Server Staffing
in Real-World Telephone Service Systems

by

Shawniqua T. Williams

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Bachelor of Science and Master of Engineering
in
Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1995

Author .....
Department of Electrical Engineering and Computer Science
May 12, 1995

Certified by .....
Vien Nguyen
Assistant Professor of Management Science
Thesis Supervisor

Accepted by .....
F. R. Morgenthaler
Chairman, Departmental Committee on Graduate Theses

Server Staffing
in Real-World Telephone Service Systems

by

Shawniqua T. Williams

Submitted to the Department of
Electrical Engineering and Computer Science
May 12, 1995
in partial fulfillment of the requirements for the degrees of
Bachelor of Science and Master of Engineering
in
Electrical Engineering and Computer Science

Abstract

This thesis compares methods of solving the *operator staffing problem* as relates to a real-world telephone service system. The system studied is the Cardmember Services Department at First USA Bank. The operator staffing problem asks for he minimum number of operators to be staffed as a function of time so that a certain level of performance is maintained. The performance level can be measured by the probability that a caller must wait in queue before speaking to an operator. I model part of the network as an infinite capacity queue with a Poisson arrival process and exponentially distributed service times. Both the arrival and service processes are time-dependent. I show that approximating the queue using an infinite server model yields similar results to approximating it as a stationary queue with additional variability in the arrival process.

Thesis Supervisor: Viên Nguyen
Title: Assistant Professor of Management Science

Acknowledgments

Special thanks to Derrick Forbes, Jr. for his help and support, and to Prof. Viên Nguyen for her patient guidance and vision.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The purpose of this thesis is to investigate the *operator staffing problem* as it relates to a real-world telephone service system. Given a service center to which incoming calls arrive according to a probabilistic arrival process, the operator staffing problem asks for the minimum number of operators to be staffed so that a desired level of service is achieved. This level of service can be characterized in several ways, one of which is the percentage of calls to the service center that are delayed. It is generally assumed that the staffing level must be determined based on forecasts of number of calls and handling time, and not in response to real-time system measurements.

1.1 The Subject: First USA

The real-world system to be analyzed is the Cardmember Service center of First USA Bank, a credit card company. The company receives account-related inquiries via its toll free telephone numbers 24 hours a day. Its objective is to effectively service as many calls as possible at lowest cost.

Salaries of the telephone operators (also called agents or representatives) constitute a major cost, as well as communications services (toll-free numbers and telecommunications equipment). Salaries are a function of the number of operators staffed,

while telephone bills are calculated according to the number and length of calls. The equipment costs, which come from amortized expenses and maintenance fees, are relatively fixed. Calls that arrive while all operators are busy wait in a first come, first served queue. The time these calls spend in queue adds to the length of the calls.

The planning problem at First USA can be divided into four components: call forecasting, operator staffing, operator scheduling and call routing. Call forecasting is the problem of estimating the number of calls throughout the day as well as the processing time requirements of each call. The operator staffing component uses the forecasted data to determine the number of representatives to staff at each time interval. Scheduling refers to actual allocation of operators under constraints like the total number of representatives available and the maximum consecutive number of hours an individual can be asked to work. The issue of call routing arises from the fact that there are two separate locations that can handle the calls.

The scope of this research is limited to the second component: server staffing. Scheduling and routing constraints are ignored, and it is assumed that forecasting results are reasonably accurate.

1.2 Methodology

The steps taken in the course of this research are as follows:

1. First I observed First USA's system to determine its constraints and characteristics. This step culminated in the development of a model that is specific enough to represent the major characteristics of the system and general enough to allow mathematical analysis.

2. The next step was to study various methods of solving this and related models. By *solving* I mean developing formulas or approximations for the performance parameters of the model. The set of parameters might include the probability that a caller must wait before speaking to an operator (probability of delay), or

the probability density function (pdf) for the number of calls being handled at any given time. These concepts are further explored in Chapter 2.

3. Third, I chose the methods that were most applicable to this problem and determined how they could be applied to First USA's system. I used these methods to calculate staffing levels for randomly chosen days for which forecast data was available.

4. Finally, I compared the performances of the methods and developed recommendations as to how First USA should go about solving this problem.

1.3 Overview of Thesis

The remainder of the thesis is structured as follows: Chapter 2 gives an explanation of basic queuing theory concepts necessary for understanding this research. It also discusses prior research related to the subject. Chapter 3 describes First USA's system and presents a queueing model of the problem. Chapters 4 and 5 discuss two methods for solving this and similar models. Finally, the conclusion restates the purpose of this research, the methodology and findings.

Chapter 2

Background and Literature Review

In this Chapter I will acquaint the reader with some basic queueing theory that is necessary for understanding this document, and discuss related prior research.

2.1 Introduction to Queueing Theory

Queueing theory is the study of systems in which a stream of customers or *jobs* arrive to be serviced. The jobs may be people arriving to a gas station, automobiles to be processed on an assembly line, or in this case telephone calls to be handled by operators. In some systems there are many *servers* (the gas pumps, assembly stations or operators) available to process the jobs, and in others there may be only one. Jobs that arrive while all servers are busy may wait in a queue or be rejected.

Queueing systems are characterized by a string of characters of the following form: $G_1/G_2/s/k/Disp$. Here G_1 represents the type of probabilistic process which governs the arrival of customers and G_2 represents the type of service time distribution. The parameter s is the number of servers, and $k-s$ is the size of the waiting area ($k=\infty$ if the queue size is unlimited). A zero-capacity waiting area may be denoted by $k=s$

or $k = 0$. *Disp* refers to the service discipline, or the rules used to determine the order of service. This thesis will be concerned only with the first come first served (FCFS) discipline, which means customers are accepted to service in the order that they arrive. Other possible disciplines include last come first served (LCFS), where the next customer to be serviced is the one that arrived most recently, and processor sharing (PS), where servers divide their energy equally among the customers present. For the purposes of this research, I will assume the service time distribution to be identical for all servers. In cases where the network capacity and the service discipline are omitted, an infinite capacity FCFS queue is implied.

The purpose of studying queueing systems is to determine formulas and approximations for such performance parameters as probability of delay, expected wait time and blocking probability. The probability of delay is the probability that an arriving customer has to wait in queue before being processed. The expected wait time is the average amount of time that a customer has to wait in queue before beginning service. When the queue size is finite, the blocking probability is the probability that an arriving customer finds the system at full capacity and is rejected.

Fundamental to the study of queues in the notion of *server utilization*, which I denote by ρ:

$$\rho = \frac{\lambda}{s\mu} \tag{2.1}$$

The server utilization measures the long-term average percentage of time each server is busy and is given by the ratio of the average arrival rate of customers (λ) and the average rate at which customers can be served ($s\mu$). It has been shown that larger systems can perform as well as smaller ones (with "good" performance characterized by a low probability of delay) at higher utilization levels (see [7]). The size of the system is characterized by the number of servers and the service rate. For example, consider two systems with Poisson arrival processes and exponential service time distributions (Table 2.1). A system with one server, a mean service rate of 5 jobs per minute and a mean arrival rate of 3 jobs per minute will have a

	Example 1	Example 2
Arrival Rate	3 jobs/min	23 jobs/min
Service Rate	5 jobs/min	10 jobs/min
# Servers	1	3
Delay Probability	0.6	0.6
Utilization	0.6	0.77

Table 2.1: Comparison of utilization and performance levels for an $M/M/1$ queue and an $M/M/3$ queue

steady state probability of delay of 0.6. A second system with 3 servers and a mean service rate of 10 jobs per minute will have a 0.6 probability of delay in steady state if the mean arrival rate is as much as 23.1 calls per minute. The server utilization in the former case is 0.6, compared to 0.77 for the latter. In general queueing systems require the utilization to be less than unity in order to be stable. If $\rho > 1$, jobs are arriving faster than they can be processed, which means the number of customers waiting in the system can only get infinitely larger as time goes on. When $\rho = 1$ the probabilistic nature of the arrival and service processes means that the number of customers waiting in queue could be increasing or decreasing at any time. However, it has infinite room to grow and only finite room in which to decrease (e.g. the number of calls in queue can never go below zero). Therefore, the number of calls in queue in the case of unit utilization is unbounded, meaning such a system is instable. For the remainder of this thesis, the utilization will be assumed to be less than unity.

2.2 The $M/M/1$ Queue

Perhaps the simplest of all queueing systems is the $M/M/1$ queue, characterized by a Poisson arrival process, an exponential service time distribution, and a single server (Figure 2.2). The omission of the last two parameters implies that the queue has infinite capacity and FCFS service discipline. The use of the letter "M" refers to the memorylessness of the arrival and service distributions. This property implies that

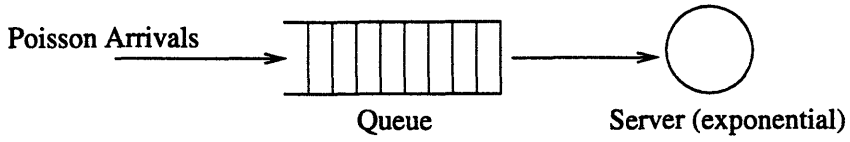

Figure 2-1: The $M/M/1$ Queue

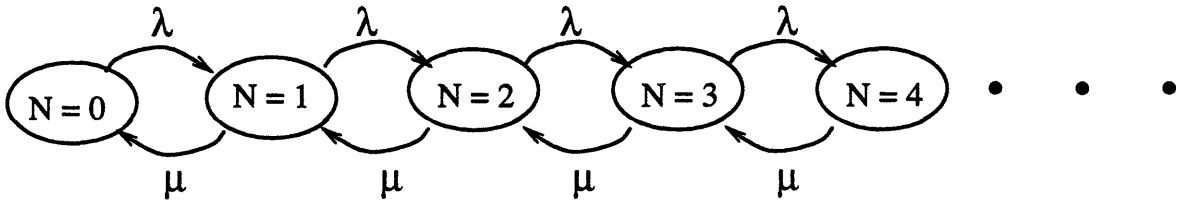

Figure 2-2: Markov chain governing an M/M/1 queue

the number of customers in the system (which includes customers in the queue as well as those being serviced), denoted by N, can be represented by a Markov birth-death chain as shown in Figure 2.2. The steady state probability density function of this Markov Chain is as follows:

$$P(N = n) = \frac{\rho^n}{1 + \sum_{k=1}^{\infty} \rho^k} = \rho^n(1 - \rho). \tag{2.2}$$

(Recall that $\rho < 1$.)

The probability of delay is simply the probability that an arriving customer sees more than one customer already in the system. Because "Poisson arrivals see time averages" (PASTA) (see [4]) the probability of delay is equal to the probability that there are one or more customers in the system *at any time*:

$$P_{delay} = 1 - P(N = 0) = \rho \tag{2.3}$$

2.3 The $M/M/s$ Queue

The $M/M/s$ queue is a generalization of the $M/M/1$ queue where there may be more than one server. As discussed earlier, this queue can offer the same level of

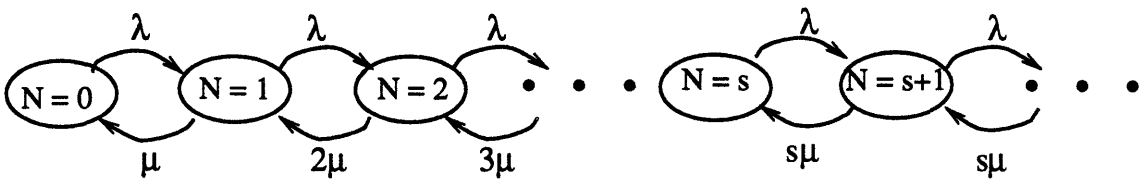

Figure 2-3: Markov chain governing an M/M/s queue

performance as an $M/M/1$ queue at a higher level of utilization, however analysis is more complicated because the death rate in the governing Markov birth-death chain varies depending on the number of jobs in the system (see Figure 2.3): it is $N\mu$ if $N < s$ and $s\mu$ if $N \geq s$. The resulting steady state pdf for N is as follows for $n \leq s$:

$$P(N = n) = \frac{(s\rho)^n(\frac{1}{n!})}{1 + \sum_{k=1}^{s-1}(s\rho)^k(\frac{1}{k!}) + (s\rho)^s(\frac{1}{s!(1-\rho)})} \tag{2.4}$$

and for $n \geq s$:

$$P(N = n) = \frac{(s\rho)^n(s^{s-n})\frac{1}{s!}}{1 + \sum_{k=1}^{s-1}(s\rho)^k(\frac{1}{k!}) + (s\rho)^s(\frac{1}{s!(1-\rho)})}. \tag{2.5}$$

The probability of delay is given by:

$$P_{delay} = 1 - \sum_{k=0}^{s-1} P(N = k). \tag{2.6}$$

2.4 General Arrival and Service Processes

A general arrival and/or service distribution is denoted by the letter "G". Current literature does not offer any closed-form solutions for the distributions of the performance parameters of a $G/G/s$ queue, as its interarrival and service time distributions are not made explicit by its description. The $G/G/\infty$ queue, however, is more tractable because the effects of queueing are eliminated: customers arriving to the infinite server queue go straight to service and are not affected by those already in

the system.

The $G/G/\infty$ queue is rarely observed in real life, however it is often used to approximate the more common $G/G/s$ queue. In cases where the finite server system under consideration has a low level of utilization, this *infinite server approximation* performs well because arriving customers are more likely to find a server free, thus queueing is diminished.

In the $G/G/\infty$ queue, the steady state pdf of the number of customers in the system approaches a normal distribution as the arrival rate increases with respect to the service rate (see [7]):

$$\lim_{\lambda \to \infty} P(N \geq s) = 1 - \Phi[\frac{s - \rho s}{\sqrt{z \rho s}}] \tag{2.7}$$

In Equation (2.7) z represents the heavy traffic peakedness, which is a measure of congestion when the system is relatively full:

$$z = 1 + \mu(c_a^2 - 1) \int_0^\infty [1 - G(t)]^2 dt \tag{2.8}$$

Heavy traffic peakedness will be discussed in further detail in Chapter 5. $\Phi[x]$ is the standard normal density function, defined by:

$$\Phi[x] = P[\mathcal{N}(0,1) \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(\frac{1}{2})x^2} dx \tag{2.9}$$

2.5 Nonstationary Queues

All of the queues discussed above are time-invariant, i.e. the parameters by which they are described (arrival and service rates, number of servers, etc.) do not change with time. In *nonstationary* queueing systems, some parameters may be functions of time. This is the case at First USA, for example, where the call arrival rate at midday is much larger than at 2:00am. Time dependence is typically denoted by a subscript 't' in the system description (e.g. $M_t/G_t/s_t$). Nonstationary systems are extremely difficult

to analyze and offer no closed-form solutions such as Equations (2.2) through (2.9). Indeed, they are the subject under investigation in this thesis.

Chapter 3

First USA

In this chapter I provide a description of the telephone service system to be studied, develop the model to be used in this research, and discuss the assumptions and simplifications of the model.

3.1 Overview

Figure 3-1 shows a diagram of First USA's Cardmember Services telephone network. There are two locations where cardmember calls are handled: Wilmington, Delaware and Austin, Texas. Each site is equipped with two *Voice Response Units*, or VRUs. The VRUs are automated computer systems that answer calls and run prerecorded scripts. Through touch-tone technology callers can communicate with the VRUs, navigating through menus and retreiving information about their own accounts. Each VRU can handle twelve calls at one time.

Calls to the Cardmember Services Department are routed to one of two sites (the methodology behind the routing procedure is beyond the scope of this thesis). When a call arrives at a site it is first directed to a VRU. There are some exceptions to this rule, for example calls that require TTY (telephone typewriter) communication for hearing-impaired cardmembers are routed directly to an agent. These and other

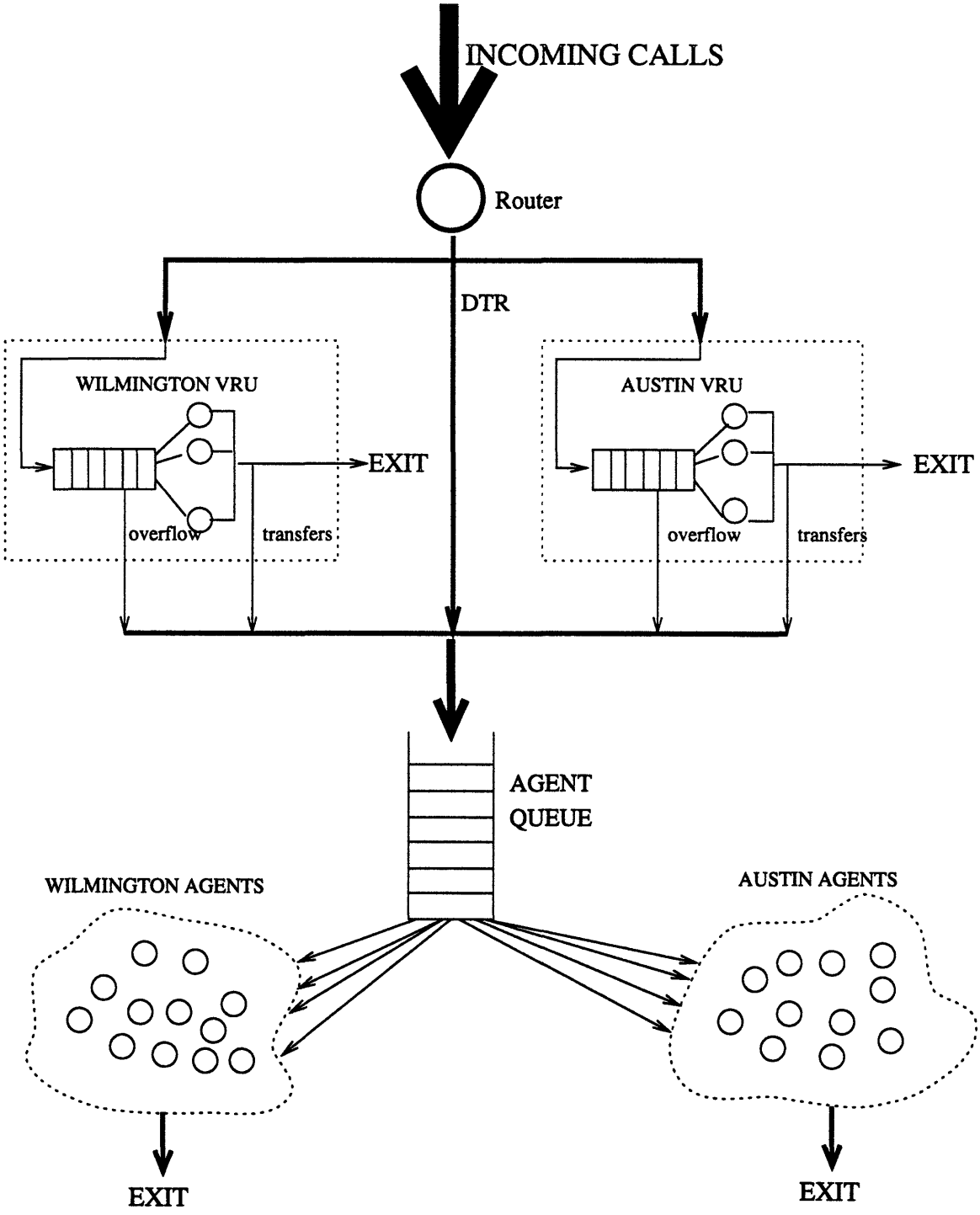

Figure 3-1: First USA's Cardmember Services telephone system

direct to rep (DTR) calls make up approximately 6.5% of the total calls received.

If all 24 VRU ports at a site are busy with other calls, normal incoming calls wait in the *VRU queue.* Each location has a FCFS VRU queue with a capacity of 30 calls. Calls that arrive while the VRU queue is full go straight to the *agent queue.* This phenomenon is referred to as *VRU overflow.* Calls that have been in the VRU queue for more than 10 seconds also contribute to VRU overflow, but this is a very rare occurrence. VRU overflow calls make up about 0.5% of the total call volume.

Callers that interact with the VRU choose from several menu options such as "Balance Inquiry," "Filing a Dispute" and "Status of Application." Transferring to a representative and repeating the menu are also options. Forty-four percent of the callers that interact with the VRU eventually transfer to an agent.

There is a single FCFS agent queue of unlimited capacity that feeds operators at both sites. When a call begins service (is connected to an agent), the agent creates a database record of the conversation through a tracking system called First Assist. First Assist also allows agents to retrieve and alter information about cardmembers' accounts. The information is stored in a large database operated by First Data Resources (FDR), a company located in Omaha, Nebraska. Most calls require that information be transferred to and from FDR.

The administrators of the system keep constant vigil over its performance. Statistical information is gathered every half hour and compiled every day. It includes the total number of calls received, the average time required to service a call and the number of calls transferred from the VRU to the agent queue. The performance of the system is measured by the *service level* (SL), the *average speed of answer* (ASA) and the *average handle time* (AHT). The service level is the percentage of calls that wait in the agent queue for less than 20 seconds. The average speed of answer is the average amount of time a call waits in the agent queue. The average handle time measures the average amount of time a caller spends talking to an agent. The latter two performance measures can also be applied to the VRU, but unless this is explicitly stated, they refer to the agent queue.

Day of Week	Average # Calls
Monday	31077
Tuesday	24993
Wednesday	23075
Thursday	22681
Friday	21377
Saturday	11688
Sunday	6869

Table 3.1: Average call volumes to First USA for each day of the week

3.2 Characteristic Behavior

3.2.1 General Statistics

Table 3.1 shows average call volumes by day of the week. These values were computed from the measured call volumes of approximately five weeks. Mondays are the busiest days, while Saturday and Sunday contact rates are very low. Many calls result from cardmembers receiving their account statements in the mail. Those receiving statements on Friday and Saturday often wait until the next business day (Monday) to call. Analysts forecast the number of calls per day based on past history and recent relevant events (such as statement mailings). Historical forecast errors for the number of calls per day range from 13% under actual rates to 20% above actual rates.

3.2.2 Agent Interval Statistics

Figure A-1 shows the typical trajectory of call arrival rates to the agent queue throughout the day for each day of the week. The number of calls reaches its peak between 10:00am and 6:00pm and drops to under 1.5 calls/minute between 2:00am and 6:00am. Figure A-2 compares average arrival rate trajectories for Wednesday, Saturday and Sunday of a given week.

Interval statistics are also given for average service rates over the course of a day. As can be seen in Figure A-1 the service rate is considerably less volatile than the

arrival rate. The arrival rate for a typical Wednesday may range from one call every four minutes to as many as 32 calls per minute, while the service rate for the same day may range from one call every two minutes to one call every four minutes.

3.2.3 VRU Interval Statistics

The call arrival rates to the VRU are approximately twice the rates to the agents, The service process for the VRU is even more stable than the agent service process, averaging to slightly under 1 call/minute (VRU AHT $\approx$ 63 seconds).

3.2.4 Performance

Service Level. At First USA the service level is defined to be the percentage of calls that wait in the agent queue for less than 20 seconds. One of the main goals is to consistently perform at 90% service level, meaning approximately 10% of calls that arrive to the gent queue wait 20 seconds or more. Figure A-3 shows service level graphs on an interval basis for several randomly chosen days. The service level is very inconsistent, often dipping below the target. Comparison with the call arrival forecast performance (Figure A-4) for the same days shows that this inconsistency in performance does not always coincide with ineffective forecasting, which suggests that more efficient staffing would be helpful. One of the goals of this thesis is to offer a method of staffing that will provide a more consistent level of service.

Average Speed of Answer. Another established measure of performance is the average speed of answer, which is the mean wait time in the agent queue. First USA compiles aggregate data on the average speed of answer daily. The average ASA computed over a period of two weeks is 7.34 seconds. There is no established target performance level with regard to ASA.

Probability of Delay. While First USA does not actively use probability of delay (equivalent to percentage of calls delayed in empirical terms) to measure the performance of their system, it is mentioned here because it is a measure of quality from the caller's perspective, much as is the service level. Also, calculations for probability of delay are simpler than those for service level, which make it a more desirable measure of performance from the researcher's point of view.

In order to get an idea of the target performance as measured by the probability of delay, I approximated the trajectory of the planned probability of delay $\hat{P}_{delay}$ for several randomly chosen days (see Figure A-5). These values were calculated using forecasted arrival rates, forecasted service rates, and "required agents", the number of servers as determined by First USA (Their method of calculation will be discussed in Section 3.4.). For each interval, the forecasted numbers were applied to the $M/M/s$ equations (Equations (2.4) and (2.6)). For each day shown, $\hat{P}_{delay}$ reaches an approximate peak of 0.25 between 4:00am and 6:00am and almost immediately decreases to a minimum of 0.08 or 0.09. The curve is jagged throughout the rest of the day, but the daily averages are consistently between 0.14 and 0.17. The overall average planned probability of delay for the dates analyzed was 0.154.

3.3 Narrowing the Focus

Since calculating the required number of agents is the primary focus of this work, I will limit the model to the diagram depicted in Figure 3-2, which consists solely of the agent queue and the agents at both sites. Arrivals to this queue consist of DTR and VRU overflow calls from both sites as well as calls transferred from the VRUs. This modified arrival process can no longer be accurately modelled as having a Poisson distribution. However, I assume it to be Poisson for simplicity and in the absence of detailed information. Likewise, the service time distribution for agents will be assumed exponential.

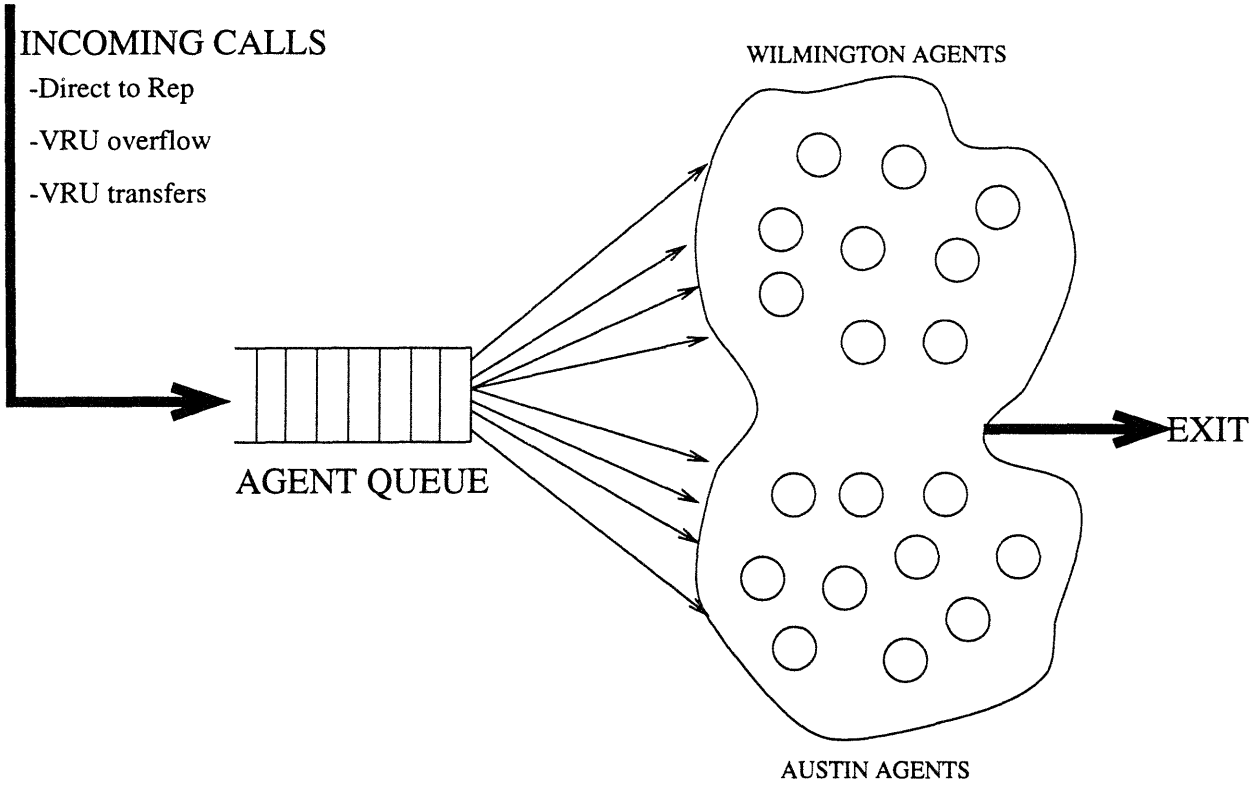

Figure 3-2: First USA's agent queue

3.4 Current Methodology

This section briefly discusses the methodology currently in use to determine staffing levels at First USA. The administrators of the system use a software package called Cybernetics, to which they input the forecast number of calls that the agents will receive each day. Cybernetics determines how the call volume will be distributed over the course of the day based on previous similar days. For example, the distribution of calls for a Monday would be based on the distribution from the last three Mondays, unless one of the four is a holiday.

The administrators were unable to explain the process Cybernetics uses to forecast average handle time. Likewise, they were not familiar with the mathematical models used to determine the appropriate server staffing level as a function of time. Figure A-6 compares Cybernetics' required agent calculations with the results of simple intuitive calculations using the forecast arrival and service rates in the following equation:

$$s(t) = \frac{\lambda(t)}{\mu(t)} \tag{3.1}$$

The above equation is what might be employed if the arrival and service processes were stationary and deterministic because only under these circumstances would a utilization rate ($\rho = \frac{\lambda}{s\mu}$) of 1 result in a stable system [7]. The ratio of servers calculated by Cybernetics to servers calculated according to Equation (3.1) ranges from 1.2 at peak arrival rates to 2.7 during the least busy times.

Table A.1 compares Cybernetics' required agents with staffing levels calculated using Equations (2.4) through 2.6 using a target delay probability of 0.154 for a randomly selected Monday. The values shown are very similar to the values derived by Cybernetics, however there is no reason to believe, from this and similar tables, that Cybernetics approximates First USA's system as a stationary M/M/s queue. The method compared to Cybernetics here is called a pointwise stationary approximation and will be discussed in detail in Chapter 4.

Chapter 4

The JMMW Method

The purpose of this chapter is to outline the server staffing solution provided in the paper by Jennings, Mandelbaum, Massey and Whitt [3], specifically as relates to First USA's problem. Section 4.1 will present the solution itself, followed by a discussion of its application to First USA's system in Section 4.2. Finally, the results of using the method to calculate the appropriate number of servers as a function of time will be presented in Section 4.3.

4.1 Methodology

The purpose of this model is to determine the appropriate number of servers as a function of time in a nonstationary $G_t/G_t/s_t$ queueing system. It requires as input the mean and variance of the arrival process ($\lambda(t)$ and $\sigma_a^2(t)$, respectively) and of the service process ($\mu(t)$ and $\sigma_s^2(t)$, respectively), as well as a target performance measure such as probability of delay. The goal is to provide a near constant quality of service over time. It assumes all servers are fed by a single infinite-capacity queue, and that the service time pdf is the same for all servers. It also assumes that the number of servers cannot be changed in real-time, in response to actual loads.

Jennings, Mandebaum, Massey and Whitt begin by discussing two simple methods

of analyzing such a system: the *pointwise stationary approximation* (PSA) and the *simple stationary approximation* (SSA). They then develop a method based on the infinite server approximation, designed for systems in which PSA and SSA may not adequately estimate the performance measures.

4.1.1 Pointwise Stationary Approximation

PSA approximates the performance measures of a queueing system at time t by their steady state distributions given the instantaneous parameters at that time. For example, using PSA one would estimate the probability of delay of an $M_t/M_t/s$ queueing system at time τ to be the steady state probability of delay in a stationary M/M/s queue with arrival rate $\lambda(\tau)$ and service rate $\mu(\tau)$; from Equations (2.4) through (2.6) we have:

$$P_{delay} = \frac{(s\rho(\tau))^s \frac{1}{s!(1-\rho(\tau))}}{1 + \sum_{k=1}^{s-1}(s\rho(\tau))^k \frac{1}{k!} + (s\rho(\tau))^s \frac{1}{s!(1-\rho(\tau))}} \tag{4.1}$$

where $\rho(\tau) = \frac{\lambda(\tau)}{s\mu(\tau)}$.

PSA's quick reaction to fluctuating arrival rates makes it desirable when the arrival rate changes slowly (or does not change at all) with respect to the mean service time, so that at each point in time the system is close to steady state. When this is not the case, PSA may underestimate the congestion of the system because the system does not rest long enough to approach steady state. Therefore, approximations based on steady state values may be inaccurate. Jennings, Mandelbaum, Massey and Whitt [3] offer an example of such a system, where the arrival process is Poisson and service times are exponentially distributed. Figure 4-1 shows the arrival rate function $\lambda(t) = 30 + 20\sin(5t)$ calls/min and the server staffing function that results from applying Equation (4.1). Compared to the constant service rate ($\mu(t) = 1$ for all t), the arrival rate goes from 10 to 50 jobs per minute in less than 38 seconds. Using a target delay probability of 0.13, the calculated number of servers oscillates between

15 and 60 in the same amount of time. The resulting actual probability of delay is unacceptable because it has a mean of 0.46 and oscillates nearly over the whole interval between 0 and 1.

4.1.2 Simple Stationary Approximation

SSA also uses the corresponding stationary model to approximate performance measures. In this case, however, the parameters of the approximation are derived from their long term average values in the system. Consider the previous example. SSA would also approximate this $M_t/M/s_t$ system using Equations (2.4) through (2.6), this time using $\lambda = \frac{1}{T}\int_0^T \lambda(t)dt = 30$ and $\mu = 1$. Here T represents the period of the arrival rate function; if it were not periodic one would choose some suitably large T so as to calculate the average value of $\lambda(t)$ over the length of time with which we are concerned. The resulting s(t) would be a constant of 38 for the target delay probability of 0.13. This results in an actual delay probability as shown at the bottom of Figure 4-2. The oversimplification of SSA causes P_{delay} to fluctuate from 0.05 to 0.28, because it ignores the effects of changing offered load. This oversimplification is desirable when the arrival rate oscillates rapidly (with a period considerably less than the mean service time), but otherwise is not adequate.

The PSA method is effective for systems where the arrival rate does not change significantly with respect to the service rate. The SSA method is useful for systems where the arrival rate fluctuates very rapidly with respect to the service rate. However, neither method seems appropriate when the system's arrival rate changes significantly but not extremely rapidly with respect to the service rate. Jennings, Mandelbaum, Massey and Whitt propose a third method of analysis to be utilized in such cases.

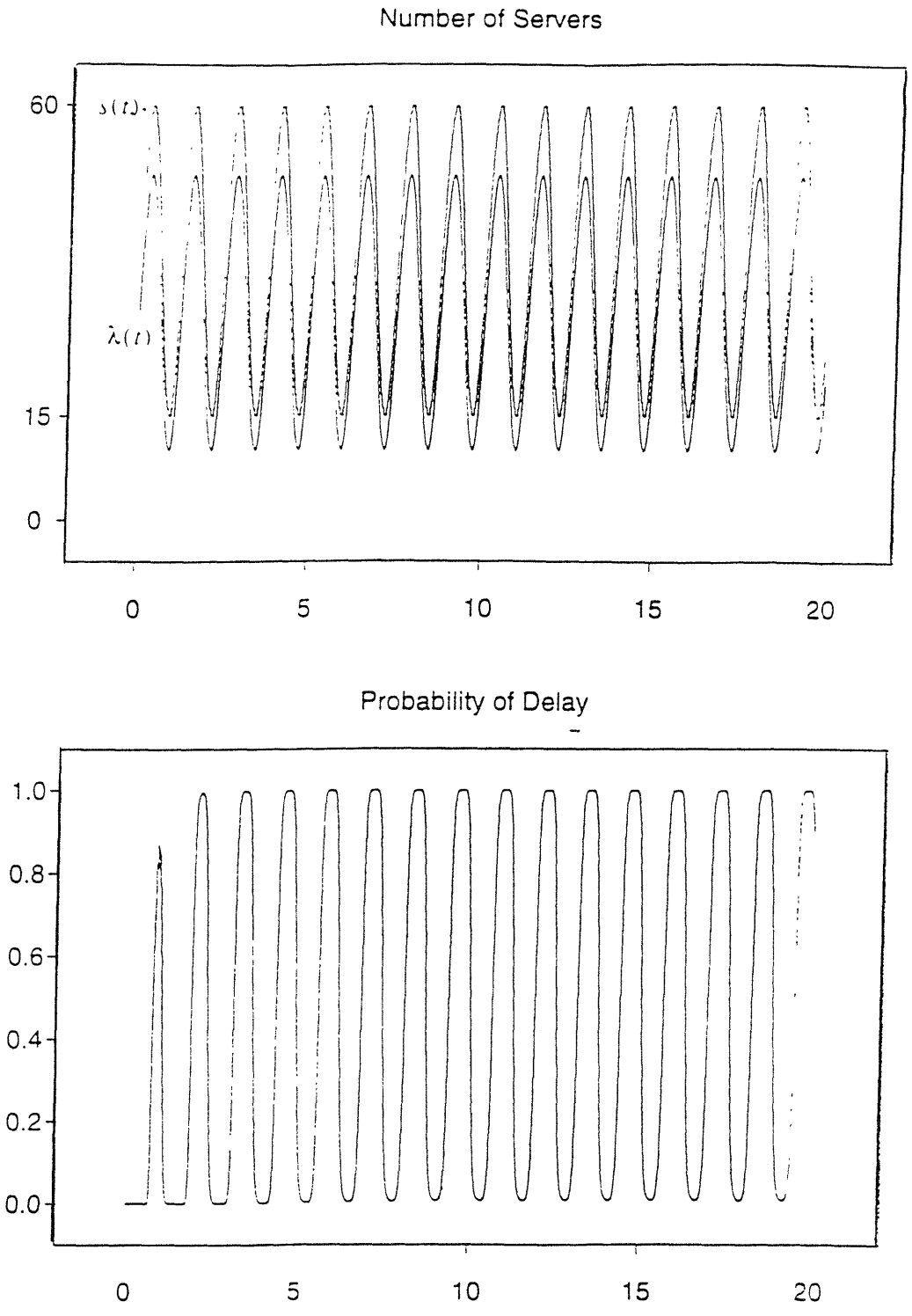

Figure 4-1: PSA applied to an $M_t/M/s_t$ queue with $\lambda(t) = 30 + 20\sin(5t)$, $\mu(t) = 1$ and target $P_{delay} = 0.13$ (example from Jennings, Mandelbaum, Massey and Whitt)

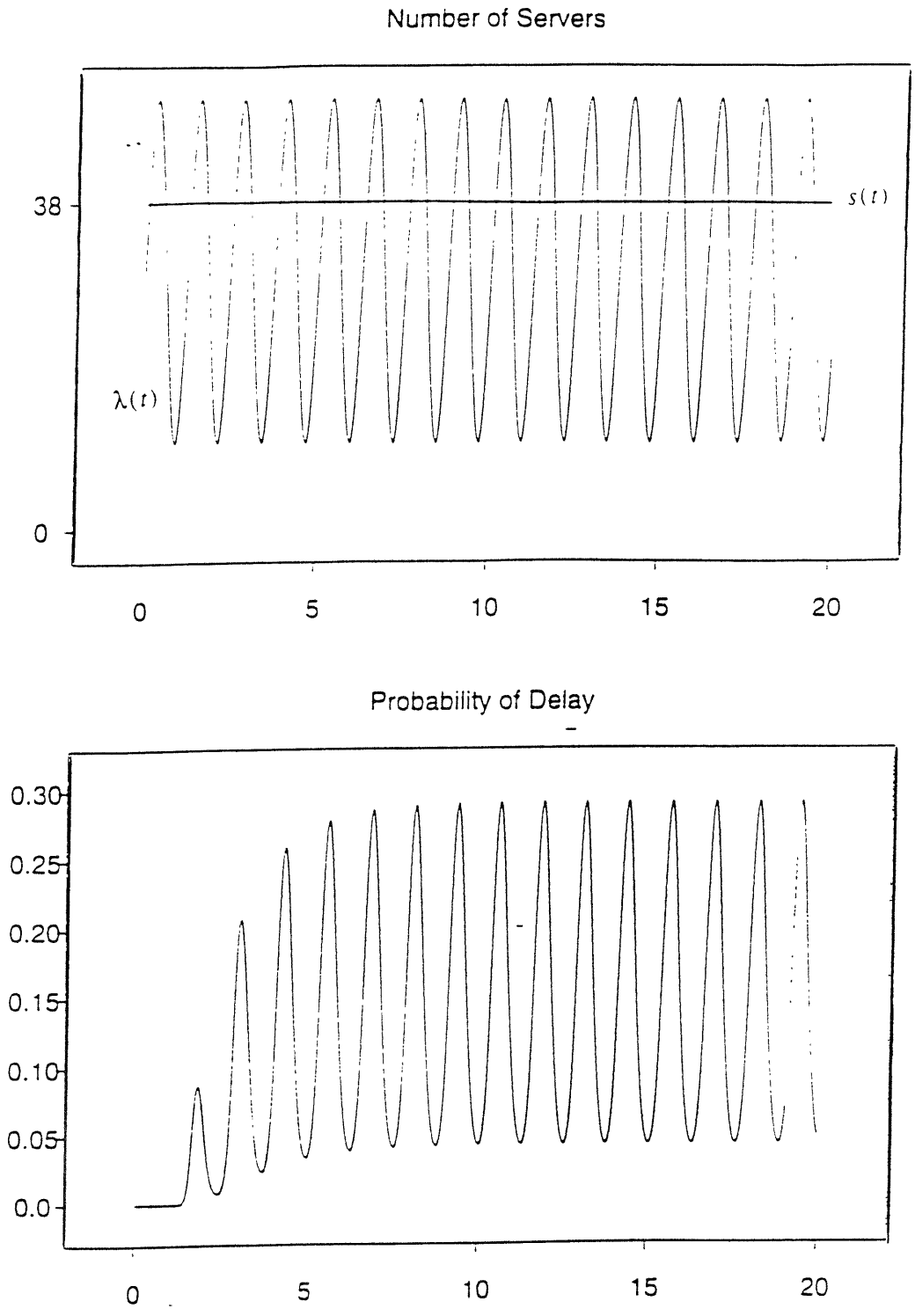

Figure 4-2: SSA applied to an $M_t/M/s_t$ queue with $\lambda(t) = 30 + 20\sin(5t)$, $\mu(t) = 1$ and target $P_{delay} = 0.13$ (example from Jennings, Mandelbaum, Massey and Whitt)

4.1.3 Infinite Server Approximation–The JMMW Method

Throughout this paper, I will refer to the method proposed by Jennings, Mandelbaum, Massey and Whitt [3] as the JMMW method. It is based on an infinite server normal approximation, in which the authors suggest approximating a $G_t/G_t/s_t$ queue by the corresponding $G_t/G_t/\infty$ queue. The number of servers is then chosen so that

$$P(Q(t) \geq s(t)) \leq \epsilon \tag{4.2}$$

and

$$P(Q(t) \geq s(t) - 1) > \epsilon \tag{4.3}$$

for some target probability ϵ, where Q(t) is the number of busy servers at time t in the $G_t/G_t/\infty$ queue. In systems where the probability of delay is very small (e.g. $P_{delay} \leq 0.01$), ϵ is a good approximation for the actual probability of delay in the $G_t/G_t/s_t$ queue. For systems with larger probability of delay the effects of queuing are nonnegligible. Thus a better approximation for P_{delay} would be:

$$P_{delay} = \epsilon + \sum_{k=s(t)+1}^{\infty} P(Q(t) = k) \tag{4.4}$$

Mathematical analysis has established that the distribution of Q(t) is approximately normal with mean and variance that I will denote by m(t) and v(t), respectively (see [7]). Therefore, $\frac{Q(t)-m(t)}{\sqrt{v(t)}}$ is approximately normally distributed with zero mean and unit variance, from which we obtain the following formula for the required number of operators:

$$s(t) = \lceil m(t) + \beta_\epsilon \sqrt{v(t)} \rceil \tag{4.5}$$

where $\lceil x \rceil$ denotes the smallest integer greater than x, and β_ϵ satisfies

$$P(\mathcal{N}(0,1) > \beta_\epsilon) = \epsilon. \tag{4.6}$$

s - server queue P(s servers busy) = P_{delay}	∞ - server queue P(s servers busy) = ϵ	Normal tail percentiles β_ϵ
0.001	0.001	3.115
0.005	0.005	2.614
0.010	0.009	2.375
0.050	0.041	1.740
0.100	0.078	1.420
0.250	0.175	0.936
0.500	0.306	0.506
0.750	0.413	0.221
0.900	0.467	0.083
1.000	0.500	0.000

Table 4.1: Example values of ϵ and β_ϵ in an M/M/s system for various P_{delay}'s

In a stationary M/M/s model with s as specified in Equation (4.5), the asymptotic probability of delay as the arrival rate increases is given in [3] as follows:

$$P_{delay} \equiv \frac{1}{1 + \sqrt{2\pi}\beta_\epsilon(1-\epsilon)e^{\frac{1}{2}\beta_\epsilon^2}} \tag{4.7}$$

The dependence on λ and μ is hidden in the relationship between β_ϵ and ϵ. Example values for ϵ and β_ϵ for various delay probabilities are given in Table 4.1.

To make this model more effective, an adjustment is made to Equation (4.5):

$$s(t) = \lceil m(t) + 0.5 + \beta_\epsilon\sqrt{v(t)} \rceil \tag{4.8}$$

The additional 0.5 is to account for the discreteness of the final server staffing function, i.e. to add a "buffer" server. This insures that systems for which the infinite server approximation might suggest $\lceil 3.01 \rceil$ servers (for example) do not get treated the same as systems for which the calculated number of servers is $\lceil 3.99 \rceil$.

Approximating m(t)

In order to make use of Equation (4.8), the mean $m(t)$ and variance $v(t)$ of the number of busy servers must be determined. If we assume an $M_t/M_t/s_t$ system, the discussion can be limited to $m(t)$ because as Jennings, Mandelbaum, Massey and Whitt point out, $v(t) \approx m(t)$ in such a system.

The steady state mean number of busy servers in a $G_t/G_t/\infty$ queue is:

$$m(t) = \int_0^t (1 - G_u(t-u))\lambda(u)du \tag{4.9}$$

where $G_u(t)$ is the cumulative distribution of the service time of an arrival at time u ([6]). There are several approximations and assumptions we can employ to simplify the expression for $m(t)$. For example, if $G_u(t)$ is independent of u (i.e. the service process is stationary) Equation (4.9) reduces to

$$m(t) = E[\int_{t-S}^t \lambda(u)du] = E[\lambda(t - S_e)]E[S] \tag{4.10}$$

where S is the service time and S_e is a *service time stationary excess random variable*:

$$P(S_e \leq t) = \frac{\int_0^t P(S > u)du}{E[S]}. \tag{4.11}$$

A logical approximation for systems where the service time changes slowly with respect to arrival rate is shown in Equation (4.12):

$$m(t) = E[\lambda(t - S_e(t))]E[S(t)] \tag{4.12}$$

where the excess service time S_e may be time-dependent. The above approximations are based on the idea that the average number of busy servers at time t will be approximately the product of the expected instantaneous service time and the expected arrival rate at a time t' prior to t. The time difference $t - t'$ is such that the average job arriving at time $t - t'$ will still be in the system at time t. This results in what

Jennings, Mandelbaum, Massey and Whitt refer to as "smoothing" of $m(t)$, so that it is dependent upon the arrival and service rates during the time immediately prior to t.

Jennings, Mandelbaum, Massey and Whitt recognize that these approximations are not useful for efficient computation, as they must be recomputed at each interval. They note that assuming some special structure for the arrival rate function may simplify computations. For example if the service time distribution is exponential, the rate of change of the number of busy servers is the difference between the arrival rate and the total rate at which customers are being served;

$$m'(t) = \lambda(t) - m(t)\mu(t). \tag{4.13}$$

In addition, the quadratic approximations in [1] are cited:

$$m(t) \approx \lambda(t - E[S_e(t)]) * E[S(t)] + 0.5\lambda''(t) * Var[S_e(t)] * E[S(t)]. \tag{4.14}$$

The reasoning behind this approximation is similar to that of Equation (4.12). However, it is simpler to calculate because there are no complicated functions of random variables. Equation (4.14) introduces a time lag $E[S_e(t)]$ and a space shift (the λ'' term). It can be further simplified to the pointwise stationary approximation (PSA2):

$$m(t) \approx \lambda(t) * E[S(t)] \tag{4.15}$$

This is the simplest form of Equation (4.10), in which each part of the integrand is evaluated separately.

This approximation is not to be confused with the pointwise stationary approximation discussed in Section 4.1.1. Equation (4.15) provides an estimate of the mean number of busy servers in the $G_t/G_t/\infty$ queue based on its instantaneous $G/G/\infty$ counterpart. The $G_t/G_t/\infty$ queue is then used as an infinite server approximation to the $G_t/G_t/s_t$ queue. Thus, using the $M_t/M_t/s_t$ queue as an example, the number of

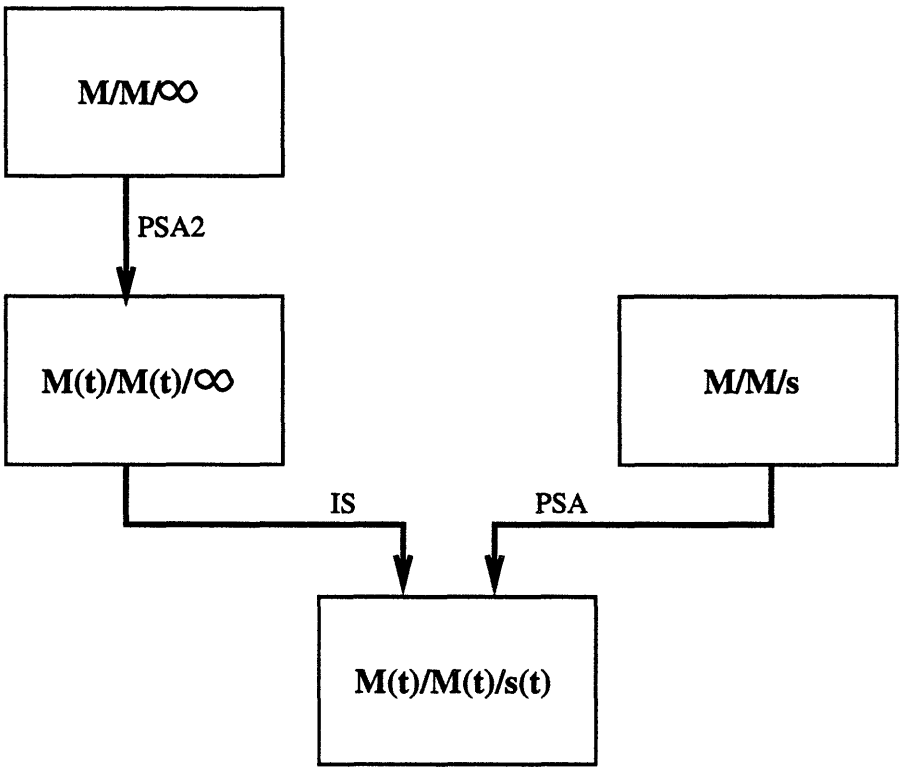

Figure 4-3: Comparison of PSA and PSA2

servers is calculated according to the following:

$$s(t) = \lceil \frac{\lambda(t)}{\mu(t)} + 0.5 + \beta_\epsilon \sqrt{\frac{\lambda(t)}{\mu(t)}} \rceil \tag{4.16}$$

where given a target delay probability, β_ϵ is defined according to Equations (4.6) and (4.7). By contrast, the use of PSA in Section 4.1.1 approximates the $G_t/G_t/s_t$ queue by its $G/G/s$ counterpart, treating it as if it were in steady state at each time interval. Using this for the $M_t/M_t/s_t$ queue, one would choose the appropriate number of servers so that Equation (4.1), with P_{delay} set to a target delay probability, is satisfied. Figure 4-3 shows a graphical comparison between the two methods.

4.2 Application

In this section I will describe the application of the JMMW method to First USA's system. The specific assumptions and approximations will be stated, as well as the reasoning behind them.

I model the system at hand as an $M_t/M_t/s_t$ queue, where all servers are fed from the same queue. I do not expect the PSA and SSA methods of analysis to be effective in this problem because in a typical day, the arrival rate may change from 0.60 calls/minute ($\approx$ 1 call every 2 minutes) to 36 calls/minute within 6 hours, while the service rate may change from 6.6 calls/minute to 7.2 calls/minute within the same time frame. The relative change in arrival rate seems too small to be supported by PSA and too large to be supported by SSA. Since my model for First USA includes a Poisson arrival process and exponentially distributed service times, I will assume $v(t) = m(t)$. For simplicity I also start with the PSA2 approximation for $m(t)$. Thus, the combined formula for $s(t)$ according to my application of the JMMW method is as described in Equation (4.16), where β_ϵ is calculated from Equations (4.6) and (4.7) given a target delay probability. These key equations are reproduced below:

$$s(t) = \lceil (\frac{\lambda(t)}{\mu(t)}) + 0.5 + \beta_\epsilon \sqrt{\frac{\lambda(t)}{\mu(t)}} \rceil \tag{4.17}$$

$$P(\mathcal{N}(0,1) > \beta_\epsilon) = \epsilon \tag{4.18}$$

$$P_{delay} \equiv \frac{1}{1 + \sqrt{2\pi}\beta_\epsilon(1-\epsilon)e^{\frac{1}{2}\beta_\epsilon^2}} \tag{4.19}$$

For First USA, the target probability of delay would likely range from 0.05 to 0.25, generating values of β_ϵ from 1.740 to 0.936.

4.3 Results

Appendix B shows the results of applying the above formulas to the forecast data of two randomly chosen days at First USA. I used a target P_{delay} of 0.154; this is the average planned P_{delay} over several days surrounding the example dates according to the Cybernetics solutions. The results are very close to the server staffing levels computed by Cybernetics, and nearly identical to the results of using the pointwise stationary approximation (see Tables B.2 and A.1). These traits were consistent throughout all the days for which server staffing levels were calculated.

The similarity with Cybernetics' results is of interest because the planned P_{delay} (calculated from Cybernetics' solutions and forecasted arrival and service rates) over the course of a day ranges from 0.05 to 0.25, while the target P_{delay} for the JMMW method was a constant 0.154. This demonstrates that the server staffing solution is insensitive to small changes in target P_{delay}, because the number of operators staffed must take on discrete integer values. The effect of this quantization is more noticeable where the system (number of servers employed) is smaller.

The similarity between results using this application of the JMMW method and those using PSA suggest that a pointwise stationary approximation of an infinite-server approximation of the $M_t/M_t/s_t$ queue does not perform much differently from a pointwise approximation of the queue by its $M/M/s$ counterpart. For general interarrival and service time distributions the results may not be so similar; this is a topic for possible future research.

Chapter 5

The SPA Method

The *Stationary Process Approximation* (SPA) method developed by Massey and Whitt in [5] is a method of analyzing nonstationary Erlang loss models to measure various aspects such as blocking probability and congestion. In this chapter I will present the method and explain how it can be used to solve First USA's problem. The solution itself will be described in Section 5.1. Section 5.2 will discuss the assumptions and approximations necessary for its application to First USA's problem, and the results of such application will be presented in Section 5.3.

5.1 Methodology

The goal of the SPA method is to properly characterize a loss system so as to effectively approximate its average performance measures over discrete intervals of time. Specifically, Massey and Whitt consider the average blocking probability in the $M_t/G/s_t/0$ queue. Recall that this queue has a nonstationary Poisson arrival process, a general stationary service time distribution, s_t servers and no extra waiting room. The findings offered in [5] can be extended to systems with general arrival rates and nonzero capacity queues.

A natural approach to approximating the $M_t/G/s_t/0$ system would be to analyze

it over each time interval as if it were stationary. For example, over the interval $(0, T]$, one would estimate the blocking probability by calculating it for a stationary $M/G/\bar{s}$ model with arrival rate $\bar{\lambda}$ and number of servers $\bar{s}$ equal to their time-averaged values over the given interval:

$$\bar{\lambda} = \frac{1}{T}\int_0^T \lambda(u)du \tag{5.1}$$

$$\bar{s} = \frac{1}{T}\int_0^T s(u)du \tag{5.2}$$

While simple and straightforward, this method ignores the variability caused by time-dependence, thus underestimating the blocking probability. Massey and Whitt suggest using a G/G/s/0 model instead of the stationary M/G/s/0 model mentioned above. The service process and the number of servers would remain the same, but extra stochastic variablility would be introduced into the arrival process. The extra variability is based on time fluctuations in the arrival rate and is characterized by the *heavy traffic peakedness* (z). The peakedness of a system is a means of measuring the system's congestion. It is formally defined as the ratio of the variance to the mean of the steady state number of customers in the associated infinite server model. In [5], Whitt and Massey give the the limiting behavior of the peakedness as the arrival rate grows with respect to the service rate as follows.

$$z = 1 + \mu(c_a^2 - 1)\int_0^\infty [1 - G(t)]^2 dt \tag{5.3}$$

The heavy traffic peakedness (z) is a function of the service rate (μ), the cumulative distribution function (cdf) of the service time $(G(t))$, and the squared coefficient of variation c_a^2 of the arrival process:

$$c_a^2 = \lim_{t\to\infty} \frac{VarA(t)}{E^2[A(t)]} \tag{5.4}$$

If the service times are exponentially distributed, then the integral in Equa-

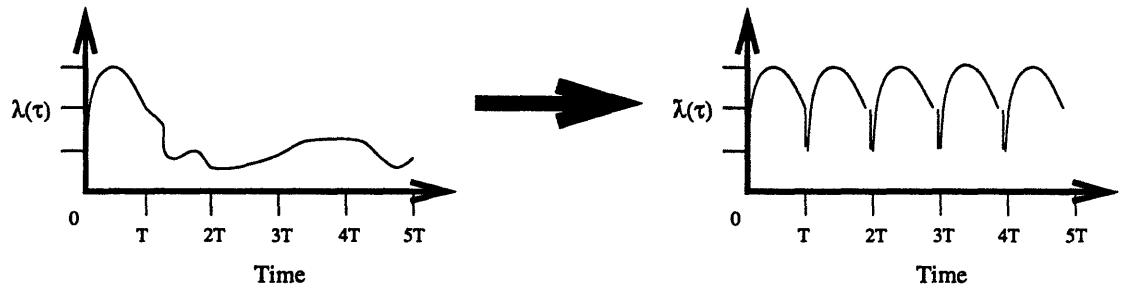

Figure 5-1: Periodic extension of the arrival rate function over the interval (0, T]

tion (5.3) reduces to $\frac{1}{2\mu}$, and the heavy traffic peakedness is:

$$z = 1 + \frac{(c_a^2 - 1)}{2} \tag{5.5}$$

If the interarrival times were also exponentially distributed, c_a^2 (and therefore z) would reduce to unity.

The service rate, service time cdf and squared coefficient of variation are needed to determine the heavy traffic peakedness. It is assumed that the first two are known; the latter can be found as follows. Modify the arrival process by forcing the arrival rate function to repeat itself periodically over intervals of length T. This is depicted in Figure 5-1. (I am assuming that the interval that we are working with is from time 0 to time T.) Create a stationary point process $N(t)$ where each increment corresponds to an event in the altered arrival process. The *index of dispersion for counts* is the ratio of the variance and the expected value of $N(t)$:

$$I(t) = \frac{VarN(t)}{\bar{\lambda}t} = \frac{E[N(t)^2 - (\bar{\lambda}t)^2}{\bar{\lambda}t} \tag{5.6}$$

The second moment of $N(t)$ can be written as:

$$E[N(t)^2] = \frac{1}{T}\int_0^T [\Lambda_t(s) + \Lambda_t(s)^2]ds \tag{5.7}$$

where for all $(0 \leq s \leq T-t)$,

$$\Lambda_t(s) = \int_s^{s+t} \lambda(u)du \tag{5.8}$$

and for all $(T-t \leq s \leq T)$,

$$\Lambda_t(s) = \int_s^{T} \lambda(u)du + \int_0^{s-T+t} \lambda(u)du. \tag{5.9}$$

Massey and Whitt suggest approximating the squared coefficient of variation by I(E[S]), since the mean service time indicates the time scale of interest. If the arrival rate is assumed to be linear over the interval under consideration, Equations (5.6) through (5.9) eventually result in $\Lambda_t(s)$ as follows:

$$\Lambda_t(s) \approx (a+rs)t \qquad \forall(0 \leq s \leq T) \tag{5.10}$$

and after much calculation,

$$c_a^2 \equiv I[E(S)] \approx 1 + \frac{r^2T^2E(S)}{6(2a+rT)} \tag{5.11}$$

It remains to incorporate the extra variability into the blocking probability calculations. The Hayward approximation suggests using the Erlang blocking formula for $M/M/s/0$ systems, replacing the values of $\bar{s}$ and $\frac{\bar{\lambda}}{\mu}$ with $\frac{\bar{s}}{z}$ and $\frac{\bar{\lambda}}{z\mu}$, respectively. The Erlang Blocking formula is as follows:

$$P_{blocking} = \frac{(\frac{\bar{\lambda}}{\mu})^s(\frac{1}{s!})}{\sum_{k=0}^{s}(\frac{\bar{\lambda}}{\mu}^k)(\frac{1}{k!})}. \tag{5.12}$$

5.2 Application

In this section I will discuss the assumptions and approximations necessary to apply the SPA method to First USA's system. First, the approximation must be extended

to include delay models. As discussed in Chapter 2, the probability of delay in a G/G/s system can be approximated by an infinite server normal approximation [7]:

$$P_{delay} \approx 1 - \Phi[\frac{s - \rho s}{\sqrt{z\rho s}}] \tag{5.13}$$

Recall that $\rho = \frac{\lambda}{s\mu}$. I use Equation (5.13) to numerically determine the number of servers as a function of time for a given target P_{delay}. The value of λ for any given interval will be the average arrival rate over that interval, and μ will be the average service time over that interval. The natural interval length to use is 30 minutes, since most of First USA's statistical data is available in half-hour intervals. This is consistent with the provisions of the work by Massey and Whitt [5], which requires the size of the interval to be between six and twenty times the average service length. At First USA, the average service time is approximately three minutes, so the ratio of interval length to average service time is approximately ten.

I also assume for this method that the arrival rate is piecewise linear, instead of piecewise constant, as is assumed at First USA. The reasoning behind this is that perhaps an arrival rate function which is everywhere connected can more accurately model the real system. The SPA method easily accomodates piecewise linear arrival rate functions, as seen in Section 5.1. In order to derive a piecewise linear function from the discrete arrival rates measured at at First USA, I assumed the arrival rate function is everywhere connected and that it is constant during the interval where the average arrival rate is at a minumum (this would be approximately 4:00am each day). I then extrapolated backwards from that time to the beginning of the day and forward to the end of the day, using the average arrival rates as midpoints to determine a linear function for each interval. To illustrate, Figure 5-2 shows the average forecast arrival rates and the resulting extrapolated arrival rate function for a sample day. The function used was not as "smooth" as I had expected, however it was not volatile enough to significantly affect the heavy traffic peakedness, as results will show.

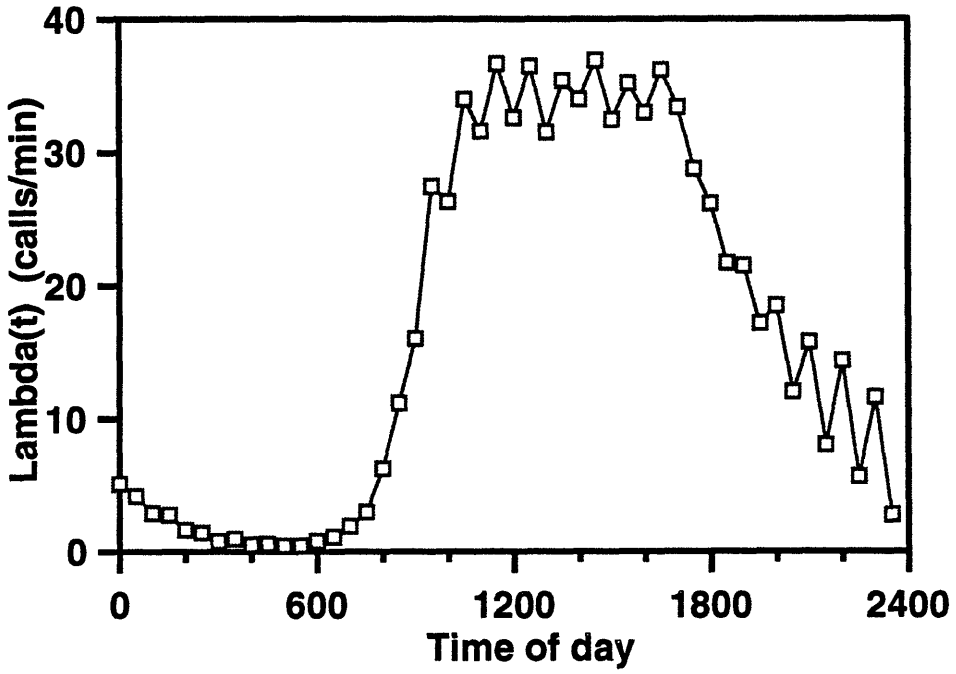

Figure 5-2: Extrapolated arrival rate function for use with the SPA method.

5.3 Results

Equation (5.13) can be used to numerically determine the optimal number of servers as a function of time, given a target P_{delay}. Refer to Appendix C for server calculations for two randomly chosen days, using 0.154 as a target delay probability. These staffing levels are generally 1 to 2 servers less than those levels calculated using the JMMW method. At first glance, this seems to be a rather surprising fact since both JMMW calculations and SPA calculations are carried out on an interval basis using an infininte server normal approximation. In fact, SPA is designed to assume additional variability in the arrival process, which means more servers would be needed to attain the same delay probability. There are several reasons why this does not hold true here.

First, as pertains to this application, the additional variability introduced into the arrival process is negligible. This fact can be seen most clearly in the calculated values for z, the heavy traffic peakedness. For most intervals, $1.01 < z < 1.05$. This shows

numerical proof that little variability is gained by assuming $\lambda(t)$ to be linear instead of constant over each interval even though the contrived arrival rate function was more volatile than expected. The result can be attributed to the relative shortness of the intervals within the possible range of lengths, i.e. The arrival rate changes sufficiently slowly that assuming it to be piecewise constant over each of these intervals is not unreasonable.

Another reason why the SPA method generated lower staffing levels is the fact that the JMMW method is not based *entirely* on the infinite server normal approximation. Equation (4.8) includes an extra 0.5 as a "buffer" server. In addition, the SPA method does not assume as much congestion in the system because the original derivation assumes the service process to be stationary. This assumption is not necessary for the JMMW method, which would therefore generate higher staffing levels to accomodate such congestion.

Chapter 6

Conclusion

The purpose of this work was to compare two different methods of analyzing a real-word telephone service system, namely that of First USA Bank. It was necessary to determine if and how each method could be used to solve the operator staffing problem as it relates to First USA.

It was assumed that First USA's agent queue could be approximated using a nonstationary delay model with Poisson arrivals, exponential service times, and a queue of infinite capacity. For the SPA model we also assumed the service time distribution was stationary and that the arrival rate was piecewise linear. For the JMMW method the arrival rate was considered to be piecewise constant instead. performance level.

Calculations showed that my application of the JMMW method results in staffing levels very similar to those calculated by Cybernetics, and almost identical to those calculated using the pointwise stationary approximation. The SPA method results in slightly lower staffing levels. The similarity between the results of these methods suggests that for a nonstationary $M_t/M_t/s_t$ queue any of these aproximations will perform equally well. The proven inconsistency of Cybernetics' performance, however, suggests that there is indeed room for improvement. It may be more accurate to model First USA's agent queue as a $G_t/M_t/s_t$ system, or as a $G_t/G_t/s_t$ system. As discussed

earlier, literature does not yet provide any exact solutions for these systems. One possibility may be to apply the JMMW method without assuming equality between the mean $m(t)$ and variance $v(t)$ of the average number of busy servers. One could also choose a different approximation for $m(t)$. As stated earlier, this is a possible subject for further research.

Appendix A

Analysis of First USA's System

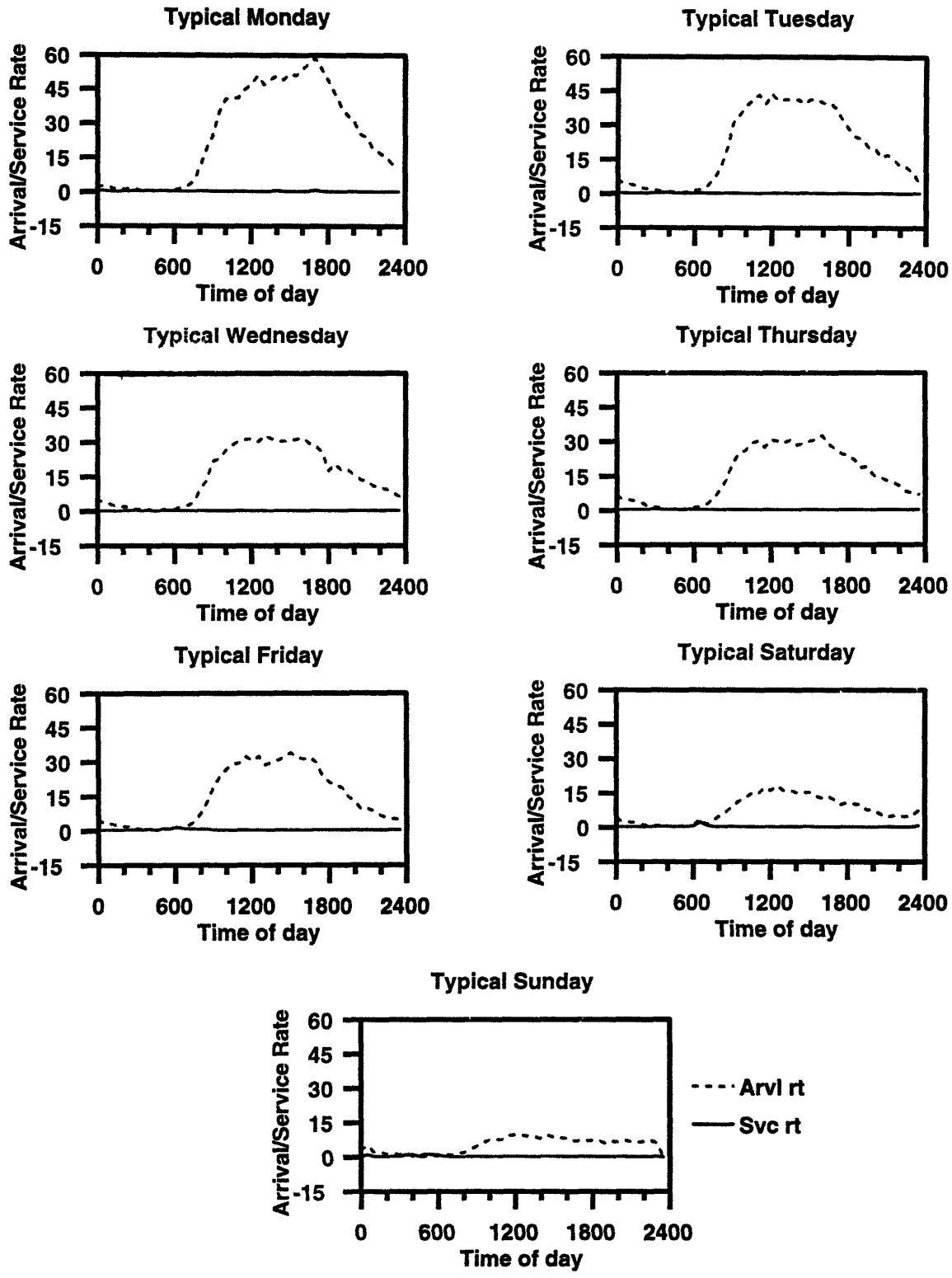

Figure A-1: Average trajectories of call arrival and service rates for the agent queue (calls/min)

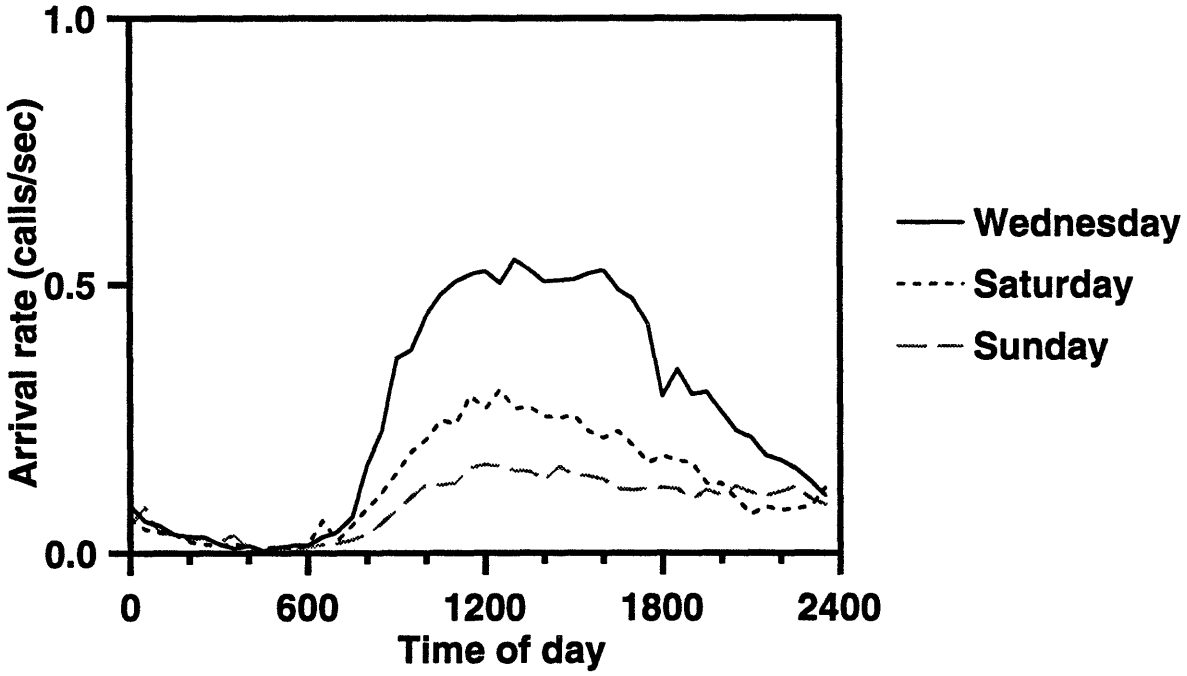

Figure A-2: Comparison of call arrival rates for an average Wednesday, Saturday and Sunday

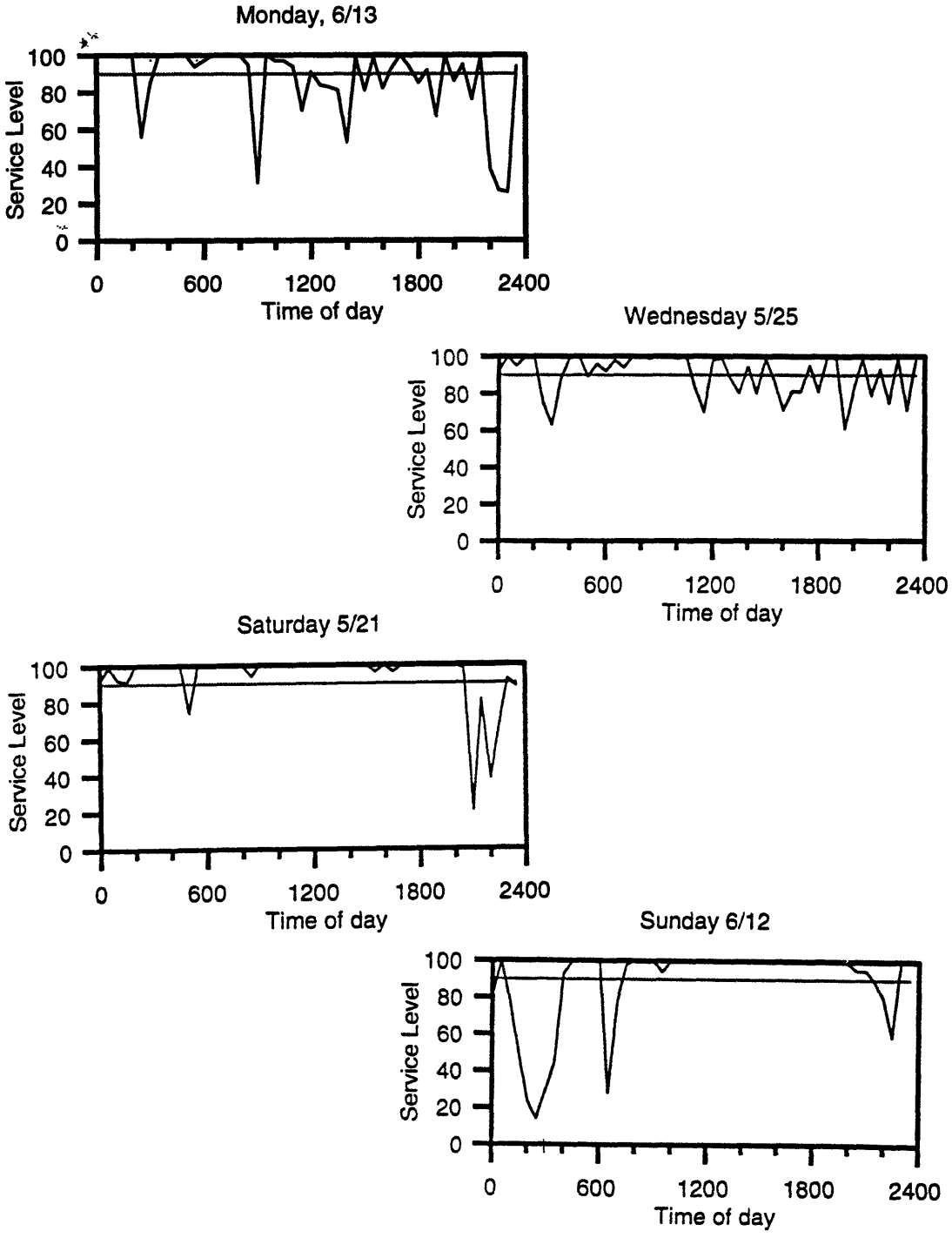

Figure A-3: Trajectory of the Service Level for several randomly selected days (target = 90%)

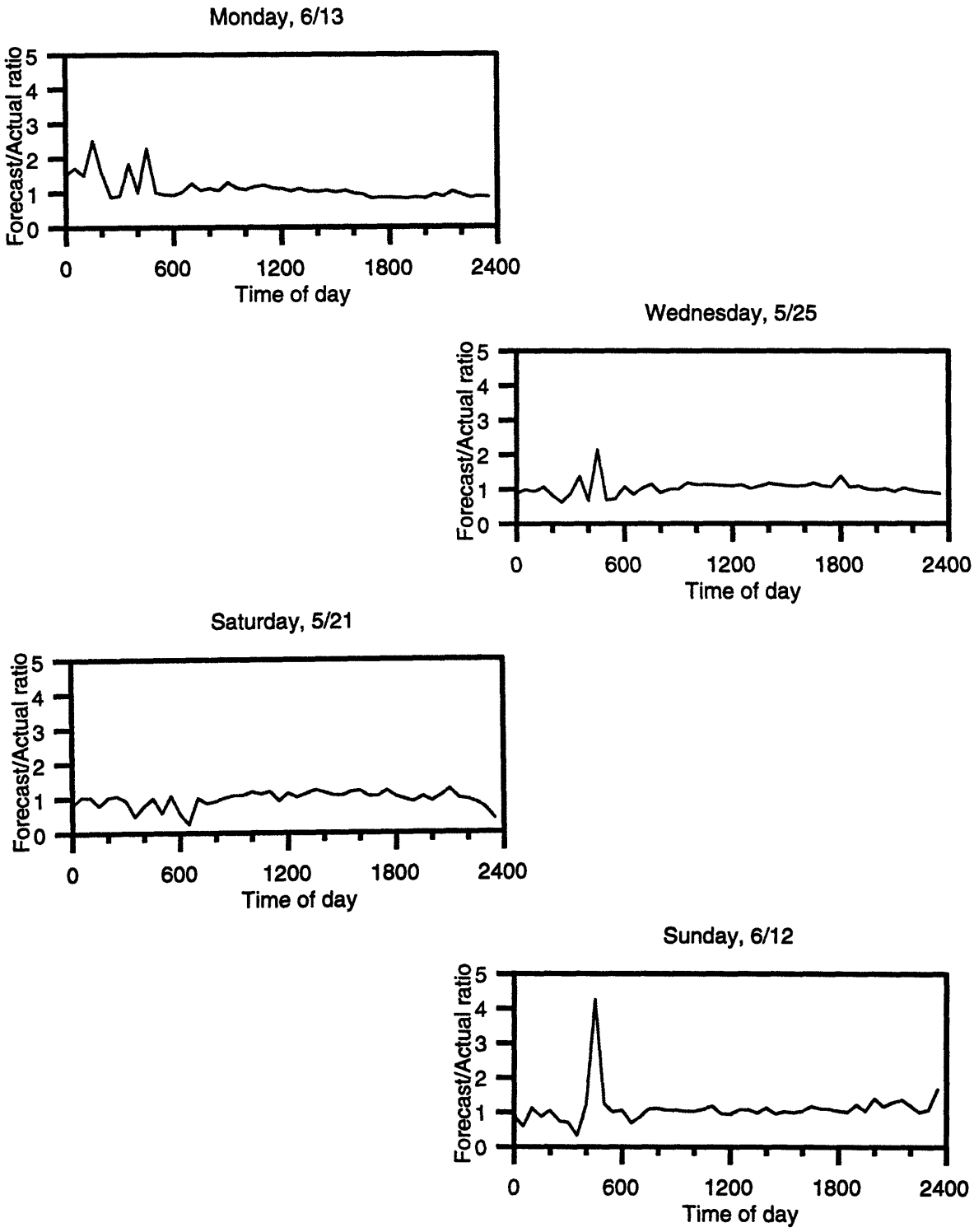

Figure A-4: Call arrival forecast performance, measured by the ratio of forecast to actual arrival rate (target = 1)

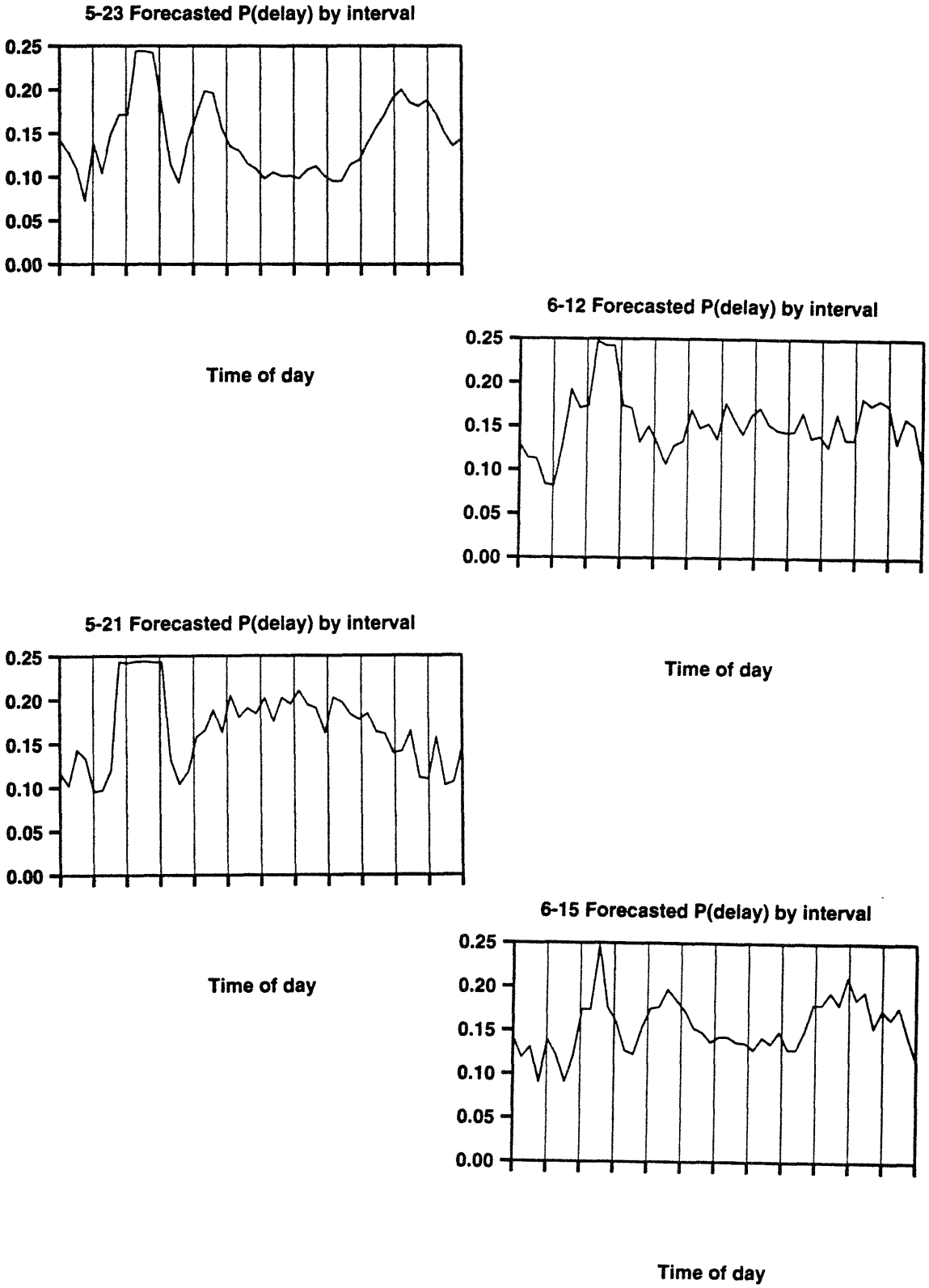

Figure A-5: Trajectory of "planned" probability of delay according to computations by Cybernetics.

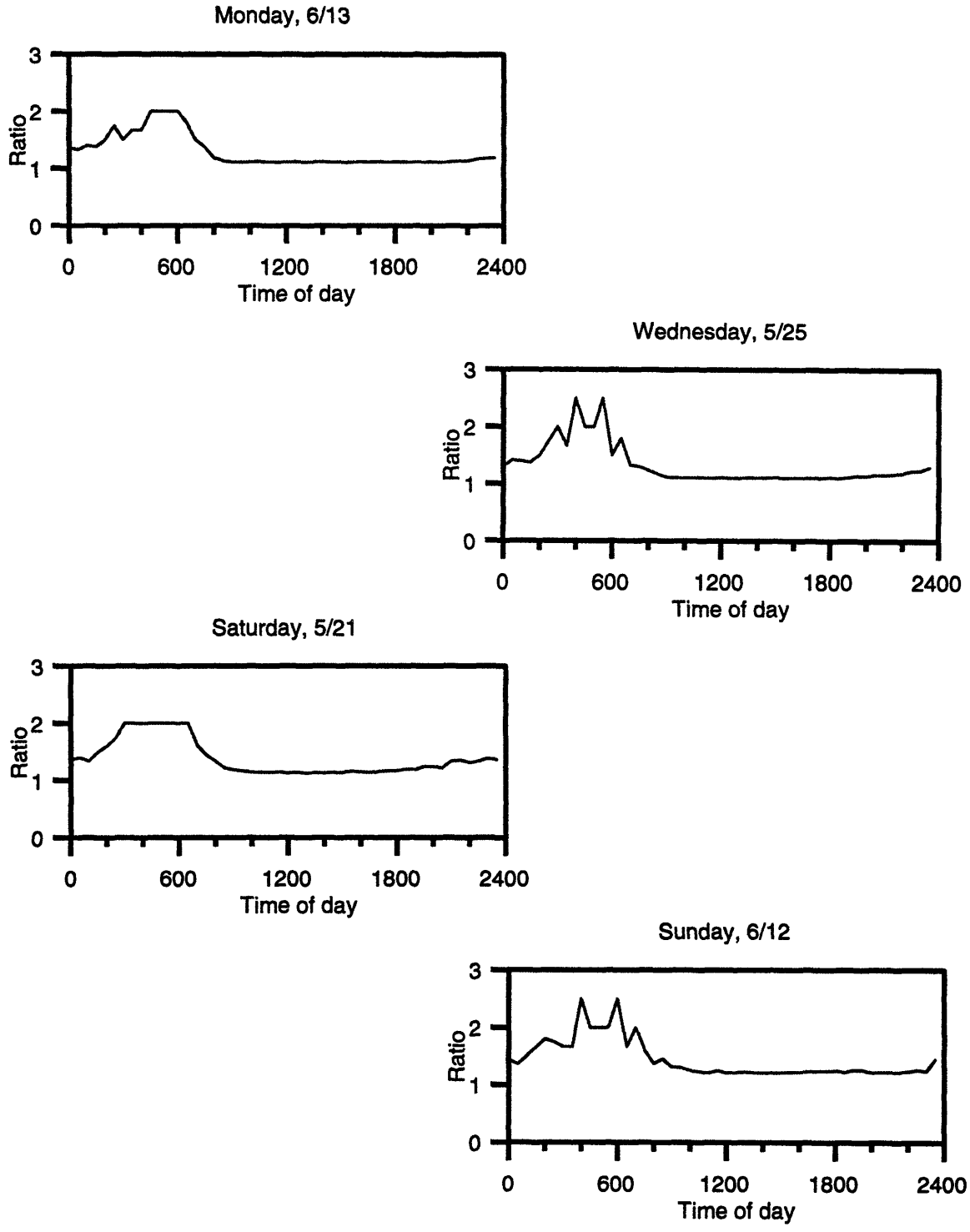

Figure A-6: Ratio of Cybernetics required agents to number of servers according to a simple calculation: $s = \frac{\lambda}{\mu}$.

	Cybernetics	M/M/s
INTVL	Req Agents	Appx'n
0	19	19
30	16	16
100	14	14
130	11	11
200	9	9
230	7	7
300	6	6
330	5	5
400	5	5
430	4	4
500	4	4
530	4	4
600	6	6
630	9	9
700	15	15
730	22	22
800	45	46
830	69	71
900	112	114
930	141	141
1000	162	161
1030	176	174
1100	181	178
1130	184	182
1200	192	189
1230	198	195
1300	195	192
1330	198	195
1400	201	197
1430	204	200
1500	198	194
1530	198	195
1600	204	200
1630	210	206
1700	195	191
1730	181	178
1800	168	166
1830	148	148
1900	121	122
1930	111	112
2000	108	109
2030	101	103
2100	89	91
2130	80	82
2200	68	70
2230	55	56
2300	46	47
2330	38	39

Table A.1: Comparison of Cybernetics required agents with number of servers calculated as if the system were $M/M/s$. (Monday, 6/13)

Appendix B

Results Using the JMMW Method

This appendix contains results of the application of the JMMW method to the data from two randomly chosen days at First USA. It is referred to in Chapter 4.

	Cynerbetics	JMMW
INTVL	Req Agents	Req Agents
0	13	13
30	15	15
100	12	12
130	10	10
200	9	9
230	7	7
300	5	5
330	5	5
400	5	4
430	4	4
500	4	4
530	4	4
600	5	4
630	5	5
700	6	6
730	8	8
800	11	11
830	16	16
900	21	21
930	26	26
1000	30	31
1030	32	32
1100	35	36
1130	35	35
1200	35	36
1230	39	40
1300	38	38
1330	34	35
1400	35	36
1430	35	36
1500	34	34
1530	33	33
1600	33	33
1630	32	33
1700	31	31
1730	31	31
1800	30	30
1830	28	29
1900	29	29
1930	29	29
2000	34	35
2030	33	34
2100	33	34
2130	33	34
2200	32	32
2230	29	30
2300	26	27
2330	13	13

Table B.1: Server staffing levels calculated using the JMMW Method for Sunday, 6/12/94

	Cybernetics	JMMW
INTVL	Req Agents	Req Agents
0	19	19
30	16	16
100	14	14
130	11	11
200	9	9
230	7	7
300	6	6
330	5	5
400	5	5
430	4	4
500	4	4
530	4	4
600	6	6
630	9	9
700	15	15
730	22	22
800	45	46
830	69	71
900	112	114
930	141	141
1000	162	161
1030	176	174
1100	181	178
1130	184	182
1200	192	189
1230	198	195
1300	195	192
1330	198	195
1400	201	197
1430	204	200
1500	198	194
1530	198	195
1600	204	200
1630	210	206
1700	195	191
1730	181	178
1800	168	167
1830	148	148
1900	121	122
1930	111	112
2000	108	109
2030	101	103
2100	89	91
2130	80	82
2200	68	70
2230	55	56
2300	46	47
2330	38	39

Table B.2: Server staffing levels calculated using the JMMW Method for Monday, 6/13/94

Appendix C

Results Using the SPA Method

This appendix contains results of the application of the SPA method to the data from three randomly chosen days at First USA. It is referred to in Chapters 5 and 6.

	Cynerbetics	SPA
INTVL	Req Agents	Req Agents
0	13	12
30	15	14
100	12	11
130	10	9
200	9	8
230	7	6
300	5	5
330	5	4
400	5	4
430	4	4
500	4	3
530	4	3
600	5	4
630	5	4
700	6	5
730	8	8
800	11	10
830	16	15
900	21	20
930	26	25
1000	30	29
1030	32	31
1100	35	34
1130	35	34
1200	35	35
1230	39	38
1300	38	37
1330	34	33
1400	35	35
1430	35	34
1500	34	33
1530	33	32
1600	33	32
1630	32	32
1700	31	30
1730	31	30
1800	30	29
1830	28	27
1900	29	28
1930	29	28
2000	34	35
2030	33	34
2100	33	34
2130	33	34
2200	32	32
2230	29	30
2300	26	26
2330	13	17

Table C.1: Server staffing levels calculated using the SPA Method for Sunday, 6/12/94

	Cybernetics	SPA
INTVL	Req Agents	Req Agents
0	19	18
30	16	15
100	14	13
130	11	10
200	9	8
230	7	6
300	6	5
330	5	4
400	5	4
430	4	3
500	4	3
530	4	4
600	6	5
630	9	8
700	15	14
730	22	21
800	45	48
830	69	69
900	112	119
930	141	139
1000	162	162
1030	176	172
1100	181	179
1130	184	182
1200	192	190
1230	198	195
1300	195	192
1330	198	196
1400	201	199
1430	204	202
1500	198	196
1530	198	195
1600	204	199
1630	210	204
1700	195	188
1730	181	177
1800	168	164
1830	148	147
1900	121	119
1930	111	110
2000	108	107
2030	101	101
2100	89	90
2130	80	80
2200	68	70
2230	55	55
2300	46	46
2330	38	37

Table C.2: Server staffing levels calculated using the SPA Method for Monday, 6/13/94

Bibliography

[1] S. Eick, W. Massey, and W. Whitt. The physics of the $m_t/g/\infty$ queue. *Operations Research*, 41, 1993.

[2] W.K. Grassmann. Finding the right number of servers in real-world queuing systems. *Interfaces*, 18(2):94–104, 1988.

[3] O. Jennings, A. Mandelbaum, W. Massey, and W. Whitt. Server staffing to meet nonstationary demand. Technical report, AT&T Bell Laboratories, 1994.

[4] Leonard Kleinrock. *Queuing Systems*, volume 1. John Wiley and Sons, Inc., New York, 1975.

[5] W. Massey and W. Whitt. Stationary process approximations for the nonstationary erlang loss model. Technical report, AT&T Bell Laboratories, 1992.

[6] W. Massey and W. Whitt. Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems*, 13, 1993.

[7] W. Whitt. Understanding the efficiency of multi-server service systems. *Management Science*, 38, 1992.