. β



Real Call Center: Empirical waiting time, given positive wait

(1) α=0.1 (*QD*) **(2)** α=0.5 (*QED*) **(3)** α=0.9 (*ED*)



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 $N_{i+1}(\bullet)$ is our staffing level;

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

(∞) Last iteration. The algorithm converges to a

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

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