ALTERNATIVE METHODS OF INVESTIGATING THE TIME DEPENDENT M/G/k QUEUE

by

PEETER A. KIVESTU

B.S. BROWN UNIVERSITY

(1974)

SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September, 1976

Signature of Author _____
Department of Aeronautics and Astronautics
August 16, 1976

Certified by _____
Thesis Supervisor

Received by _____
Chairman, Departmental Graduate Committee

ARCHIVES

MASS. INST. TECH.
OCT 13 1976
LIBRARIES

ALTERNATIVE METHODS OF INVESTIGATING THE TIME DEPENDENT M/G/k QUEUE

by

PEETER A. KIVESTU

Submitted to the Department of Aeronautics and Astronautics on August 16, 1976, in partial fulfillment of the requirements for the degree of Master of Science.

ABSTRACT

The time dependent M/G/k queue is studied with the aim of obtaining good numerical approximations and descriptors of system behavior rather than exact closed form solutions. Five models that have been used to investigate this problem are presented: Simulation; first order models: "fluid approximation" and equilibrium analysis; second order models: "diffusion approximation" and Koopman's model. The assumptions used in postulating these models and their consequences are evaluated. The impracticality of direct numerical solution is reviewed.

The second order models are investigated in detail. From the diffusion approximation information about the transient behavior of stationary M/G/1 queues is obtained. Exact closed form expressions for the transient state probabilities of the stationary M/M/1 queue (Morse) are given and the time constants for this system derived. The exact value of the time constant of the M/M/1 system is compared to the corresponding result from the approximating diffusion model. The general properties of the transient behavior of the M/G/1 queue are discussed. An application where knowledge of these time constants is imperative is given in a model due to Gupta. Also, a new model for solving the time dependent M/M/1 queue, using closed form expressions for the transient behavior of the M/M/1 queue, is presented.

The models are evaluated and the importance of the time constant is discussed in the context of a case study on airport runway queuing systems. Special emphasis is given to explaining the reasons for the success of Koopman's model. Other numerical results for specific cases of the transient behavior of various M/G/k queues are provided to further describe the time constants and supplement theoretical results.

Thesis Supervisor: Amedeo R. Odoni

Title: Associate Professor of Aeronautics and Astronautics

ACKNOWLEDGEMENTS

The author wishes to express his sincere gratitude to Professor Amedeo R. Odoni, who devoted his time unfailingly, giving constructive advice in countless discussions, and reading several drafts of this thesis; to the Flight Transportation Laboratory and its director, Professor Robert W. Simpson, without whose support this work would not have been possible; and to Patricia Nero, Anne Clee and Steven Breitstein, all of whom exhibited great patience in doing an excellent typing job.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

This thesis investigates a service facility with a strongly time varying Poisson arrival rate and service times governed by some general pdf. This is referred to in Queueing Theory as the M/G/k queueing system. The simultaneously probabilistic and time-dependent nature of the arrival process, as well as the "general" character of the service process render this problem impossible to solve by analytical means. The aim here, therefore, is to obtain good approximations and descriptors of system behaviour rather than exact closed form solutions.

The time-dependent M/G/k queue is studied here in the context of airport runway usage. It is clear that this is a situation where there are very large social and economic costs associated with providing either excess or less than adequate capacity. Either large tracts of land occupied by underutilized runways or long queues of aircraft in the air waiting landing clearance are costly, as well as socially unacceptable conditions.

We are concerned here with the busiest of airports - handling large number of scheduled aircraft every day of the year. The time dependency of the demand in these airports extends beyond the presence of morning and evening (or similar) travel peaks common to all transportation systems. Weekly and seasonal demand variations are also very noticeable and extremely important. Further time-dependence is also introduced by the airport authorities themselves who may schedule additional capacity during traffic peaks, by utilizing more runways than during off-peak periods. Weather, as well, in the case of airports, has a highly varying effect on capacity. Consequently it is evident that no single queueing analysis of any airport

system can be expected to model this situation accurately except for a fraction of the time. The need for efficient computational methods is therefore apparent.

There are two common criteria for judging the level of service provided at a facility, both of which are extensively used by aviation authorities. These are the average waiting time per aircraft (clearly as a function of the time of day) and the fraction of aircraft delayed for an amount of time greater than some time t_0, (also as a function of time of day). The first of these is the readily available from existing queueing models and in addition forms a base for many other frequently used statistics - total daily delay, annual delay and annual delay cost. Chapter 2 therefore discusses the numerous approaches that have been suggested for determining the expected waiting time in a time-dependent M/G/k queue. In approximately their historical order of appearance, 5 models are presented. The first is the traditional approach of simulation. The remaining four models are obtained by relaxing various conditions of the time dependent M/G/k queueing system. Of these, the first two (Newell [11]) are first order models in the sense that they either ignore the probalistic nature of both the arrival and service processes or their time-dependence. Second order models, (Gaver [1], Koopman [8]) to these are constructed by including both of these conditions but making other assumptions on their probabilistic characteristics. With increased accuracy, however, comes a substantial increase in computational effort required.

As these latter two models are developed the concept of the transient behaviour of the expected waiting time becomes important as a unifying concept between them. This is because the strong time dependence of the utilization ratio ρ (ratio of demand to service capacity) rarely allows the system to remain in equilibrium. Chapter 3 therefore addresses the problem of seeking the descriptors of system behaviour (in our case the expected waiting time) when the system is not in equilibrium.

We start by presenting exact models (Morse [9], [10]) for the transient behaviour of both the finite and infinite queue capacity M/M/1 systems. When the forms of the two results are compared we conclude that, as expected, large finite capacity queues behave in the transient state not much differently from infinite capacity queues. The major outcome of this analysis is a single time constant for the M/M/1 system that is valid for all values of $\rho \geq 0$.

At this point in the analysis, we recall the approximating model of the M/G/1 queue given by Gaver and presented in Chapter 2. The time constant of the M/M/1 system as determined from this model is shown in fact to coincide with the time constant from the exact model. This allows at least some theoretical justification for accepting the time constants from the approximating M/G/1 model as the true values for the M/G/1 queue. Based on this we then relate the general properties of the relaxation time of the expected waiting time of an M/G/1 queue to its equilibrium value.

Chapter 3 concludes with the presentation of two other models. The first is an approximating model (Gupta [4]) of the time-dependent M/G/k queue that expressly needs analytic forms of the time constant just described. The other model (Clarke [15]) is just an alternate form of the exact transient behaviour model of $M/M/1/\infty$ given by Morse [9]. After we indicate how to modify this model to remain valid for $\rho > 1$ this form becomes just as useful and computationally much more efficient than the model for M/M/1/m presented at the beginning of the chapter.

In Chapter 4 we first give a case study to which we apply three of the models discussed: equilibrium analysis (a first order method), Koopman's model and the last model given in Chapter 3. We discuss the relative merits of the models and the specific applications to which each is likely to be most useful. The transient behaviour of queues is ultimately shown to be responsible for the considerable success of Koopman's method.

Then in the next two sections, we pass to numerical analysis of the transient behaviour of the expected waiting time for single and multiserver queues under specific conditions. For the single server queue we first compare the observed time constant to its exact value. Then the observed relations between the transient behaviour of the expected waiting times of various M/G/1 queues are compared to the behaviour predicted by the approximating M/G/1 model. For the multiserver queue we first give some analytical results (Morse [9]) from which we obtain an approximate evaluation of the time constant for these systems. Again numerical comparisons are provided to verify these. A brief comparison is also

made between the relaxation times of M/M/k and M/D/k systems as k increases from 1.

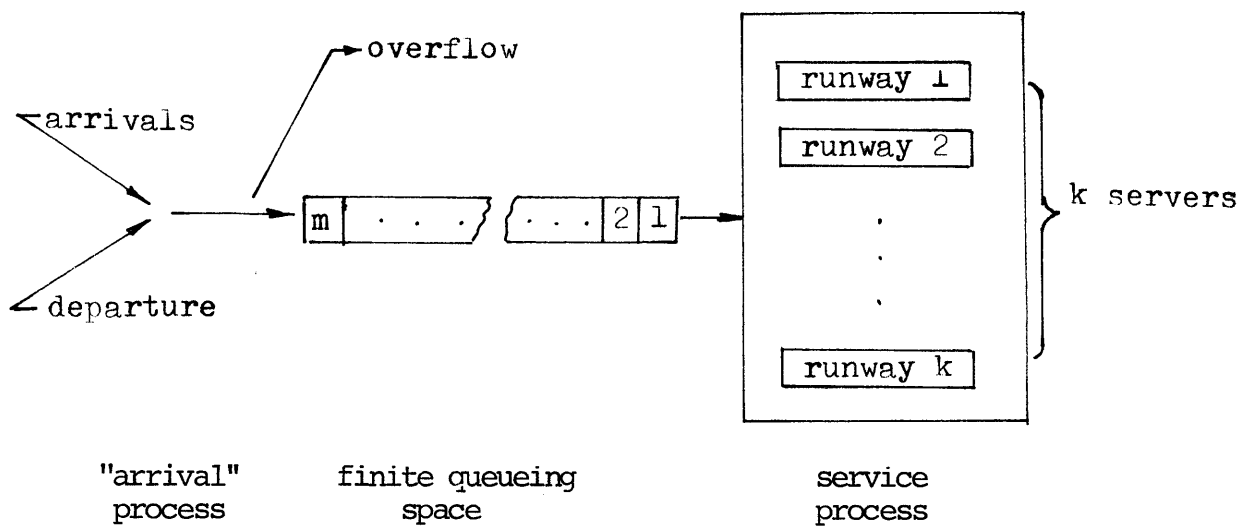
As will be indicated all along, the transient behaviour of queues is a controlling factor in the modelling of time dependent M/G/k queues. The analytic and approximating expressions for the time constant as given in Chapter 3 and the numerical experience of Chapter 4 should provide a sound basis for extending the modelling and numerical analysis of time-dependent M/G/k queues.

Chapter 2

FIVE MODELS FOR THE TIME DEPENDENT M/G/k QUEUE

2.1 Airport Runway Queuing Systems

We illustrate the generic attributes of queuing systems in the context of airport runway use in Figure 2.1



Schematic diagram of airport runway use

Figure 2.1

We define the arrival process with a class of time dependent demand profiles described by the inhomogeneous Poisson pdf with average (time dependent) arrival rate $\lambda(t)$. The Poisson assumption on the arrival process which has been extensively used in the study of airports accounts for the randomization of actual "arrival" times at the server due to

various ATC factors. The time dependence for the larger more congested airports reflects the traffic peaks common to transportation systems and, generally, is cyclic with a period of 24 hours.

In contrast to the arrival process, we will not assume a "neat" probabilistic characterization of the service process. This is motivated by the fact that aircraft service times are neither completely regular (different aircraft have different service times) nor completely random (there do exist standard ATC practices). Also, since the mean rate of service changes infrequently during the course of the day, we assume for simplicity that the service process operates at constant rate $\mu(t) = \mu$. The service process defined by these attributes is called the "general" homogeneous service process.

Throughout the mathematical analysis, whenever the number of servers k (runways) is greater than one, they will be assumed independent and identical. They will operate with a single queue governed by a "first-in, first-out" service discipline. Except in very special cases,[1] this is not a serious misrepresentation of airports.

In summary, we have outlined a time dependent M/G/k/m ($k \leq m \leq \infty$) queuing system in the context of airport operations. All of our subsequent examples will draw from airport situations. The arrival process which is simultaneously probabilistic and time dependent, as well as the "general" character of the service process render this problem impossible to solve by analytical means. The aim, therefore, is to obtain good approximations and

---

1 Such as the case where a certain class of users (e.g., general aviation) commands exclusive use of a particular runway.

descriptors of system behavior rather than exact closed form solutions.

To this purpose, we shall discuss or describe in this and the following chapters a number of mathematical models and approaches. Five of the total eight models will be discussed here in Chapter 2. We will refer to them by number from the list as follows:

Model 1: Computer Simulation, M/G/k

Model 2: Time Dependent "Fluid" Approximation, D/D/k

Model 3: Equilibrium Analysis M/D/k, M/M/k, $M/E_c/1$

Model 4: Stationary Diffusion Approximation, M/G/1

Model 5: Time Dependent Chapman Kolmogorov Equations, M/M/k, M/D/k

The first model is the traditional and only non-analytic approach to the time dependent queuing problem. The second is a deterministic model - it ignores the probabilistic nature of both the arrival and service processes, relying entirely on the time dependency for the estimate of the delay. We contrast this immediately to Model 3, equilibrium analysis, which uses fully the probabilistic characterizations of the queuing system processes but ignores the time dependency. The estimates from Models 2 and 3 will be valuable for both the lower and upper bounds they give for delay as well as for providing the groundwork for more sophisticated models. Models 4 and 5 are the extensions of 2 and 3, respectively. The key to these extensions turns out to be the transient behavior of queues which is the central concern for the models to be discussed in later chapters. (Models 6, 7 and 8 are heavily dependent on transient concepts and hence will be omitted from this chapter.)

2.2 Model 1: Computer Simulation, M/G/k

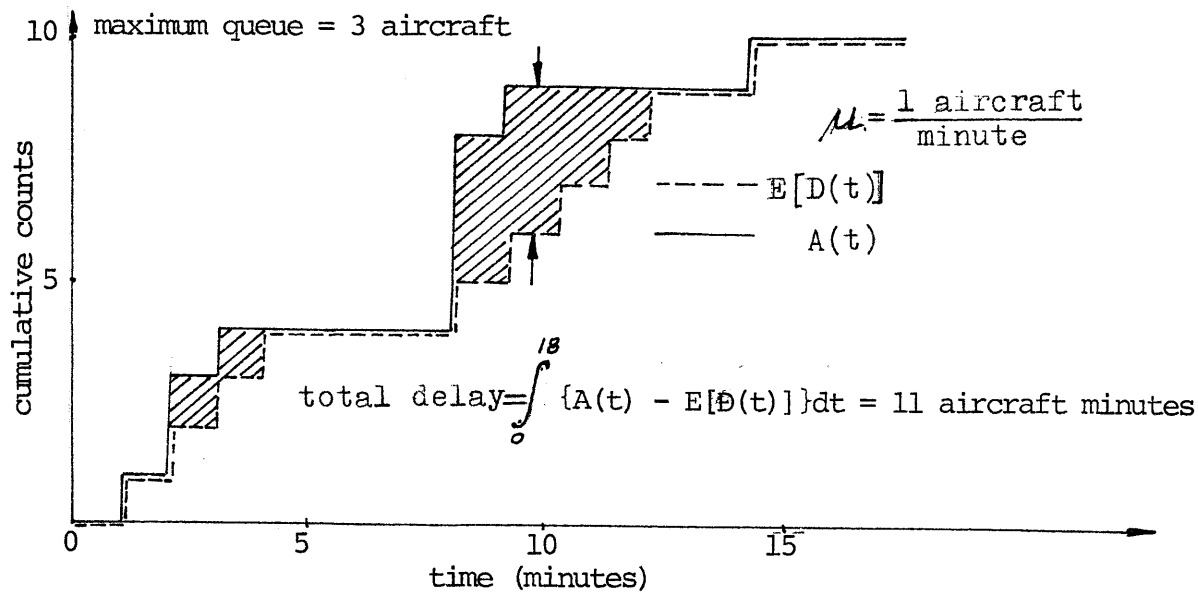
A most accurate representation of a runway queueing system can be obtained through use of a simulation model. The simulation user though pays heavily in terms of computational effort for the additional detail. Each simulation run includes essentially the same computations as those for a single Model 2 estimate, with the addition only of sampling from probability distributions for interarrival and service times. (The assumption of course is that these probability distributions are available and empirically or otherwise verifiable.) For practical purposes, however, because of the sampling, one simulation run provides no more statistical evidence than a single day spent in observing the actual airport. The number of simulation runs needed to provide statistically reliable data must then be large and depends on the stochastic properties of the interarrival and service time probability distributions. In cases where many airport configurations and demand patterns have to be considered this approach is then quite impractical. Moreover, Koopman [8] showed that the number of computations in Monte Carlo simulation increases as the square of the desired statistical precision. This is to be contrasted with an increase which is proportional to the logarithm of the relative precision for the direct solution approach (Model 5).

2.3 Model 2: Deterministic "Fluid" Approximation, D/D/k

We have mentioned that the airport queueing problem exhibits a strongly time dependent arrival rate $\lambda(t)$. Frequently, although not always, this $\lambda(t)$ contains "rush-hours" where $\lambda(t)$ increases to a point exceeding the service rate of the facility and then subsides. In these situations, the analysis of delays has frequently been conducted with the aid of deterministic queuing models of the type D/D/k. Such a model essentially needs two variables, the actual cumulative number of customers entering the queuing system, $A(t)$, and the actual cumulative number of customers leaving the system, $D(t)$. If the interarrival and service times were deterministic then the model would be exact. Since, however, in most real systems the arrival and service processes are stochastic, the "deterministic" approximation is obtained by assuming that the actual number of operations is exactly equal to the expected number of operations, i.e. that

$$A(t) = E[A(t)] \qquad D(t) = E[D(t)]$$

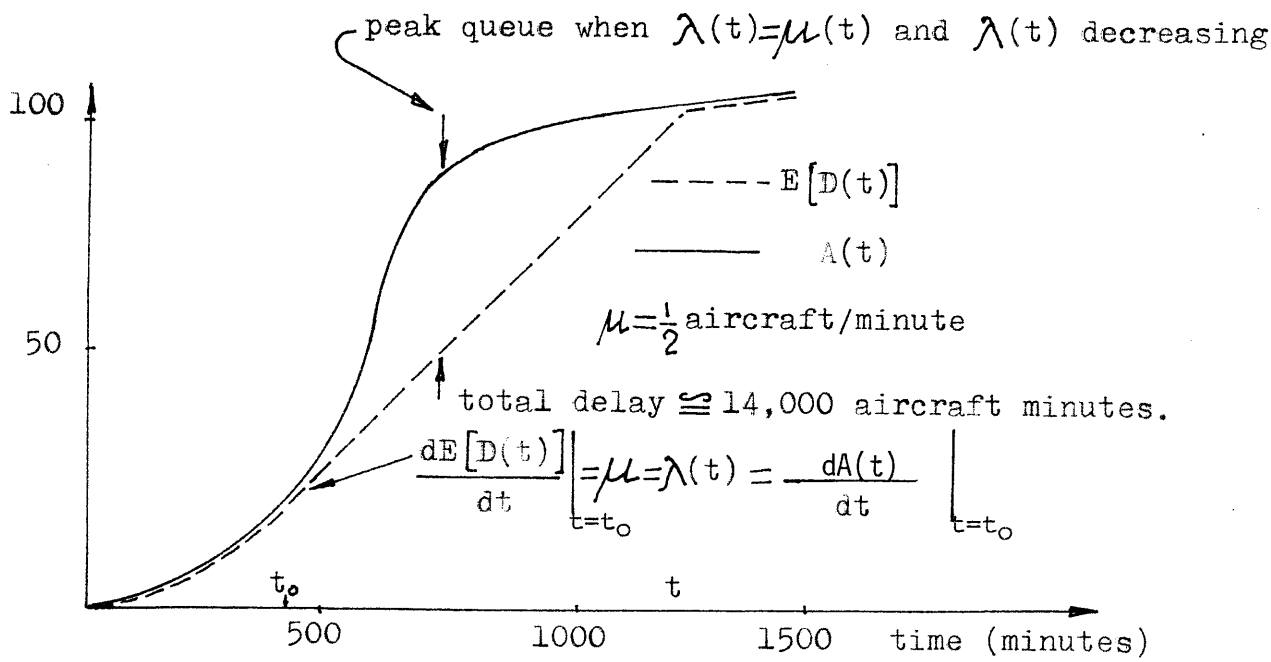
D/D/k analysis is customarily done graphically (Figure 2.2). The curve $E[D(t)]$ is superimposed on the curve $A(t)$. Cumulative delay is then easily calculated by the time integral of the vertical difference of the two curves over the period of interest.



An example of graphical analysis to solve D/D/k systems

Figure 2.2

This method is particularly attractive, computationally, if the number of customers is large. Then on a coarse scale of measurement the queueing



Example of "fluid approximation" in graphical analysis of D/D/k systems

Figure 2.3

model treats customers as a continuous fluid that is flowing into a reservoir at some rate $\frac{dA(t)}{dt} = \lambda(t)$ (Figure 2.3). The maximum flow rate out is the service rate μ:

$$\frac{dE[D(t)]}{dt} \leq \mu^{1}$$

As long as $\frac{dA(t)}{dt}$ remains less than the maximum service rate, i.e. $\lambda(t) < \mu$, we approximate the queue to be zero. A queue exists between time t_0 when $\lambda(t)$ first exceeds μ, and time t_1 when $A(t)$ first again equals $E[D(t)]$.

The consequences of ignoring the statistical fluctuations inherent in the random functions of time, $A(t)$ and $D(t)$, are of course non-negligible. The deterministic model assumes that the server works at a rate μ from the beginning of the rush hour until $A(t)$ and $E[D(t)]$ are again equal. In a real queuing process, even under heavy traffic conditions, there is a significant finite probability that the server remains idle. Service in reality proceeds at a rate μ whenever a queue exists, and is interrupted otherwise. Thus, the real average service rate over the

---

1. The method does not actually require a constant maximum flow rate μ. The maximum flow rate, just like the average arrival rate,can be time dependent, i.e. $\mu = \mu(t)$. Nothing would change in the following discussion under the assumption of a time dependent maximum flow rate.

rush hour can never exceed μ, and as we have pointed out, is expected to be somewhat less. On account of this loss of system capacity we expect a nonvanishing queue even when $\lambda(t) < \mu$ for all t. The D/D/k model, however, ignores the presence of a queue for $\rho(t) = \lambda(t)/\mu < 1$, and this underestimation of cumulative delay is nondecreasing in time for the duration of the period of interest.

The simplicity of the D/D/k approach, however, has led to its adoption by the FAA in a real time system for anticipating and preventing large delays to scheduled traffic due to inclement weather.[1] As in these situations the system will almost always be in the oversaturated, $\rho(t) > 1$, state, the queue is likely to grow rapidly in time, making it less likely that statistical fluctuations could allow the queue to vanish. This would tend to support empirical evidence that the model performs adequately in these situations. However, in other than rush hour situations, by ignoring the stochastic nature of the arrival process, we disregard, in effect, all of the known queuing phenomena which actually take place on an everyday basis at major airports.

---

1. Developed at the Transportation Systems Center of the DOT for the Flow Control Facility of the FAA in Washington D.C.

2.4 Model 3: Equilibrium Analysis, M/D/k/∞, M/M/k/∞

The governing integro-differential equations for M/G/k/∞ queuing systems are very complex. Therefore, even the simplest case when the system is in equilibrium does not yield queuing theoretic statistics in closed form. A single exception to this involves the expected steady state delay E[W] for an M/G/1/∞ system, which is given exactly by the Pollaczek-Khintchine formula:

$$E[W] = \frac{\rho}{2(1-\rho)} \left[1 + \frac{Var[s]}{g_1^2} \right] g_1 \qquad \rho < 1 \tag{2.1}$$

where s is the random variable described by the service time distribution, and g_i is the ith moment of the random variable s.

For some special cases of the M/G/k/∞ queue, with well-behaved service processes, the governing equations do turn out to be relatively easy write and solve analytically. For these cases the state probabilities and other statistics obtainable from the state probabilities (expected waiting times, etc.) can be derived from the governing equations. In particular, two cases that have received much attention in this type of analysis are the M/D/k/∞ and M/M/k/∞ queuing models. These are in fact extreme cases since service times are constant for the first while "perfectly random" for the latter. (The negative exponential pdf, governing the service process in the M/M/k systems, is frequently called "perfectly random" because of its property that at no matter what point in time we observe the process, the time until the next completion of a service is completely independent of the past history of the system). Furthermore, it turns out from the evaluation of (2.1) that E[W] for the M/D/1/∞ system is exactly one

half of that of the $M/M/1/\infty$ system, independent of ρ or μ. If we accept the intuitive proposition that individual aircraft service times must be "less regular" than the perfect regularity of constant service time, yet "less random" than the "perfect randomness" of the negative exponential service time, then we have shown that the delays for $M/D/1/\infty$ and $M/M/1/\infty$ systems provide lower and upper bounds on the true value of delay in an $M/G/1/\infty$ system. Although no similar relation to (2.1) is available for the multiserver queues $M/D/k/\infty$ and $M/M/k/\infty$, the closed form formulas for $E[W]$ of these systems are available and again provide bounds on the $M/G/k/\infty$ delay.

2.4.1 Poisson Arrivals, Deterministic Service ($M/D/k/\infty$)

When service time is of a deterministic nature, for the purpose of making the analysis tractable, the servers are assumed to be identical and to always commence and terminate service simultaneously at equally spaced discrete points in time. These separating time intervals are exactly equal to the service time of $\frac{1}{\mu}$, where μ is the service rate (operations/unit time) of a single one of the identical k servers. At the designated points in time the model will discharge k or fewer aircraft from service, depending on the number of aircraft in the system, and observe n new aircraft arriving into the queue with probability $a(n)$ given by the Poisson probability distribution

$$a(n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \qquad n = 0,1,\ldots$$

We then define the state probability $p_i(t)$ as the probability that at time t the queue is of length i, $(i=0,1,2,\ldots)$.

The recursion equations relating the state probabilities, known in the literature as the Chapman-Kolmogorov (C-K) equations, for any and all discrete instants of time are given by:

$$p_0(t + \frac{1}{\mu}) = e^{-\lambda_s} q_k(t)$$
$$p_j(t + \frac{1}{\mu}) = \frac{\lambda_s^j e^{-\lambda_s}}{j!} q_k(t) + \frac{\lambda_s^{j-1} e^{-\lambda_s}}{(j-1)!} p_{k+1}(t) + \ldots + e^{-\lambda_s} p_{k+j}(t) \tag{2.2}$$
$$1 \leq j$$

where the following notational conventions have been adopted:

$$\lambda_s = \lambda(t) \cdot [\text{one service interval}] = \lambda(t) \cdot \frac{1}{\mu}$$
$$q_k(t) = \sum_{i=0}^{k} p_i(t)$$

It is possible for stationary λ to solve the system of equations (2.2) for the steady state probabilities $p_i(t)$ and other queueing statistics. Of particular interest here is the steady state expected waiting time E[W]. Saaty [12] gives for the arbitrary server M/D/k/∞ queue:

$$E[W] = \sum_{i=1}^{\infty} e^{-i\rho k} \left[\sum_{j=ik}^{\infty} \frac{(i\rho k)^j}{j!} - \frac{1}{\rho} \sum_{j=ik+1}^{\infty} \frac{(i\rho k)^j}{j!} \right]$$
$$\rho = \frac{\lambda}{\mu k} < 1$$

2.4.2 Poisson Arrivals, Negative Exponentially Distributed Service, (M/M/k/∞)

Alternatively consider the case of a probabilistic service process with a negative exponential pdf for service time duration (maintaining Poisson arrivals). The C-K equations describing the state probabilities $p_i(t)$ for

the M/M/k/∞ system are as follows:

$$p_0(t + \Delta t) = (1 - \lambda\Delta t - \mu\Delta t)\ p_0(t) + \mu\Delta t p_1(t)$$

$$p_j(t + \Delta t) = \lambda\Delta t p_{j-1}(t) + (1 - \lambda\Delta t - j\mu\Delta t)\ p_j(t) + (j+1)\ \mu\Delta t p_{j+1}(t)$$

$$1 \leq j \leq k-1$$

$$p_i(t + \Delta t) = \lambda\Delta t p_{j-1}(t) + (1 - \lambda\Delta t - k\mu\Delta t) p_j(t) + k\mu\Delta t p_{j+1}(t)$$

$$k \leq j$$

Passing to the limit, as $\Delta t \to 0$, we obtain from the difference equations Kolmogorov's forward differential equations:

$$\frac{dp_0(t)}{dt} = \lambda p_0(t) + \mu p_1(t)$$

$$\frac{dp_j(t)}{dt} = \lambda p_{j-1}(t) - (\lambda + j\mu)\ p_j(t) + (j + 1)\ \mu p_{j+1}\ (t)$$

$$1 \leq j \leq k-1$$

$$\frac{dp_j(t)}{dt} = \lambda p_{j-1}(t) - (\lambda + k\mu)\ p_j(t) + k\mu p_{j+1}(t) \qquad k \leq j \qquad (2.3)$$

As for the M/D/k queue, closed form expressions for the p_i and E[W] are available. For the arbitrary server M/M/k/∞ queue:

$$E[W] = \frac{\rho(k\rho)^k}{\lambda k!(1-\rho)^2} \cdot p_o \qquad \rho = \frac{\lambda}{k\mu} < 1$$

where

$$p_o = \left[\sum_{n=0}^{k-1} \left[\frac{(k\rho)^n}{n!}\right] + \frac{(k\rho)^k}{k!(1-\rho)} \right]^{-1}$$

2.4.3 Application of Equilibrium Analysis

The main problem with the application of equilibrium analysis to the airport queuing problem is that although it correctly recognizes the problem as stochastic, it really is valid only when $\lambda(t)$ is invariant, or varies extremely slowly with time. Even then, the equilibrium approximation with a constant λ must be made over a sufficiently long period of time so as to dominate any transient effects. Examination of the demand profiles of the airports most likely to exhibit uniform traffic throughout the day (which are generally the most congested airports, such as New York's LaGuardia, where a quota system on operations is in effect) indicates significant peaks and valleys that seem to invalidate the steady state approach except as a very rough first order approximation.

Furthermore, situations where demand equals or exceeds capacity require special treatment. Up until now, we have been using infinite queue systems. Equilibrium analysis is invalid for these systems when $\rho \geq 1$ because the queue length is not approaching any limiting steady state value. We could partially circumvent this by considering finite queue systems of large maximum length m.[1] This ensures a finite value of delay for all values of ρ and furthermore, it is a good approximation of the infinite queue system when $\rho < 1$. We illustrate this property for the M/M/1/m and M/M/1/∞ cases.

---

1 It is by choosing m to be large that we are, later in this study, able to model systems with infinite queue capacity on the computer (i.e., to "solve" an "infinite" number of C-K equations).

The closed form equilibrium state probabilities for these systems, P_i^m and P_i respectively, are known and simple to compare:

$$\text{finite queue:} \quad P_i^m = \frac{(1-\rho)\rho^i}{1-\rho^{m+1}} \qquad \rho \gtrless 1$$

$$\text{infinite queue:} \quad P_i = (1-\rho)\rho^i \qquad \rho < 1$$

Clearly, for $\rho < 1$, the $\lim_{m\to\infty} (1-\rho^{m+1}) = 1$ and

$$\lim_{m\to\infty} P_i^m = P_i \tag{2.4}$$

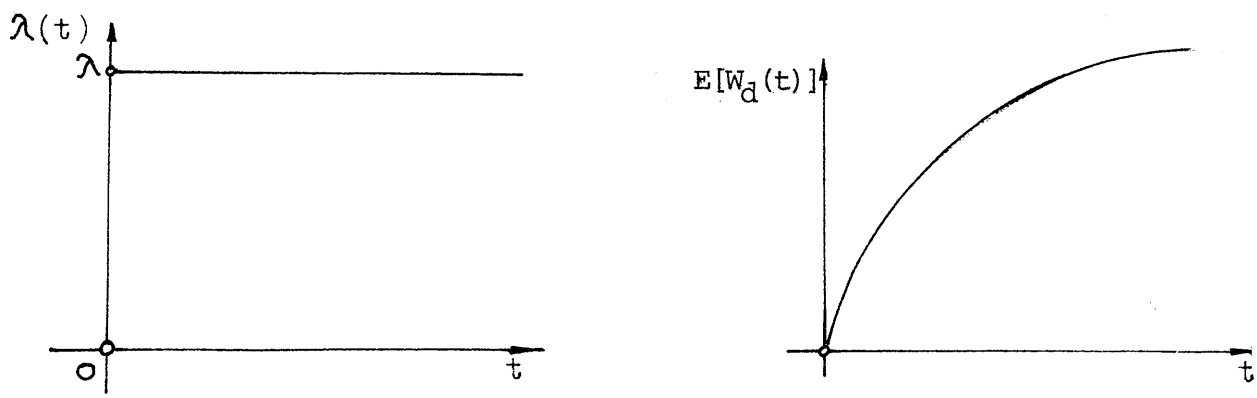
On the other hand, when $\rho > 1$ in a finite queue capacity system, the delay analysis hinges on the fact that there is a large probability that the queue is saturated. When this happens, aircraft "arriving" into the queue are being forced to cancel operations altogether. As a result, the waiting time obtained from finite queue analysis will be less than the true value of delay in a system where actually all aircraft are served. We conclude that waiting time in the queue may then no longer be a sufficient descriptor of system performance.

Unfortunately, the approximation of an infinite queue capacity system by finite queue capacity system cannot be extended to the M/D/k case, simply because there are no closed form results. Therefore, inevitably, in cases where temporary oversaturation does occur, the estimating of delays with equilibrium analysis reduces to educated guesswork on the period of time and level of oversaturation. It appears that for the very intervals of time when the most significant delays occur at our major airports, the steady state models provide only a very crude approximation.

2.5 Model 4: Stationary Diffusion Approximation M/G/1

2.5.1 Motivation and Assumptions

The deterministic or "fluid" approximation treated in Model 2 should be considered a "first-order" approximation. A "second-order" model is the diffusion approximation model which is the main topic of this section. Whereas the deterministic approximation relied only upon the expected values of the parameters $\lambda(t)$ and $\mu(t)$, the diffusion approximation requires also the second moment of the service process. We review here the analysis by Harris [6] to obtain $E[W_d(t)] \geq 0$, the expected waiting time in queue for a customer arriving a time t after a step demand has been turned on at t = 0. The service facility consists of a single server characterized by a "general" stationary service process, with first and second moments of service time g_1 and g_2.



(a) Step demand at a service facility applied at t = 0

(b) Response of system to given input.

Figure 2.4

Model 4 will differ substantially from the previous three models in the sense that we will define a new state variable. Previously, we used q(t), the queue length, or the probabilities $p_i(t)$ of queue length i, as the state variables. For systems with the particular service processes we have discussed so far, these were state variables because knowledge of their values alone was sufficient to determine probabilistically all future states of the system. For example: (a) With the deterministic service process we discretize time into intervals of single service time length. Then the departure time of a customer from service, given that the customer is in service now, is always the beginning of the next interval, which is entirely independent of the system history. (b) With service time governed by the "perfectly random," negative exponential pdf, we have the Markov memoryless property. Therefore, by definition, the time to next departure from service is completely independent of system history at all times.

In the case of the "general" service process, the probability of a departure from the system in any instant is no longer independent of the system history. We seek then to construct a simple new state variable, one which hopefully would possess the Markov property. Our experience with the differential equations of Models 3 for systems possessing the Markov property indicated that treatment of such systems was considerably simplified by this property.

Therefore, let W(t) be the waiting time in queue for a customer

arriving at time t. This is the virtual waiting time proposed by Takacs as a state statistic. The basic shape in time of a W(t) path of a queuing system (with infinite waiting room) is a random sawtooth. W(t) increases with vertical jumps of independent identically distributed (iid) service time magnitudes at the instants of customer arrival while continuously decreasing deterministically at slope -1 whenever the queue is nonzero.

The W(t) path so described is both additive and Markovian. Since the arrivals follow the Poisson process, the next arrival will, independently of the past history, occur on the basis of an interarrival time with negative exponential distribution with (possibly nonstationary) mean $\frac{1}{\lambda}$. All arrivals, independent of the state of the system, cause jumps in W(t) of height given by iid random variables which are distributed as the service times of the system.

In addition to the additive and Markovian properties of W(t) we make two key assumptions on W(t) prior to presenting the diffusion model. The first assumption requires that changes in W(t) in a small time interval be negligible with respect to W(t). Due to the jumps observed in W(t) this requires that the queue always contain many customers as well as $W(t) \gg g_1$. In general, W(t) remains large, and the first assumption is satisfied when ρ is "close" to unity. As the model parameters are defined, the implications of, and measures for "closeness" will be made more precise.

The second assumption hypothesizes the existence of an infinitesimal time interval τ, short enough to allow W(t) to change only by a negligible amount, but simultaneously long enough for many events to take place in τ. (The mean interarrival time must therefore be short when compared to the time scale of the approximation.) This allows us to invoke the Central Limit approximation of sums of random variables. Were it not for the condition $W(t) > 0$, then W(t) could easily be obtained by adding sums of independent normal random variables (of waiting time) corresponding to arrivals and departures from the queue. Also it is clear that invoking the Central Limit Theorem causes the discrete component of W(t), the number of teeth in the random sawtooth, corresponding to the Poisson arrivals, to be modelled only approximately.

2.5.2 Model Development

In the following discussion we will describe the (conditional) transition pdf $f(x,t \mid x_0,t_0)$ for the waiting time x in the system at time t, given that it was x_0 at t_0 a short time earlier. As postulated, x is described by a continuous transition Markov or diffusion process. This approximation method to the waiting time problem is governed by the partial differential equation of the Weiner process:

$$\frac{\partial f}{\partial t} = -a(t)\frac{\partial f}{\partial x} + \frac{b(t)}{2}\frac{\partial^2 f}{\partial x^2} \tag{2.5}$$

This process has the property that it is completely specified by the parameters a(t) and b(t) once initial boundary conditions are set. a(t) and b(t) correspond to the infinitesimal mean and variance of the motion of the process. They are given by:

$$a(t) = \tau^{-1}E[W(t + \tau) - W(t)]$$

$$b(t) = \tau^{-1}\ \mathrm{Var}[W(t + \tau) - W(t)]$$

Now a(t), b(t) can be readily computed because of the Markovian property of W(t). Only two events are possible in time τ - no arrival (and the waiting time decreases by τ), or an arrival (and the waiting time increases by random variable s, from the distribution of the service times, less service completed on the customer in service, τ)

$$E\left[W(t + \tau) - W(t)\right] = -\tau(1 - \lambda(t)\tau) + \lambda(t)\tau(g_1 - \tau) = \tau a(t)$$

$$\Rightarrow a(t) \quad g_1\lambda(t) - 1$$

$$\mathrm{Var}\left[W(t + \tau) - W(t)\right] \simeq \lambda(t)\tau g_2 = \tau b(t)$$

$$\Rightarrow b(t) = \lambda(t)g_2$$

The boundary condition is simply $f(0,t) = 0 \qquad t \geq 0$

Although there are many possible solution methods to this problem, the most commonly cited approach is the method of images from physics (Harris [6], Gaver [5], Newell [11]). For the

stationary[1] parameter case $\lambda(t) = \lambda$:

$$E[W_d(t)] = \begin{cases} \dfrac{\lambda g_2}{2(1-g_1\lambda)} - \dfrac{(g_2\lambda)^{\frac{3}{2}}}{2(1-g_1\lambda)^2 t^{\frac{1}{2}}} e^{-\frac{(1-g_1\lambda)^2}{2g_2\lambda} t} & g_1\lambda < 1 \quad (2.6) \\[2ex] \sqrt{\dfrac{2g_2\lambda t}{\pi}} & g_1\lambda = 1 \quad (2.7) \\[2ex] (g_1\lambda-1)t + \dfrac{g_2\lambda}{2(g_1\lambda)^2} - \dfrac{(g_2\lambda)^{\frac{3}{2}}}{2(1-g_1\lambda)^2 t^{\frac{1}{2}}} e^{-\frac{(1-g_1\lambda)^2}{2g_2\lambda} t} & g_1\lambda > 1 \quad (2.8) \end{cases}$$

2.5.3 Characteristics of $E[W_d(t)]$ (Model 4)

It is clear that the expressions (2.6) - (2.8) for $E[W_d(t)]$ from Model 4 differ substantially from the values obtained by the methods of Model 2. The most obvious improvement over Model 2 is that the diffusion approximation predicts a delay for all $\rho > 0$, not just $\rho > 1$. Another aspect is that $E[W_d(t)]$ agrees at least in form with the behaviour we would expect by solving directly the integro-differential for the M/G/1

---

1. Harris shows that straight forward application of this method will not work for nonstationary parameters.

system: namely $E[W_d(t)]$ is a composition additively of terms which are functions of time and a term which is a constant independent of time - the transient and steady state terms - respectively.

Before examining the transient and steady state terms we pause briefly to observe why Model 4 is in fact a "second-order" approximation with respect to Model 2. The argument centers on assumption two (and the Central Limit Theorem) which implies that events during a small interval of time τ are normally distributed. Equating the two incremental parameters $a(t)$ and $b(t)$ of the real and approximating $(f(x,t))$ processes ignores all higher moments of the real transition pdf. Contrast this with Model 2 which neglects the possibility of fluctuations in events completely, through a law of large numbers argument. This corresponds to the total neglecting of the term $\frac{b(t)}{2} \frac{\partial^2 f}{\partial x^2}$ in (2.5).

Model 2 had no transient or even steady state component. The only term appearing in the waiting time for both Models 2 and 4 is the linear growth with time: $(g_1\lambda - 1)t$ for $\rho > 1$. The much more intuitive behaviour of Model 4 will be illustrated in the following. The $E[W_d(t)]$ (2.6) - (2.8) are valid for the case of a step demand λ beginning at time t=0, with initially no customers in the queue. We expect and observe the time dependent waiting time $E[W_d(t)]$ approaching its limiting steady state behaviour from below.

The case $\rho < 1$ is the only case that approaches an equilibrium solution with the decay of the transient. For the overloaded queue,

$\rho > 1$, not all the customers can be served, causing the queue to grow with time. There is still, however, a transient term. As for $\rho < 1$, the transient decays exponentially, and $E[W_d(t)]$ approaches growth as $(g_1\lambda - 1)t$. The relation for the special case of demand exactly equaling the mean rate of throughput also agrees with intuition. Neither an equilibrium solution as for $\rho < 1$, nor linear growth with time, as for $\rho > 1$, is expected for $E[W_d(t)]$ when $\rho = 1$. The observed growth as $\sqrt{t}$, slower than for $\rho > 1$, indicates the basically unstable nature of the queue, as expected.

The behaviour of the transient term should be closely examined. Apparently for both $\rho \gtrless 1$ the transient term decreases as $t^{-\frac{1}{2}}e^{-t/t_0}$, where t_0 is a constant. Because of the dominance of the exponential component in t, t_0 has been termed the <u>relaxation time</u>, defined as the length of time required for the transient part of the queuing statistic to relax to $\frac{1}{e}$ of its original value. Chapter 3 will be devoted entirely to analytic and empirical treatments of the transient behaviour. Therefore, this behaviour will not be discussed further here. We will, however, quote known steady state results in the next few paragraphs to show that the asymptotic behaviour $E[W_d] = \lim_{t\to\infty} E[W_d(t)]$ does reasonably approximate these results.

The expected steady state delay E[W] for an M/G/1 system is given exactly by the Pollacek Khintchine formula (2.1). With the diffusion results Gaver[5] shows:

$$E[W_d] = \frac{\lambda g_2}{2(1-\rho)} = \frac{\rho}{2(1-\rho)} \frac{g_2}{g_1} = \frac{\rho}{2(1-\rho)} [g_1 + \frac{Var[s]}{g_1}] = E[W] \tag{2.9}$$

The expected values are therefore exactly equal. The exact expression for the variance of the waiting time is

$$Var[W] = \left[\frac{\lambda g_2}{2(1-\rho)}\right]^2 + \frac{\lambda g_3}{6(1-\rho)}$$

To obtain $Var[W_d]$ we return to the original differential equation for the diffusion process:

$$\frac{\partial f}{\partial t} = a(t) \frac{\partial f}{\partial x} + \frac{b(t)}{2} \frac{\partial^2 f}{\partial x^2}$$

For the stationary case $a(t) = a_0$ = constant, $b(t) = b_0$ = constant, we can set $\frac{\partial f}{\partial t} = 0$ and solve for f, the density of the waiting time in queue.

The equation is satisfied by:

$$f(x) = \frac{2(1-\rho)}{g_2\lambda} \exp\left[-\frac{2(1-\rho)x}{g_2\lambda}\right] \qquad \rho < 1$$

Recognizing this as an exponential distribution we obtain the variance

$$Var[W_d] = E^2[W_d] = \left[\frac{\lambda g_2}{2(1-\rho)}\right]^2 < Var[W]$$

Therefore, we obtain a smaller variance for $\rho < 1$. Gaver [1] was able to prove that these do coincide for the case $\rho = 1$.

2.6 Model 5: Numerical Solution of the Time Dependent Chapman Kolmogorov Equations, M/D/k, M/M/k

2.6.1 Method and Assumptions

Koopman [8], rather than relying on steady state results, has sought and obtained numerical solutions for the Chapman-Kolmogorov equations for M/D/k and M/M/k systems. Since the queuing equations are valid for all values of ρ (not just $\rho < 1$), iterative numerical solution of the equations yields the distribution of the queue length $p_i(t)$ $i = 0,1,\ldots$ from which the expected delay $E[W(t)]$ is calculated. Further it is possible to adjust the average arrival or service rates at any iteration of the numerical solution of the equations to reflect any corresponding variations in the "real world." (Frequently airports operate with fewer runways at nights or during bad weather. This situation can be accomodated in the numerical solution.)

Koopman has shown that, in the absence of steady state conditions, periodicity of the $\lambda(t)$ and $\mu(t)$ usually observed at airports (diurnal 24 hour schedule pattern) guarantees the existence and uniqueness of the state probabilities $p_i(t)$. These solutions for the state probabilities will be valid as long as (a) the average service time $\frac{1}{\mu}$ of an aircraft is small enough such that no significant change in arrival rates or general conditions is observed during a single service time and (b) there exists during the day a period when the level of operations is negligible compared to the airport capacity. This point must be chosen as the starting point

so that transient considerations remain negligible. In the event that such a point does not exist the $p_i(t)$ should be evaluated over a period of time equal to twice the cycle of the $\lambda(t)$ to ascertain that periodicity is indeed achieved. For practical purposes, however, this issue is not particularly important. The remaining basic assumptions of this approach are identical to those presented in section 2.4. Koopman developed the one runway case assuming a finite maximum queue length (since numerically it is possible to solve only a finite number of state equations). No distinction was made between arriving and departing aircraft, although possible extensions to multiqueue situations were illustrated.

2.6.2 Computational Considerations

Minor changes must be made to the last k equations in (2.2) and (2.3) when there are only a finite number of states m. These are to ensure that the maximum queue length is not violated by the number of new aircraft arriving into the queue, as well as to ensure that the probabilities add up to 1. For the M/D/k system the last k equations are:

$$p_j(t + \frac{1}{\mu}) = \frac{\lambda_s^j e^{-\lambda_s}}{j!} q_k(t) + \frac{\lambda_s^{j-1} e^{-\lambda_s}}{(j-1)!} p_{k+1} + \ldots + \frac{\lambda_s^{j-(m-k)} e^{-\lambda_s}}{[j-(m-k)]!} p_m(t)$$

$$m - k < j \leq m-1$$

$$1 = q_k(t + \frac{1}{\mu}) + p_{k+1}(t + \frac{1}{\mu}) + \ldots + p_m(t + \frac{1}{\mu})$$

and for the M/M/k system:

$$\frac{dp_j(t)}{dt} = \lambda p_{j-1}(t) - (\lambda + k\mu)p_j(t) + k\mu\, p_{j+1}(t) \qquad m - k < j \leq m-1$$

$$1 = p_0(t) + p_1(t) + \ldots + p_m(t)$$

According to the discussion of Model 3 we would like m to be large enough such that queue saturation never takes place. Bearing in mind the strongly time dependent $\lambda(t)$ (and/or $\mu(t)$), we realize that in the course of an iterative solution the value of m for which $p_m(t)$ becomes negligibly small can vary appreciably. Much computational effort may be saved by recognizing that negligibly small probabilities need not be calculated. Thus a fictitious "maximum allowable" queue length can be set up to vary at each iteration in a fashion to ensure that the probability of queue saturation does not exceed a negligibly small predetermined tolerance.

Given then a set of initial values $p_i(0)$ $i = 0,1,\ldots,m$, and the average arrival and service rate profiles $\lambda(t)$ and $\mu(t)$ for a time period of interest [0,T], the equations can be solved for the queue length distributions at discrete points in time for the interval. For the M/D/k case the distributions obtained correspond to points in time $\frac{1}{\mu(t)}$ (=1 aircraft service time) apart. The M/M/k equations must be solved as difference equations and consequently yield the solutions at arbitrarily small, but regular, intervals of time Δt.

2.6.3 Application

The statistic usually of greatest interest, the expected waiting time in queue, $E[W(t)]$, is readily obtained once the $p_i(t)$ are known:

$$E[W(t)] = \frac{1}{k\mu(t)} \sum_{i=k}^{m} (i-k+1)\, p_i(t) \tag{2.10}$$

A minor change to (2.10) must be made for the M/D/k case to account for the modelling hypothesis that all arrivals into, and all departures from the system occur at discrete instants of time:

$$E[W(t)_{M/D/k}] = \frac{1}{k\mu(t)} \sum_{i=k}^{m} (i-k+\tfrac{1}{2})\, p_i(t)$$

Other measures of importance, the expected queue length,

$$E[L_q(t)] = \sum_{i=k}^{m} (i-k)p_i(t)$$

and probability that an aircraft is delayed prior to service,

$$P(W > 0) = 1 - \sum_{i=0}^{k} p_i(t)$$

are easily computable.

Koopman discovered that delays observed in systems with $\lambda(t)$ strongly time dependent were remarkably <u>insensitive</u> to the precise

stochastic properties of the service process.

Also, we have noted in our discussion of steady state results, from the Pollaczek-Khintchine formula for the single server queue, and closed form expressions for many server systems, that M/M/k and M/D/k systems appear to provide upper and lower bounds on the average equilibrium delay. Together these observations imply a tremendous simplification of the study of the M/G/k queue. Koopman's results suggest that judicious interpolation of nonequilibrium M/D/k and M/M/k solutions is all that is needed to obtain a good approximation of actual delays in a nonstationary M/G/k queuing system. Similar closeness of the variances of the expected queue lengths (the variances tend to be large compared to the difference of the expected queue lengths for the two different service policies) further strengthens the case for the validity of interpolation.

2.6.4 An Alternative to Interpolation

We conclude this section with a brief investigation of a mathematically tractable alternative to interpolation between M/M/k and M/D/k delay statistics in obtaining the delay for M/G/k systems. At least part of the success of Koopman's solution technique is due to the fact that these two extremes have the simplest forms of the C-K equations (2.2) and (2.3). We will now investigate the efficacy of numerically solving C-K equations when more general service processes are introduced. In particular we write the C-K equations for a class of service processes that includes both the deterministic and negative exponentially distributed

service times as its extremes. The outcome will be that even for the simplest case of a single server, numerical solution of these equations fails to be practical when compared to Koopman's method.

The class of service processes we mean is the cth order Erlang distribution whose pdf f(s), for an arbitrary integer $c > 0$, is given by:

$$f(s) = \frac{(\mu c)^c}{(c-1)!} s^{c-1} e^{-c\mu s} \qquad s \geq 0 \qquad (2.11)$$

By definition f(s) is the distribution of the sum of c iid exponential random variables x, $f(x) = c\mu e^{-c\mu x}$, $x > 0$, each x having mean $\frac{1}{c\mu}$. Therefore, the mean service time and variance of s are:

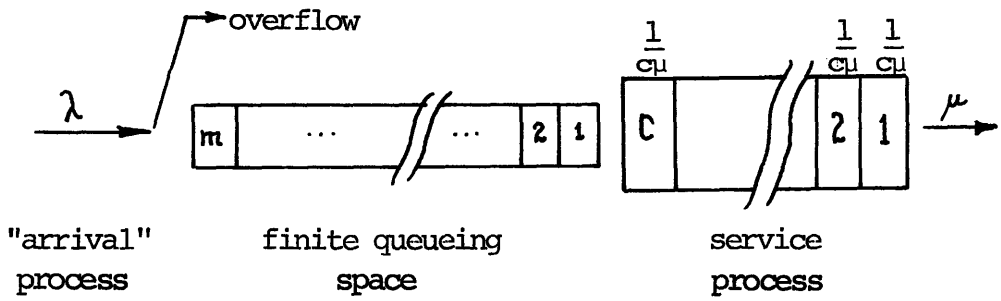
$$g_1 = c \cdot \frac{1}{c\mu} = \frac{1}{\mu}$$

$$\mathrm{Var}[s] = c\sigma_x^2 = c\,\frac{1}{(c\mu)^2} = \frac{1}{c\mu^2} \qquad (2.12)$$

From (2.11) and (2.12) it is obvious that both negative exponentially distributed and deterministic service times are special cases of service time described by the cth order Erlang pdf, with c given by 1 and ∞ respectively.

For the purpose of writing the C-K equations of a single server queueing system with service process defined by the cth order Erlang pdf (still with Poisson arrivals) abbreviated as $M/E_c/1$, we note that although the Erlang distribution itself is not Markovian, the distribution

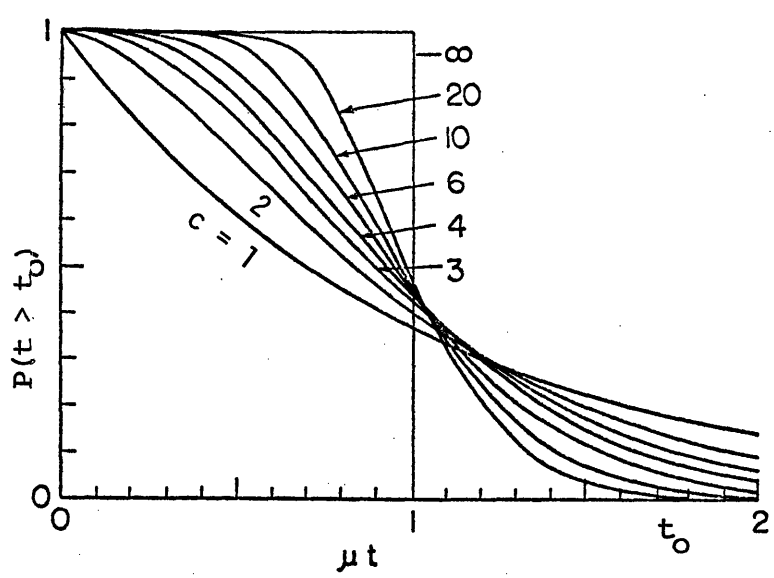
is composed of c iid negative exponential units. We call these units "phases" and imagine a queueing system with c^{th} order Erlang service time distribution as the "c phase service" queue illustrated in Figure 2.5:



A schematic diagram of the cth order Erlang type service queue

Figure 2.5

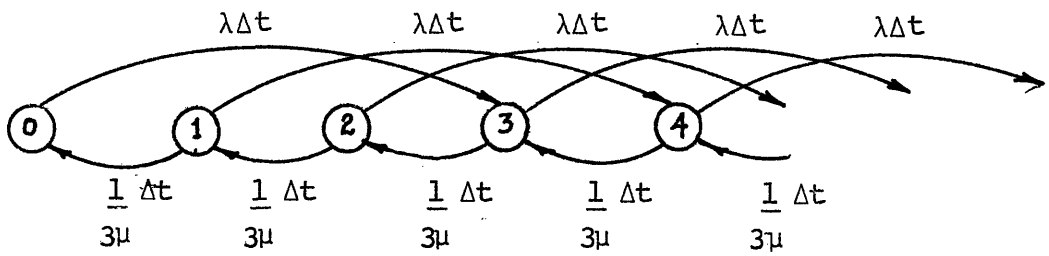
The effect of increasing the number of phases required by each customer, i.e. going from "perfectly random" to deterministic type service, on the distribution of the service times is shown in Figure 2.6.



(1-cumulative distribution function) of the cth order Erlang pdf(Morse[10])

Figure 2.6

We take advantage of the description of service as "phased" by measuring the queue length in terms of phases remaining to be completed, rather than the total number of waiting customers. This is easily done as the arrival of every customer corresponds to the bulk arrival of c phases of service. The state diagram for the $M/E_3/1$ queue, in terms of the phases is given in Figure 2.7. As each phase possesses the



State diagram for an $M/E_3/1$ system

Figure 2.7

Markovian property the state transition probabilities are independent of time. The differential equations for $p_i(t) = 0,1\ldots,$ the probability of i phases of service remaining to be completed at time t, are then written as in (2.3):

$$\frac{dp_0}{dt} = -\lambda p_0 + c\mu\, p_1$$

$$\frac{dp_i}{dt} = -(\lambda + c\mu)p_i + c\mu\, p_{i+1} \qquad 0 < i < c \qquad (2.13)$$

$$\frac{dp_i}{dt} = \lambda p_{i-c} - (\lambda + c\mu)p_i + c\mu\, p_{i+1} \qquad i \geq c$$

The numerical solution of the equations (2.12) is obtained the same way as for (2.3) although such a direct solution is computationally disadvantageous. The reason, of course, is that the number of states, and consequently the number of differential equations, is c times that of the "equivalent" (in the sense of customers in the queue) M/M/1 system. As these equations must be solved simultaneously the increase in computational effort is greater than linear. Solving, for instance, $M/E_3/1$ systems takes about an order of magnitude more computation time than the M/M/1 system. However, mumerically solving the $M/E_c/1$ system will provide the exact results for the transient behaviour of yet other special cases of the M/G/1 queue (besides M/M/1 and M/D/1). Therefore, many results will be presented for this case in the section on numerical results.

2.7 Summary

In this chapter we have reviewed the five main types of models which have been used in the analysis of nonstationary M/G/k queues. The main objections to three of the models were:

Model 1, Computer Simulation: requires excessive computation time and poses serious statistical difficulties in interpreting results.

Model 2, D/D/k "Fluid Approximation": ignores known and observed statistical fluctuations in both the demand profile and service times. Predicts no waiting time unless $\rho > 1$, and always underestimates waiting time. When compared to exact solutions Model 2 fares worst when ρ is frequently less than, but close to, 1. Performance is better when $\rho > 1$ and is best for lengthy periods of oversaturation.

Model 3, Equilibrium Analysis: fluctuations in airport demand profiles are very significant with the resultant conclusion that the system is never truly in the steady state. No solution for common "rush hour" situations, $\rho > 1$, for infinite queue systems, and difficulties in interpreting the values of E[W] for finite queues (because of queue saturation).

The fourth model (stationary diffusion approximation) was introduced as a natural extension of the deterministic approximation Model 2. Although analysis of delays for nonstationary demand profiles has been conducted with this method (Harris [6]), it is limited to particular

demand profiles and single servers, and hence was not discussed here. However, the results of Model 4 are invaluable in providing approximate closed form results for the transient behaviour of (stationary) M/G/1 queues, since exact results exist only for M/M/1.

The most promising method so far is Koopman's method of interpolating between the results of numerical solutions of the M/M/k and M/D/k systems because:

(1) apparently in time dependent M/G/k systems E[W(t)] is not strongly dependent on the exact form of the general service process. Therefore, interpolation between the extreme values of E[W(t)], based on the properties of the service process, is generally sufficient to approximate the true value of E[W(t)]. Besides, direct solution of the C-K equations for the particular service process was shown to require excessive computation time for even relatively "well behaved" service process such as the Erlang k.

(2) it requires substantially less computational effort than a simulation, and overcomes the difficulties of both Model 2 (when $\rho < 1$) and Model 3 (when $\rho > 1$) by calculating exact results for all values of ρ.

The following chapters will offer an explanation, based on theoretical considerations, for the success of Koopman's method. It will also provide a framework for investigating other models that can be developed for the M/G/k system with this knowledge.

Chapter 3

DERIVATION OF THE TIME CONSTANT OF THE M/G/1 SYSTEM AND APPLICATIONS

3.1 Motivation

Until now we have mentioned the term "transient behaviour" infrequently and each time in a slightly different connection. The first mention of it was in Model 3 where we said that the equilibrium approximation could be safely applied only when $\lambda(t)$ has been (nearly) constant for a sufficiently long period of time so as to overshadow any transient effects. Although constant demand profiles are currently rarities for airports, this might change. Setting of hourly quotas on runway movements, and marginal cost pricing of runway use are two of the policy decisions that may at least partly bring this about. If the resultant policy creates (near) constant daily demand profiles, then the application of Model 3 would seem to provide adequate rough approximations to the delays. Consideration of transient effects will show that this can in fact be misleading. Steady state delays increase exponentially (2.1), as the utilization ratio approaches 1. However, the length of time required to reach the steady state also increases exponentially.[1] Therefore, both the relaxation time and the actual transient component of $E[W(t)]$ must be known before steady state results can be of any value.

Our second mention of the transient concerned the response of an M/G/1 queue, Model 4, to a step demand applied at time 0 when the queue was

---

1. For average steady state delays per operation in the tolerable range for airports this period of time approaches a length equal to that of the operating day.

initially empty. Closed form results available from the governing equations of Model 4 can be used in creating approximating models of the more complex time dependent queue. For instance, the expression for the transient behaviour is the controlling factor in a computationally efficient methodology developed by Gupta. Based on knowledge of the steady state behaviour and some (incomplete) knowledge of the duration of nonnegligible transient effects, he was able to construct and solve an approximating differential equation for the queue length of time dependent M/M/K and M/D/K queues.

The third reference to the transient was in Model 5 in connection with initial starting conditions. Besides this, the transient plays another extremely important role in Model 5, and we take this opportunity to mention it. The reader will recall that the results of equilibrium analysis indicated that the steady state E[W] for the M/M/K and M/D/K queues differed by approximately a factor of 2 for small k. Koopman, however, observed that in the numerical solution of the equations the two systems provided relatively tight bounds on the system delay. It is of interest then to investigate the transient behaviour of the M/M/K and M/D/K systems as a cause of this behaviour.

It is clear from the discussion that any progress we make in analyzing transient behaviour will be extremely useful. We will start by presenting yet another model, Model 6, comparable to Model 4 in the sense that it deals with a stationary queue, with the differences that it is an exact model and limited to the M/M/1 case (as opposed to the approximating

M/G/1 analysis). Our aim is to obtain some workable expression for the transient behaviour that can be considered exact. We will note the similarity of the queueing statistics obtained from Model 6 and Model 4, and make comparisons. Since Model 4 will turn out to be a very good approximation of the stationary M/M/1 queue, we make attempts to stretch our knowledge of the "exact" transient behaviour to the gamut of M/G/1 queues via the parameters of Model 4. We conclude with Models 7 and 8. The former is an approximation model of the time dependent M/M/k and M/D/k queues which expressly needs the kinds of closed form results that this chapter is concerned with. Model 8 uses the development of Model 6 to formulate a new model for the time dependent M/M/1 queue. In particular we suggest Model 8 as an alternative to Model 5 when quick approximations are needed.

3.2 Model 6: Transient Solutions for Stationary Queues, M/M/1

3.2.1 Morse's Methodology

This section develops series-form expressions for the transient state probabilities $p_i(t)$ for stationary finite queues. The methodology is due to Morse [10], and is perfectly general provided the arrival and service processes constitute a birth and death Markovian model.

Drawing upon the birth and death process property of the model we are able to write the linear Chapman-Kolmogorov equations. When the system is not in equilibrium we have the forward Kolmogorov differential equation for the time rate of change of the state probabilities:

$$\frac{dp_i(t)}{dt} = \sum_{n=0}^{m} E_{in} p_n(t) \tag{3.1}$$

where the E_{in} are the rates of transition into state i given the system was in state n an arbitrarily short time earlier. The E_{in} are the mean arrival rates of the exponential units, which may in their most general form be themselves functions of time. For the transient analysis, however, all parameters will be assumed stationary.

The analysis begins with the assumption that the transient state probabilities we seek, $p_i(t)$, are completely described by the form (3.2):

$$p_i(t) = \sum_{s=0}^{m} B_{is} e^{-\gamma_s t} \tag{3.2}$$

where: (a) the queue is of finite maximimum length m

(b) B_{is} $i,s = 0,1,2,\ldots,m$ are constants to be adjusted to meet initial conditions

(c) γ_s $s = 0,1,2,\ldots,m$ are rates of decay of transient components

Substituting (3.2) into the differential equations (3.1) we obtain:

$$\gamma_i B_{ii} e^{-\gamma_i t} + \sum_{n=0}^{m} E_{in} p_n = 0 \qquad i = 0,1,\ldots,m$$

$$(\gamma_i + E_{ii}) p_i + \sum_{\substack{n=0 \\ n \neq i}}^{m} E_{in} p_n = 0 \qquad i = 0,1,\ldots,m \tag{3.3}$$

Equations (3.3) must have solutions for the p_i, for all i, yielding the following secular equation (3.4) in the γ_i:

$$\begin{vmatrix} (\gamma_0 + E_{00}) & E_{01} & \cdots & E_{0m} \\ E_{10} & (\gamma_1 + E_{11}) & \cdots & E_{1m} \\ & & & \\ E_{n0} & E_{n1} & & (\gamma_n + E_{nm}) \end{vmatrix} = 0 \tag{3.4}$$

From linear algebra the system (3.3) has a nontrivial solution if and only if the determinant of coefficients (of the p_i) (3.4) equals zero. We show

this to be the case for the general birth and death queueing system from the properties of a Markov process. The rate of transition into a state always equals the rate of transition out. By definition (from equation (3.3) above), $-(E_{ii} + \gamma_i)$ equals the rate of transition into state i. Now the column sum of the determinant over all rows j, $j \neq i$ is:

$$\sum_{\substack{j=0 \\ j \neq i}}^{m} E_{ji} = \text{ rate of transition out of state } i$$

Consequently, the column sum of (3.4) is 0 for every column. It follows then that the determinant (3.4) is zero.

The secular equation is of the $(m + 1)^{st}$ order in the variable γ. Solution for the m + 1 roots γ_s yields:

(1) the zero root, $\gamma_0 = 0$: the determinant of the coefficients is zero and consequently the coefficient of the zeroth power of γ, i.e. the constant, in the secular equation is zero. Therefore, one root of the equation in these systems will always be 0.

(2) n positive roots γ_s, $s = 1,\ldots m$

When the coefficients $B_{i,s}$ are adjusted to fit the initial conditions the $p_i(t)$ are readily computable for (3.2) The presence of the zeroth root implies that the first term in the expression (3.2) must be a constant independent of time, i.e. the steady state probability P_i. The remaining n roots are therefore associated with the transient behaviour of the

system. With a constant term added to an exponential transient term, it should be noted that this behaviour coincides at least in form with the approximating diffusion result given in section 2.5.

3.2.2 Application of Morse's Methodology to the Queue M/M/1/m

For small m the roots γ_i are readily obtainable from the secular equation (3.4). In the case of large m though, Morse avoids writing (3.4) explicitly. Instead, he obtains the roots by appropriate manipulations of the m + 1 state equations.

The forward Kolmogorov differential equations, written for the single channel case are:

$$\frac{dp_0}{dt} = \mu p_1(t) - \lambda p_0(t)$$

$$\frac{dp_i}{dt} = \lambda p_{i-1}(t) - (\lambda + \mu)p_i(t) + \mu p_{i+1}(t) \qquad i = 1,\ldots,m-1 \tag{3.5}$$

$$\frac{dp_m}{dt} = \lambda p_{m-1}(t) - \mu p_m(t)$$

Assume the transient to be of the form

$$p_i(t) = \rho^{\frac{i}{2}} B_{i,s} e^{-\gamma_s t} \tag{3.6}$$

where ρ is the utilization ratio, $\frac{\lambda}{\mu}$.

Substitute (3.6) for each s, s = 1,...,m into the differential equations (3.5) to obtain the following algebraic equations:

$$\sqrt{\rho}\, B_{i,s} + (x_s - \rho)\, B_{0,s} = 0 \tag{3.7}$$

$$\sqrt{\rho}\, (B_{i-1,s} + B_{i+1,s}) + (x_s - 1 - \rho)\, B_{i,s} = 0 \qquad i = 1,\ldots,m-1 \tag{3.8}$$

$$\sqrt{\rho}\, B_{m-1,s} + (x_s - 1)\, B_{m,s} = 0 \tag{3.9}$$

where $x_s = \gamma_s/\mu$

Writing the determinant of the coefficient and solving for the m+ 1 roots of the system (3.7)-(3.9) would be tedious. Instead Morse provides simple algebraic and trigonometric manipulations that will reduce the consideration of the m + 1 equations (3.7)-(3.9) to the single equation for the s nonzero roots:

$$\gamma_s = \lambda + \mu - 2\sqrt{\lambda\mu}\, \cos\left(\frac{s\Pi}{m+1}\right) \qquad s = 1,\ldots,m \tag{3.10}$$

The derivation of (3.10) proceeds by first finding the value of the coefficients B_{is} $i = 1,\ldots,m-1$ $s = 1,\ldots m$ satisfying the m-1 equations (3.8). Fundamental to the reduction is the trigonometric identity:

$$\sin[(i-1)y] + \sin[(i+1)y] = 2\sin(iy)\cos(y)$$

Trying $B_{i,s} = \sin(iy)$, substituting in equations (3.8) yields the following equation which is independent of the state i, i = 1,..., m - 1.

$$2\sqrt{\rho}\cos(y) = 1 + \rho - x_s \tag{3.11}$$

Now, although $B_{i,s} = \sin(iy)$ does not satisfy the first boundary condition (3.7), the form $B_{i,s} = \sin(iy) - \sqrt{\rho}\sin[(i+1)y]$ does. Moreover, substitution of the latter form of $B_{i,s}$ into equation (3.7) yields the identical equation (3.11) in x_s. The second boundary condition, equation (3.9), is also satisfied by this value of $B_{i,s}$ provided we can set $\sin[(m + 1)y] = 0$. Any multiple of $\frac{\Pi}{m+1}$ would do. In particular, let $y = \frac{s\Pi}{m+1}$ where s is integer valued s = 1,...,m corresponding to the s in the subscript of $B_{i,s}$ and the m district roots $\gamma_s = x_s/\mu$ of the secular equation.

The zero root corresponds to the steady state term P_i to which the queue will relax to regardless of initial conditions. Then, except for a constant C_s chosen to suit the initial conditions, P_i^o, the m+1 time dependent state probabilities are given by the sum of the constant and transient parts:

$$p_i(t) = P_i + \rho^{\frac{i}{2}} \sum_{s=1}^{m} C_s[\sin(\frac{si\Pi}{m+1}) - \sqrt{\rho}\sin(\frac{s[i+1]\Pi}{m+1})]\, e^{-\gamma_s t}$$

For initial conditions of the following type:

$$p_i(0) = \delta_{ij} \quad i = 0,\ldots,m$$

i.e. for when there are exactly j customers in the system at t = 0, the

$$p_i^j(t) = P_i + \frac{2}{m+1}\rho^{\frac{1}{2}(i-j)} \sum_{s=1}^{m} \left(\frac{1}{x_s}\right) \left\{\sin\left[\frac{sj\Pi}{m+1}\right] - \sqrt{\rho}\,\sin\left[\frac{s(j+1)\Pi}{m+1}\right]\right\}$$

$$\left\{\sin\left[\frac{si\Pi}{m+1}\right] - \sqrt{\rho}\,\sin\left[\frac{s(i+1)\Pi}{m+1}\right]\right\} e^{-\gamma_s t} \tag{3.12}$$

For arbitrary initial queue length distributions p_j^o $j = 0,1,\ldots m$ the transient state probabilities are obtained by forming the appropriate combinations of (3.12).

$$p_i(t) = \sum_{j=0}^{m} p_j^o p_i^j(t) \qquad i = 0,\ldots,m \tag{3.13}$$

Since the γ_s can be shown to be positive for all values of ρ, it is clear that after some period of time the $p_i^j(t)$ will be composed purely of the constants P_i. Now the systems of equations (3.5) are valid for all $\rho > 0$, and furthermore the system is finite. Therefore P_i exist for all $\rho > 0$ and consequently the $p_i^j(t)$ (3.12) are similarly valid for all $\rho > 0$.

3.2.3 Closed Form Transient Solutions for M/M/1/∞ Queues

Much the same type of analysis,[1] leading to a form similar to (3.10) can be used for infinite queue systems.

We note first of all that letting the queue length $m \to \infty$ eliminates

---

1. Alternative derivations exist that use generating functions, although the resulting form of the transient state probability is such that its convergence to the known steady state results is not as obvious as before.

one of the boundary conditions (3.9) from the previous problem as the system now consists of infinitely many differential equations. Keeping the same basic presumed form of the solution as (3.2):

$$q_i(\theta,t) = \rho^{\frac{i}{2}} B_{i,\theta} e^{-\gamma_\theta t} \tag{3.14}$$

θ replacing s (as $\frac{s}{m+1}$ becomes a continuous variable for m -> ∞) and performing the same manipulation to fit the remaining boundary conditions, (3.7), (3.8), we obtain

$$B_{i,\theta} = \sin(i,\theta) - \sqrt{\rho}\,\sin[(i+1)\theta] \tag{3.15}$$

and let

$$\gamma_\theta = \omega = \lambda + \mu - 2\sqrt{\lambda\mu}\,\cos\theta$$

The initial conditions which are as before,

$$p_i(0) = \delta_{ij} \qquad i = 0,1,2,\ldots$$

can be met through the following two step process. First define the function

$$Q_i(j,t) = \frac{1}{\Pi}\left(\frac{\mu}{\lambda}\right)^{\frac{j}{2}} \int_0^{2\Pi} \sin(j\theta) q_i(\theta,t)\,d\theta \tag{3.16}$$

It can be shown that this function satisfies the following initial conditions:

$$Q_i(j,0) = \begin{cases} -1 & i = j-1 \\ 0 & i \neq j-1,j \\ +1 & i = j \end{cases}$$

The original conditions can then be met from combinations of $Q_i(j,t)$ to give, as before, the time dependent component $p'(t)$ of the transient state probabilities

$$p'^{j}_{i}(t) = \sum_{k=1}^{j} Q_i(k,t) - \sum_{k=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^{k} Q_i(k,t) \qquad j,i = 0,1,2\ldots$$

or equivalently the true time dependent state probability $p^{j}_{i}(t) = P_i + p'^{j}_{i}(t)$. Morse claims that this reduces to:

$$p^{j}_{i}(t) = P_i + \left(\frac{\mu}{\Pi}\right)\left(\frac{\lambda}{\mu}\right)^{\frac{1}{2}(i-j)} \int_0^{2\pi} [\sin(j\theta) - \sqrt{\rho}\sin[(j+1)\theta]] \tag{3.17}$$

$$[\sin(i\theta) - \sqrt{\rho}\sin[(i+1)\theta]] \frac{e^{-\omega t}}{\omega} d\theta$$

(3.17) of course resembles very closely the series representation for the M/M/1/m queue (3.12) when the integral is replaced by a finite sum of the identical function evaluated at increments of $\frac{2\Pi}{m+1}$.

Retracing a few steps to the definition (3.16) of the function $Q_i(j,t)$, there is an alternative way of expressing (3.16) that will become significant later on. For this we need two of the properties of the hyperbolic Bessel function $I_n(z)$:

$$I_n(z) = I_{-n}(z) = i^{-n}J_n(iz) = \frac{1}{2}\Pi \int_0^{2\Pi} \cos(n\theta)e^{z\cos\theta}\, d\theta \qquad n = 0,1,2,\ldots$$

Using the trigonometric identity on the product of sines an equivalent representation of $Q_i(j,t)$ (3.13) is:

$$Q_i(j,t) = \left(\frac{\lambda}{\mu}\right)^{\frac{1}{2}(i-j)} e^{-(\lambda+\mu)t}\Big\{I_{j-i}(z) - I_{i+j}(z) - \sqrt{\rho}\,[I_{j-i-1}(z) - I_{j+i+1}(z)]\Big\} \tag{3.18}$$

with $z = 2t\sqrt{\lambda\mu}$

3.3 Transient Behaviour of Model 6, M/M/1

In the previous section we developed two exact expressions for the transient behaviour of the M/M/1 queue. An approximate expression for transient behaviour was also discussed in section 2.5 for Model 4. So far our analysis has shown that for both the exact expressions of Model 6 and the approximation of Model 4 this behaviour in time is determined essentially by an exponential function of time, additive to the (constant) expected steady state value. Section 3.3 will attempt to characterize, based largely on Model 6, the parameters that are the natural descriptors of this evolution of a queue in time. The following section 3.4 will be a more general extension based on Model 4.

3.3.1 Exact Time Constant for M/M/1

The expressions for the transient state probabilities (3.12) and (3.17) immediately reveal that the time dependent terms approach 0 exponentially at a rate no slower than the smallest coefficient of t in the exponential, i.e. the smallest γ_s for (3.12), and the smallest $\omega(\theta)$ for (3.17). Knowledge of the actual rate of decay will enable us to determine the length of time T after which the transient is reduced to a certain fraction of its initial value. In particular, the time for the transient to reach $\frac{1}{e}$ of its original value is defined by Morse and others as the relaxation time. An appropriate root γ_s or value of $\omega(\theta)$ is then referred to as the time constant of the system. The

remainder of this section will be devoted to finding an expression for the relaxation time.

We will examine first the (finite) M/M/1/m queue,arriving at its time constant by examining the values of the roots γ_s from (3.10). Of all the γ_s, the smallest nonzero one is γ_1:

$$\gamma_1 = \lambda + \mu - 2\sqrt{\lambda\mu}\cos\frac{\Pi}{m+1} \approx (\sqrt{\mu} - \sqrt{\lambda})^2$$

Thereafter, as s increases, γ_s increases to its largest value $\gamma_m \approx (\sqrt{\mu} + \sqrt{\lambda})^2$. By definition (3.2), each transient state probability $p_i(t)$ is of the form $\sum_s B_{is} e^{-\gamma_s t}$. Therefore each $p_i(t)$ will have components decaying at varying rates, from the slowest $e^{-\gamma_s t}$, to the fastest $e^{-\gamma_m t}$. However, by virtue of the product $B_{is}e^{-\gamma_s t}$, the relative magnitudes of the constants B_{is} are major determinants of the rate of decay of the sum (3.2) and consequently the relaxation time. Without explicit evaluation of the constants B_{is} of (3.12) for <u>particular initial conditions</u> we have no way of knowing which of these are the largest in magnitude. We can, therefore, only infer from the above that the true value of the relaxation time is limited to being somewhere between the reciprocal of the smallest and largest roots, γ_1 and γ_m respectively:

$$\gamma_1^{-1} \approx \frac{1}{(\sqrt{\mu}-\sqrt{\lambda})^2} = \frac{(\sqrt{\mu} + \sqrt{\lambda})^2}{(\mu - \lambda)^2} = \frac{1 + \rho + 2\sqrt{\rho}}{\mu(1-\rho)^2}$$

$$\gamma_m^{-1} \approx \frac{1}{(\sqrt{\mu} + \sqrt{\lambda})^2} = \frac{1 + \rho - 2\sqrt{\rho}}{\mu(1-\rho)^2}$$

By assuming simply that the relaxation time T_r may be obtained from the arithmetic mean of the γ_1 and γ_m, we obtain:

$$T_r = \frac{1 + \rho}{\mu(1-\rho)2} \approx \frac{2\rho}{\mu(1-\rho)2} \quad \text{for } \rho \approx 1 \tag{3.19}$$

To complete the analysis of the relaxation time from the "exact" expressions of Model 6, we will discuss briefly the extensions of the above discussion to the queue M/M/1/∞. Since in deriving (3.19) we were dealing with large finite queues, we note that as m approaches infinity $\frac{s\Pi}{m+1}$ approximates more and more the continuous variable θ in (3.15). The time constant for the finite system, γ_s, therefore converges to its counterpart in the infinite queue system, $\omega(\theta)$. To show that the magnitudes of the constants $B_{i\theta}$ do not change the relaxation behaviour, it suffices to inspect the sum (3.12) and integral (3.17) as we did in the previous section to see that they do approach each other as $m \to \infty$.

Since the $p_i(t)$ of the M/M/1/m and M/M/1/∞ queues appear to converge for large m, we expect T_r to do the same. To determine T_r for the M/M/1/∞ queue, Morse obtained the auto-correlation function $\Psi(t)$

for the queue length:

$$\Psi(t) = \left[\frac{\lambda^2}{(\mu - \lambda)^2} \right] + \left[\frac{\lambda\mu}{(\mu - \lambda)^2} \right] \exp \left[-(\mu - \lambda)^2 \frac{t}{\lambda}\right]$$

and from $\Psi(t)$ the frequency spectrum of the fluctuations about the mean queue length. The relaxation time of these fluctuations T_r is described as a measure of the transient behaviour of the queue by Goddard [2], and is given by:

$$T_r = \frac{2\rho}{\mu(1-\rho)^2}$$

which is identical to (3.19).

3.3.2 Approximate Time Constant for M/M/1 from Model 4

We now have, for the very specific case M/M/1, an "exact"[1] expression (3.19) for the relaxation time. The aim now will be to see how (3.19) compares with the relaxation time that we could derive from the approximate Model 4.

---

1. We use "exact" in quotations since it is really an approximation from the exact closed form transient $p_i(t)$.

Consider the transient component of $E[W_d(t)]$ for Model 4:

$$\frac{(g^2\lambda)^{\frac{3}{2}}}{2(1-g_1\lambda)^2 t^{\frac{1}{2}}} \; e^{\frac{-(1-g_1\lambda)^2 t}{2g_2\lambda}} \tag{3.20}$$

Substituting the values of g_1 and g_2 for the M/M/1 case into (3.20) we obtain the expression in t, which we define as R_d (ignoring the the constant coefficient):

$$t^{-\frac{1}{2}} e^{-\frac{(1-\rho)^2\mu^2 t}{2\lambda 2}} = t^{-\frac{1}{2}} e^{-\frac{(1-\rho)^2\mu t}{4\rho}} = R_d \tag{3.21}$$

Basically R_d differs from the form (3.2) by the presence of the leading term $t^{-\frac{1}{2}}$. Fortunately, however, we will be able to show that this behaviour does indeed still coincide with the "exact" results.

For the purpose of obtaining an expression R_e similar to (3.21) from Model 6 we return to the Bessel function form (3.18) of the $p_i(t)$. The two forms (3.16) and (3.18) are equivalent by construction and must clearly exhibit the same behaviour. To show, however, that we obtain a somewhat different relaxation time, consider the asymptotic behaviour of the hyperbolic Bessel function:

$$I_n(Z) \to \frac{e^Z}{(2\Pi Z)^{\frac{1}{2}}}\left[1 - \frac{n^2 - \frac{1}{4}}{2Z} + \ldots\right] \to \frac{e^Z}{(2\Pi Z)^{\frac{1}{2}}} \quad \text{as } Z \to \infty$$

Therefore, the $Q_n(m,t)$ used to build up the $p_n^m(t)$ behave as

$$e^{-(\lambda+\mu)t} \frac{2\sqrt{\lambda\mu}\,t}{(2\Pi \cdot 2\sqrt{\lambda\mu}\,t)^{\frac{1}{2}}} \propto t^{\frac{1}{2}} e^{-(\sqrt{\mu}-\sqrt{\lambda})^2 t} \equiv R_e \tag{3.22}$$

The result R_e,(3.22), is another expression for the "exact" transient behaviour of the M/M/1 queue. As ρ is comparable with 1, and $\sqrt{\rho} \approx \rho$, we see from (3.21) and (3.22) that

$$R_e = t^{\frac{1}{2}} e^{-(\sqrt{\mu} - \sqrt{\lambda})^2 t} \approx t^{-\frac{1}{2}} e^{-\left[\frac{(\mu-\lambda)^2}{\mu+\lambda+2\sqrt{\mu\lambda}}\right]t} \approx t^{-\frac{1}{2}} e^{-\frac{(1-\rho)^2\mu}{4\rho}t} = R_d \tag{3.23}$$

We conclude from (3.23) the remarkable fact that for the M/M/1 queue the approximating Model 4 exhibits the same transient behaviour as the exact Model 6. This is tremendously important as it leads us to expect that the transient behaviour is modelled "exactly" by Model 4 for all types of M/G/1 queues.

Before we can attempt to prove that Model 4 does indeed correctly

predict transient behaviour, we need workable expressions for the relaxation time. This task is complicated by the presence of the $t^{-\frac{1}{2}}$ term in (3.21). We noted in (3.23), however, that T_d , the Model 4 relaxation time, given by the inverse of the coefficient of t in the power of e,

$$T_d = \frac{4\rho}{(1-\rho)^2\mu}$$

was the same for both (3.21) and (3.22). For the M/M/1 queue the relation of T_d to T_r is given in (3.24)

$$T_d = \frac{4\rho}{\mu(1-\rho)^2} = 2 \cdot \frac{2\rho}{\mu(1-\rho)^2} = 2T_r \tag{3.24}$$

We might believe that a similar simple relation holds between the Model 4 relaxation time T_d and the time value T_r for all queues of the type M/G/1. This would considerably simplify the analysis of transient behaviour for M/G/1 queues other than M/M/1. Section 3.4 will present a result indicating that T_r does indeed describe, within a constant, the transient behaviour of Model 4. Numerical results presented in Chapter 4 will also confirm (3.24) for a few different M/G/1 queues.

3.3.3 Properties of the Relaxation Time

We conclude the section with a few notes on the basic properties of the relaxation time we have obtained. The first concerns the fact that its value, depends not only upon the dimensionless utilization ratio, but also the mean service rate. The second is that (3.19) for T_r is valid for all $\rho > 0$, exhibiting similar behaviour for both underloaded and overloaded queues. The relaxation time is therefore approximately symmetrical around $\rho = 1$. As ρ increases T_r grows rapidly, peaking at $\rho = 1$, and then gets smaller as ρ becomes much greater than 1. The actual peak value of T_r has been approximated by Morse

$$T_r \Big|_{\rho=1} = \frac{(m+1)^2}{\Pi^2} \quad \text{for } M/M/1/m \text{ queue}$$

Since we said earlier that the T_r is related to the autocorrelation function of the queue length we present a quick summary (for M/M/1) of the expected number of coustomers in the system Q, and the mean square fluctuation $(\Delta Q)^2$ about Q (for the steady state, M/M/1/m queue)

$$Q = \sum_{i=0}^{m} ip_i \rightarrow \begin{cases} \rho + \rho^2 & \rho \ll 1 \\ \frac{1}{2} m(m+2) & \rho \rightarrow 1 \\ m - (\frac{1}{\rho}) & \rho \gg 1 \end{cases}$$

$$(\Delta Q)^2 = \sum_{i=0}^{m} i^2 p_i - L^2 \rightarrow \begin{cases} \rho + 2\rho^2 & \rho \ll 1 \\ \frac{1}{12}(m+2) & \rho \rightarrow 1 \\ (\frac{1}{\rho}) + (\frac{2}{\rho^2}) & \rho \gg 1 \end{cases}$$

we observe that ΔQ for the large values of m that we are concerned with becomes very large as $\rho \rightarrow 1$. Furthermore, these fluctuations are approximately symmetric around $\rho = 1$. It would seem that as the fluctuations ΔQ around Q become large the relaxation time of the queue length to Q would likewise become larger. This was born out by (3.19).

Since T_r is strictly valid only for finite queues (although m may be very large) when $\rho > 1$, it is interesting to observe the transient behaviour of Model 4 which is an infinite queue system. The $E[W_d(t)]$ (2.6)-(2.8) show that the transient part for $\rho < 1$ is identical to the transient part for $\rho > 1$. This not only confirms that Model 4 approximates the exact transient behaviour very well, i.e. one expression T_d identical to T_r independent of ρ, but also the assumption that large finite queues behave very much like infinite queues. The latter, which has been emphasized throughout this section will be a key assumption in Model 8.

3.4 Transient Analysis of Model 4, M/G/1

Model 4 actually offers much more information than is presented in section 2.5. Both the underlying steady state and transient behaviours of the M/G/1 queue become clear when we reduce the diffusion equation (2.5) to dimensionless form. We will show that there exist natural descriptors inherent to this equation that justify further (3.24) as an alternative characterization of the transient of Model 4.

We recall that in setting up the parameters a(t) and b(t) of the governing partial differential equation (2.5) of Model 4, only two items of information about the pdf of the service process were actually used - the mean rate and mean square rate of service. We point out here that by a simple rescaling of these coordinates we can obtain a dimensionless equation depending neither on a(t) nor b(t). Solution of the dimensionless equation then provides, under appropriate variable transformations, $E[W_d(t)]$ for any choice of service process. These variable transformations will be performed by choosing new units of time and waiting time in queue:

$$t' = \frac{t}{T_0} \text{ and } x' = \frac{x}{W_0}$$

where T_0 and W_0 are defined respectively as the characteristic time and waiting time in queue. The differential equation (2.5) then becomes

$$\frac{\partial f}{\partial x'} = \frac{T_0 a(t)}{W_0} \frac{\partial f}{\partial x'} + \frac{T_0 b(t)}{2W_0^{\,2}} \frac{\partial^2 f}{\partial x'^2} \tag{3.25}$$

We render the coefficients of the equation (3.25) dimensionless by choosing:

$$\frac{T_0 a(t)}{W_0} = 1 \quad \text{and} \quad \frac{T_0 b(t)}{W_0^{\,2}} = 1 \tag{3.26}$$

i.e. $W_0 = \dfrac{b(t)}{a(t)}$ $\qquad T_0 = \dfrac{b(t)}{[a(t)]^2}$

The resultant equation

$$\frac{\partial f}{\partial t'} = \frac{\partial f}{\partial x'} + \frac{1}{2} \frac{\partial^2 f}{\partial x'^2} \tag{3.27}$$

has been solved by Gaver [1], Harris [6], and Newell [11], with an extensive tabulation provided by Harris. As it is much faster to scale variables and look up tables, rather than solve a new partial differential equation each time, there is considerable computational advantage to using the equation (3.27) to solve for $E[W_d(t)]$ of the M/G/1 queue. In Chapter 4 we will make use of this property to compare the exact solutions of various special cases of the M/G/1 queue to the results of Model 4.

Returning to the nondimensional equation (3.27), since there are

no parameters in it other than constants of order 1, its qualitative behaviour can be described by inspection. Given an initial value of the nondimensional waiting time in the queue of order $x' = 1$, or some non-equilibrium distribution of waiting times over a range of x' of order 1, then the relaxation to the equilibrium waiting time distribution must occur within a time t' of order 1.

Using this qualitative observation we can proceed to an interpretation of what this means in terms of our original values of time t and waiting time x. In order to do that though we must take a closer look at the variable transformations (3.26). We already have from the Pollaczek-Khintchine formula (2.1) the relation for steady state waiting time for E[W] for the M/G/1 queue. Now

$$W_0 = \frac{b(t)}{a(t)} = \frac{g_2\lambda}{1-g_1\lambda} = 2E[W]$$

indicating that what we described earlier as the characteristic waiting time for (3.25) is exactly, except for a constant, equal to E[W]. Therefore, since we showed in section 2.5 that $E[W_d(\infty)] = E[W]$, we have

$$W_0 = 2E[W] = 2E[W_d(\infty)]$$

$$T_0 = \frac{b(t)}{[a(t)]^2} = \frac{g_2\lambda}{(1-g_1\lambda)^2} \tag{3.28}$$

Similarly, in the previous section 3.3 we already showed that the characteristic value T_0 of (3.25) coincided with the "exact" relaxation time result (3.19) for the special case of the M/M/1 queue. On the basis of the relation of W_0 to (2.1), T_0 to (3.19), we have cause to believe that there exists an analogous form for the relaxation time of the M/G/1 system as exists for E[W] in the form of (2.1). Again, the probable validity of this inference will be shown by numerical comparisons in Chapter 4.

Having now made W_0 and T_0 precise by setting $x' = t' = 1$, we obtain the result that the equilibrium waiting time

$$x = W_0 \text{ is proportional to } (1-g_1\lambda)^{-1} = (1-\rho)^{-1}$$

whereas the relaxation time

$$t = T_0 \text{ is proportional to } (1-g_1\lambda)^{-2} = (1-\rho)^{-2}$$

The relaxation time grows much faster than the equilibrium queue length. It is evident that in any situation with ρ close to 1 the transient component is significant for a very long time.

3.5 Model 7: Approximating Method for the Time Dependent Queues, M/M/k, M/D/k

The following method for solving the time dependent queuing problem depends strongly on the knowledge of transient behaviour. It was suggested by Gupta [4] as an approximate scheme for cases where the values of the individual state probabilities were not needed. The method proceeds by defining a differential equation in the time dependent expected queue length q(t). It then draws upon intuition and steady state queuing results to fit the solution of the differential equation to the transient behaviour determined analytically. Exactly as in Koopman's method, two equations are derived for M/M/k and M/D/k respectively.

In addition to the usual "quantity" $\rho(t)$ we also need to define the following quantity s(t), which plays a key role in the development of this approach:

$$s(t) = \frac{S(t)}{\mu(t)}$$

where $S(t)$ is the time dependent actual service rate of the system. In defining these terms we note that the parameter of the service process, $\mu(t)$ represents the average service rate when the server is busy. Even when the expected queue length is large though there remains a finite probability that the server remains idle . Therefore, in an expected value sense the time dependent throughput of the server, s(t), will always be less than the maximum rate $\mu(t)$. From this we infer the first

property of s(t) - that it is always less than 1. Dropping the t dependence for notational simplicity we have the differential equation in the state variable of the model:

$$\frac{dq}{dt} = \lambda - s \quad \text{or equivalently} \tag{3.29}$$

$$\frac{dq}{dt} = \mu(\rho - s)$$

The only other obvious property of s is observed from the differential equation under steady state conditions:

$$\frac{dq}{dt} = \mu(\rho - s) = 0 \qquad \Rightarrow \qquad \rho \equiv s$$

From steady state results presented earlier it is easy to show that for infinite queue systems, independent of the queuing process, the steady state probability of having the server idle is $p_0 = 1 - \rho$. Therefore, s is at least a function of ρ. Now analytic expressions show that in the steady state the queue length is a function of ρ and the number of servers, k:

$$q_{\text{steady state}} = Q(\rho,k)$$

Gupta assumes that the steady state queue length is independent of the number of servers,

$$Q(\rho,k) = Q(\rho) \qquad\qquad k = 1,2,\ldots$$

Assume then that the time dependent $s(t)$ is some function of the instantaneous values of ρ and q. (This is an approximation much the same as the diffusion approximation makes - s is in general dependent on the entire history of the model. However, when the queue lengths are large compared to 1, then the success of the diffusion equation verifies the effectiveness of the assumption that the movement of the differential equation is governed by "local" behaviour). Gupta suggests the following as a possible form of s:

$$s = s(\rho,q) = p\xi(q) + (1 - p)\rho \tag{3.30}$$

where p and $\xi(q)$ are two functions yet to be determined.

To suit the boundary condition $s(\rho,Q(\rho)) = \rho$, define $\xi(q) = \rho$ when $\frac{dq}{dt} = 0$. $\xi(q)$ then just turns out to be the inverse function of $Q(\rho) = q$. The steady state solution $Q(\rho) \geq 0$ exists and is monotone increasing for all $0 \leq \rho < 1$ with $\lim_{\rho \to 1} Q(\rho) = \infty$. By the property of inverse functions then $\xi(q)$ exists and is monotone increasing for all $q \geq 0$ with $\lim_{q \to \infty} \xi(q) = 1$. Therefore by construction, for $\rho < 1$ in the steady state, $s(\rho,q) = s(\rho,Q(\rho)) = \rho$ as required. On the other hand, a very large queue need not be associated with the steady state result for some $\rho < 1$. In rush hour situations it is common for ρ to exceed 1 temporarily . A large queue could then be temporary for some $\rho > 1$.

s is now given by

$$s(\rho,q) = p\xi(q) + (1 - p)\rho \lesseqgtr 1 \qquad \rho \geq 1$$

This may violate the first condition that $s \leq 1$. To correct for this we arbitrarily modify the form (3.30) of s to:

$$s = p\xi(q) + (1 - p)\rho' \qquad \rho' = \begin{cases} \rho & \rho \leq 1 \\ 1 & \rho > 1 \end{cases} \tag{3.31}$$

For stationary $\rho \geq 1$, $\lim_{t\to\infty} s(t) = 1$ which satisfies the intuitive notion that if ρ has been large for a long time then the probability of the server being idle converges to 0.

So far it is not clear what the form of p should be. We expect however, that s exhibits a transient behaviour similar to that of $p_0(t)$, the time dependent probability of having the server idle. Therefore, p must exhibit the behaviour of the time constant for the system. Substitution of (3.31) into the differential equation (3.29) yields:

$$\frac{dq}{dt} = \mu p \left\{ \rho - \xi(q) \right\} \tag{3.32}$$

Let $q = Q(\rho)$, then $\xi(q) = \rho$ and the steady state condition is of

course

$$\frac{dq}{dt} = \mu p \left\{\rho - \rho\right\} = 0$$

Otherwise, assume that the queue length q is sufficiently close to q_0, $q = q_0 + q'$ such that the following approximation holds:

$$\xi(q_0 + q') \simeq \xi(q_0) + q' \left.\frac{d\xi}{dq}\right|_{q=q_0} = \rho + \frac{q'}{\left.\left(\frac{dQ}{d\xi}\right)\right|_{\xi=\rho}}$$

Substituting this in the differential equation (3.32) now for q',

$$\frac{dq'}{dt} = \mu p\left\{\rho - \rho - \frac{q'}{\left.\left(\frac{dQ}{d\xi}\right)\right|_{\xi=\rho}}\right\} = - \frac{\mu p q'}{\left.\left(\frac{dQ}{d\xi}\right)\right|_{\xi=\rho}}$$

The solution for this equation is

$$q' = e^{-\frac{\mu p t}{\left.\frac{dQ}{d\xi}\right|_{\xi=\rho}}}$$

Thus, transient effects of the expected queue length q' decay exponentially to 0 with a relaxation time T_G:

$$T_G = \left(\frac{dQ}{d\xi}\right)\Bigg|_{\xi=\rho} \Big/ \mu p \qquad \text{where p is still an unknown}$$

quantity. Further consideration of the model now depends on forehand knowledge of the two unknown quantities in the expression for p:

1) $\frac{dQ}{d\xi}$ the behaviour of the steady state queue length with the utilization ratio ρ.

2) T_G, the analytically or empirically determined values of the relaxation time.

The first set of conditions is easily met. We already have from Chapter 2 the Pollacek Khintchine formula (2.1) which yields $Q(\rho,1)$ for arbitrary service processes and in particular the result that

$$Q(\rho,1)_{M/D/1} = \frac{1}{2} Q(\rho,1)_{M/M/1} \qquad [E[W] = \frac{Q}{\lambda}]$$

Although no similar result is available for comparisons of M/D/k and M/M/k systems, the steady state expressions E[W] are available and were given in Chapter 2.

The description of the transient behaviour is not as clearly determined. We have already seen that the relaxation time $T_r = T_r(\mu,\rho)$ and from (3.12) and (3.17) that this may be dependent on initial conditions. The conclusion of the former sections was that $T_{r_{M/M/1}}$ is approximated by

$$T_{r_{M/M/1}} \simeq \frac{2\rho}{\mu(1-\rho)^2}$$

and we can infer from (3.28) that the corresponding value for the deterministic service case is:

$$T_{r_{M/D/1}} = \frac{1}{2} T_{r_{M/M/1}} = \frac{\rho}{\mu(1 - \rho)^2}$$

Now at least for the one runway case, the Gupta model calibration may be completed. Noticing that $Q_{M/D/1} = \frac{\rho^2}{(1-\rho)} = \frac{1}{2} Q_{M/M/1}$ and $T_{r_{M/D/1}} = \frac{1}{2} T_{r_{M/M/1}}$ we find that substitution into the expression for p yields the same form independent of the service process

$$p = \frac{\left(\frac{dQ}{d\xi}\right)\Big|_{\xi=\rho}}{\gamma_t \mu} = 1 - .5\xi(q) \qquad (3.33)$$

Gupta assumed that all the relationships holding for single server queues held for multiple server systems. When programmed and compared to exact solutions from Model 5, the Gupta model was found to corroborate these assumptions except under certain conditions. It was found that whenever a large expected queue length existed, and conditions were such that q was decreasing, then the queue length in Gupta's model always decreased at a slower rate than the exact solution. Gupta supplied an arbitrary fix for this discrepancy by modifying (3.33) as follows: whenever $\xi(q) \geq 0.85$ (i.e. queue length is large) <u>and</u> q is decreasing, use p = 0.9, otherwise use p as in (3.33). This was found to work well.

Although the fix is somewhat arbitrary it does have the effect of reducing the relaxation time T_G and thus causing the queue length to decrease at a faster rate.

3.6 Model 8: A New Model for the Time Dependent Queue, M/M/1

3.6.1 Introduction

The contention of this section is that the behavioural aspects of the relaxation time presented up to now could be helpful in developing approximating schemes for the time dependent M/G/k system. Model 7 is one such example. The advantage of that model is its significant reduction of computation time over the exact method of Model 5. The disadvantages, however, are due to the fact that the model does not provide the individual (time dependent) state probabilities which are needed to calculate any of the following:

(a) prob (queue length $< x_0$ at t_0)

(b) expected number of operations cancelled due to lack of waiting space (queue saturation)

(c) the (time dependent) variance of the queue length

Therefore, the technique we will suggest here aims to determine an alternative method to Model 5 for determining the time dependent state probabilities by successively evaluating the transient state probabilities over the period of interest.

3.6.2 Methodology

In evaluating the transient state probabilities we can only use results for the queues for which the $p_i(t)$ are available in closed form,

namely M/M/1. The method to be proposed here evaluates an equivalent form of the $p_i(t)$ (3.17) for the M/M/1/∞ queue, given by Clarke[15]. The form given by Clarke, as we were able to show with Morse (3.18), contains Bessel functions, but the two are not the same, and Clarke's form is considerably easier to compute.

Let $I_k(v)$ be the modified Bessel function as given in the alternative form (3.18) in terms of Bessel functions for (3.17). The probabilities $p_i^j(t)$ have the same interpretation as in (3.12) and (3.17), and are given by:

$$p_i^j(t) = e^{-(\lambda + \mu)t} \left[(\mu/\lambda)^{j-i} I_{j-1}(v) + (\mu/\lambda)^{j+i+1} I_{j+i+1}(v) + P_i \sum_{k=j+1+2}^{\infty} (\mu/\lambda)^k I_k(v) \right]$$

where $v = 2\sqrt{\lambda\mu}t$ $\qquad$ $i,j = 0,1,2...$ $\qquad$ (3.34)

and $P_i \equiv$ M/M/1/∞ steady state probability $= (1 - \rho)\rho^i$.

As given earlier, arbitrary initial queue length distributions can be met by computing

$$P_n(t) = \sum_{i=0}^{\infty} P_i^0 \, P_n^i \, (t) \qquad n = 0,1,2...$$

We note, however, the presence of the steady state probabilities P_i in (3.34). This is the greatest obstacle to the implementation of (3.34) as the P_i do not exist for M/M/1/∞ when $\rho > 1$. Since a large part of the

discussion so far has included situations with temporary oversaturation, it is crucial that we show how to overcome this.

Fortunately, we have the developments of the two cases M/M/1/m and M/M/1/∞ given in Model 6. There we were able to show that for large m the form (3.12) converges to (3.17), separately in both the steady state and transient components. We believe therefore that the substitution of P_i^m (state probabilities for the finite queue M/M/1/m, in the steady state) for P_i into (3.34), provided we make m large enough, yields a justifiable approximation of M/M/1/∞ in the transient state. The actual concept of making m large enough is not difficult to justify either, because there exists a k such that for all i > k the $p_i(t)$ remain less than a specified small probability and contribute only a negligible amount to E[W(t)]. We call Clarke's form with large finite m, and P_i^m substituted for P_i, the <u>modified Clarke's form</u>.

3.6.3 Computation Requirements

The computational aspects of the modified Clarke's form are much more attractive than either of the series form for M/M/1/m (3.12) or the integral form for M/M/1/∞ (3.17) because of the expression in the $I_k(v)$, and the need for considerably fewer multiplications. There exists for the modified Bessel function $I_k(v)$ a simple recursion formula

$$I_{k-1}(v) - I_{k+1}(v) = \frac{2k}{v} I_k(v) \tag{3.35}$$

and the relation $I_{-k}(v) = I_k(v)$

In order to use the recursion relation (3.35), it would seem that we need to compute at least two of the infinite sums $I_k(v)$ to get the recursion started. We show, however, that in fact there are no infinite sums to be computed.

The reason for this is that there are severe roundoff problems incurred by using (3.35) in the increasing k direction. Instead that for $k > v$, the $I_k(v)$:

$$I_k(v) = \sum_{i=0}^{\infty} \frac{(v/2)^{k+2i}}{i!(k+i)!}$$

is a decreasing function, $\lim_{k\to\infty} I_k(v) = 0$. Therefore, the following method for computing $I_k(v)$, due to Morse [9], should be employed. Let q be the number of significant digits required in the final result, and m be the largest value of k for which $I_k(v)$ is needed. Set

$$n = m + q, \qquad I'_n(v) = 0, \qquad I'_{n-1}(v) = 10^{-q} \tag{3.36}$$

where $I_k(v)$ is defined as $I'_k(v) \equiv AI_k(v)$, where A is some constant. Using the recursion formula (3.35), and the initial conditions (3.36), compute $I'_k(v)$, $k = n-2, n-3, \ldots 0$. To obtain from the $I'_k(v)$ the value of $I_k(v)$ consider the normalizing equation:

$$e^v = I_0(v) + 2I_1(v) + 2I_2(v) + \ldots$$

to obtain

$$e^v = \frac{1}{A} \; I_0'(v) + 2I_1(v) + 2I_2'(v) + \ldots$$

Consequently,

$$I_k(v) = \frac{I_k'(v)e^v}{I_0(v) + 2I_1'(v) + 2I_2'(v) + \ldots} \tag{3.37}$$

In computing the values (3.37) it is possible to simultaneously compute

$$(\sqrt{\mu/\lambda})^k I_k(v)$$

After that, essentially all that remains to obtain a numerical value of (3.34) is a sequence of additions which clearly are more easily performed than the multiplications needed in (3.12) and (3.17).

3.6.4 The Time Dependent Model

We have shown that it is computationally feasible to evaluate the transient state probabilities (3.34). The application of this process yields for constant ρ,μ the values of the $p_i(t)$ at an arbitrary period of time τ after t_0, under arbitrary initial conditions:

$$p_i(t_0) \quad i = 0,1,\ldots,m \quad \text{satisfying} \sum_{i=0}^{m} p_i(t_0) = 1 \tag{3.38}$$

To use this method in a time dependent situation we propose the discretization of the time period of interest into intervals of lengths t_i (not necessarily all equal). Once the time intervals are selected, a representative value of ρ and μ for each interval should be chosen and the transient state probabilities for the end of the interval evaluated by (3.34). The outcomes (3.39)

$$p_i(t_0 + t_i) \qquad i = 0,1,\ldots m \tag{3.39}$$

for each interval form the initial conditions (3.38) for the next. Note that if we make the intervals small enough, e.g. 1 minute, we are in effect obtaining the same figures as those from a direct solution of the C-K equations. It is clear that it is not desirable to do this, though, for the value of the model lies not in high accuracy estimates but rather in its ability to sketch out quickly the nature of the queuing statistics. This is done by assuming λ is constant for longer time intervals and estimating the delay from the observed transient behaviour (based on constant λ). In comparison to Model 5, Model 8 overestimates delay in the increasing λ situations and underestimates it for decreasing λ. We expect, therefore, the total delay for a typical demand profile of an airport to be approximately the same whether analyzed by Model 5 or Model 8. By employing Model 8 it is possible to save much computation time in the evaluation of queuing statistics for periods of the day not likely to

need much scrutiny. However, if more accurate figures are needed, Model 8 provides the queue length distribution at the end of each time interval. This is the essential piece of information needed if it is desired to pass to more accurate study of certain periods by using, say, Model 5.

Chapter 4

NUMERICAL EVALUATION OF THE MODELS AND THE TIME CONSTANT

4.1.1 Introduction

In this chapter specific numerical results illustrating the behaviour of transient terms and its implications on modelling are presented. There will be two areas of concentration:

(1) A case study for which all of the models discussed in the thesis are applicable is presented. Some attention is first given to the construction of the input demand profile from the data. Then the relative merits of the different methods are evaluated. The new Model 8 is tested on the case and compared to the results obtained with Model 5. Results of a first order method, Model 3 provide an opportunity to see how the effects of transient behaviour can be used in constructing simpler models.

(2) Analysis of the transient behaviour in stationary queues. We start by considering the single server queue. The theoretical results available for T_r from Model 6 for M/M/1 and from Model 4 for M/G/1 are reviewed and compared to the exact results obtained by numerical evaluation of the C-K equations of these systems with Model 5. To justify use of T_r derived from the approximating Model 4 we provide numerous examples comparing Model 4 to the exact results of Model 5. Having the theoretical value of T_r and using Model 5, we then study the behaviour of the transient under varying initial conditions.

We then study the multiserver queue and observe the deviations of T_r for the multiserver queue from that of the single server queue. This will be based entirely on numerical results from Model 5.

4.2 A Case Study and Comparisons of Results

4.2.1 Introduction

To illustrate the applicability of the models we elaborate on the following case study from Scalea [13]. Schiphol is a large international airport in Amsterdam, Holland servicing mainly the scheduled airlines. Currently the airport operates two independent parallel runways, one exclusively for departures and the other for arrivals. The data used in the case study comes from the Schiphol airport authority who maintain precise records on each operation.[1]

There are two central issues to be resolved before passing to analyses with the models: (a) whether the assumptions of the models are satisfied, and (b) what method should be used to construct a demand profile from the raw data.

The major assumptions for the stochastic models are that the arrival process is Poisson and that the servers are independent and identical. We indicated in section 2.1 that the "arrivals" into the queue of landing aircraft satisfy the Poisson assumption to a much greater degree than the pattern of "arrivals" into the takeoff queue. Therefore, we concentrate in the case study on the single independent runway for landings. Also, it is expected that the aircraft service time for arrivals is more consistent with the negative

---

1. In the United States the F.A.A. currently maintains on-line, all schedule data for airlines listed in the OAG (Official Airline Guide) for domestic airports. Additionally, the F.A.A. provides general aviation (GA) factors where applicable. The GA factor is a percentage which relates estimated GA activity to air carrier activity.

exponential service process than the departures' would be. Therefore, as a predominantly M/M/1 case, the Schiphol arrivals runway is also an ideal situation for testing the validity of Model 8.

Issue (b) is closely related to (a) because of the assumption of Poisson arrivals. The probabilistic models need the average arrival rate $\lambda(t)$ of what is assumed to be the governing inhomogeneous Poisson process. The raw airport data give the estimated arrival time of the aircraft. For all the reasons given in section 2.1 we might expect that the actual arrival time will have a substantial variance, Steuart [14]. The deterministic models ignore this, of course, and need nothing but a cumulative arrival count over the period of interest. This is trivial to obtain from the data of Table 4.1 (where the first column lists the arrivals at Schiphol).

STA	Flight	Route	RMP/GTE	REG	Remarks	
0120	KL017	STR HAJ	VR1/	DNOIG	FL	K
0235	SR798	BSL	VR2/	D98	FL	K11
0240	KL055	GOT ARN	VR2/	PHDNN	FL	K
0250	SK051	CPH	VR2/	LI88	FL	K11
0315	DG807	LBV	VR2/	D855	FC	A
0340	SM100	LGW	VR1/	DC3	FL	K11

Sample of Data Maintained by Schiphol Airport

Table 4.1

To compute an average $\lambda(t)$ though, the simplest method involves splitting the day into a mesh of equal segments (such as hourly intervals), counting up the number of aircraft arriving in each interval, and letting this be the mean arrival rate, constant throughout the corresponding time interval. We call this Method α. Method α may be undesirable if we believe that the observed demand peaks are more frequently gradual changes rather than step changes. Therefore, the alternative is to compute the average hourly arrival rate, fix $\lambda(t)$ as this value at some point in the interval, and interpolate between the $\lambda(t)$ in adjacent intervals for all other values of t. We call this Method β.

It is still possible, though, that the selection of the mesh size is inappropriate for either method, i.e. hourly intervals might simply be too long to indicate all the traffic peaks actually observed. Certain scheduling practices or desirable arrival times may cause activity peaks that are hidden because of the averaging over a large mesh.[1] On the other hand, shrinking the time interval affects the assumptions of the model. Precisely for the reason that the demand fluctuates wildly, arrivals in small time intervals may no longer be statistically independent.

For the purposes of the case study we will assume that meshes of 15, 30, and 60 minutes do satisfy the statistical independence assumption. Decreasing the mesh size from 60 to 15 minutes will give us a handle on the delays caused by convenience scheduling practices. In the event that statistical independence does not hold, we still have an upper bound on the delay. The substantial differences that can exist between the demand profiles when 10 and

---

1. This is particularly true of departure situations. "On the hour" departure times, for instance are conveniently remembered by passengers. The "quarter hours" also exhibit these demand peaks.

60 minute meshes are used are shown in Figure 4.1 for the case of the Atlanta airport (one of the busiest in the world).

We wish then to perform comparisons of delays for the single runway with the following models:

(4.2.2) Model 3: Equilibrium Analysis for M/M/1

(4.2.3) Model 8: Numerical Evaluation of the Closed Form Transient Solution, M/M/1

(4.2.4) Model 5: Numerical Evaluation of the Time Dependent C-K Equations, M/M/1

4.2.2 Model 3: Equilibrium Analysis for M/M/1

The 60 minute mesh demand profile for Schiphol is given in Figure 4.2.[1] It is clear that since equilibrium analysis is based on the assumption that demand is constant for substantial periods of time, no mesh size less than one hour could be adequate (except in cases where ρ is always much less than one). Even so, from the strong time dependency of the data, it is unlikely that equilibrium analysis is at all amenable to delay calculation in these situations. However, in view of our progress in understanding the transient, we now possess a means to interpret correctly steady state results, should that be our only recourse in delay calculation.

E[W] for each hour can easily be evaluated for $\rho < 1$ with the Pollaczek-Khintchine formula (2.1). This yields for the M/M/1 queue of the case study:

---

1. Actually, Scalea elected to give results for twice the observed demand. The real level of demand would have been too low to exhibit significant queueing delay.

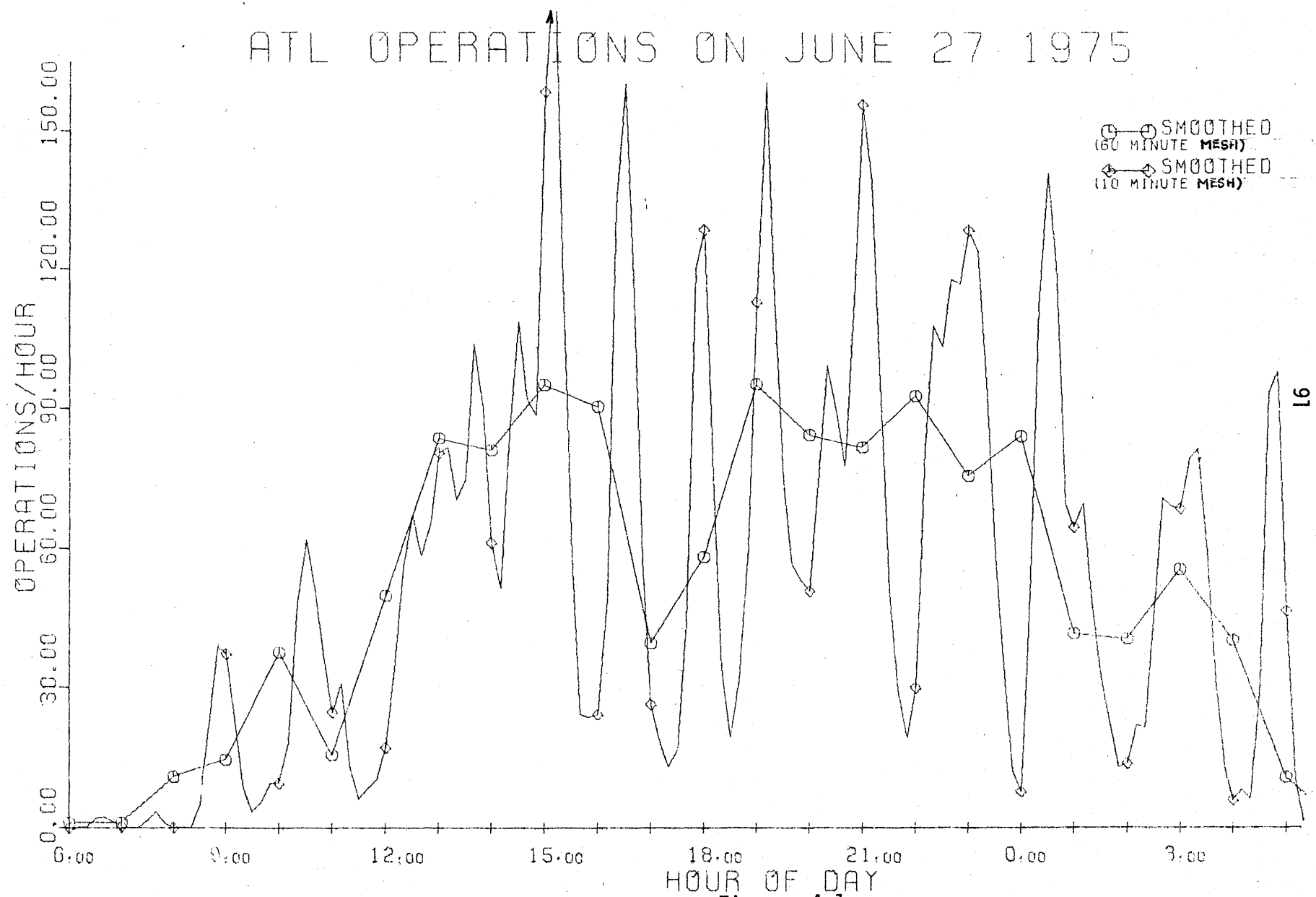
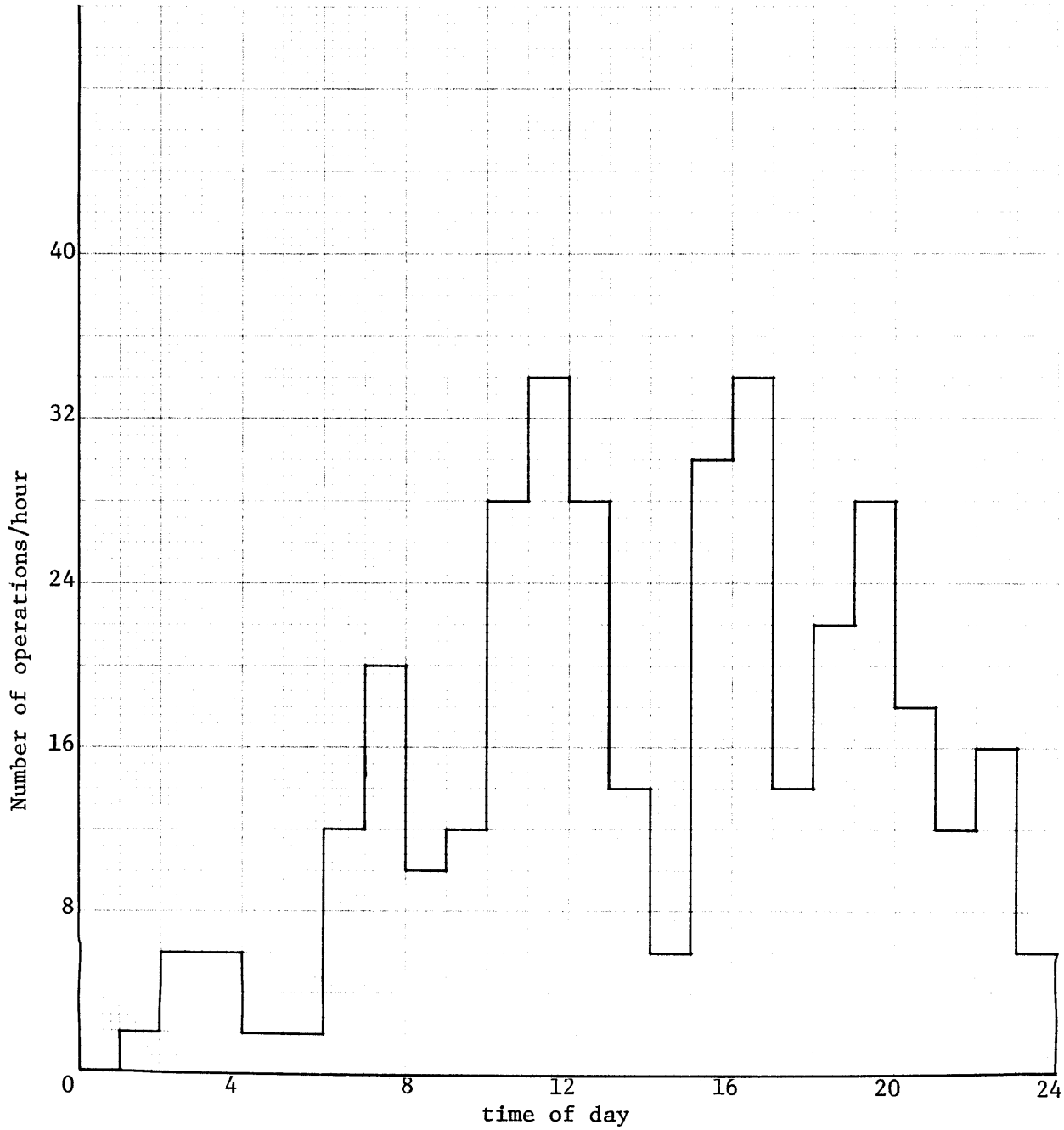

ATL OPERATIONS ON JUNE 27 1975
SMOOTHED
(60 MINUTE MESH)
SMOOTHED
(10 MINUTE MESH)
OPERATIONS/HOUR
0.00
30.00
60.00
90.00
120.00
150.00
6:00
9:00
12:00
15:00
18:00
21:00
0:00
3:00
HOUR OF DAY

Figure 4.1



Twice the number of arrivals recorded at Amsterdam's Schiphol airport on March 11, 1976

Figure 4.2

$$E[W] = \frac{\rho}{\mu(1-\rho)} \qquad \rho < 1$$

Alternatively, if it is felt that service time deviates significantly from the negative exponential pdf the following relations for steady state waiting times can be used:

$$E[W]_{M/D/1} = \frac{1}{2} E[W]_{M/M/1}$$

$$E[W]_{M/E_c/1} = \frac{c+1}{2c} E[W]_{M/M/1}$$

When $\rho > 1$ no equilibrium results can exist. For the purpose of illustration we construct a procedure that parallels the diffusion equation results for $\rho > 1$ and t very large. Let the delay in an interval of oversaturation be given by the steady state delay for the immediately preceding interval (if this is "close" to saturation) plus the term $(\rho-1)\Delta t$, where Δt is the mesh size. If the hour immediately preceeding has very low delay use the diffusion result directly.

$$E[W] = \begin{cases} \dfrac{\rho}{\mu(\rho-1)} + (\rho-1)\Delta t & \rho > 1 \\ \sqrt{\dfrac{4\rho\Delta t}{\mu\Pi}} & \rho \simeq 1 \end{cases}$$

Using this procedure we obtained the results of Table 4.2 and the expected daily delay of 4757 minutes. We would expect this, however, to be way too high. We know that delays increase as $(1-\rho)^{-1}$ but that the relaxation time increases even faster as $(1-\rho)^{-2}$. During the peak hours

hour beginning	one hour mesh		two hour mesh	
	opns/hr	E[W] mins/op	opns/hr	E[W] mins/op
0	0	0.000	1	0.069
1	2	0.143	1	0.069
2	6	0.500	6	0.500
3	6	0.500	6	0.500
4	2	0.143	2	0.143
5	2	0.143	2	0.143
6	12	1.333	16	2.250
7	20	4.000	16	2.250
8	10	1.000	11	1.158
9	12	1.333	11	1.158
10	28	28.00	31	9.653*
11	34	33.00*	31	15.95*
12	28	28.00	21	4.667
13	14	1.750	21	4.667
14	6	1.500	18	3.000
15	30	9.495*	18	3.000
16	34	17.36*	24	8.000
17	14	1.750	24	8.000
18	22	5.500	25	10.00
19	28	28.00	25	10.00
20	18	3.00	15	2.000
21	12	1.333	15	2.000
22	16	2.250	11	1.158
23	6	0.500	11	1.158

* =mean value

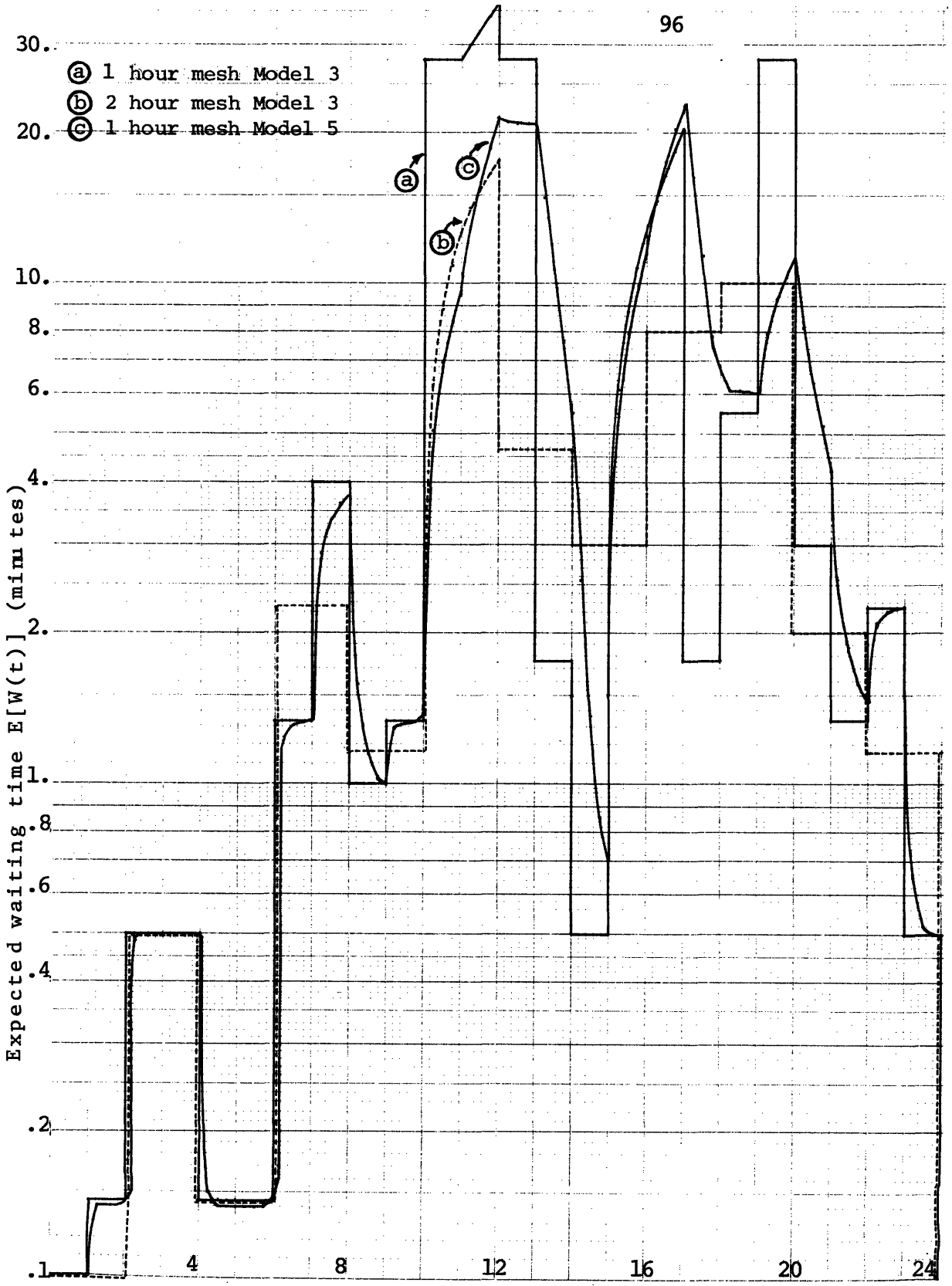
The arrival rates and corresponding delays obtained by equilibrium analysis for the demand profile Figure 4.2

Table 4.2

then the transient component remains large, and the expected delay far from equilibrium even after 60 minutes. Therefore, the error incurred by equilbrium analysis is rapidly increasing as $\rho \to 1$. As often is the case in these demand profiles the following hour has a drastically different number of operations and the transient again has a large component.

We propose a correction of this by choosing a larger mesh size which will have the effect of simultaneously reducing the number of hours with both very great and very low arrival rates. Choosing a two hour mesh and computing E[W] based on the arrival rate averaged over two hours is then a method of crudely estimating the average value of the transient. This correction of the mesh size also follows an accepted practice in the calculation of airport delays from FAA handbooks. The practice has been to compute the sum of the top two consecutive demand hours in the profile, and then to do a table lookup based on $\hat{\rho}$ calculated from the averaged arrival rate.

When the calculations with the increased mesh size were performed the delay turned out to be 2282 minutes or approximately one half of the original estimate. Results of the other models will show in fact that this is the more reasonable estimate. However, it is clear that equilibrium analysis coupled with a knowledge of the transient behaviour can be used to derive meaningful delay estimates using even the hourly mesh. To illustrate this we solved the M/M/1 C-K equations, using Model 5, an hourly mesh and Method α (λ constant for each hour) for the same Schiphol demand profile. Since we specify that μ does not change, ρ is constant for each hour,then by definition each hourly interval represents a stationary queueing situation with some arbitrary ini-



Comparison of E[W(t)] obtained with Models 3 & 5 for profile Figure 4.2

Figure 4.3

tial conditions. These initial conditions are given by the distribution of the queue length at the end of the previous one hour interval, and the initial $E[W(t)]$ computed with (2.10). The transient component of $E[W(t)]$ is then the difference between the initial waiting time and the equilibrium value of $E[W]$ for the particular value of ρ.

The delay calculated with Model 5 and Method α is 3149 minutes and much less than Model 3 delay with a one hour mesh, yet more than with the two hour mesh. The comparison of Model 3 delay with both hour and two hour meshes to the "exact" Model 5 delay is illustrated in Figure 4.3. When we observe from the diagram the exponential decay (with rate given by the time constant) of the transient component, and thus compare $E[W(t)]$ from Model 5 with the constant value of delay $E[W]$ given by Model 3, it is clear why equilibrium analysis is not suited to strongly time dependent demand profiles. Furthermore, for those periods when $\rho>1$, when no Model 3 results exist, it is notable that the diffusion approximation results which were used duplicated surprisingly closely the exact behaviour. In fact, this example shows the inconsistencies that occur when equilibrium analysis is used on time dependent profiles with periods of oversaturated conditions: when ρ is close to 1 ($.9<\rho<1$), we predict with equilibrium analysis a much higher value of delay than when $\rho>1$ and the diffusion approximation is used. Since this is clearly wrong, it does indicate how valuable more information about the transient behaviour would be.

4.2.3 Model 8: Numerical Evaluation of the Closed Form Transient Solution

This section will describe the straightforward application of Model 8 to delay analysis of time dependent M/M/1 queues. We conclude that in exchange for a probably insignificant variation of delay from Model 5 results, a substantial reduction in computation time is achieved.

We first point out the reasons for the differences in delay values obtained. These are a function of the interpretation of the demand profile by two methods that we earlier labelled Method α and Method β. We recall Figure 4.2 in which we give an hourly mesh demand profile typical of airport runway situations. For this demand we give in Figure 4.3, among other curves, the behaviour of E[W(t)] using Method α. Now the property we proved and indeed observe in Figure 4.3 is that as $\rho \to 1$ the relaxation time becomes large. Therefore in the presence of peak periods, with ρ either increasing to, or decreasing from a very large value ($\rho \geq 0.9$) the choice of Method α or Method β will not affect E[W(t)] greatly. For instance, in the increasing ρ direction Method α will yield a slightly higher delay than Method β because the constant large value of ρ is being applied for the entire 60 minute period. On the other hand, even using Method β,ρ is changing sufficiently rapidly during the 60 minute period that E[W(t)] always has a substantial transient component. The opposite will hold true in the decreasing ρ sense. Some of the error between the two methods therefore, will cancel.

The situation is somewhat different for small values of ρ ($\rho \leq 0.7$). In most cases (where initial delay is not very large) relaxation time for small values of ρ is very short. Method α will in this case predict virtually a constant delay across the 60 minute interval (i.e. the steady state value). For Method β we can assume, because of the short relaxation time (when ρ is small) that E[W(t)] is mostly given by the steady state E[W] for the instantaneous value of ρ. The E[W(t)] we observe over the 60 minute interval is then essentially the straight line connecting the values of E[W] corresponding to average values of ρ for the adjacent hours before interpolation. However, even though the delay behaviours for Method α and Method β differ

substantially in form, when ρ is small, the delays during these periods are frequently insignificant when compared to the total delay.

Hence we obtained for the profile given in Figure 4.2 the delay

Method α: 3149 minutes

Method β: 2862 minutes

The two methods therefore yield delay values only 10% apart.

In view of the other assumptions made in modelling the time-dependent queue, a 10% error is probably not significant. This is strong motivation to devise a technique that could estimate the delay at some point(s) in the mesh intervals without having to go through the whole lot of intermediate computation required by numerical solution of the C-K equations. Evaluation of the closed form transient state probabilities, Model 8 is precisely such a technique (i.e. given any initial conditions at $t = t_0$, we can evaluate the (transient) state probabilities, in one step, an arbitrary period of time hence).

We reiterate at this point that Model 8 <u>can reproduce</u> exactly the Model 5 results for the M/M/1 queue. That would involve setting the mesh size at one minute, computing the expected delay for that minute and using the resultant queue length distribution as the initial conditions for the next one minute step. This, however, is computationally impractical. Empirically it turns out that it is faster to solve the differential equations than to use Model 8. On the other hand the results are encouraging, since a mesh size of 1 minute proves to be far less than required. Apparently as few as 2 evaluations per hour (30 minute mesh) can be used to approximate within 10%

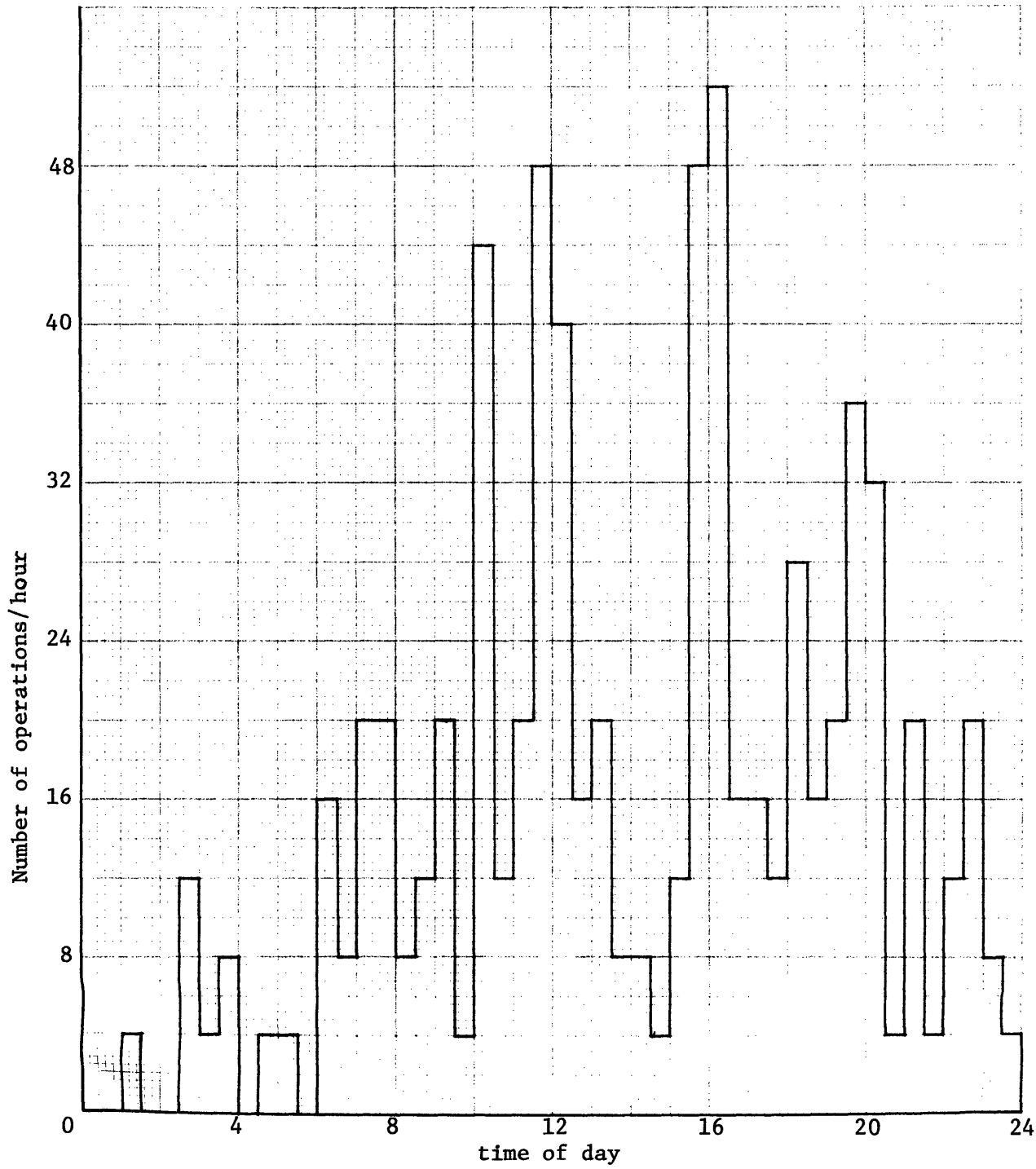
the value of the delay obtained via Method β. Additionally, in so doing the computer time needed to evaluate the total daily delay has been reduced seven fold.

mesh size	estimated delay	computer time
60 minutes	3590 minutes	.021 minutes
30 minutes	3161 minutes	.025
continuous solution of C-K equations	2862 minutes	.188

μ ops/hr	mesh size →	15	30	60
		total daily delay (minutes)		
25	M/M/1	-	8580	7008
	M/D/1	-	7560	5256
	ratio	-	1.13	1.33
30	M/M/1	4986	4092	2862
	M/D/1	4176	3288	1908
	ratio	1.19	1.24	1.50
35	M/M/1	-	2358	1458
	M/D/1	-	1812	846
	ratio	-	1.30	1.72
40	M/M/1	2148	1482	852
	M/D/1	1782	1056	456
	ratio	1.21	1.40	1.87

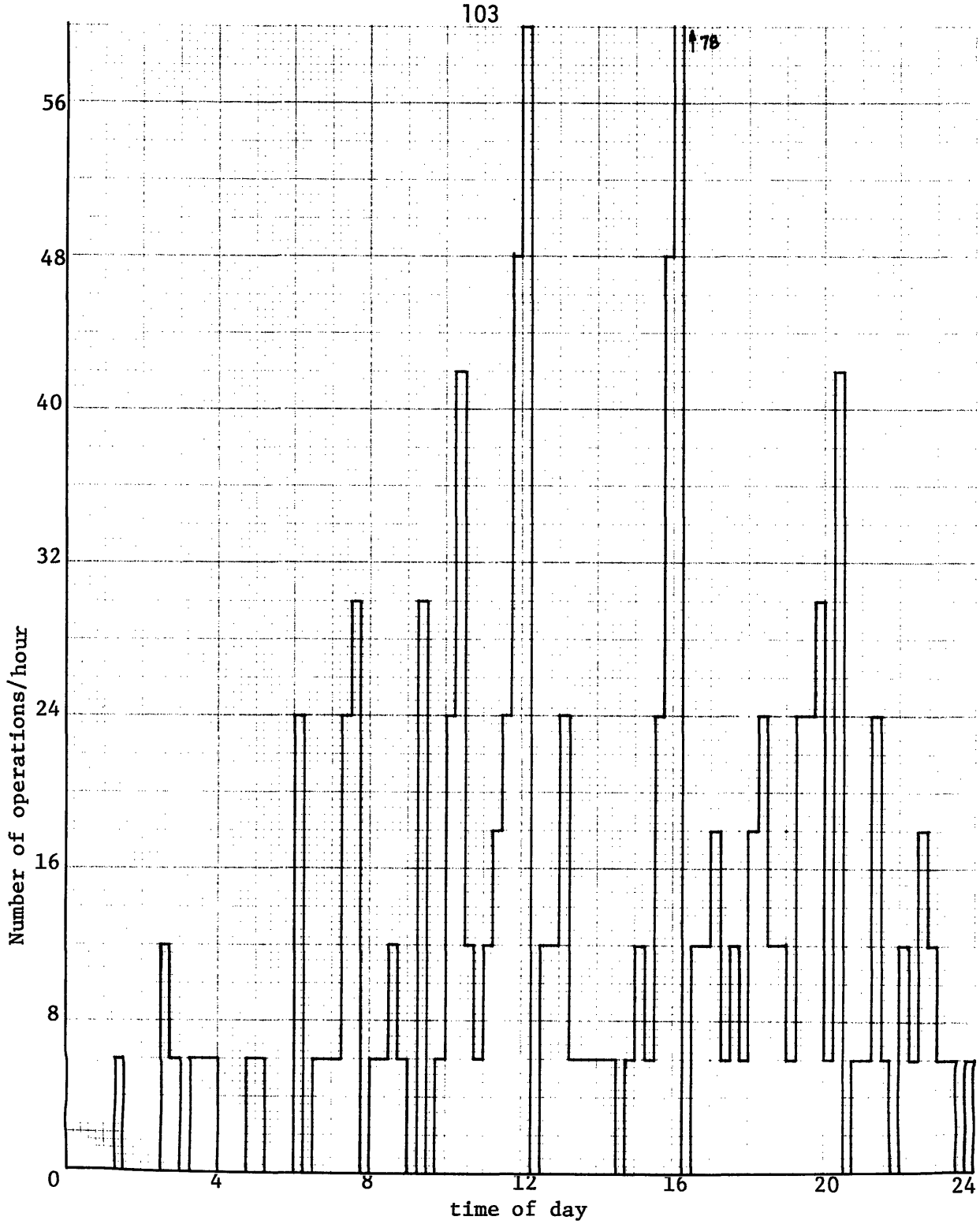
Expected total daily delays on Schiphol arrivals runway for both negative exponentially distributed and deterministic service times. The delays and their ratios are given as μ varies: 25,30,35,40 operations/hour, and as the mesh size of the demand profile varies: 15,30 and 60 minutes.

Table 4.3



Twice the number of arrivals recorded at Amsterdam's Schiphol airport on March 11, 1976, given by half hour intervals.

Figure 4.4



Twice the number of arrivals recorded at Amsterdam's Schiphol airport on March 11, 1976, given by quarter hour intervals.

Figure 4.5

4.2.4 Model 5: Numerical Solution of the C-K Equations, M/M/1, M/D/1

This section discusses the numerical solution of the C-K equations for the M/M/1 and M/D/1 systems. Two issues (which will be shown to be closely inter-related) will become important here: (1) the selection of the mesh size, and (2) the error incurred by not knowing the exact form of the service time pdf. We will use the results for expected daily delay for the demand profiles, illustrated in Figures 4.2, 4.4 and 4.5, obtained by Scalea [13] and reproduced in Tables 4.3.

To each of the Tables 4.3 we have added in the bottom row the ratio of the M/M/1 expected delay to that of the M/D/1. What we observe is that the ratios $\frac{E[W(t)]_{M/M/1}}{E[W(t)]_{M/D/1}}$ increase with increasing mesh size and/or service rate. This behavior is summarized in Table 4.4:

μ \ mesh size	15	30	60
40	1.21	1.40	1.87
35	-	1.30	1.72
30	1.19	1.24	1.50
25	-	1.16	1.33

Ratios of the daily delays observed for M/M/1 and M/D/1 queue

TABLE 4.4

Large ratios are troublesome because they leave more room for error in the interpolation process. If service time can be no more regular than a constant, nor more random than a negative exponentially distributed random variable, then we expect the true value of the M/G/1 queue delay to be between these bounds. Therefore, it is clear that the smaller the ratio,

the smaller the maximum possible error in our daily delay estimate, regardless of the true service time pdf. Fortunately, we will be able to conclude that we associate with large ratios, demand profiles that are not of particular interest to this study, i.e. (a) time dependent demand profiles with $\rho << 1$ at all times such that equilibrium analysis yields very much the same results as direct solution of the equations or (b) virtually constant demand profiles to which application of equilibrium analysis is fully justified. The general conclusion will be that the excellent performance of Model 5 is explained largely by the behavior of the time constants of the M/M/k and M/D/k systems.

In order to present the simple analysis that follows, we will assume Method α pertains. All of Scalea's results, given in Tables 4.3, were actually obtained through consideration of demand profiles constructed with Method β. As our analysis is aimed more at developing an understanding of the behavior rather than describing it exactly, this will not be a significant problem.

The basis for the analysis is the hypothesized form of the transient state probability from Model 6 (ignoring the dependence on initial conditions):

$$p_i(t) = P_i + \sum_s B_{is}\, e^{-\gamma_s t}$$

From this, we can compute the expected value of the queue length, and hence the expected waiting time:

$$E[W(t)] = E[W] - g e^{-t/T_r} \tag{4.1}$$

where the time dependent part has been greatly simplified to have a single constant g and a single time constant T_r. T_r of course is the relaxation time of the system as determined from the autocorrelation function of the

expected queue length. The only information available about g is that when $E[W(0)] = 0$, then $g \equiv E[W]$. With the extreme simplification we have introduced in (4.1), we will be able to show, <u>in very general terms</u>, that the key concept in comparing the M/M/1 and M/D/1 time dependent delay estimates for a given profile is the time constant, or its inverse, the relaxation time.

It is clear that M/M/1 and M/D/1 system delays, when the systems are in transient state, would always be related by a factor of 2 <u>if their relaxation rate was the same</u>. This is easy to show. Let the equilibrium values of delay for ρ_i be given by $E_i[W]_{M/D/1}$ and $E_i[W]_{M/M/1}$. Suppose $E[W(0)] = 0$, and furthermore suppose the (common) rate of relaxation of $E[W(t)]$, after the step demand of ρ_o begins, is $\frac{1}{T}$. After time t_o, we observe $E[W(t_o)]_{M/D/1}$, (call it A_o), and $E[W(t_o)]_{M/M/1}$, (call it B_o). Substituting values in (4.1):

$$A_o = E_o[W]_{M/D/1} - E_o[W]_{M/D/1}\, e^{-t_o/T} = \frac{1}{2} E_o[W]_{M/M/1} - \frac{1}{2} E_o[W]_{M/M/1}\, e^{-t_o/T}$$

$$= \frac{1}{2} B_o$$

The reader can easily verify that the next time interval t_1, which starts with initial waiting times of A_o in the M/D/1 system and B_o in the M/M/1 system, finishes with the waiting times still related by a factor of 2, independent of t_o and t_1. The factor of 2, of course, is a consequence of the Pollaczek Khintchine result (2.1).

The relaxation times, however, are not the same for M/M/1 and M/D/1 systems. In fact, we showed with (3.28) that $T_{r_{M/D/1}} = \frac{1}{r} T_{r_{M/M/1}}$. With this information, let us now consider a time dependent demand profile consisting of two time intervals of length t_o and t_1 during which the values of ρ_o and ρ_1, respectively. Let T_o denote the value $T_{r_{M/M/1}}$ for ρ_o. Then:

$$A_o = E[W]_{M/D/1} \left(1 - e^{-2t_o/T_o}\right)$$

$$B_o = 2E[W]_{M/D/1} \left(1 - e^{-t_o/T_o}\right)$$

$$\frac{B_o}{A_o} = \frac{2\left(1 - e^{-t_o/T_o}\right)}{1 - e^{-2t_o/T_o}} \tag{4.2}$$

Evaluating (4.2), for the purposes of illustration, at $t_o = T_o$:

$$\frac{B_o}{A_o} = \frac{2\left(1 - e^{-1}\right)}{1 - e^{-2}} = \frac{2 \cdot 0.632}{.865} = 1.462 \tag{4.3}$$

This means that the M/M/1 system delay at a time t_o (given in this case by the relaxation time T_r of the M/M/1 system) is much less than twice the M/D/1 system delay. Next, assume that at $t_o = T_o$, ρ changes to ρ_1 (with new relaxation time $T_1 = T_{r_{M/M/1}}$). At the end of the next interval, we observe the system delays:

$$A_1 = E_1[W]_{M/D/1} - \left(E_1[W]_{M/D/1} - A_o\right) e^{-2t_1/T_1}$$

$$B_1 = E_1[W]_{M/M/1} - \left(E_1[W]_{M/M/1} - B_o\right) e^{-t_1/T_1}$$

and their ratio:

$$\frac{B_1}{A_1} = \frac{2 - \left(2 - \dfrac{B_o}{E_1[W]_{M/D/1}}\right) e^{-t_1/T_1}}{1 - \left(1 - \dfrac{A_o}{E_1[W]_{M/D/1}}\right) e^{-2t_1/T_1}} \tag{4.3}$$

For the purposes of illustration, if $t_1 = \frac{1}{2} T_{r_{M/M/1}}$ (which we happen to know to be the true value of the relaxation time under these initial conditions), then by definition we expect $B_o = .63\ E_o[W]_{M/M/1}$, and if $\rho_o = 0.8$, $\rho_1 = 0.9$,

μ = 1/minute, $E_0[W]_{M/M/2}$ = 4.0 minutes and $E_1[W]_{M/D/1}$ = 4.5 minutes, then evaluation of (4.3) yields

$$\frac{B_1}{A_1} = \frac{2 - (2 - \frac{2.528}{4.500} e^{-t_1/T_1}}{1 - (1 - \frac{.684 \cdot 2.528}{4.500}) e^{-2t_1/T_1}} = \frac{2 - 1.438e^{-t_1/T_1}}{1 - 0.616e^{-2t_1/T_1}} \qquad (4.4)$$

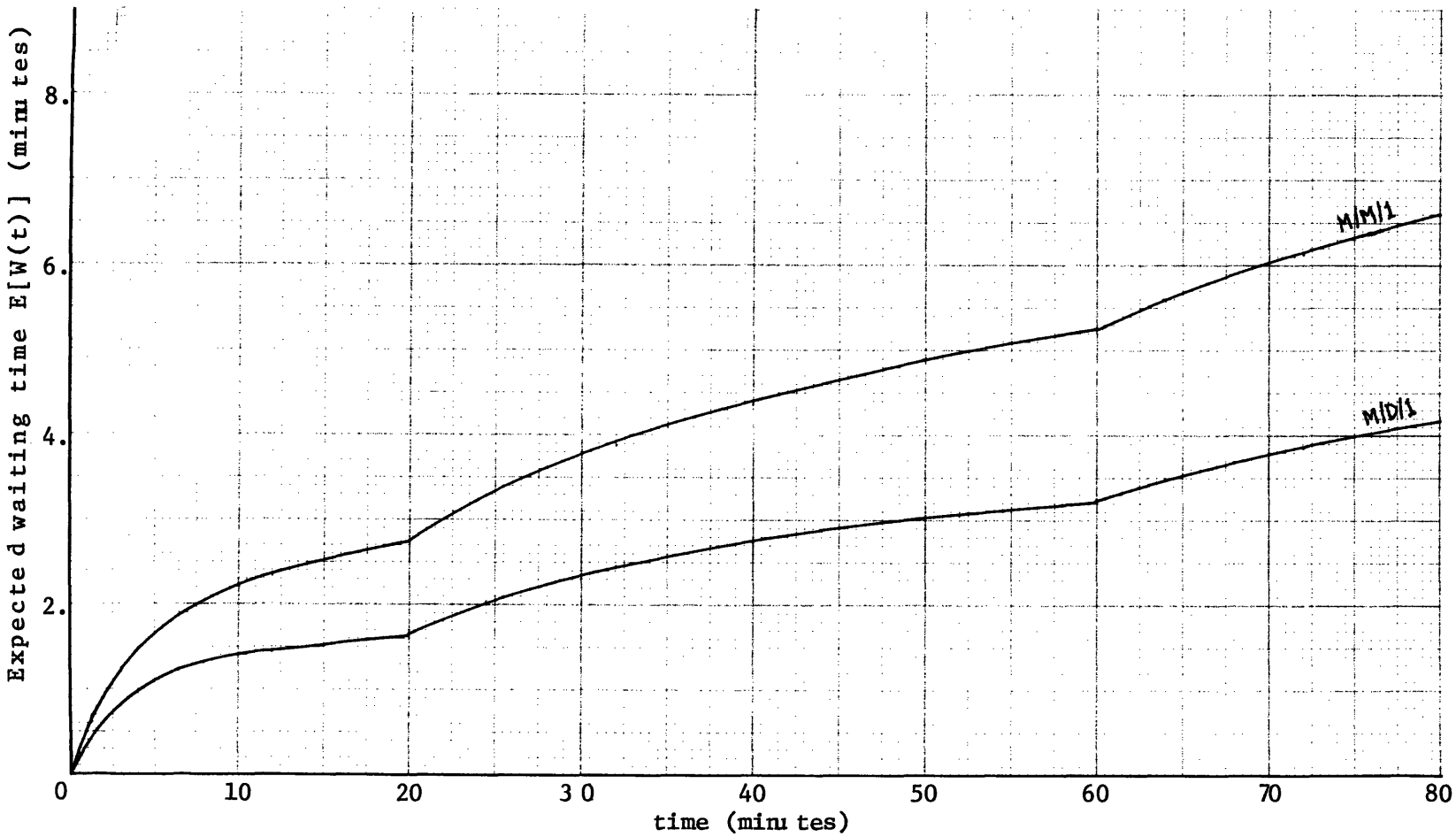
Evaluating (4.4) with $t_1 = T_1 = T_{r_{M/M/1}}\Big|_{\rho = .9,\ \mu = 1/\text{minute}}$

$$\frac{B_1}{A_1} = \frac{1.471}{0.917} = 1.60$$

In spite of the broad simplifications introduced, this example presents fairly representative numbers. To prove this, we provide an example of a three interval demand profile with ρ = 0.8, 0.9 and 0.95 in intervals of 20, 40 and 40 minutes, respectively, and μ constant at 1/minute. Listed in Table 4.5 are the ratios of the values of E[W(t)] illustrated in Figure 4.6 for the M/M/1 and M/D/1 systems as t goes from 0 to 100. We observe that the ratios increase rapidly for ρ = 0.8, less rapidly for ρ = 0.9. For ρ = 0.95, the relaxation time is so long that after 40 minutes, the ratios were still ,in fact ,decreasing. For comparison, we show in Table 4.6 the values of the ratios for ρ = 0.9 μ = 1/minute when E[W(0)] = 0. It is clear by comparing the first 20 minute intervals of both examples that there are substantial differences in the ratios observed for ρ = 0.8 and ρ = 0.9.

To show that the three interval example coincides with the somewhat more analytical treatment given earlier, we need some results from the section on the single server queue, 4.3. We show there two properties:

(1) that when E[W(0)] = 0, the observed value of the relaxation time is 1/2 of its theoretical value.



Comparison of E[W(t)] of M/M/1 and M/D/1 for a three interval demand profile, $\rho = 0.8$ for 20 minutes, $\rho = 0.9$ for 40 minutes, $\rho = 0.95$ for 20 minutes, $\mu = 1$/minute, $E[W(0)] = 0$

Figure 4.6

ρ	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.95	0.95	0.95	0.95	0.95
time	4	12	20	24	32	40	52	60	64	72	80	92	100
M/M/1	1.46	2.30	2.74	3.21	3.89	4.39	4.97	5.28	5.60	6.13	6.60	7.20	7.55
M/D/1	0.99	1.45	1.65	2.00	2.45	2.75	3.08	3.24	3.48	3.87	4.18	4.58	4.80
ratio	1.48	1.59	1.66	1.61	1.59	1.60	1.61	1.63	1.61	1.59	1.58	1.57	1.57

Ratios of the values of E[W(t)] for M/M/1 and M/D/1 systems for a three interval time dependent demand profile beginning with E[W(0)]=0, μ=1/min,ρ = .8,.9,.95

Table 4.5

time	5	10	15	20	40	60	90
M/M/1	1.75	2.53	3.08	3.52	4.71	5.47	6.24
M/D/1	1.35	1.83	2.16	2.41	3.06	3.43	3.77
ratio	1.30	1.38	1.43	1.46	1.54	1.60	1.66

Ratios of the values of E[W(t)] for M/M/1 and M/D/1 systems for stationary demand $\lambda \doteq .9$, μ = 1/minute, with E[W(0)] = 0.

Table 4.6

(2) the relaxation time T_r, even when the above property is considered, cannot be substituted into (4.1) to obtain a reliable value of $E[W(t)]$ outside the region $\frac{1}{2} T_r < t < 2T_r$.

Property (1) tells us therefore that the value 1.426 derived in (4.3) for the relation of the $E[W(t)]$ for M/M/1 and M/D/1 systems should hold at values of t given by $t = \frac{1}{2} T_{r_{M/M/1}}$. Therefore, we compare 1.426 to the observed value of 1.661 for $\rho = 0.8$ at $t = 20$ and 1.656 for $\rho = 0.9$ at $t = 90$. Whereas it might not seem at the outset that (4.3) is very accurate, we recognize the bounds that it provides us. For this, we recall from Chapter 3 that whereas $E[W]$ increases as $(1 - \rho)^{-1}$, T_r increases as $(1 - \rho)^{-2}$. Equation (4.3) tells us that when initial waiting time is 0 or very small we can expect the M/M/1 system delay to be approximately 146% of the M/D/1 delay at the value of time t that grows as $(1 - \rho)^2$. Since in our demand profile we are concerned with values of very high ρ of one half or one hour's duration, we are in effect expecting even smaller ratios. Furthermore, from (3.28)

$$T_r = \frac{2\rho}{\mu(1 - \rho)^2}$$

which implies that as μ is reduced, T_r also increases exponentially. Again using (4.3), we see that the time t at which we expect a ratio of 1.46 grows exponentially.

Property (2) is important because it means that, especially as ρ becomes large, $E[W(t)]$ increases much faster in its initial stages ($\tau = \frac{t}{T_r} < 0.5$) than specified by the time constants. The consequence of this is, unfortunately, that our rough analytical model (4.1) cannot be extended to consider the ratios of $E[W(t)]$ at values of $t < \frac{1}{r} T_{r_{M/M/1}}$ without much more information

about the true relaxation time. However, at values of t comparable to T_r, as with (4.3), our calculated ratio of 1.6 compares well with the pattern of the ratios from t = 20 to t = 60 in Table 4.5, as well as with the ratio of $\sim$ 1.75 observed in a separate example with similar numbers.

We conclude therefore that large values of the ratios are only possible in time dependent situations where either the transient components are always very small, and the exponential terms therefore negligible, as in the case of demand profiles with nearly constant levels of demand, or in highly time dependent situations where $\rho < \sim 0.7$ at all times such that the relaxation time is very short and the exponential term becomes negligible very quickly. Neither of these cases is usually of much interest for the purpose of time dependent delay evaluation. On the other hand, as congestion increases (ρ increases) the importance of accurate knowledge of delay in the system is magnified. We have shown here that the values of E[W(t)] of M/D/k and M/M/k systems in exactly these cases move closer to each other, thus providing tight bounds on the true value of delay.

4.3 The Transient Behaviour of the Single Server Queue

4.3.1 Introduction

This section will illustrate numerically the important properties of the transient behaviour of the M/G/1 queue for some special cases. For the most part the analysis will be concerned with the relaxation time, with many comparisons of exact numerical results with theoretical results. Except as indicated the conditions μ = 1/minute and $E[W(0)] = 0$ apply to the illustrations.

The first concern will be to select from the M/G/1 queues specific cases for numerical study. The most obvious candidates, the ones for which we were easily able to write the C-K equations, have already been suggested in sections 2.4 and 2.7. Since we would definitely like to study the extreme cases, the queues M/D/1 and M/M/1 are acceptable choices. Inclusion of the queue $M/E_c/1$, as pointed out in section 2.7 is desirable because not only does it include the above extremes, as c goes from 1 to infinity, but it also represents or approximates extremely well many service time pdfs that lie, in terms of probabilistic characteristics (Figure 2.6), between "perfectly random" and deterministic.

We will decide on which value of c (in $M/E_c/1$) to study on the basis of past models. Odoni [7] has studied the time dependent queueing delays with models of Type 5. For estimation purposes (and based purely on intuition) Odoni has assumed that the delay is given by the

relation:

$$E[W(t)] = \frac{2}{3} E[W(t)]_{M/D/1} + \frac{1}{3} E[W(t)]_{M/M/1} \tag{4.5}$$

Examining (4.1) more closely, assuming <u>steady state conditions</u>, and using (2.1):

$$\frac{2}{3} E[W]_{M/D/1} + \frac{1}{3} E[W]_{M/M/1} = \frac{2}{3} \cdot \frac{1}{2} E[W]_{M/M/1} + \frac{1}{3} E[W]_{M/M/1}$$

$$= \frac{2}{3} E[W]_{M/M/1} \tag{4.6}$$

But we can show for a single server, using (2.1):

$$E[W]_{M/E_c/1} = \frac{\rho}{2(1-\rho)} \left[1 + \frac{\frac{1}{c\mu^2}}{\frac{1}{\mu^2}} \right] \frac{1}{\mu} = \frac{(c+1)}{2c} \cdot E[W]_{M/M/1} \tag{4.7}$$

Evaluating (4.7) for $c = 3$ ($M/E_3/1$) yields:

$$E[W]_{M/E_3/1} = \frac{2}{3} E[W]_{M/M/1} \tag{4.8}$$

implying that (4.5) and (4.7) coincide. Thus, calculating (4.5) is equivalent to finding $E[W]_{M/E_3/1}$ when the <u>system is in the steady state</u>. It does not seem unreasonable therefore to suggest the $M/E_3/1$ queue for study in this section in addition to M/D/1 and M/M/1.

4.2.2 <u>Comparison of Relaxation Times for M/M/1, $M/E_3/1$ and M/D/1, with $E[W(0)] = 0$</u>

We give first for the three queues the theoretical relaxation times

as hypothesized in sections 3.3 and 3.4, then pass to Model 5 for numerical evaluations of the systems. Also this part will examine the relationship (3.28) between the relaxation times of various M/G/1 queues as predicted by Model 4 and show that (3.28) is an excellent approximation to the relationship observed from the exact numerical solution of Model 5.

For the theoretical T_r, we begin with (3.19) to obtain T_r for M/M/1:

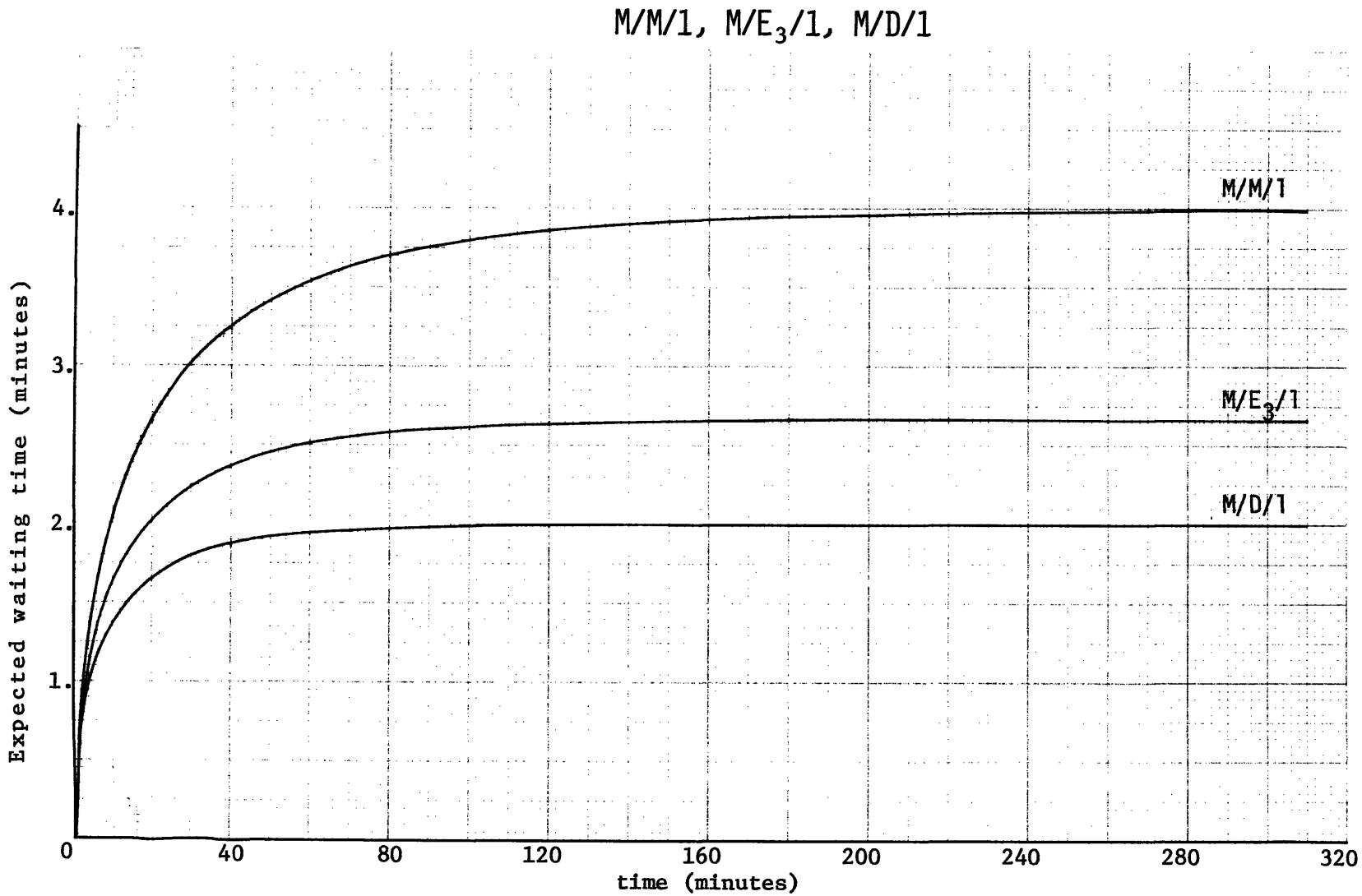
$$T_{r_{M/M/1}} = \frac{2\rho}{\mu(1 - \rho)^2}$$

For example, when $\rho = .8$, $\mu = 1$/minute, $T_{r_{M/M/1}}$ is 40 minutes. For the other systems we assume (3.28) holds, substituting $c = 3$ and $c = \infty$ to obtain

$$\frac{T_{r_{M/E_3/1}}}{T_{r_{M/M/1}}} = \left.\frac{c+1}{2c}\right|_{c=3} = \frac{2}{3}$$

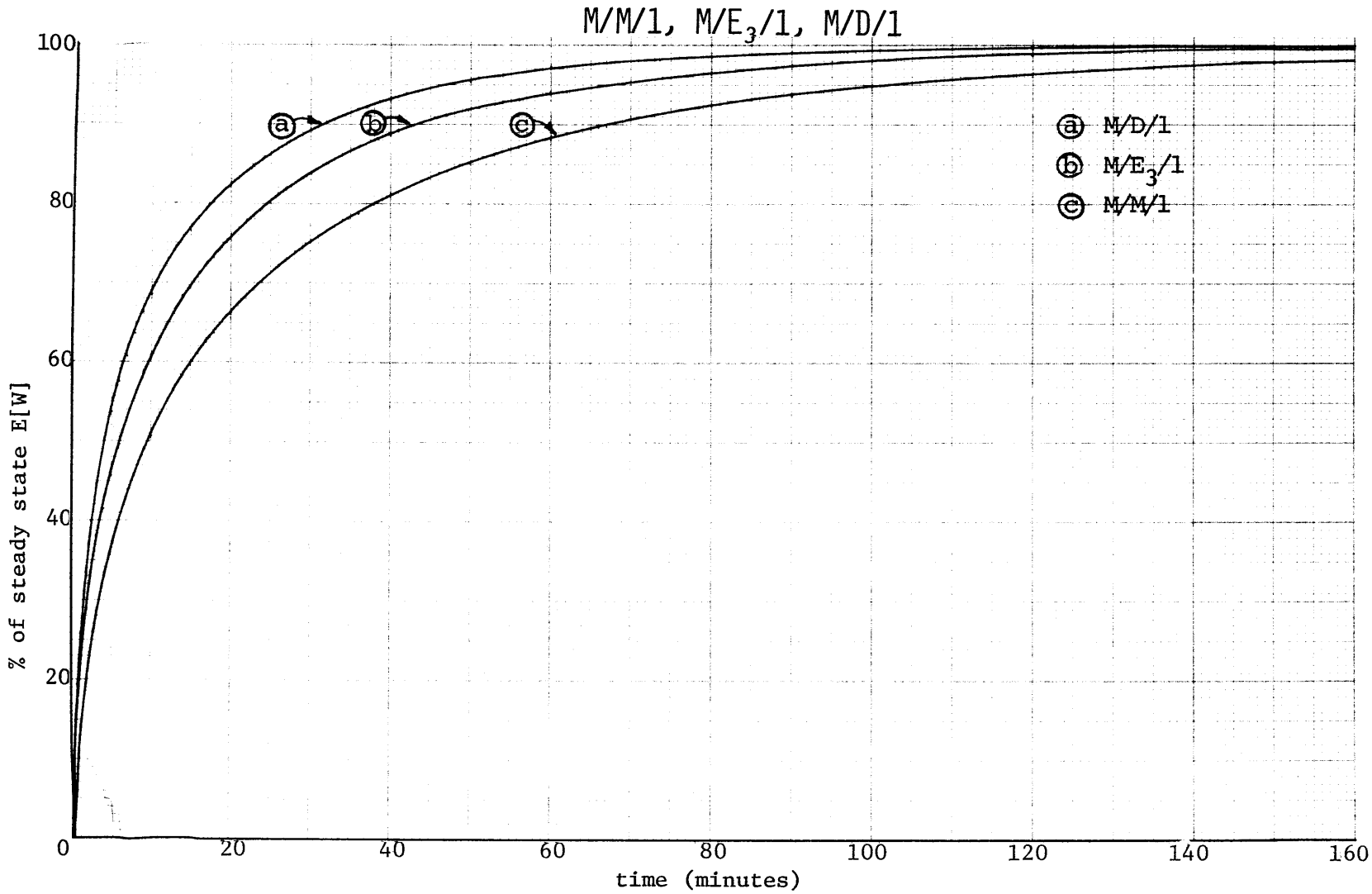
$$\frac{T_{r_{M/D/1}}}{T_{r_{M/M/1}}} = \lim_{c \to \infty} \frac{c+1}{2c} = \frac{1}{2}$$

Hence T_r for $M/E_3/1$ and M/D/1 are theoretically 26.7 minutes and 20 minutes respectively for the same values of ρ and μ.



Response of E[W(t)] to the conditions ρ=.8, μ=1/min, E[W(0)]=0

Figure 4.7



Response of E[W(t)] (as a percent of E[W]) to the conditions ρ=.8, μ=1/min,E[W(0)]=0

Figure 4.8

time (mins)	T_r (theoretical value)			O_r (observed value)		
queue \ ρ	.7	.8	.9	.7	.8	.9
M/M/1	15.6	40.0	180	8.0	17.5	67.0
$M/E_3/1$	10.4	26.7	120	5.1	11.3	44.0
M/D/1	7.8	20.0	90.0	3.3	7.5	32.0
queues \ ratio	Theoretical ratio			Observed ratio		
$\frac{M/D/1}{M/M/1}$	.500			.413	.429	.478
$\frac{M/E_3/1}{M/M/1}$	.667			.638	.646	.657

Theoretical and empirical values of the relaxation time T_r of the queues M/M/1, $M/E_3/1$, and M/D/1 from an initial value of $E[W(0)] = 0$ with $\rho = .7, .8, .9$, $\mu = 1$/minute. Also the theoretical and empirical ratios of the relaxation time of these systems.

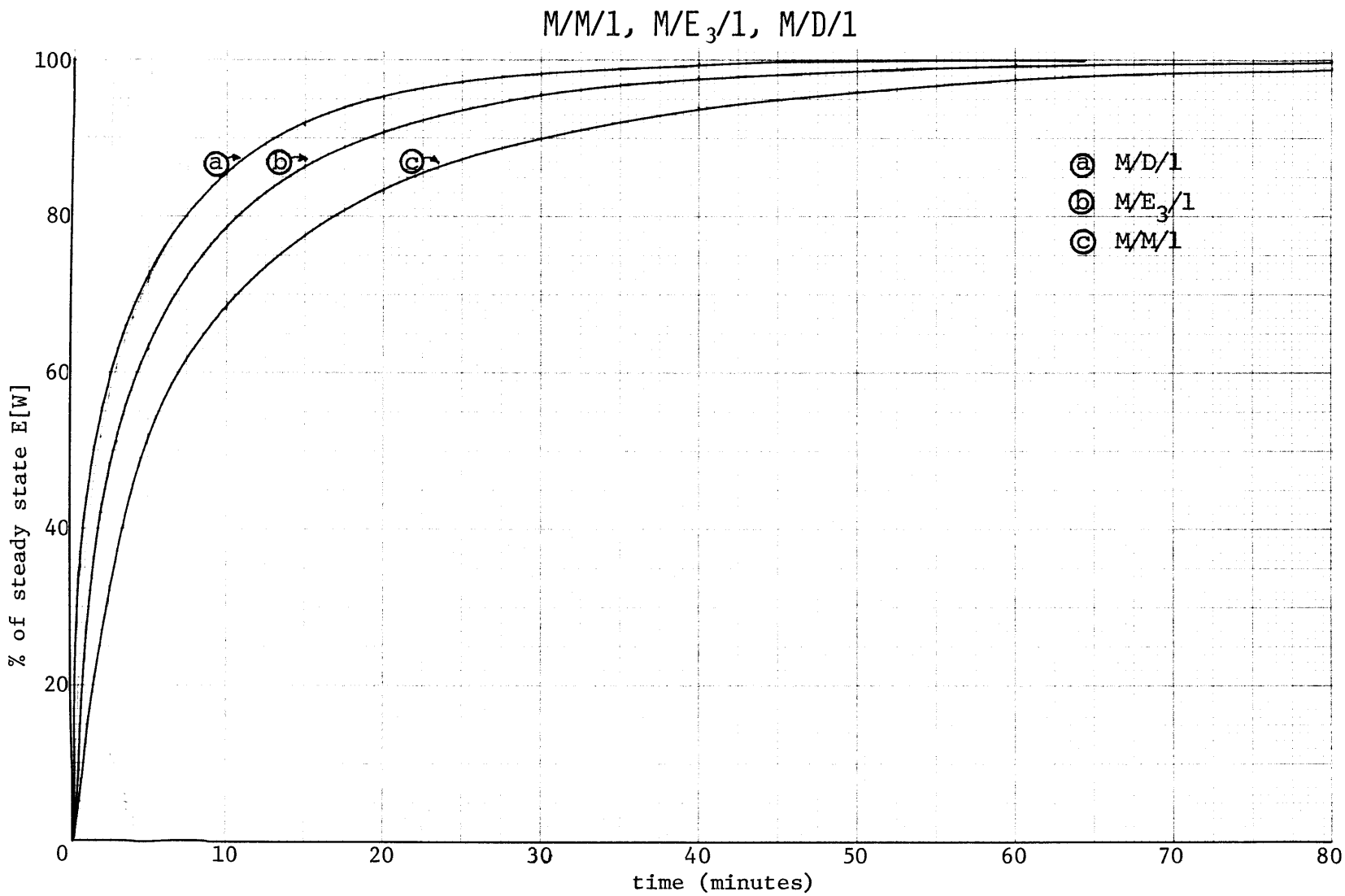
Table 4.7

The empirical values of T_r for $\rho = .8$, $\mu = 1$/minute are found from Figure 4.7 as the time at which the transient component has been reduced to $\frac{1}{e}$ of its original value. Since we assumed for these cases the specific initial condition $E[W(0)] = 0$, the transient component is clearly equal to the entire steady state value, $E[W]$ itself. We replot Figure 4.7 in Figure 4.8, this time having the vertical axis denoting the "% of $E[W]$". The empirical values of T_r are then found from the time axis to be the values corresponding to the points $(1 - \frac{1}{e}) \cdot 100\% = 63.2\%$ on the vertical axis of Figure 4.8 . After similar analysis for $\rho = .7$ and $\rho = .9$, $\mu = 1$/minute for both, we obtain the theoretical and empirical values of T_r listed in Table 4.7.

It appears that the observed values of the relaxation time, O_r are far from the theoretical values T_r. However, we recognize that our value T_r (3.19) from Model 6 was based on the mean (an arbitrary weighting of the fastest and slowest decay rates, and is highly dependent on the initial conditions. Since we calibrated T_d of Model 4 by (3.19) and (3.24), the deviation of O_r from T_r for $M/E_3/1$ and $M/D/1$ is not unexpected either.

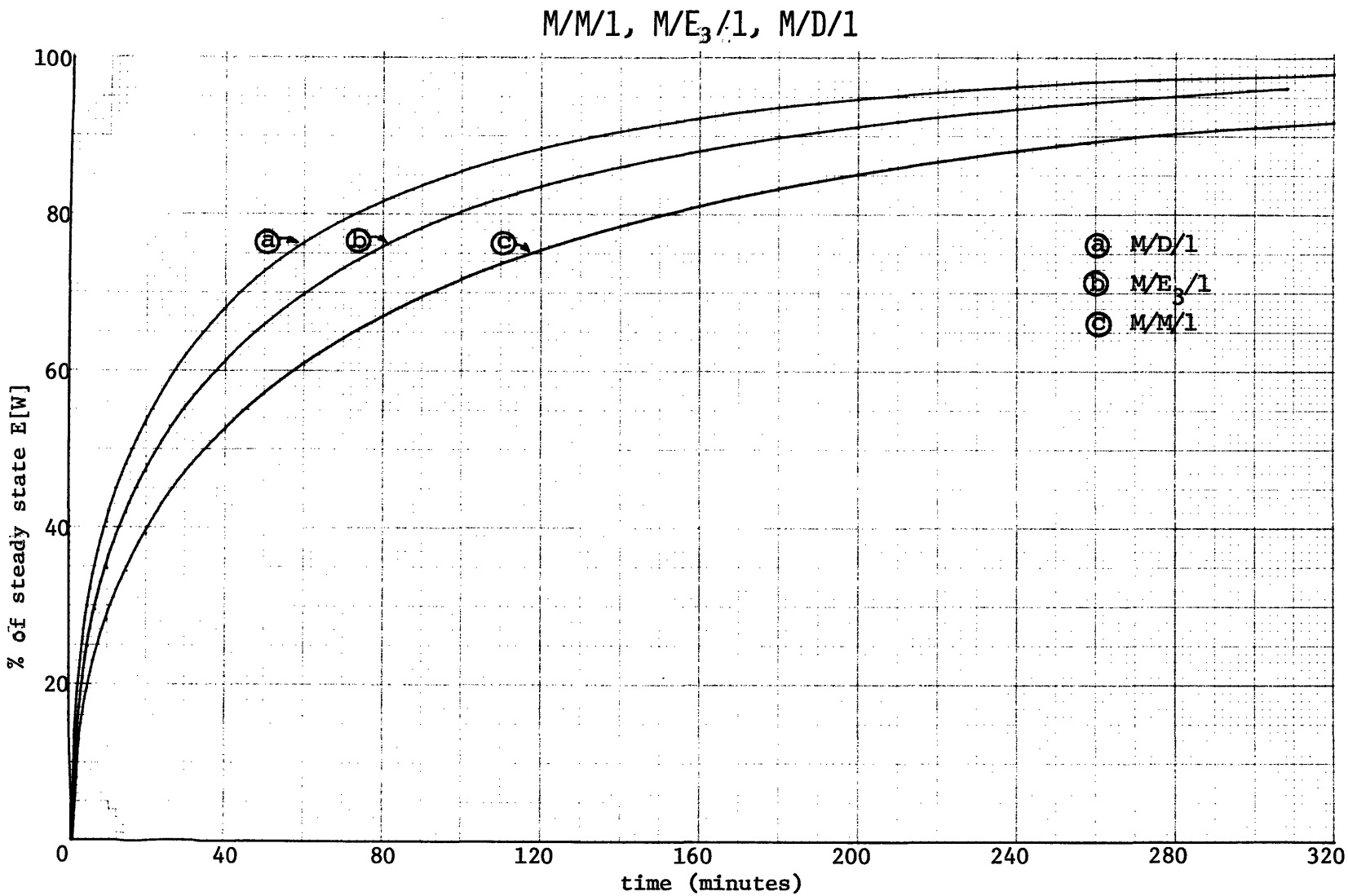
The significance of the deviations of the observed ratios from the expected ratios is difficult to interpret. While exhibiting somewhat the behaviour expected from the theoretical results, the magnitudes of the errors appear to be large. Therefore we propose to examine closely the transient behaviour of the queues for longer periods of time.

To investigate the relaxation time as a valid measure of the dissipation of the transient, consider the dimensionless quantity τ given by $\frac{t}{T_r}$. Then let $f(\tau)$ be the fraction of the transient component of $E[W(t)]$ dissipated at time $t = T_r\tau$. If the relaxation time T_r given by (3.19) is



Response of E[W(t)] (as a percent of E[W]) to the conditions $\rho = 0.7$, $\mu = 1$/minute, E[W(0)]=0

Figure 4.9



Response of E[W(t)] (as a percent of E[W]) to the conditions $\rho = 0.9$, $\mu = 1$/minute, E[W(0)]=0

Figure 4.10

queue	0.20	0.25	0.50	1.0	2.0	3.0	E[W]
			*** ρ = .7 ***				(mins)
M/M/1	37.9	44.5	63.0	78.2	90.3	95.2%	2.333
$M/E_3/1$	43.0	48.5	64.2	79.2	91.2	95.9	1.556
M/D/1	47.0	51.0	66.0	81.0	92.3	96.7	1.167
			*** ρ = .8 ***				
M/M/1	46.5	50.8	66.2	81.0	92.5	96.4	4.000
$M/E_3/1$	48.8	52.5	67.5	81.8	93.0	96.8	2.667
M/D/1	49.5	53.6	68.6	82.5	93.2	97.1	2.000
			*** ρ = .9 ***				
M/M/1	50.7	55.0	69.4	83.2	93.4	97.0	9.000
$M/E_3/1$	50.8	55.3	69.8	83.5	93.6		6.000
M/D/1	51.8	55.9	70.5	83.8	93.7	97.1	4.500

Comparison of the percentages $f(\tau)$ of the transient component of the expected waiting time dissipated after time $\tau = \frac{t}{T_r}$, for the systems M/M/1, $M/E_3/1$, and M/D/1 for the conditions ρ = .7, .8, .9, μ = 1/minute and $E[W(0)] = 0$.

Table 4.8

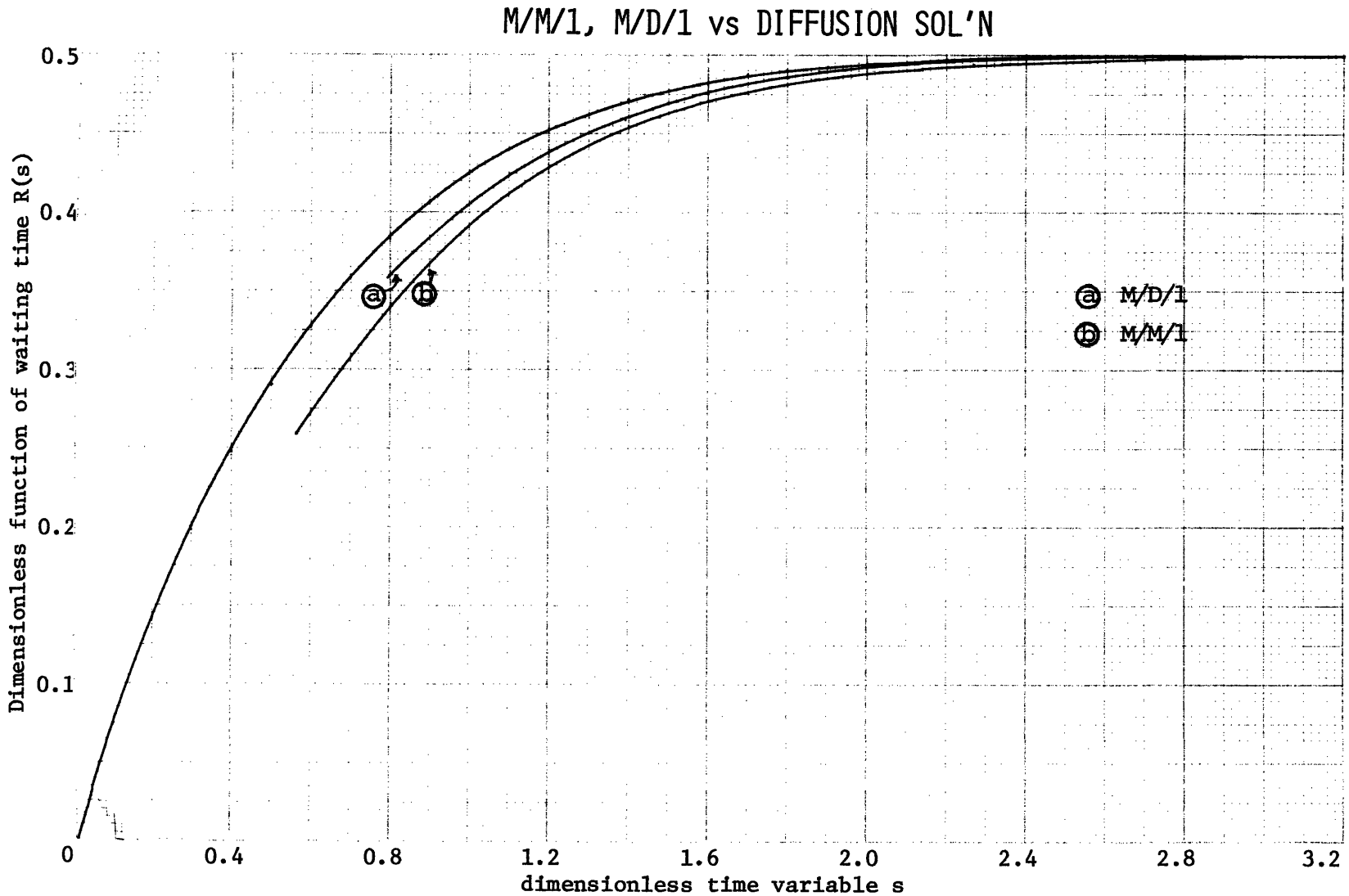
indeed correct for all queues of type M/G/1, then $f(\tau)$ should be the same, or close, for M/M/1, $M/E_3/1$ and M/D/1 for all values of τ. This is approximately true for each value of ρ individually, although three distinct patterns of deviations from this postulated behavior are observed in Table 4.8:

(1) for constant τ, for all values of τ, the differences between $f(\tau)$ for the systems M/M/1, $M/E_3/1$ and M/D/1 decrease as ρ increases. As an example, when $\tau = 1$, the differences in values of f(1) for $\rho = .7$, 18 and .9 are 2.8%, 1.5% and 0.6%. Small values of τ have larger differences in values of $f(\tau)$ and vice versa for large values of τ.

(2) for constant τ, for all values of τ, the value of $f(\tau)$, for any of the systems M/M/1, $M/E_3/1$ and M/D/1, is increasing as ρ increases. Considering the M/M/1 system as an example, we observe f(0.2) increasing from 37.9% for $\rho = .7$ to 50.7% for $\rho = .9$.

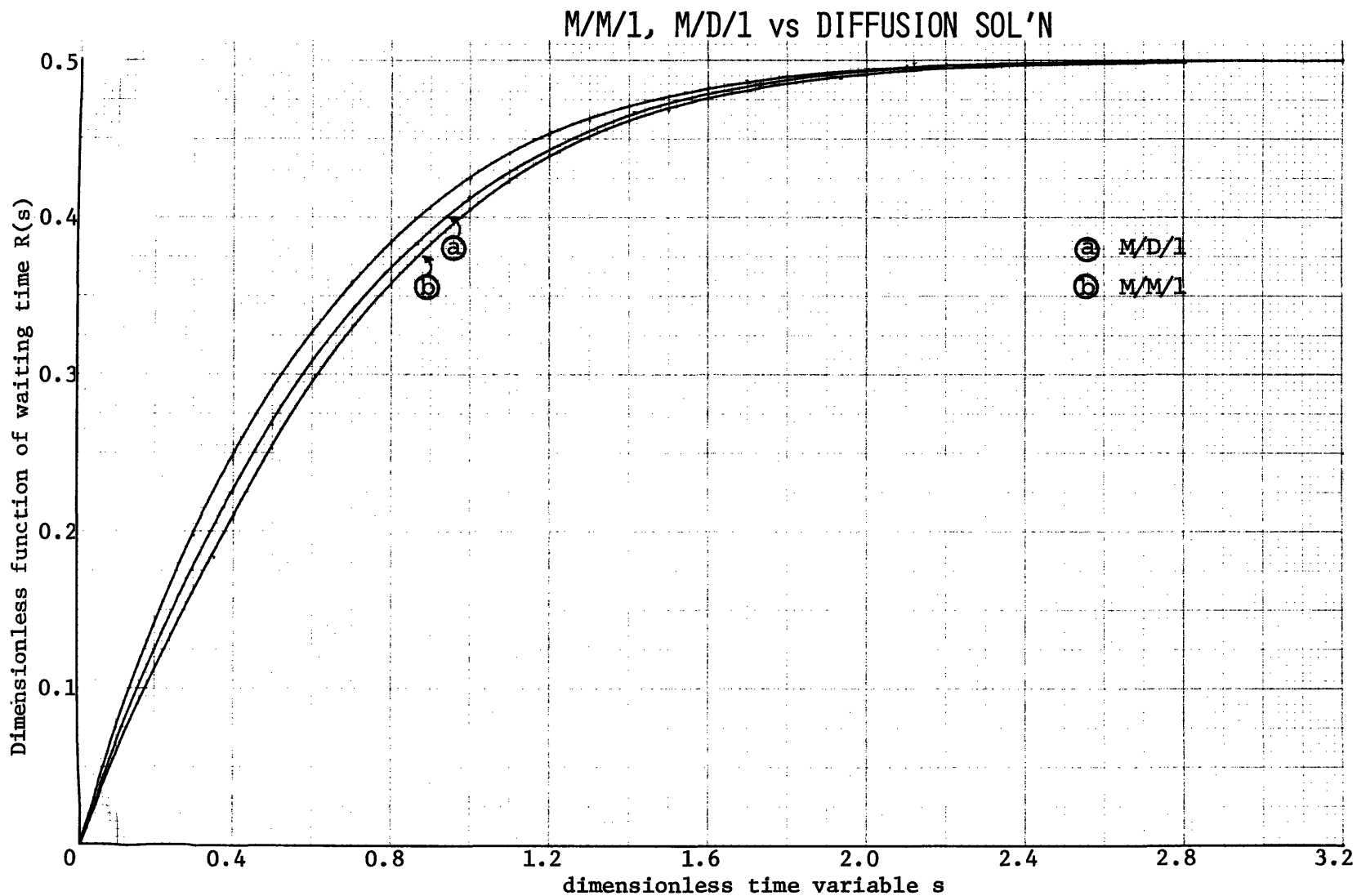
(3) the mean value of $f(\tau)$ of the three systems M/M/1, $M/E_3/1$ and M/D/1, as ρ varies, are closer for large values of τ than for small values of τ. For $\tau = 0.2$, the mean values of f(0.2) for $\rho = .7$, .8, .9 are 42.6%, 48.3% and 51.1%. The corresponding values for $\tau = 3.0$ are much closer: 95.9%, 96.8%, and 97.1%.

Of these (2) is very significant in terms of the modeling we presented briefly in Section 4.2.4. The variable values of $f(\tau)$ for any system for small values of τ mean that systems with large values of ρ relax in their initial stages much faster than systems with low values of ρ. Since $\tau = 1.0$ for large values of ρ represent very long values of elapsed time, in the

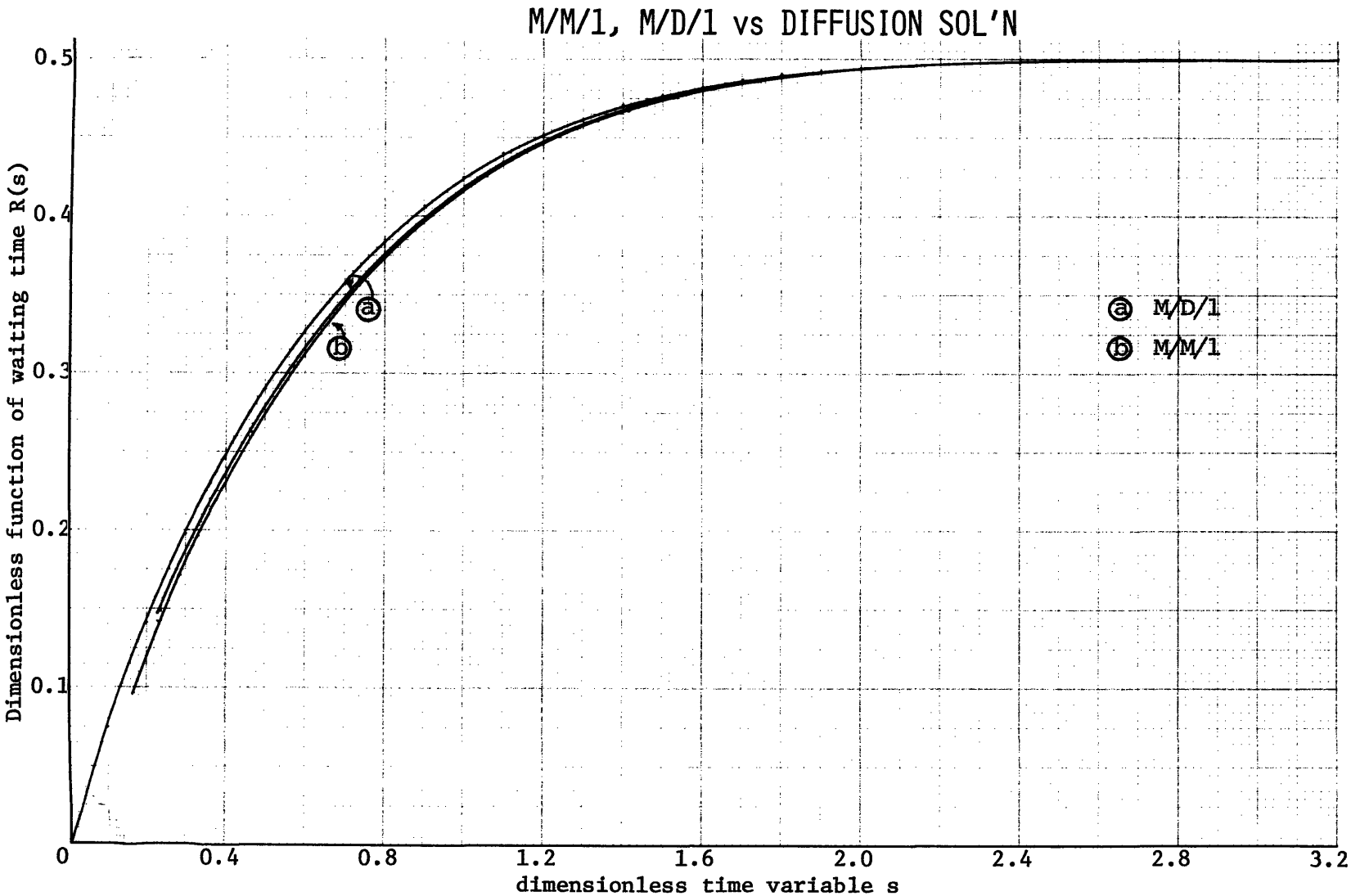


Comparison of the dimensionless response of E[W(t)] to the diffusion solution for the conditions ρ=.7, μ=1/min, E[W(0)]=0

Figure 4.11

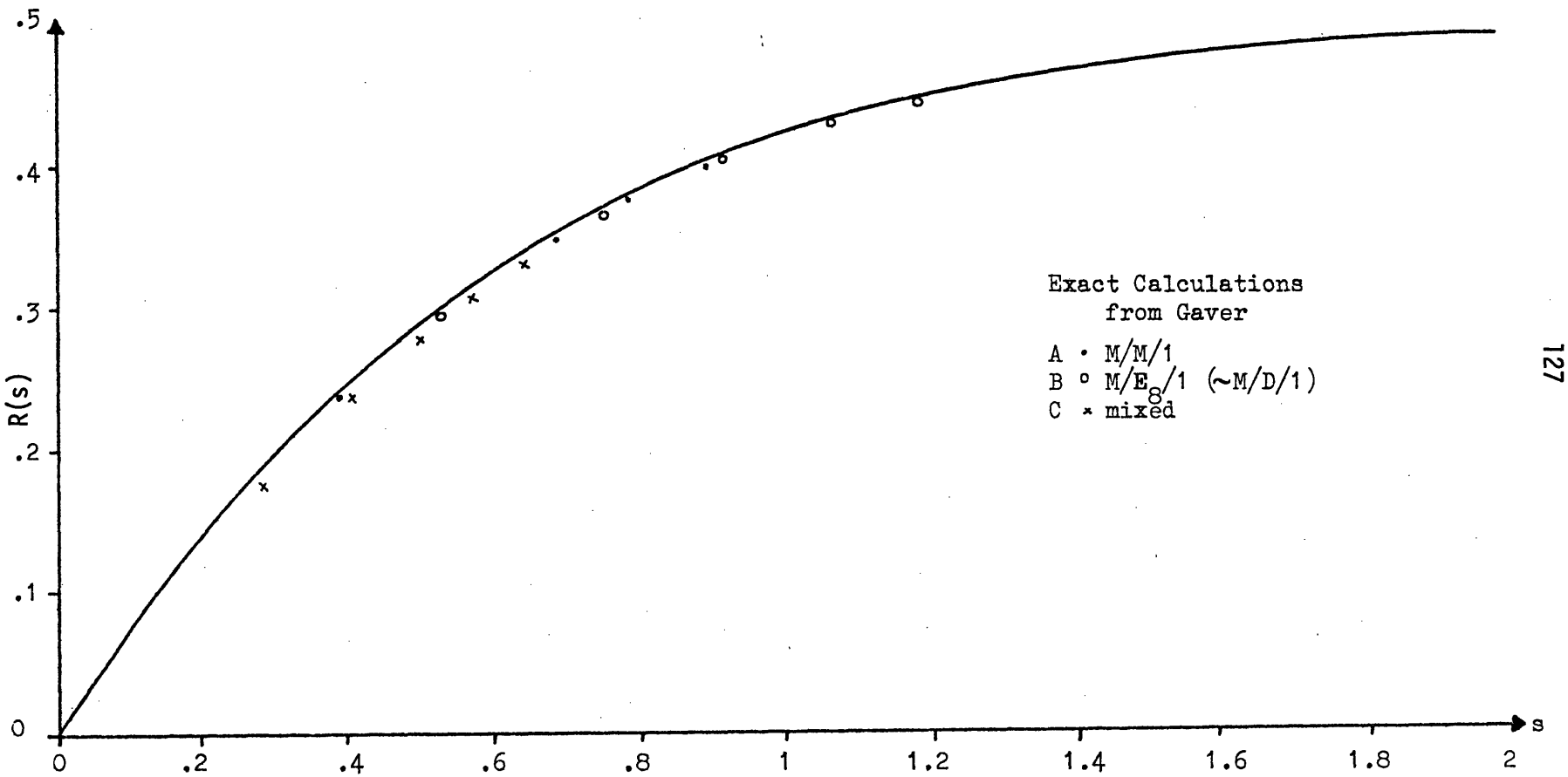


Comparison of the dimensionless response of E[W(t)] to the diffusion solution for the conditions ρ=.8, μ=1/min, E[W(0)]=0
Figure 4.12



Comparison of the dimensionless response of E[W(t)] to the diffusion solution for the conditions ρ=.9, μ=1/min, E[W(0)]=0

Figure 4.13



Exact Calculations versus Diffusion, $\rho = 0.95$, $\mu = 1/\text{min}$
Figure 4.14

modeling of time dependent gueues we are almost always concerned with small values of τ, where, unfortunately, the least consistent behavior is observed.

To show the data diagrammatically, and possibly highlight the reasons for the differences in $f(\tau)$ observed in Table 4.8, we compare the exact results to the plot R(s) of the solution of the dimensionless diffusion equation (3.27). (The reader will recall that the T_r (3.19) used to compute τ in Table 4.8 are based on (3.27) for queues other than M/M/1.) The coordinates of the points plotted in the Figures 4.11, 12, 13, 14 are calculated from ρ, t, g_2, λ and E[W(t)], obtained with Model 5, by:

$$\frac{E[W(t)]}{(g_2\lambda)/(1-\rho)}$$

on the vertical axis and:

$$\sqrt{\tau} = \sqrt{\frac{t}{T_r}} = (1 - \rho)\sqrt{\frac{t}{g_2\lambda}}$$

on the horizontal axis.

Figures 4.11-4.13 illustrate the comparison to R(s) of M/M/1 for ρ = .7, .8, .9, μ = 1 and M/D/1 for ρ = .2, .7, .8, .9 μ = 1. Results for ρ = .95 (Table 4.10) with service as described in Table 4.9 were provided by Gaver [1] and are shown in Figure 4.14.

Case	A	B	C	D	E
Queue Parameter					
g_1(minutes)	1	1	1	1	1
g_2(minutes^2)	2	1.25	3.8	2	1.25
λ (per minutes)	.95	.95	.95	1.1	1.1
Queue Type	M/M/1	M/E_8/1	mixed	M/M/1	M/E_8/1

Table 4.9
Parameters of the Various Service Processes Used by Gaver [1]

Time (in hours)		2	4	6	8	10	∞
Case	expected waiting time						
A	$E[W(t)]$	9.02	11.64	13.23	14.34	15.17	19
	$E[W_d(t)]$	9.14	11.92	13.4	14.54	15.26	19
B	$E[W(t)]$	6.33	7.85	8.69	9.22	9.58	10.68
	$E[W_d(t)]$	6.43	7.91	8.75	9.25	9.6	10.68
C	$E[W(t)]$	12.57	16.99	19.91	22.07	23.77	36.1
	$E[W_d(t)]$	13.78	18.25	21.0	23.4	24.4	36.1
D	$E[W(t)]$	19.3	32.7	45.2	57.5	69.7	
	$E[W_d(t)]$	20.14	33.4	46.06	58.46	70.6	
E	$E[W(t)]$	16.9	29.4	41.6	53.7	65.7	
	$E[W_d(t)]$	17.2	30.0	42.0	54.3	66.1	

Comparison of $E[W(t)]$ as obtained by (a) (lower rows) Model 4 and (b) (upper rows) exact, explicit numerical inversion of the Laplace transform.

Table 4.10

From Figures 4.11, 12, 13, and to a lesser extent Figure 4.14, it appears that the M/M/1 and M/D/1 values of $E[W(t)]$ are very close together in their nondimensional forms, and (for Figures 4.12,13) in the range $0 \leq s < 1.4$ they are closer to each other than to $R(s)$. The apparent consistency of this type of behavior would tend to explain the failure of relations (4.7)

and (4.6) to hold exactly for small values of τ. On the other hand, it does suggest that the M/G/1 queue is indeed governed by a dimensionless differential equation, such as (3.27). This observation is supported by Gaver [1] who indicates that although it hasn't been proven it appears that the solution to (3.27) provides an upper bound to E[W(t)].

From Figures 4.11-4.14, it can be seen that the bound becomes successively less accurate as ρ decreases. This is expected because of the underlying assumptions of the model given in Section 2.5. However, we note that the coordinate on the horizontal axis is $\sqrt{\tau}$, which is just a multiple of $\sqrt{T_r}$. Consequently, for a case such as ρ = .2, the value of T_r is very small, and from Figure 4.11, even though for large values of s the error we observe may be large, the value of $t = \tau T_r = s^2 T_r$ will still be comparatively small. As a result, after any significant period of time, the exact solution will be very close to R(s). We observed this in Table 4.8:as τ increases the differences in $f(\tau)$ for the three cases diminish. The precise form of the nondimensional equation though is still unknown.

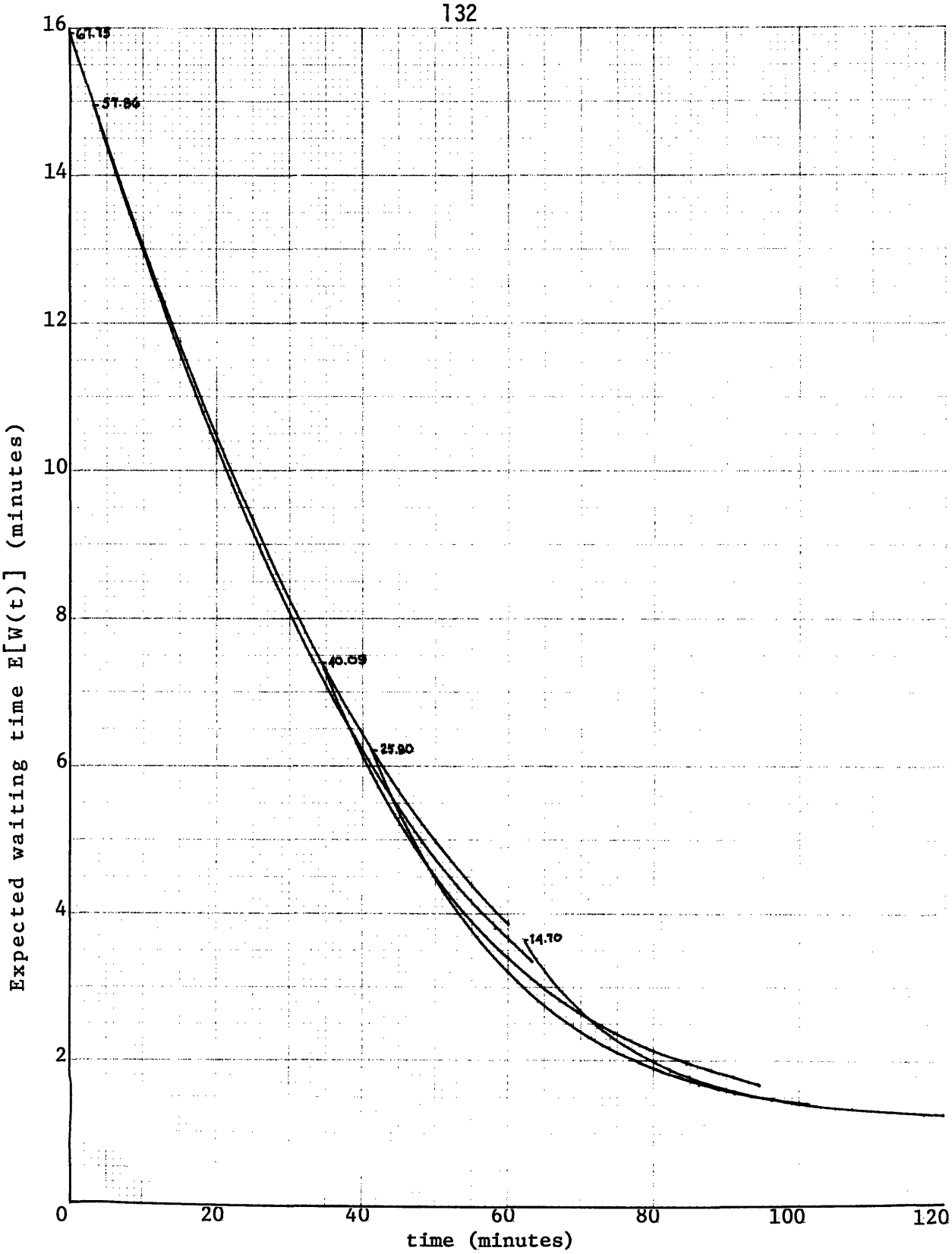
4.2.3 Comparisons of Relaxation Times for Initial Conditions other than E[W(0)] = 0

Up to now, we have investigated the relationships that exist for the relaxation time of the M/G/1 queue(3.28). We have said nothing about how we expect to modify T_r to account for varying initial conditions, as seems necessary from the results of Section 3.3. Studying the analytical expressions (3.12) and (3.17) is difficult as it necessitates studying the constants B_{is}, which vary as the initial conditions, for each $p_i(t)$. Therefore, as a substitute, we show the behavior by simply examining the values

of E[W(t)] obtained from Model 5.

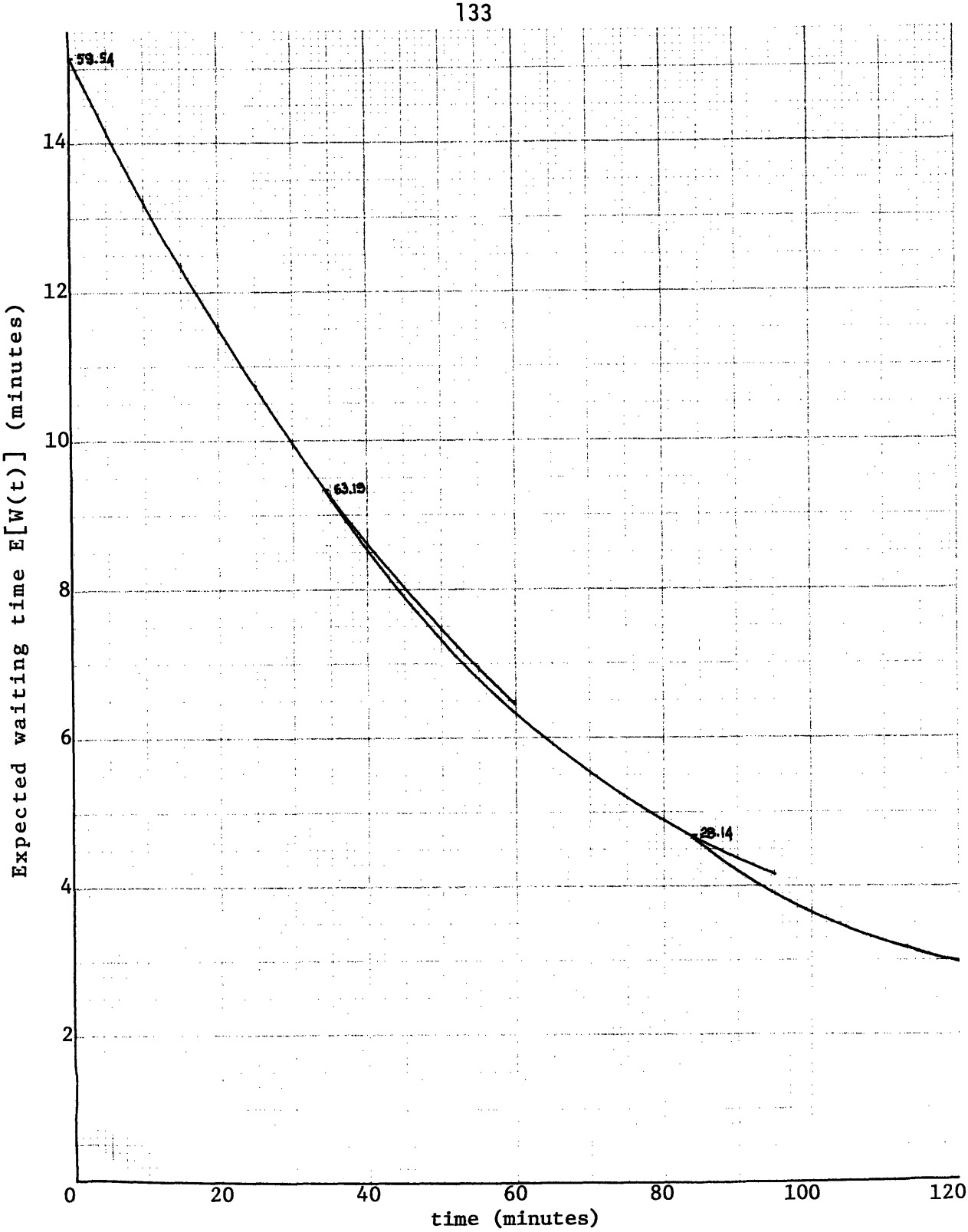
At this point, it is imperative to separate two concepts: the relaxation time and the transient behavior. The relaxation time is defined as the time required to dissipate $(1 - \frac{1}{e})$ or some other specified fraction of the transient component. Transient behavior on the other hand refers to the shape of the E[W(t)] curve in time, i.e., the path described by the relaxation of the transient component to equilibrium. Even though both terms are essentially specified by the same parameter, the time constant ω, we show that whereas the relaxation time is highly dependent on the initial value E[W(0)], the transient behavior varies much less. We show the meaning of this with an example.

Suppose we apply at time 0 with E[W(0)] = 0 a step demand equivalent to $\rho = 0.8$ to an M/D/1 system resulting in a stationary queuing system in transient state. Call this System A. Further suppose we observe System A at t_0, when the expected waiting time is given by $E[W(t_0)]$, and specify the state of System A at t_0 as the initial conditions for a separate, otherwise identical, M/D/1 system which we call System B. Then if, neglecting the interval $[0, t_0]$ of System A, we compare the two curves of E[W(t)] for Systems A and B, naturally they would be identical. The relaxation time of course would be completely different because the transient component of System A is 2 (from equation (2.1)) and that of System B is $2 - E[W(t_0)]$. Since T_r is defined by the autocorrelation function of the queue length, we expect to find situations where the time relaxation time is either shorter or longer than T_r (as we have already seen at the beginning of this section). What is not clear,however,is how significant the initial conditions are to the transient behavior when,contrary to our example with System A and



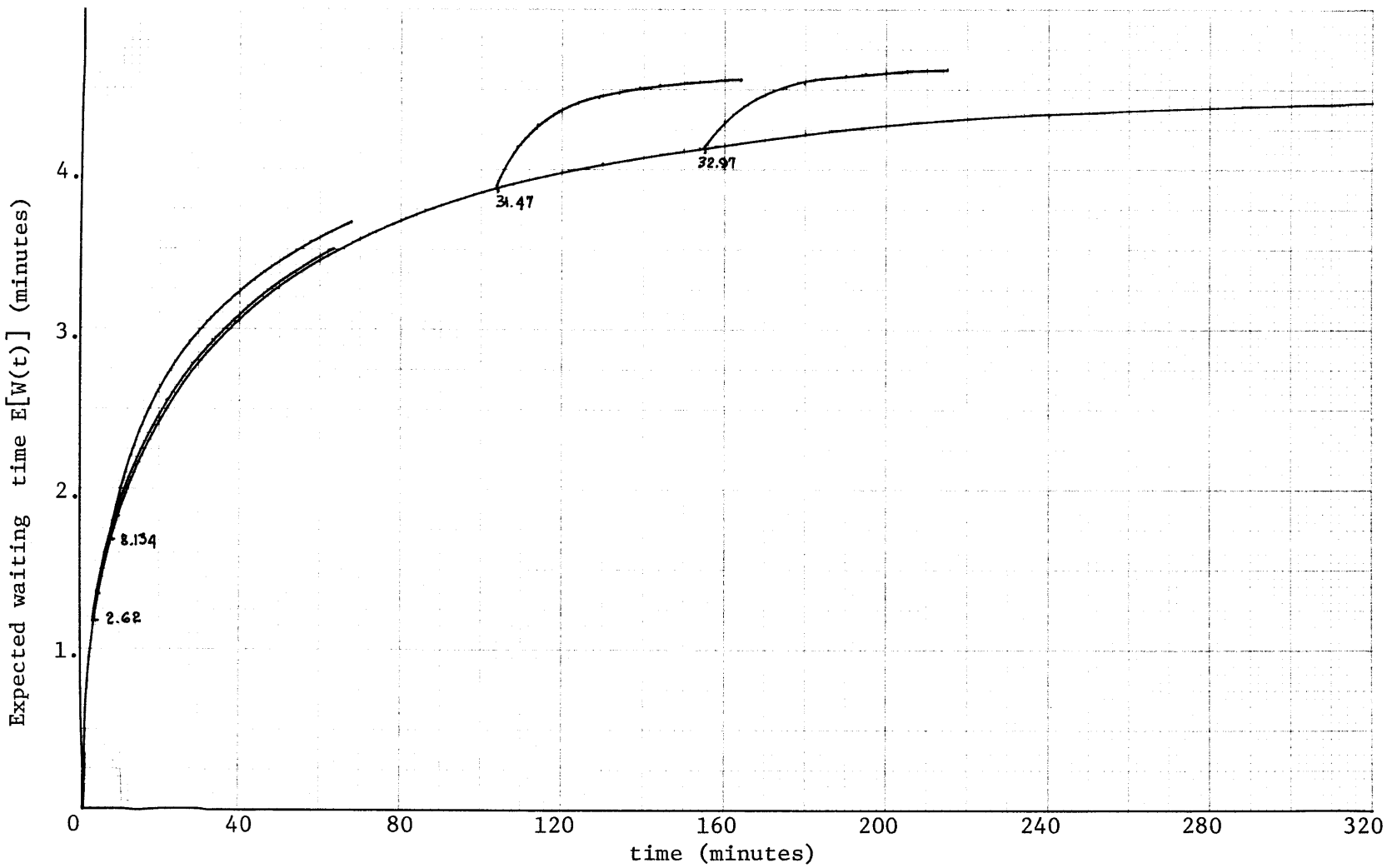
Relaxation of E[W(t)] from various initial conditions for ρ = .7, μ = 1/min (M/D/1)

Figure 4.15



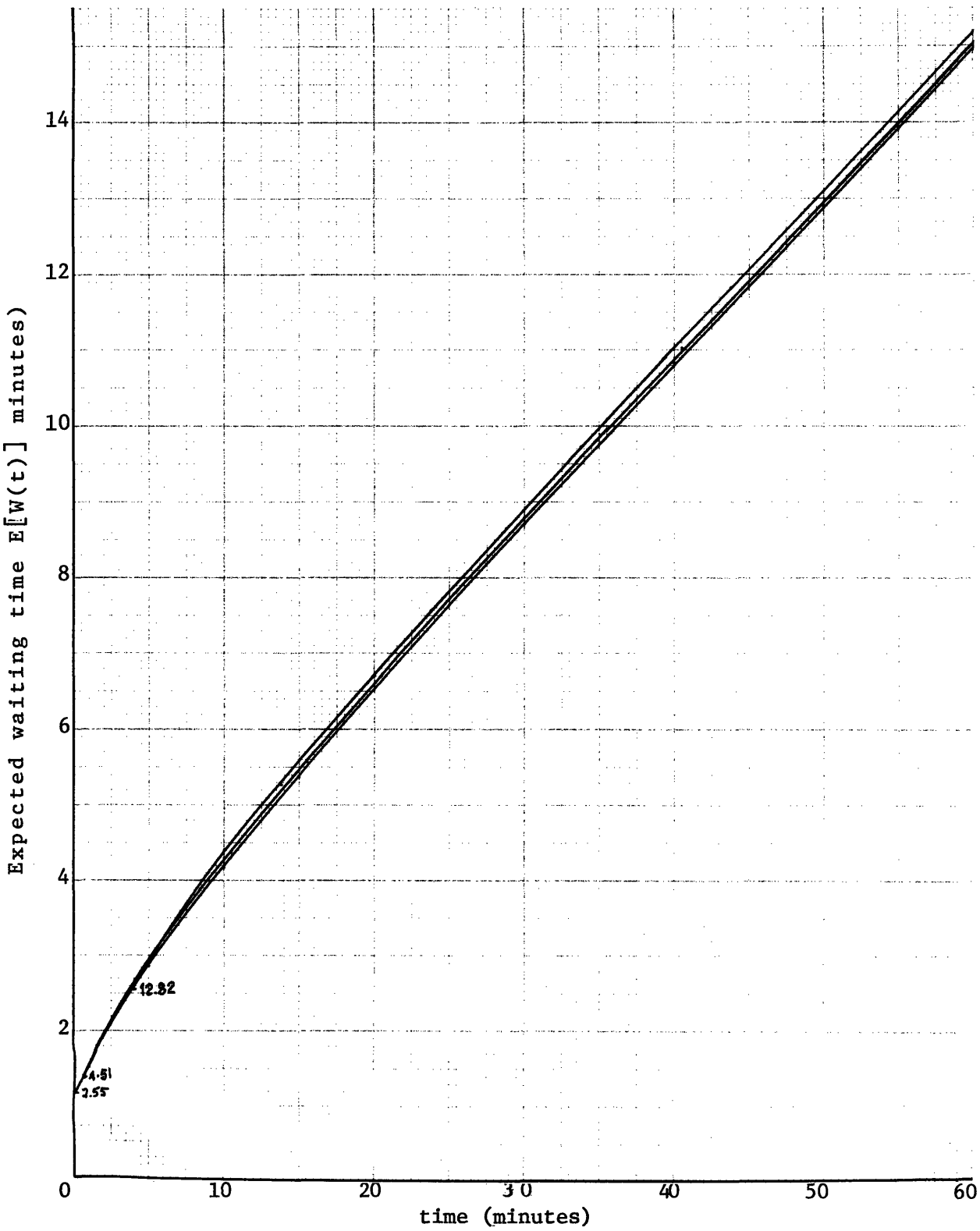
Relaxation of E[W(t)] from various initial conditions for ρ = .8, μ = 1/min (M/D/1)

Figure 4.16



Relaxation of E[W(t)] from various initial conditions for $\rho = .9$, $\mu = 1$/minute (M/D/1)

Figure 4.17



Relaxation of E[W(t)] from various initial conditions for $\rho = 1.2$, $\mu = 1/\text{min}$ (M/D/1/∞)

Figure 4.18

System B, the initial queue length distributions are not identical.

In each of Figures 4.15-4.18 we illustrate the transient response of E[W(t)] for M/D/1 to a step demand applied to a variety of initial queue length distributions and waiting times E[W(0)]. For the purpose of comparison, for each ρ we select one set of initial conditions and draw the curve for that set as the datum. All other curves for varying initial conditions are then compared to the datum for their particular value of ρ.

Perhaps the most important thing to reiterate about these curves is that the initial conditions are specified not only by E[W(0)], but also by the queue length distribution, which, as has been shown in Section 3.5, can have a substantial variance. We contend that it is plausible, therefore, that,in the presence of large queues in nonstationary systems with the initial variance "close" to the variance of the equilibrium distribution, the relaxation occurs faster than would be expected from (3.19) or (3.24). Therefore, alongside each curve in Figures 4.15-4.18, we give the variance of the queue distribution when the step demand was applied.

To a large extent our hypothesis is verified, although the differences do not appear to be very large in most cases. With regard to the effects of the differences in the initial variance, one pattern of transient behaviour is especially visible in Figure 4.17. The systems are relaxing to equilibrium with $\rho = 0.9$, in which case the equilibrium variance is given approximately as 24. We observed that in two cases where E[W(t)] is less than E[W], but the initial variance was greater than the equilibrium value ($\sigma^2 = 31$ and 33), the relaxation occurred extremely quickly (compared to the slowly increasing tail of the datum. (Here the datum is the response of E[W(t)] to a step demand

of $\rho = 0.9$ when $E[W(0)] = 0$). Also on Figure 4.17, consider two other cases: one in which $E[W(t)] = 1.681$, $\sigma^2 = 8.134$ and the other with $E[W(t)] = 1.174$, $\sigma^2 = 2.616$. Whereas the initial expected waiting times are close, the relaxation to the equilibrium value of delay is noticeably faster for the initial conditions with the higher variance. It appears therefore that in the three cases where the variance was the closest to the equilibrium variance that relaxation of $E[W(t)]$ to $E[W]$ proceeded noticeably faster.

The observed consistency of the transient behavior lends credibility to the fact that some reliable time constants can be given for transient behavior in certain strongly time dependent systems. This is the concept on which Gupta built his model. The point to remember when considering anomalies such as the behavior for $\rho = 0.9$ in our last example is that although the forms for the datum and the other curve were radically different, the difference in expected waiting times was only about 10% of the $E[W]$. Lots of other pathological cases can undoubtedly be created: for instance, the transient behavior of relaxation for $\rho = 0.8$ from the initial condition of, say, exactly 60 a/c in queue will almost certainly be much different from the behavior for $\rho = 0.8$ seen on Figure 4.16. However, in most of the solutions for time dependent profiles that we have seen, the probability that there are 60 a/c in the queue is usually small and almost always negligible. Thus, <u>for strongly time dependent profiles</u>, where no unusual congestion takes place, consideration of transient effects can probably provide reliable estimates of $E[W(t)]$.

4.4 The Multi-Server Queue in Transient State

Comparatively little is known about the transient behaviour of multi-server queues. Saaty [12] has obtained a form for the Laplace transform of the transient state probabilities $p_i(t)$ for the $M/M/k/\infty$ queue with arbitrary k. However, in view of the successful treatment of the transient behaviour based on Model 6, the most useful form is available from Morse [9]. Although he does not complete the derivation of the $p_i(t)$ for the multiserver queue, we will be able to deduce the theoretical transient behaviour from the form of the components of $p_i(t)$ given. Again, as for the single server queue, the theoretical results will be compared to exact numerical results provided by Model 5.

The components given by Morse are the integrands $q_i(\theta,t)$ in the form (3.16) and are given by (4.9) - (4.11)

for the single server case:

$$q_i(\theta,t) = \left(\frac{\lambda}{\mu}\right)^{\frac{i}{2}} \left[\sin(i\theta) - \sqrt{\frac{\lambda}{\mu}}\ \sin(i+1)\theta\right] e^{-\omega t} \tag{4.9}$$

$$\omega = \mu + \lambda - 2\sqrt{\lambda\mu}\ \cos\theta$$

for the two channel case:

$$q_0(\theta,t) = \frac{1}{3}\sqrt{\lambda/2\mu}\ \sin\theta\ e^{-\omega t}, \quad (\omega = \lambda + 2\mu - 2\sqrt{2\lambda\mu}\ \cos\theta)$$

$$q_i(\theta,t) = (\lambda/2\mu)^{\frac{1}{2}i} \left[\frac{1}{3}\sin(i-2)\theta - \frac{1}{3}(\sqrt{\lambda/2\mu} + \sqrt{2\mu/\lambda})\ \sin(i-1)\theta \right. \tag{4.10}$$
$$\left. + \sin(i\theta) - \frac{2}{3}\sqrt{\lambda/2\mu}\ \sin(i+1)\theta\right] e^{-\omega t}, \quad (i > 0)$$

and for the three channel case:

$$q_0(\theta,t) = -\frac{1}{9}\sqrt{\lambda/3\mu}\,\sin\theta\, e^{-\omega t}, \quad (\omega = \lambda+3\mu - 2\sqrt{3\lambda\mu}\cos\theta)$$

$$q_1(\theta,t) = \frac{1}{3}\sqrt{\lambda/3\mu}\,(\sin\theta - \sqrt{\lambda/3\mu}\,\sin 2\theta)e^{-\omega t},$$

$$q_i(\theta,t) = (\lambda/3\mu)^{\frac{1}{2}i}[\frac{1}{6}\sin(i-4)\theta - \frac{1}{6}(\sqrt{\lambda/3\mu} + \frac{5}{3}\sqrt{3\mu/\lambda})\sin(i-3)\theta \tag{4.11}$$

$$+ \frac{1}{9}(7 + 3\mu/\lambda)\sin(i-2)\theta - \frac{1}{2}(\sqrt{\lambda/3\mu} + \frac{11}{9}\sqrt{3\mu/\lambda})\sin(i-1)\theta$$

$$+ \sin(i\theta) - \frac{1}{2}\sqrt{\lambda/3\mu}\,\sin(i+1)\theta]e^{-\omega t}, \qquad (i > 1)$$

These $q_i(\theta,t)$ can be used to obtain expressions similar to (3.17) for $M/M/k/\infty$ with arbitrary k. We showed in section 3.3 that the coefficient ω of t in the integrand (4.9) is a descriptor (to within a constant) of the relaxation time for the single server queue. By analogy to the $q_i(\theta,t)$ and $\omega(\theta)$ for the single server case we can make some inferences about the behaviour of the relaxation time for 2 and 3 server cases. Note especially that if we maintain the total service capacity μ of a facility constant at some value μ_0 (operations/unit time), varying only the number of (independent, identical) servers k, then for the multiserver cases:

$$\omega = \lambda + k\mu - 2\sqrt{\lambda k\mu} = \lambda + k\,\frac{\mu_0}{k} - 2\sqrt{\lambda k\,\frac{\mu_0}{k}} = \lambda + \mu_0 - 2\sqrt{\lambda\mu_0} \tag{4.12}$$

Apparently, ω, the time constant and hence the determinant of the relaxation time remains independent of k. We do recall, however, that

ω was not the sole determinant of the relaxation time. The relaxation time is actually bounded above and below by the inverse of the smallest and largest values of ω respectively. The precise value of the relaxation time depends on the coefficients $B_{i\theta}$ of $e^{-\omega(\theta)t}$, which are dependent on the initial conditions (the initial queue length distribution) and vary for each $p_i(t)$. Ignoring for the moment the dependence on initial conditions, the transient component of each $p_i(t)$ is composed of terms of the form

$$\int_0^{2\pi} B_{i\theta}\, e^{-\omega(\theta)t}\, d\theta$$

(for infinite queues; for finite queues we replace the integral by a sum). Let the integrands for the single server be given by $B^1_{i\theta}$ and let $B^2_{i\theta}$ be the comparable integrands for the two server case. Then, even for a single one of the $p_i(t)$, and in spite of the fact that $\omega \neq \omega(t)$, it is not immediately obvious how the quantity

$$\frac{\int_0^{2\pi} B^1_{i\theta}\, e^{-\omega(\theta)t}\, d\theta}{\int_0^{2\pi} B^2_{i\theta} e^{-\omega(\theta)t} d\theta}$$

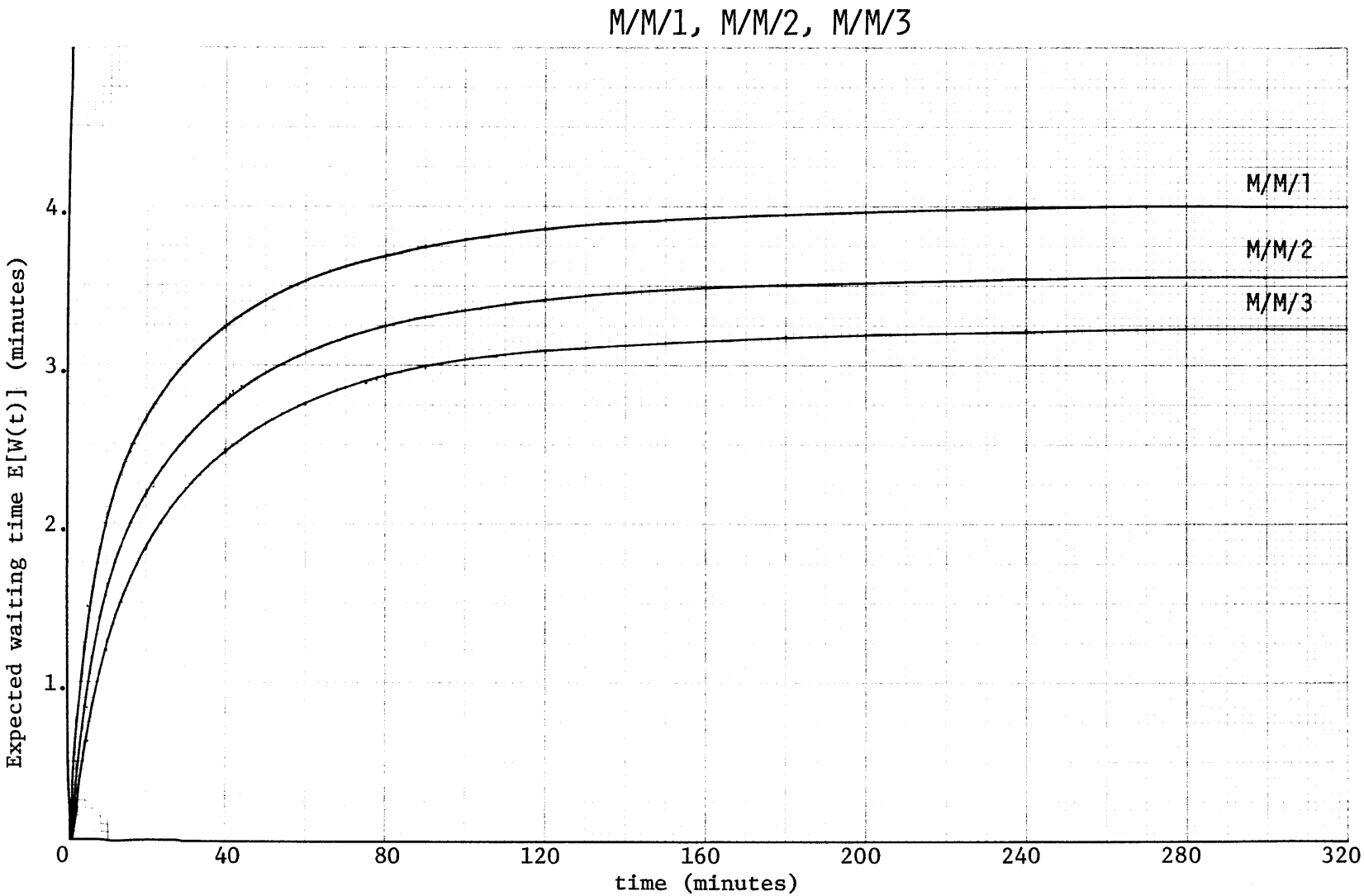
will change as t goes from 0 to ∞.

We pass therefore to numerical analysis of the transient behaviour of multiserver queues. First, the systems M/D/k and M/M/k, k = 1,2,3 were evaluated for ρ = .8, μ = 1/minute, E[W(0)] = 0, using Model 5. The resultant values of E[W(t)] are plotted vs. time in Figures 4.9 and 4.20 and the ratios $\frac{E[W(t)]}{E[W]} \cdot 100\%$ listed in Tables 4.11 and 4.12.

From these tables it is immediately obvious that the relaxation time for multiserver queues is greater than that for single server queues. To further investigate the relationships between the transient behaviours of queues as k varies, we observe the time t_f needed to attain the ratio $f = \frac{E[W(t_f)]}{E[W]}$ for the following values of f: 0.6, 0.8, 0.9.

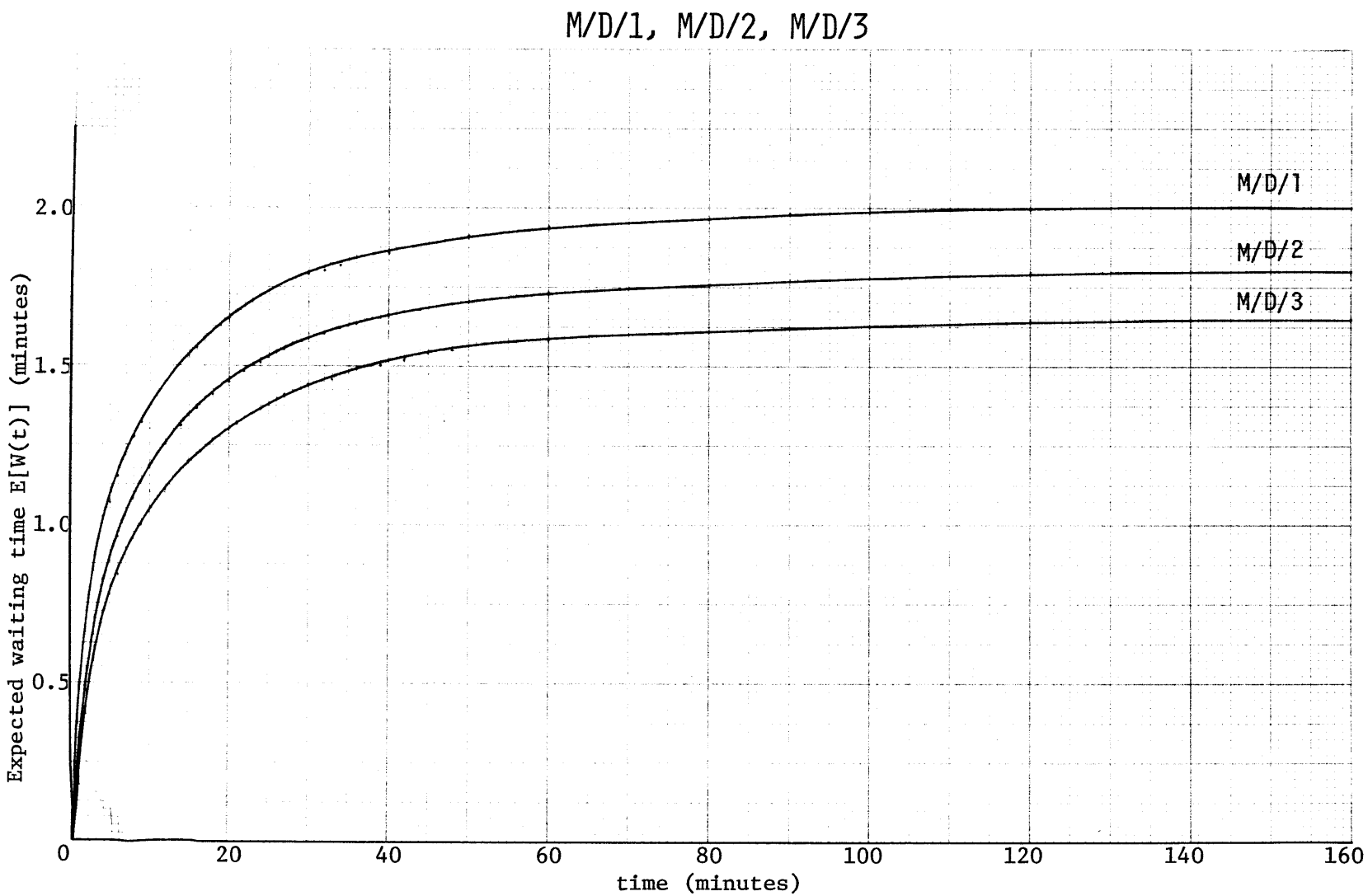
In Table 4.13 record for the M/M/k system with ρ = 0.8, μ = 1/minute, for each of the three values of f, the value $E[W(t_f)]$, t_f, and the ratio $r_f = \frac{t_{f_{M/M/a}}}{t_{f_{M/M/b}}}$. The same analysis was performed for M/D/k ρ = 0.8, μ =1 in Table 4.14 In addition, we observed two other cases of the M/D/k queue. Shown in Table 4.15 is the same case as in Table 4.14 i.e. ρ = 0.8, except with the service rate reduced to μ = 0.5 /min. Then we show in Table 4.16 the case where ρ is increased substantially to ρ = 0.9, maintaining μ = 1/min.

The trend for the multiserver queues that we observe from Tables 4.13 - 4.16 is that the time required to reach any two fractions f_1 and f_2



Response of E[W(t)] to the conditions ρ=.8, μ=1/min, E[W(0)]=0

Figure 4.19



Response of E[W(t)] to the conditions ρ=.8, μ=1/min, E[W(0)]=0

Figure 4.20

time→ / queue	5	10	15	20	40	60	90	120	180	240	300	360	420	Steady State
						(mins)								
M/M/1	36.7	50.8	59.8	66.2	81.0	88.2	93.7	96.5	98.7	99.6	99.8	100	100%	4.000 mins
M/M/2	27.7	43.9	54.2	61.6	78.5	86.7	92.9	96.0	98.6	99.5	99.8	99.9	100	3.556
M/M/3	19.3	37.0	48.6	57.0	76.0	85.1	92.1	95.6	98.5	99.4	99.8	100	100	3.236

M/M/k - % of steady state E[W] attained as a function of time, ρ =.8, μ=1/min, E[W(0)]=0

Table 4.11

time→ / queue	5	10	15	20	40	60	90	120	180	240	Steady State
					(mins)						
M/D/1	53.6	68.6	76.7	82.5	93.2	97.1	99.0	99.5	100	100%	2.000 mins
M/D/2	44.8	65.2	72.4	80.3	91.9	95.7		98.9	99.4	100	1.796
M/D/3		59.8	71.9	76.5	90.7	95.3		98.8	99.4	100	1.646

M/D/k - % of steady state E[W] attained as a function of time, ρ =.8, μ=1/min, E[W(0)]=0

Table 4.12

	f = 0.6			f = 0.8			f = 0.9		
queue	E[W(t)]	$t_{.6}$	$r_{.6}$	E[W(t)]	$t_{.8}$	$r_{.8}$	E[W(t)]	$t_{.9}$	$r_{.9}$
	(mins)								
M/M/1	2.400	15.2	1.243	3.200	38.1	1.129	3.600	67.7	1.086
M/M/2	2.133	18.9	1.158	2.844	43.0	1.102	3.200	73.5	1.068
M/M/3	1.942	22.2	1.461	2.589	47.4	1.244	2.912	78.5	1.160
M/M/1	2.400	15.2		3.200	38.1		3.600	67.7	

M/M/k - Comparison of times t_f to reach $f = \frac{E[W(t_f)]}{E[W]}$

for $f = .6, .8, .9$ for $\rho = .8$, $\mu = 1$/minute

Table 4.13

	f = 0.6			f = 0.8			f = 0.9		
queue	E[W(t)]	$t_{.6}$	$r_{.6}$	E[W(t)]	$t_{.8}$	$r_{.8}$	E[W(t)]	$t_{.9}$	$r_{.9}$
	(mins)								
M/D/1	1.200	6.7	1.179	1.600	17.5	1.100	1.800	31.7	1.064
M/D/2	1.078	7.9	1.114	1.437	19.2	1.075	1.616	33.7	1.050
M/D/3	0.988	8.8	1.313	1.317	20.7	1.183	1.481	35.4	1.125
M/D/1	1.200	6.7		1.600	17.5		1.800	31.7	

M/D/k - Comparison of times t_f to reach $f = \frac{E[W(t_f)]}{E[W]}$

for $f = .6, .8, .9$ for $\rho = .8$, $\mu = 1$/minute

Table 4.14

	f = 0.6			f = 0.8			f = 0.9		
queue	E[W(t)]	$t_{.6}$	$r_{.6}$	E[W(t)]	$t_{.8}$	$r_{.8}$	E[W(t)]	$t_{.9}$	$r_{.9}$
	(mins)								
M/D/1	2.400	13.5	1.170	3.200	35.0	1.100	3.600	63.2	1.066
M/D/2	2.156	15.8	1.114	2.874	38.5	1.068	3.232	67.4	1.050
M/D/3	1.976	17.6	1.304	2.634	41.3	1.180	2.962	70.8	1.120
M/D/1	2.400	13.5		3.200	35.0		3.600	63.2	

M/D/k - Comparison of times t_f to reach $f = \frac{E[W(t_f)]}{E[W]}$

for $f = .6, .8, .9$ for $\rho = .8$, $\mu = .5$/minute

Table 4.15

	f = 0.6			f = 0.8			f = 0.9		
queue	E[W(t)]	$t_{.6}$	$r_{.6}$	E[W(t)]	$t_{.8}$	$r_{.8}$	E[W(t)]	$t_{.9}$	$r_{.9}$
	(mins)								
M/D/1	2.700	27.3	1.088	3.600	73.1	1.048	4.050	133.3	1.030
M/D/2	2.573	29.7	1.064	3.430	76.6	1.035	3.859	137.3	1.027
M/D/3	2.474	31.6	1.158	3.299	79.3	1.085	3.712	141.0	1.058
M/D/1	2.700	27.3		3.600	73.1		4.050	133.3	

M/D/k - Comparison of times t_f to reach $f = \frac{E[W(t_f)]}{E[W]}$

for $f = .6, .8, .9$ for $\rho = .9$, $\mu = 1$/minute

Table 4.16

of E[W] by two systems of the type M/x/a, M/x/b are not related by the same constant. Instead, we observe the ratios $r_{.6}$, $r_{.8}$, and $r_{.9}$ consistently forming a decreasing sequence for every pair of systems M/x/a, M/x/b. Since we showed that ω is independent of k for all M/M/k queues, we might expect the times required for E[W(t)] to first attain E[W], i.e., reach equilibrium, to be close for all M/M/k queues. The fact that the ratios are decreasing toward 1 would tend to support this.

We also produce one piece of evidence to support the contention that some aspects of the transient behavior of M/D/k systems are the same as for M/M/k systems. We do this by observing equations (4.9) - (4.11) for M/M/k systems and noting that the dissipation of the transient is governed by the product μt for constant values of ρ. This is because wherever μ appears in the constant coefficients of $e^{-\omega(\theta)t}$, λ appears as well, and when we keep ρ constant, nothing changes. In the time constant $\omega(\theta)$, multiplying λ and μ by the same constant c yields

$$\omega_c(\theta) = \lambda c + \mu c - 2\sqrt{\lambda c \mu c}\cos\theta = c\omega(\theta)$$

Therefore, since (a) the reduction of the transient component, with time, is governed by $e^{-\omega(\theta)t}$, (b) the constant coefficients of $e^{-\omega(\theta)t}$ do not change with μ for constant ρ, and (c) the equilibrium state probabilities P_i are independent of μ for constant ρ, it is clear that the following is true:

(a) changing μ to a new value μc changes the time required to reduce the transient component by $\frac{1}{c}$, or equivalently (b) two identical M/M/k systems (with arbitrary initial conditions, number of servers k, and value of ρ) which differ only in their values of μ, μ_1 and μ_2, will have the same queue length distribution at times t_1 and t_2 which will be related by $ct_1 = t_2$ if $\mu_1 = c\mu_2$.

We apply this in a test to see if the same behavior holds for M/D/k queues. Table 4.15 shows the behavior of the M/D/k queue when $\rho = .8$, $\mu = .5$/minute, or the same case as Table 4.14 with only the service rate halved. It turns out that, exactly as we know is true for an M/M/k queue, the relaxation time to each value of f in Table 4.15 is virtually exactly double that of the corresponding value in Table 4.14, for all values of k. All the ratios are likewise, obviously, very close.

Another statistic which we desire to investigate is one assumed by Gupta in Model 7. There he postulated that the relationship between relaxation times that holds between single server queues as the service process varies, holds also for the multiserver case. Table 4.12 shows in fact that opposite changes are observed. The steady state E[W] decreases as k increases, but the spread remains small. On the other hand, the ratio of the relaxation times increases substantially with k at f = 0.6 and somewhat less at f = 0.9. In the modeling of time dependent M/G/k queues the arrival rate is changing often and the transient component is often less than 60% dissipated. The great spread in the ratios at the small values of f are

therefore quite significant.

minutes→	Steady state E[W]			Relaxation Time to f=.6			Relaxation Time to f=.9		
servers	M/M/k	M/D/k	$\frac{M/M/k}{M/D/k}$	M/M/k	M/D/k	$\frac{M/M/k}{M/D/k}$	M/M/k	M/D/k	$\frac{M/M/k}{M/D/k}$
1	4.000	2.000	2.000	15.2	6.7	2.269	67.7	31.7	2.136
2	3.556	1.796	1.980	18.9	7.9	2.392	73.5	33.7	2.181
3	3.236	1.646	1.966	22.2	8.8	2.523	78.5	35.4	2.218

E[W] and the relaxation times to 60% and 90% of E[W] for the multi-server queue, k = 1,2,3, under the conditions ρ = .8, μ = 1/minute, E[W(0)] = 0. The ratios of the M/M/k and M/D/k system values are also provided.

Table 4.17

We have seen therefore that for at least small values of k, for the particular inital condition of no waiting time, the transient behaviour of multi-server queues as k increases is less and less like the theoretical behaviour and nondimensional form given for the behaviour of the M/G/1 queue. It does seem though that the transient behaviour of the 2 server queue differs somewhat less from the 3 server queue than does that 2 server queue from the single server one. We have also shown one result that parellels the result hypothesized for the M/G/1 queue. From analytic results the M/M/k queue preserves, for all k, the change in relaxation time for changes in the service rate μ. Numerical results with M/D/k duplicate this behaviour with almost undetectable difference.

Chapter 5

5.1 Summary and Conclusions

It became apparent in Chapter 2 that a variety of techniques for modeling the time dependent M/G/k queue exist, although many of these have severe limitations either computationally or because of the assumptions made in their formulation. The most acceptable among them is the exact solution of the C-K equations, Model 5, which has the advantage of returning exact results of upper and lower bounds on the true delay with only moderate computational requirements. Two basic questions were answered in the two chapters that followed:

(1) What is the explanation for the satisfactory performance of Model 5? and

(2) Can this explanation be used to simplify further or reduce computational requirements for the study of the time dependent M/G/k queue.

Based on evidence from the models of Chapter 2 to the effect that the transient behavior of the expected waiting time is important, we investigated the closed form expressions for the transient state probabilities of the M/M/1 queue (Model 6). We obtained from this the time constant of the M/M/1 system and showed that it was valid for all values of $\rho > 0$.

Obtaining analytical expressions for the transient behavior of the expected waiting time of the M/M/1 queue had substantial impact on our understanding of that of the M/G/1 queue. We were able to show that the time constant for M/M/1 from Model 6 and from the approximate Model 4 coincided. In addition, Model 4 leads to the conclusion that there exists an approximate dimensionless form for the expected waiting time of an M/G/1 queue in

transient state (when there is an initial waiting time of 0). Based on this result, we obtain a relationship between the relaxation times of the M/G/1 queue - similar to the Pollaczek-Khintchine formula for their E[W]'s. Also, dimensionless analysis leads to the conclusion that, whereas the values of E[W] increase as $(1-\rho)^{-1}$, the relaxation time increases as $(1-\rho)^{-2}$.

Knowledge of these two properties of the relaxation time is extremely useful in explaining the success of Model 5 and in aiding the interpretation of the results of Model 3 (equilibrium analysis). We previously knew that $E[W]_{M/D/1} = \frac{1}{2} E[W]_{M/M/1}$. From Model 4, we now know that the relaxation time of the M/D/1 system is very close to $\frac{1}{2}$ of that of the M/M/1 system. It is clear, therefore, that in the transient state $E[W(t)]_{M/D/1} > \frac{1}{2} E[W(t)]_{M/M/1}$, and, in fact, they will be very close to each other. This closeness is the basis for the success of Model 5. Furthermore, the larger ρ becomes, the closer the E[W(t)] of the M/D/1 and M/M/1 systems get, for the same value of t, because the relaxation time increases so rapidly. For the same reason, the results of Model 3 are very satisfactory for demand profiles with small ρ ($\rho \leq 0.7$), but become unreliable (generally too high) when the demand profile contains periods with ρ close to 1.

While investigating Model 6, it became apparent that an equivalent form (given as Model 8) was amenable to efficient numerical evaluation. Model 8 is a method for computing the E[W(t)] based on the solution of the exact transient state probabilities for the M/M/1/∞ queue, appropriately modified so as to yield results even when $\rho > 1$. In certain cases, it may be advantageous to use Model 8 over Model 5. The advantages of Model 8 are derived from the following two properties:

(1) Starting with an arbitrary initial queue length distribution, Model 8 gives the state probabilities of the system (for constant ρ) at an arbitrary period of time thereafter.

(2) Once the system of equations has been solved for one set of initial conditions, the state probabilities for any other arbitrary set of initial conditions are trivially easy to obtain.

Property 1 is important because it eliminates the need for intermediate computations in obtaining the delay at some point of time in the future. This is not the case with Model 5, which must solve the equations at many intermediate points. This property was exploited when Model 8 was programmed to compute the delays for time dependent demand profiles with resultant substantial savings in computer time. Should the delay at more intermediate points be required, the numerical analysis of the transient behavior under varying initial conditions given in Chapter 4 suggests the way interpolation should proceed.

Property 2 is extremely useful for sensitivity analysis. It is clear that the delays during peak periods dominate the total daily delay. In a number of instances, it may be desirable to know how the peak period delay varies when the conditions prior to the peak are varied. Property 2 makes this very simple to do by allowing the delay for additional initial distributions to be calculated with little computational penalty.

Chapter 4 provides many numerical results for various M/G/1 queues and compares these to the behavior hypothesized in earlier chapters. Some (incomplete) extensions of Model 6 to the M/M/k case are also presented and numerical comparisons made.

Further research is needed on predicting the relaxation time as the system parameters vary. We were able to show that the observed behavior of the relaxation time is exactly as predicted by theoretical results when μ varies, regardless of the transient component. However, the numerical results we present for the relaxation time as ρ varies show significant variability, especially when the system is far from equilibrium. The completion of the derivation of the closed form M/M/k state probabilities, and the computation of its autocorrelation functions would be useful for identifying more precisely the time behavior of these queues.

Also, we have yet to determine a method that provides the time dependent state probabilities for systems other than the M/M/1. The dimensionless equation for M/G/1 systems, and the success of Harris in modeling the time dependent M/G/1 system with Model 4 suggest perhaps that the results of the M/M/1 queue in transient state could be used to predict the transient waiting time distribution of the M/D/1 or $M/E_c/1$ queues.

References

1. D.P. Gaver, "Diffusion Approximations and Models for Certain Congestion Problems," Journal of Applied Probability, Vol. 5, pp. 607-623, 1968.

2. L.S. Goddard, Mathematical Techniques of Operational Research, Pergamon Press, London, 1963.

3. D. Gross and C.M. Harris, Fundamentals of Queueing Theory, Wiley, New York, 1974.

4. V.P. Gupta, "An Approximate Queueing Model for the M/M/N/∞ Problem with Nonstationary Demands," The MITRE Corporation, McLean, Virginia, Technical Memorandum D43-M3579, February 1975.

5. V.P. Gupta, "Extension of the Approximate Queueing Model to M/D/N/∞ and M/G/N/∞ Problems," The MITRE Corporation, McLean, Virginia, Technical Memorandum D43-M3579, March 1975.

6. R.M. Harris, Prediction and Optimization for the Queue M/G/1 with Non-Stationary Demands, The MITRE Corporation, McLean, Virginia, Technical Report M72-21, December 1971.

7. G. Hengsbach, and A. Odoni, Time Dependent Estimates of Delays and Delay Costs at Airports, FTL Report R75-4, January, 1975.

8. B.O. Koopman, "Air Terminal Queues under Time Dependent Conditions," Operations Research, Vol. 20, No. 6, pp. 1089-1114.

9. P.M. Morse, "Stochastic Properties of Waiting Lines," Operations Research, Vol. 3, pp. 255-251, 1955.

10. P.M. Morse, Queues, Inventories and Maintenance, Wiley, New York, 1958.

11. G.F. Newell, "Queues with Time Dependent Arrival Rates I - The Transition Through Saturation," Journal of Applied Probability, Vol. 5, pp. 436-451, 1968.

12. T.L. Saaty, Elements of Queueing Theory, McGraw-Hill, New York, 1961.

13. J. Scalea, "A Comparison of Several Methods for the Calculation of Airside Airport Delay," S.M. Thesis, M.I.T., June 1976.

14. G.N. Steuart, "Gate Position Requirements at Metropolitan Airports," Transportation Science, Vol. 8, No. 2, pp. 169-189.