DATA NETWORKS

Dimitri Bertsekas

Massachusetts Institute of Technology

Robert Gallager

Massachusetts Institute of Technology

Prentice-Hall International, Inc.

Reproduced with Publisher's Permission

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-196981-1 025

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Prentice-Hall of Southeast Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*
Prentice-Hall, *Englewood Cliffs, New Jersey*

Contents

Chapter 2 DATA LINK CONTROL AND COMMUNICATION CHANNELS 31

Chapter 6
FLOW CONTROL 423

Preface

The field of data networks has evolved over the last fifteen years from a stage where networks were designed in a very ad hoc and technology-dependent manner to a stage where some broad conceptual understanding of many underlying issues now exists. The major purpose of this book is to convey that conceptual understanding to the reader.

Previous books in this field broadly separate into two major categories. The first, exemplified by Tannenbaum [Tan81] and Stallings [Sta85], are primarily descriptive in nature, focusing on current practice and selected details of the operation of various existing networks. The second, exemplified by Kleinrock [Kle76], Hayes [Hay84], and Stuck and Arthurs [StA85], deal primarily with performance analysis. This book, in contrast, is balanced between description and analysis. The descriptive material, however, is used to illustrate the underlying concepts, and the analytical material is used to provide a deeper and more precise understanding of the concepts. We feel that a continuing separation between description and analysis is unwise in a field after the underlying concepts have been developed; understanding is then best enhanced by focusing on the concepts.

The book is designed to be used at a number of levels, varying from a senior undergraduate elective, to a first year graduate course, to a more advanced graduate course, to a reference work for designers and researchers in the field. The material has been tested in a number of graduate courses at M.I.T. and in a number of short courses at varying levels. The book assumes some background in elementary probability and some background in either electrical engineering or computer science, but aside from this, the material is self-contained.

Throughout the book, major concepts and principles are first explained in a simple non-mathematical way. This is followed by careful descriptions of modelling issues and then by mathematical analysis. Finally, the insights to be gained from the analysis are explained and examples are given to clarify the more subtle issues. Figures are liberally used throughout to illustrate the ideas. For lower-level courses, the analysis can be glossed over; this allows the beginning and intermediate-level to grasp the basic ideas, while enabling the more advanced student to acquire deeper understanding and the ability to do research in the field.

Chapter 1 provides a broad introduction to the subject and also develops the layering concept. This layering allows the various issues of data networks to be developed in a largely independent fashion, thus making it possible to read the subsequent chapters in any desired depth (including omission) without seriously hindering the ability to understand other chapters.

Chapter 2 treats the two lowest layers of the above layering. The lowest, or physical, layer is concerned with transmitting a sequence of bits over a physical communication medium. We provide a brief introduction to the subject which will be helpful but not necessary in understanding the rest of the text. The next layer, data link control, deals with transmitting packets reliably over a communication link. Section 2.4, treating retransmission strategies, should probably be covered in any course, since it brings out the subtleties, in the simplest context, of understanding distributed algorithms, or protocols.

Chapter 3 develops the queueing theory used for performance analysis of multiaccess schemes (Chapter 4) and, to a lesser extent, routing algorithms (Chapter 5). Less analytical courses will probably omit most of this chapter, simply adopting the results on faith. Little's theorem and the Poisson process should be covered however, since they are simple and greatly enhance understanding of the subsequent chapters. This chapter is rich in results, often developed in a far simpler way than found in the queueing literature. This simplicity is achieved by considering only steady-state behavior and by sometimes sacrificing rigor for clarity and insight. Mathematically sophisticated readers will be able to supply the extra details for rigor by themselves, while for most readers the extra details would obscure the line of argument.

Chapter 4 develops the topic of multiaccess communication, including local area networks, satellite networks, and radio networks. Less theoretical courses will probably skip the last half of section 4.2, all of section 4.3, and most of section 4.4, getting quickly to local area networks and satellite networks in section 4.5. Conceptually, one gains a great deal of insight into the nature of distributed algorithms in this chapter.

Chapter 5 develops the subject of routing. The material is graduated in order of increasing difficulty and depth, so readers can go as far as they are comfortable. Along with routing itself, which is treated in greater depth than elsewhere in the literature, further insights are gained into distributed algorithms. There is also a treatment of topological design and a section on recovery from link failures.

Chapter 6 deals with flow control (or congestion control as it is sometimes called). The first three sections are primarily descriptive, describing first the objectives and the problems in achieving these objectives, second, some general approaches, and finally, the ways that flow control is handled in several existing networks. The last section is more advanced and analytical, treating recent work in the area.

A topic that is not treated in any depth in the book is that of higher-layer protocols, namely the various processes required in the computers and devices using the network to communicate meaningfully with each other given the capability of reliable transport of packets through the network provided by the lower layers. This topic is different in nature than the other topics covered and would have doubled the size of the book if treated in depth.

We apologize in advance for the amount of acronyms and jargon in the book. We felt it was necessary to include at least the most commonly used acronyms in the field, both to allow readers to converse with other workers in the field and also for the reference value of being able to find out what these acronyms mean.

An extensive set of problems are given at the end of each chapter except the first. They range from simple exercises to gain familiarity with the basic concepts and techniques to advanced problems extending the results in the text. Solutions of the problems are given in a manual available to instructors from Prentice-Hall.

Each chapter contains also a brief section of sources and suggestions for further reading. Again, we apologize in advance to the many authors whose contributions have not been mentioned. The literature in the data network field is vast, and we limited ourselves to references that we found most useful, or that contain material supplementing the text.

The stimulating teaching and research environment at M.I.T. has been an ideal setting for the development of this book. In particular we are indebted to the many students who have used this material in courses. Their comments have helped greatly in clarifying the topics. We are equally indebted to the many colleagues and advanced graduate students who have provided detailed critiques of the various chapters. Special thanks go to our colleague Pierre Humblet whose advice, knowledge, and deep insight have been invaluable. In addition, Erdal Arikan, David Castanon, Robert Cooper, Tony Ephremides, Eli Gafni, Marianne Gardner, Paul Green, Ellen Hahne, Bruce Hajek, Robert Kennedy, John Spinelli, and John Tsitsiklis have all been very helpful. We are also grateful to Nancy Young for typing the many revisions and to Amy Hendrikson for computer typesetting the book using the TEX system. Our editors at Prentice-

Hall have also been very helpful and cooperative in producing the final text under a very tight schedule. Finally we wish to acknowledge the research support of DARPA under grant ONR-N00014-84-K-0357, NSF under grants ECS-8310698, and ECS-8217668, and ARO under grant DAAG 29-84-K-000.

Dimitri Bertsekas

Robert Gallager

3

Delay Models in Data Networks

3.1 INTRODUCTION

One of the most important performance measures of a data network is the average delay required to deliver a packet from origin to destination. Furthermore, delay considerations strongly influence the choice and performance of network algorithms, such as routing and flow control. For these reasons, it is important to understand the nature and mechanism of delay, and the manner in which it depends on the characteristics of the network.

Queueing theory is the primary methodological framework for analyzing network delay. Its use often requires simplifying assumptions since, unfortunately, more realistic assumptions make meaningful analysis extremely difficult. For this reason, it is sometimes impossible to obtain accurate quantitative delay predictions on the basis of queueing models. Nevertheless, these models often provide a basis for adequate delay approximations, as well as valuable qualitative results and worthwhile insights.

In what follows, we will focus on packet delay within the communication subnet (*i.e.*, the network layer). This delay is the sum of delays on each subnet link traversed by the packet. Each link delay in turn consists of four components.

1. The *processing* delay between the time the packet is correctly received at the head node of the link and the time the packet is assigned to an outgoing link

queue for transmission. (In some systems, we must add to this delay some additional processing time at the DLC and physical layers.)

2. The *queueing delay* between the time the packet is assigned to a queue for transmission and the time it starts being transmitted. During this time, the packet waits while other packets in the transmission queue are transmitted.
3. The *transmission delay* between the times that the first and last bits of the packet are transmitted.
4. The *propagation delay* from the time the last bit is transmitted at the head node of the link until the time it is received at the tail node. This is proportional to the physical distance between transmitter and receiver and is ordinarily small except in the case of a satellite link.

This accounting neglects the possibility that a packet may require retransmission on a link due to transmission errors or various other causes. For most links in practice, other than multiaccess links to be considered in Chapter 4, retransmissions are rare and will be neglected. The propagation delay depends on the physical characteristics of the link and is independent of the traffic carried by the link. The processing delay is also independent of the amount of traffic handled by the corresponding node if computation power is not a limiting resource. This will be assumed in our discussion. Otherwise, a separate processing queue must be introduced prior to the transmission queues. Most of our subsequent analysis focuses on the queueing and transmission delays. We first consider a single transmission line and analyze some classical queueing models. We then take up the network case and discuss the type of approximations involved in deriving analytical delay models.

While our primary emphasis is on packet-switched network models, some of the models developed are useful in a circuit-switched network context. Indeed, queueing theory was extensively developed in response to the need for performance models in telephony.

3.1.1 Multiplexing of Traffic on a Communication Link

The communication link considered is viewed as a bit pipe over which a given number of bits per second can be transmitted. This number is called the *transmission capacity* of the link. It depends both on the physical channel and the interface (*e.g.*, modems), and is simply the rate at which the interface accepts bits. The link capacity may serve several traffic streams (*e.g.*, virtual circuits or groups of virtual circuits) multiplexed on the link. The manner of allocation of capacity among these traffic streams has a profound effect on packet delay.

In the most common scheme, *statistical multiplexing*, the packets of all traffic streams are merged into a single queue and transmitted on a first-come first-serve basis. A variation of this scheme, which has roughly the same average delay per packet, maintains a separate queue for each traffic stream and serves the queues in sequence one packet at a time. However, if the queue of a traffic stream is empty, the next traffic stream is served and no communication resource is wasted. Since the entire transmission capacity C (bits/sec) is allocated to a single packet at a time, it takes L/C sec to transmit a packet that is L bits long.

In *time-division* (TDM) *and frequency-division multiplexing* (FDM) with m traffic streams, the link capacity is essentially subdivided into m portions—one per traffic stream. In FDM, the channel bandwidth W is subdivided into m channels each with bandwidth W/m (actually slightly less because of the need for guard bands between channels). The transmission capacity of each channel is roughly C/m, where C is the capacity that would be obtained if the entire bandwidth was allocated to a single channel. The transmission time of a packet that is L bits long is Lm/C, or m times longer than in the corresponding statistical multiplexing scheme. In TDM, allocation is done by dividing the time axis into slots of fixed length (*e.g.*, one bit or one byte long, or perhaps one packet long for fixed length packets). Again, conceptually, we may view the communication link as consisting of m separate links with capacity C/m. In the case where the slots are short relative to packet length, we may again regard the transmission time of a packet L bits long as Lm/C. In the case where the slots are of packet length, the transmission time of an L bit packet is L/C, but there is a wait of $(m-1)$ packet transmission times between packets of the same stream.

One of the themes that will emerge from our queueing analysis is that statistical multiplexing has smaller average delay per packet than either TDM or FDM. This is particularly true when the traffic streams multiplexed have a relatively low duty cycle. The main reason for the poor delay performance of TDM and FDM is that communication resources are wasted when allocated to a traffic stream with a momentarily empty queue, while other traffic streams have packets waiting in their queue. For a traffic analogy, consider an m-lane highway and two cases. In one case, cars are not allowed to cross over to other lanes (this corresponds to TDM or FDM), while in the other case, cars can change lanes (this corresponds roughly to statistical multiplexing). Restricting crossover increases travel time for the same reason that the delay characteristics of TDM or FDM are poor, namely some system resources (highway lanes or communication channels) may not be utilized while others are momentarily stressed.

Under certain circumstances, TDM or FDM may have an advantage. Suppose that each traffic stream has a "regular" character, *i.e.*, all packets arrive sufficiently apart so that no packet has to wait while the preceding packet is transmitted. If these traffic streams are merged into a single queue, it can be shown that the average delay per packet will decrease, but the variance of waiting time in queue will generally become positive (for an illustration see Prob. 3.7). Therefore, if maintaining small variability of delay is more important than decreasing delay, it may be preferable to use TDM or FDM. Another advantage of TDM and FDM is that there is no need to include identification of the traffic stream on each packet, thereby saving some overhead.

3.2 QUEUEING MODELS—LITTLE'S THEOREM

We consider queueing systems where customers arrive at random times to obtain service. The probability distribution of the time between two successive arrivals (the interarrival time), and the probability distribution of the customers' service time are given.

In the context of a data network, customers represent packets assigned to a communication link for transmission. Service time corresponds to the packet transmission time and is equal to L/C, where L is the packet length in bits and C is the link transmission capacity in bits/sec. In this chapter it is convenient to ignore the layer 2 distinction between packets and frames; thus packet lengths are taken to include frame headers and trailers. In a somewhat different context (which we will not dwell on very much), customers represent active conversations (or virtual circuits) between points in a network and service time corresponds to the duration of a conversation.

We shall be typically interested in estimating quantities such as:

1. The average number of customers in the system (*i.e.*, the "typical" number of customers either waiting in queue or undergoing service).
2. The average delay per customer (*i.e.*, the "typical" time a customer spends waiting in queue plus the service time).

We first need to clarify the meaning of the terms above. Let us denote

$$p_n(t) = \text{Probability of } n \text{ customers waiting in queue or under service at time } t$$

The typical situation is one whereby we are given the initial probabilities $p_n(0)$ at time 0 and enough statistical information is provided to determine, at least in principle, the probabilities $p_n(t)$ for all times t. Then denoting

$$\overline{N}(t) = \text{Average number in the system at time } t$$

we have

$$\overline{N}(t) = \sum_{n=0}^{\infty} n\, p_n(t)$$

Note that both $\overline{N}(t)$ and $p_n(t)$ depend on t as well as the initial probability distribution $\{p_0(0),\ p_1(0),\ldots\}$. However, the queueing systems that we will consider typically ***reach equilibrium*** in the sense that for some p_n and N (independent of the initial distribution), we have

$$\lim_{t\to\infty} p_n(t) = p_n\,, \quad n = 0, 1, \ldots \tag{3.1}$$

and

$$N = \sum_{n=0}^{\infty} np_n = \lim_{t\to\infty} \overline{N}(t)$$

We will be interested primarily in the equilibrium probabilities and the average number in the system. Note that it is possible that $N = \infty$ and this will occur whenever the arrival rate exceeds the service capacity of the system. Individual sample functions of the number of customers in the system will be denoted by $N(t)$. The time average of such a sample function in the interval $[0, t]$ is defined by

$$N_t = \frac{1}{t}\int_0^t N(\tau)\, d\tau$$

Almost every system of interest to us is ergodic in the sense that

$$\lim_{t\to\infty} N_t = \lim_{t\to\infty} \overline{N}(t) = N$$

holds with probability one. The equality of long term time average and ensemble average of various stochastic processes will often be accepted in this chapter on intuitive grounds since a rigorous mathematical justification requires technical arguments that are beyond the scope of this text.

Regarding average delay per customer, the situation is one whereby enough statistical information is available to determine in principle the probability distribution of delay of each individual customer (*i.e.*, the first, second, etc.). From this, we can determine the average delay of each customer. The average delay of the k^{th} customer, denoted $\overline{T}_k$, typically converges as $k \to \infty$ to a steady-state value

$$T = \lim_{k\to\infty} \overline{T}_k$$

The limit above is what we will call average delay per customer. (Again, $T = \infty$ is possible.) For the systems of interest to us, the steady-state average delay T is also equal (with probability one) to the long-term time average of customer delay, *i.e.*,

$$T = \lim_{k\to\infty} \overline{T}_k = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} T_i$$

where T_i is the delay of the i^{th} customer.

The average number in the system N and the average delay T are related by a simple formula that makes it possible to determine one given the other. This result, known as *Little's Theorem*, has the form

$$N = \lambda T$$

where

$$\lambda = \text{Average customer arrival rate}$$

and is given by

$$\lambda = \lim_{t\to\infty} \frac{\text{Expected number of arrivals in the interval } [0, t]}{t}$$

(We will be assuming that the limit above exists.) Phenomena reflecting Little's Theorem are familiar from everyday experience. For example, on a rainy day, traffic on a rush hour moves slower than average (large T) while the streets are more crowded (large N). Similarly, a fast-food restaurant (small T) needs a smaller waiting room (small N) than a regular restaurant for the same customer arrival rate.

Little's Theorem is really an accounting identity and its derivation is very simple. We will give a graphical proof, which assumes that customers are served in the order they arrive. A similar proof is possible for the case where the order of service is arbitrary (see Problems 3.31 and 3.32). For any sample system history let us denote:

$$\alpha(t) = \text{Number of arrivals in the interval } [0,t]$$
$$\beta(t) = \text{Number of departures in the interval } [0,t]$$

Assuming an empty system at time 0, the number in the system at time t is

$$N(t) = \alpha(t) - \beta(t)$$

Let t_i and T_i be the time of arrival and the time spent in the system, respectively, by the i^{th} customer. Consider any time t and the shaded area in Fig. 3.1 which lies between the graphs of $\alpha(\tau)$ and $\beta(\tau)$ up to time t. This area can be expressed as

$$\int_0^t N(\tau)\,d\tau$$

but also as

$$\sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i)$$

Dividing both expressions above by t and equating them, we obtain

$$N_t = \lambda_t T_t \tag{3.2}$$

where

$$N_t = \frac{\int_0^t N(\tau)\,d\tau}{t} = \text{Time average of the number of customers in the system in the interval } [0,t]$$

$$\lambda_t = \frac{\alpha(t)}{t} = \text{Time average of the customer arrival rate in the interval } [0,t]$$

$$T_t = \frac{\sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - t_i)}{\alpha(t)} = \text{Time average of the time a customer spends in the system in the interval } [0,t]$$

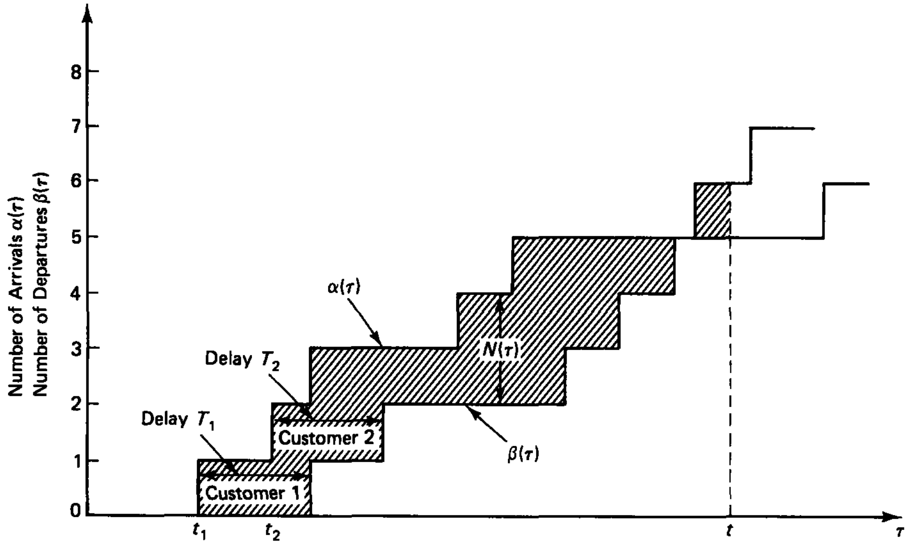


Figure 3.1 Proof of Little's Theorem. The shaded area can be expressed both as $\int_0^t N(\tau)\,d\tau$ and as $\sum_{i=1}^{\beta(t)} T_i + \sum_{i=\beta(t)+1}^{\alpha(t)} (t-t_i)$. Dividing both expressions by t, equating them, and taking the limit as $t \to \infty$ gives Little's Theorem.

Assuming that

$$N_t \to N,\ \lambda_t \to \lambda,\ T_t \to T$$

we obtain from Eq. (3.2) the desired formula.

Note that the expression T_t includes the total time spent in the system for all the arrivals from 1 to $\beta(t)$, but omits the time spent beyond t for the customers still in the system at time t. Assuming that $N_t \to N < \infty$, this end effect due to customers in the system at time t will be small relative to the accumulated time in the system of customers 1 to $\beta(t)$, and T_t for large t can be interpreted as the time average of the system time.

Strictly speaking, for the argument above to be correct, we must be assured that the time averages N_t, λ_t, T_t converge with probability one to the corresponding ensemble averages N, λ, and T. This is true in just about every case of interest to us, and in subsequent analysis, we will accept Little's Theorem without further scrutiny.

The significance of Little's Theorem is due in large measure to its generality. It holds for almost every queueing system that reaches statistical equilibrium in the limit. The system need not consist of just a single queue. Indeed, the theorem holds for many complex arrival-departure systems with appropriate interpretation of the terms N, λ, and T. The following examples illustrate its broad applicability.

Example 1

If λ is the arrival rate in a transmission line, N_Q is the average number of packets waiting in queue (but not under transmission), and W is the average time spent by a packet waiting in queue (not including the transmission time), Little's Theorem gives

$$N_Q = \lambda W$$

Furthermore if $\overline{X}$ is the average transmission time, then Little's Theorem gives the average number of packets under transmission as

$$\rho = \lambda \overline{X}$$

Since at most one packet can be under transmission, ρ is also the line's *utilization factor*, *i.e.*, the proportion of time that the line is busy transmitting a packet.

Example 2

Consider a network of transmission lines where packets arrive at n different nodes with corresponding rates $\lambda_1, \ldots, \lambda_n$. If N is the average total number of packets inside the network, then (regardless of the packet length distribution and method for routing packets) the average delay per packet is

$$T = \frac{N}{\sum_{i=1}^{n} \lambda_i}$$

Furthermore, Little's Theorem also yields $N_i = \lambda_i T_i$, where N_i and T_i are the average number in the system and average delay of packets arriving at node i, respectively.

Example 3

A packet arrives at a transmission line every K seconds with the first packet arriving at time 0. All packets have equal length and require αK seconds for transmission where $\alpha < 1$. The processing and propagation delay per packet is P seconds. The arrival rate here is $\lambda = 1/K$. Because packets arrive at a regular rate (equal interarrival times), there is no delay for queueing, so the time T a packet spends in the system (including the propagation delay) is

$$T = \alpha K + P$$

According to Little's Theorem, we have

$$N = \lambda T = \alpha + \frac{P}{K}$$

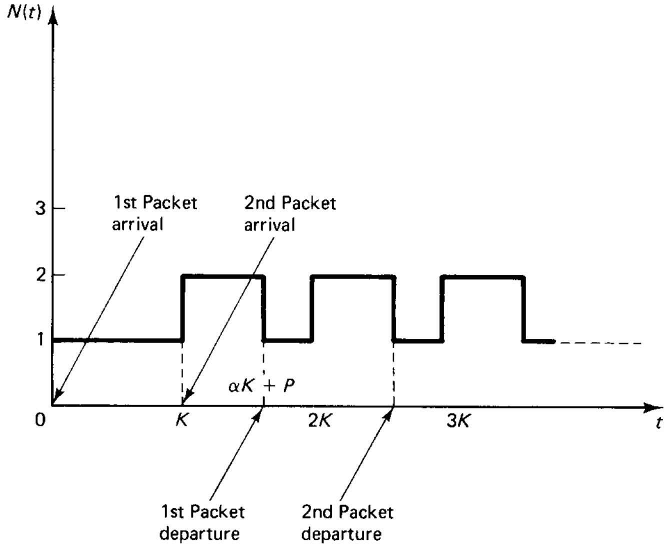


Figure 3.2 The number in the system in Example 3, $N(t)$, is deterministic and does not converge as $t \rightarrow \infty$. However, Little's Theorem holds if N, λ, and T are interpreted as time averages.

One should be careful about interpreting correctly the formula in this example. Here the number in the system $N(t)$ is a deterministic function of time. Its form is shown in Fig. 3.2 for the case where $K < \alpha K + P < 2K$, and it can be seen that $N(t)$ does not converge to any value (the system never reaches statistical equilibrium.) However, Little's Theorem is correct provided N is viewed as a long-term *time average* of $N(t)$, *i.e.*,

$$N = \lim_{t \rightarrow \infty} \frac{\int_0^t N(\tau)\, d\tau}{t}$$

Example 4

Consider a window flow control system (as described in subsection 2.8.1) with a window of size N for each session. Suppose that a session always has packets to send and that acknowledgements take negligible time; then, when packet i arrives at the destination, packet $i + N$ is immediately introduced into the network. Since the number of packets in the system per session is always N, Little's Theorem asserts that the arrival rate λ of packets into the system for each session, and the average packet delay are related by $N = \lambda T$. Thus, if congestion builds up in the network and T increases, λ must decrease. Note also that if the network is congested and capable of delivering only λ packets per unit time for each session, then increasing the window size N for all sessions merely serves to increase the delay T.

Example 5

Consider a queueing system with K servers, and with room for at most $N \geq K$ customers (either in queue or in service). The system is always full; we assume that it starts with N customers and that a departing customer is immediately replaced by a new customer. (Queueing systems of this type are called *closed.*) Suppose that the average customer service time is $\overline{X}$. We want to find the average customer time in the system T. We apply Little's Theorem twice, first for the entire system, obtaining $N = \lambda T$, and then for the service portion of the system, obtaining $K = \lambda\overline{X}$ (since all servers are constantly busy). By eliminating λ in these two relations we have

$$T = \frac{N\overline{X}}{K}$$

Example 6: Estimating throughput in a time-sharing system

Little's Theorem can sometimes be used to provide bounds on the attainable system throughput λ. In particular, known bounds on N and T can be translated into throughput bounds via $\lambda = N/T$. As an example, consider a time-sharing computer system with N terminals. A user logs into the system through a terminal, and, after an initial reflection period of average length R, submits a job that requires an average processing time P at the computer. Jobs queue up inside the computer and are served by a single CPU according to some unspecified priority or time-sharing rule.

We would like to get estimates of the throughput sustainable by the system (in jobs per unit time), and corresponding estimates of the average delay of a user. Since we are interested in maximum attainable throughput, we assume that there is always a user ready to take the place of a departing user, so the number of users in the system is always N. For this reason, it is appropriate to adopt a model whereby a departing user immediately reenters the system as shown in Fig. 3.3.

Applying Little's Theorem to the portion of the system between the entry to the terminals and the exit of the system (points A and C in Fig. 3.3), we have

$$\lambda = \frac{N}{T} \tag{3.3}$$

where T is the average time a user spends in the system. We have

$$T = R + D \tag{3.4}$$

where D is the average delay between the time a job is submitted to the computer and the time its execution is completed. Since D can vary between P (case where the user's job does not have to wait for other jobs to be completed) and NP (case

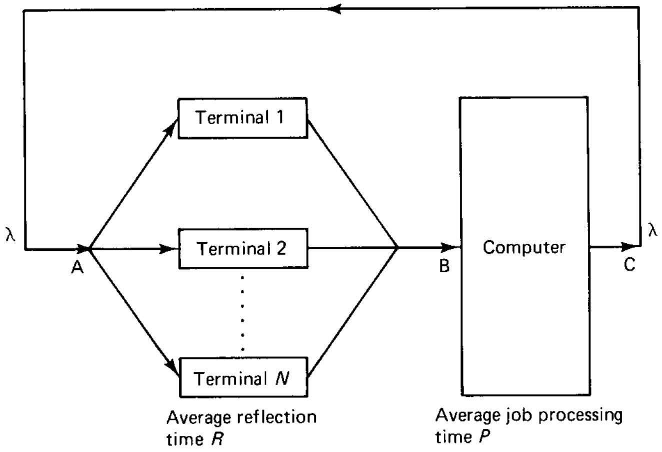


Figure 3.3 N terminals connected with a time-sharing computer system. To estimate maximum attainable throughput, we assume that a departing user immediately reenters the system or, equivalently, is immediately replaced by a new user.

where the user's job has to wait for the jobs of all the other users; compare with Ex. 5), we have

$$R + P \leq T \leq R + NP \tag{3.5}$$

Combining this relation with Eq. (3.3), we obtain

$$\frac{N}{R + NP} \leq \lambda \leq \frac{N}{R + P} \tag{3.6}$$

The throughput λ is also bounded above by the processing capacity of the computer. In particular, since the execution time of a job is P units on the average, it follows that the computer cannot process in the long run more than $1/P$ jobs per unit time, *i.e.*,

$$\lambda \leq \frac{1}{P} \tag{3.7}$$

(This conclusion can also be reached by applying Little's Theorem between the entry and exit points of the computer's CPU.)

By combining relations (3.6) and (3.7), we obtain the bounds

$$\frac{N}{R + NP} \leq \lambda \leq \min\left\{\frac{1}{P}, \frac{N}{R + P}\right\} \tag{3.8}$$

for the maximum attainable throughput. By using $T = N/\lambda$, we also obtain bounds for the average user delay when the system is fully loaded

$$\max\{NP,\ R+P\} \le T \le R+NP \tag{3.9}$$

These relations are illustrated in Fig. 3.4.

It can be seen that as the number of terminals N increases, the throughput approaches the maximum $1/P$, while the average user delay rises essentially in direct proportion with N. The number of terminals becomes a throughput bottleneck when $N < 1 + R/P$, in which case the computer resource stays idle for a substantial portion of the time while all users are engaged in reflection. In contrast, the limited processing power of the computer becomes the bottleneck when $N > 1 + R/P$. It is interesting to note that while the exact maximum attainable throughput depends on system parameters, such as the statistics of the reflection and processing times, and the manner in which jobs are served by the CPU, the bounds obtained are independent of these parameters. We owe this convenient situation to the generality of Little's Theorem.

3.3 THE M/M/1 *QUEUEING SYSTEM*

The $M/M/1$ queueing system consists of a single queueing station with a single server (in a communication context, a single transmission line). Customers arrive according to a Poisson process with rate λ, and the probability distribution of the service time is exponential with mean $1/\mu$ sec. We will explain the meaning of these terms shortly. The name $M/M/1$ reflects standard queueing theory nomenclature whereby:

1. The first letter indicates the nature of the arrival process (*e.g.*, M stands for memoryless, which here means a Poisson process (*i.e.*, exponentially distributed interarrival times), G stands for a general distribution of interarrival times, D stands for deterministic interarrival times).
2. The second letter indicates the nature of the probability distribution of the service times (*e.g.*, M, G, and D stand for exponential, general, and deterministic distributions, respectively). In all cases, successive interarrival times and service times are assumed to be statistically independent of each other.
3. The last number indicates the number of servers.

We have already established, via Little's Theorem, the relations

$$N = \lambda T, \qquad N_Q = \lambda W$$

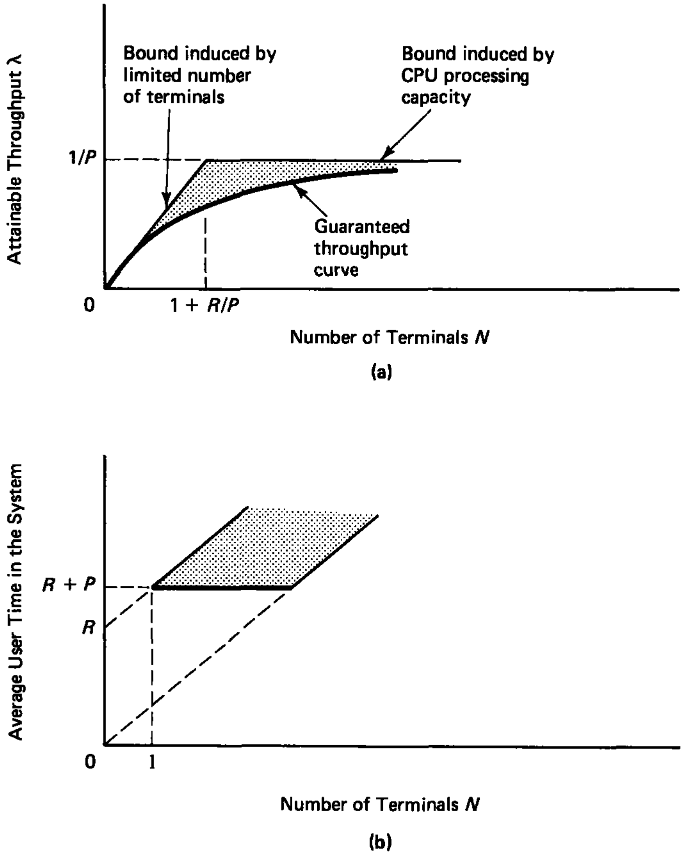


Figure 3.4 Bounds on throughput and average user delay in a time-sharing system. (a) Bounds on attainable throughput [Eq. (3.8)].
(b) Bounds on average user time in a fully loaded system [Eq. (3.9)]. The time increases essentially in proportion with the number of terminals N.

between the basic quantities,

N : Average number of customers in the system

T : Average customer time in the system

N_Q : Average number of customers waiting in queue

W : Average customer waiting time in queue

However, N, T, N_Q, and W cannot be specified further unless we know something more about the statistics of the system. Given these statistics, we will be able to derive the steady-state probabilities

$$p_n = \text{ Probability of } n \text{ customers in the system, } n = 0, 1, \ldots$$

From these probabilities, we can get

$$N = \sum_{n=0}^{\infty} n p_n$$

and, using Little's Theorem,

$$T = \frac{N}{\lambda}$$

Similar formulas exist for N_Q and W. Appendix B provides a summary of the results for the $M/M/1$ system and the other major systems analyzed later.

The analysis of the $M/M/1$ system as well as several other related systems, such as the $M/M/m$ or $M/M/\infty$ systems, is based on the theory of Markov chains summarized in Appendix A. An alternative approach is to use simple graphical arguments based on the concept of mean residual time introduced in section 3.5. This approach does not require that the service times are exponentially distributed, *i.e.*, it applies to the $M/G/1$ system. The price paid for this generality is that the characterization of the steady-state probabilities is less convenient and simple than for the $M/M/1$ system. The reader wishing to circumvent the Markov chain analysis may start directly with the $M/G/1$ system in section 3.5 after a reading of the preliminary facts on the Poisson process given in subsections 3.3.1 and 3.3.2.

3.3.1 Main Results

A stochastic process $\{A(t) \mid t \geq 0\}$ taking nonnegative integer values is said to be a *Poisson process* with rate λ if

1. $A(t)$ is a counting process that represents the total number of arrivals that have occurred from 0 to time t, *i.e.*, $A(0) = 0$, and for $s < t$, $A(t) - A(s)$ equals the number of arrivals in the interval $(s, t]$.
2. The numbers of arrivals that occur in disjoint time intervals are independent.

3. The number of arrivals in any interval of length τ is Poisson distributed with parameter $\lambda\tau$. That is, for all t, $\tau > 0$,

$$P\{A(t+\tau) - A(t) = n\} = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}, \quad n = 0, 1, \ldots \tag{3.10}$$

We list some of the properties of the Poisson process that will be of interest:

(a) Interarrival times are independent and exponentially distributed with parameter λ, *i.e.*, if t_n denotes the time of the n^{th} arrival, the intervals $\tau_n = t_{n+1} - t_n$ have the probability distribution

$$P\{\tau_n \leq s\} = 1 - e^{-\lambda s}, \quad s \geq 0 \tag{3.11}$$

and are mutually independent. (The corresponding probability density function is $p(\tau_n) = \lambda e^{-\lambda\tau_n}$. The mean and variance of τ_n are $1/\lambda$ and $1/\lambda^2$, respectively.)

(b) For every $t \geq 0$ and $\delta \geq 0$

$$P\{A(t+\delta) - A(t) = 0\} = 1 - \lambda\delta + o(\delta) \tag{3.12}$$

$$P\{A(t+\delta) - A(t) = 1\} = \lambda\delta + o(\delta) \tag{3.13}$$

$$P\{A(t+\delta) - A(t) \geq 2\} = o(\delta) \tag{3.14}$$

where we generically denote by $o(\delta)$ a function of δ such that

$$\lim_{\delta\to 0}\frac{o(\delta)}{\delta} = 0$$

These equations can be verified using Eq. (3.10) (see Prob. 3.10).

Note that if the arrivals in n disjoint intervals are independent and Poisson distributed with parameters $\lambda\tau_1, \ldots, \lambda\tau_n$, then the number of arrivals in the union of the intervals is Poisson distributed with parameter $\lambda(\tau_1 + \cdots + \tau_n)$. This follows from properties of the Poisson distribution and guarantees that the requirement of Eq. (3.10) is consistent with the independence requirement in the definition of the Poisson process (see Prob. 3.10). Another fact that we will frequently use is that if two or more independent Poisson processes $A_1, \ldots, A_k$ are merged into a single process $A = A_1 + A_2 + \cdots + A_k$, then the latter process is Poisson with a rate equal to the sum of the rates of its components (see Prob. 3.10).

Our assumption regarding the service process is that *the customer service times have an exponential distribution with parameter μ*, *i.e.*, if s_n is the service time of the n^{th} customer,

$$P\{s_n \leq s\} = 1 - e^{-\mu s}, \quad s \geq 0$$

(The probability density function of s_n is $p(s_n) = \mu e^{-\mu s_n}$, and its mean and variance are $1/\mu$ and $1/\mu^2$, respectively.) Furthermore, *the service times s_n are mutually independent and also independent of all interarrival times.* The parameter μ is called the *service rate*, and represents the rate (in customers served per unit time) at which the server operates when busy.

An important fact regarding the exponential distribution is its *memoryless*, character, which can be expressed as

$$P\{\tau_n > r + t | \tau_n > t\} = P\{\tau_n > r\}, \quad \text{for } r, t \geq 0$$

$$P\{s_n > r + t | s_n > t\} = P\{s_n > r\}, \quad \text{for } r, t \geq 0$$

for the interarrival and service times τ_n and s_n, respectively. This means that the additional time needed to complete a customer's service in progress is independent of when the service started. Similarly, the time up to the next arrival is independent of when the previous arrival occurred. Verification of the memoryless property follows from the calculation

$$P\{\tau_n > r + t | \tau_n > t\} = \frac{P\{\tau_n > r + t\}}{P\{\tau_n > t\}} = \frac{e^{-\lambda(r+t)}}{e^{-\lambda t}} = e^{-\lambda r} = P\{\tau_n > r\}$$

The memoryless property together with our earlier independence assumptions on interarrival and service times imply that once we know the number $N(t)$ of customers in the system at time t, the times at which customers will arrive or complete service in the future are independent of the arrival times of the customers presently in the system and of how much service the customer currently in service (if any) has already received. This means that $\{N(t) | t \geq 0\}$ *is a continuous-time Markov chain.*

We could analyze the process $N(t)$ in terms of continuous-time Markov chain methodology; most of the queueing literature follows this line of analysis. It is sufficient, however, for our purposes in this section to use the simpler theory of discrete-time Markov chains (briefly summarized in Appendix A).

Let us focus attention at the times

$$0,\ \delta,\ 2\delta, \ldots, k\delta, \ldots$$

where δ is a small positive number. We denote

$$N_k = \text{Number of customers in the system at time } k\delta$$

Since $N_k = N(k\delta)$ and, as discussed, $N(t)$ is a continuous-time Markov chain, we see that $\{N_k | k = 0, 1, \ldots\}$ is a discrete-time Markov chain. Let P_{ij} denote the corresponding transition probabilities

$$P_{ij} = P\{N_{k+1} = j | N_k = i\}$$

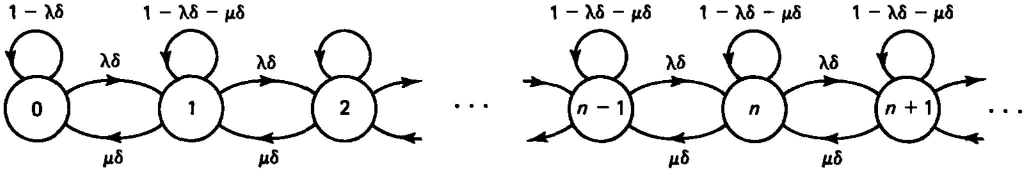


Figure 3.5 Discrete-time Markov chain for the $M/M/1$ system. The state n corresponds to n customers in the system. Transition probabilities shown are correct up to an $o(\delta)$ term.

Note that P_{ij} depends on δ, but to keep notation simple, we do not show this dependence. By using Eqs. (3.12) through (3.14), we have

$$P_{oo}= 1 - \lambda\delta + o(\delta) \qquad (3.15)$$

$$P_{ii}= 1 - \lambda\delta - \mu\delta + o(\delta), \qquad i \geq 1 \qquad (3.16)$$

$$P_{i,i+1}= \lambda\delta + o(\delta), \qquad i \geq 0 \qquad (3.17)$$

$$P_{i,i-1}= \mu\delta + o(\delta), \qquad i \geq 1 \qquad (3.18)$$

$$P_{ij}= o(\delta), \qquad i \text{ and } j \neq i,\, i+1,\, i-1$$

To see how these equations are verified, note that, when at a state $i \geq 1$, the probabilities of 0 arrivals and 0 departures in an interval $I_k = (k\delta, (k+1)\delta]$ is $(e^{-\lambda\delta})(e^{-\mu\delta})$; this is because the number of arrivals and the number of departures are Poisson distributed and independent of each other. Expanding this in a power series in δ,

$$P\{0 \text{ customers arrive and } 0 \text{ depart in } I_k\} = 1 - \lambda\delta - \mu\delta + o(\delta) \qquad (3.19)$$

Similarly, we have that

$$P\{0 \text{ customers arrive and } 1 \text{ departs in } I_k\} = \mu\delta + o(\delta)$$

$$P\{1 \text{ customer arrives and } 0 \text{ depart in } I_k\} = \lambda\delta + o(\delta)$$

These probalities add up to one plus $o(\delta)$. Thus, the probability of more than one arrival or departure is negligible for δ small. This means that, for $i \geq 1$, p_{ii}, which is the probability of an equal number of arrivals and departures in I_k, is within $o(\delta)$ of the value in Eq. (3.19); this verifies Eq. (3.16). Equations (3.15), (3.17), and (3.18) are verified in the same way.

The state transition diagram for the Markov chain $\{N_k\}$ is shown in Fig. 3.5 where we have omitted the terms $o(\delta)$.

Consider now the steady-state probabilities

$$\begin{aligned} p_n &= \lim_{k\to\infty} P\{N_k = n\} \\ &= \lim_{t\to\infty} P\{N(t) = n\} \end{aligned}$$

Note that for any $k \geq 1$, $n \geq 0$, during the time from δ to $k\delta$, the total number of transitions from state n to $n+1$ must differ from the total number of transitions from $n+1$ to n by at most 1. Thus, in steady state, the probability that the system is in state n and makes a transition to $n+1$ at the next transition instant is the same as the probability that the system is in state $n+1$ and makes a transition to n, *i.e.*,

$$p_n \lambda \delta + o(\delta) = p_{n+1} \mu \delta + o(\delta) \tag{3.20}$$

(These equations are called global balance equations corresponding to the set of states $\{0, 1, \ldots, n\}$ and $\{n+1, n+2, \ldots\}$. See Appendix A for a more general statement of these equations, and for an interpretation that parallels the argument given above to derive Eq. (3.20).) Since p_n is independent of δ, by taking the limit in Eq. (3.20) as $\delta \to 0$, we obtain

$$p_{n+1} = \rho p_n\,, \quad n = 0, 1, \ldots$$

where

$$\rho = \frac{\lambda}{\mu}$$

It follows that

$$p_{n+1} = \rho^{n+1} p_0, \quad n = 0, 1, \ldots \tag{3.21}$$

If $\rho < 1$ (service rate exceeds arrival rate), the probabilities p_n are all positive and add up to unity, so

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \rho^n p_0 = \frac{p_0}{1-\rho} \tag{3.22}$$

This equation, together with Eq. (3.21), gives finally

$$p_n = \rho^n (1-\rho), \quad n = 0, 1, \ldots \tag{3.23}$$

We can now calculate the average number of customers in the system in steady state:

$$\begin{aligned} N = \lim_{t\to\infty} E\{N(t)\} &= \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n \rho^n (1-\rho) \\ &= \rho(1-\rho) \sum_{n=0}^{\infty} n \rho^{n-1} = \rho(1-\rho) \frac{\partial}{\partial \rho}\left(\sum_{n=0}^{\infty} \rho^n\right) \\ &= \rho(1-\rho) \frac{\partial}{\partial \rho}\left(\frac{1}{1-\rho}\right) = \rho(1-\rho)\frac{1}{(1-\rho)^2} \end{aligned}$$

and, finally, using $\rho = \lambda/\mu$

$$N = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \tag{3.24}$$

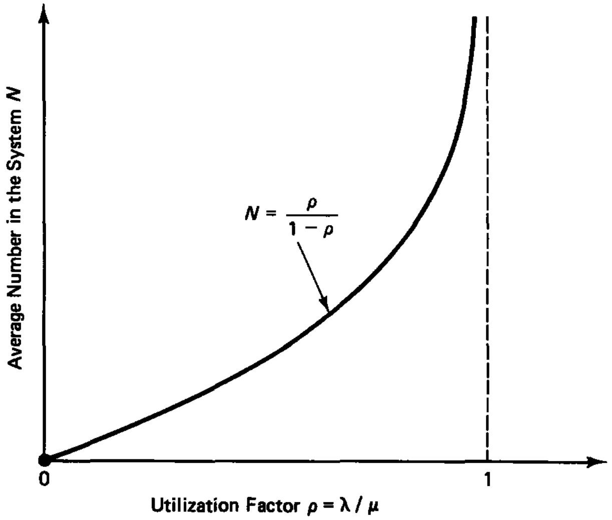


Figure 3.6 The average number in the system versus the utilization factor in the $M/M/1$ system. As $\rho \to 1$, $N \to \infty$.

This equation is shown in the diagram of Fig. 3.6. As ρ increases, so does N, and as $\rho \to 1$, we have $N \to \infty$. The diagram is valid for $\rho < 1$. If $\rho > 1$, the server cannot keep up with the arrival rate and the queue length increases without bound. In the context of a packet transmission system, $\rho > 1$ means that $\lambda L > C$, where λ is the arrival rate in packets/sec, L is the average packet length in bits, and C is the transmission capacity in bits/sec.

The average delay per customer (waiting time in queue plus service time) is given by Little's Theorem,

$$T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} \tag{3.25}$$

Using $\rho = \lambda/\mu$, this becomes

$$T = \frac{1}{\mu - \lambda} \tag{3.26}$$

The average waiting time in queue, W, is the average delay T less the average service time $1/\mu$, so

$$W = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}$$

By Little's Theorem, the average number of customers in queue is

$$N_Q = \lambda W = \frac{\rho^2}{1-\rho}$$

A very useful interpretation is to view the quantity ρ as the *utilization factor* of the queueing system, *i.e.*, the long-term proportion of time the server is busy. We showed this earlier in a broader context by using Little's Theorem (Ex. 1 in section 3.2). It follows that $\rho = 1 - p_0$, where p_0 is the probability of having no customers in the system, and we obtain an alternative verification of the formula derived for p_0 (Eq. (3.22)).

We illustrate these results by means of some examples from data networks:

Example 7: Increasing the arrival and transmission rates by the same factor

Consider a packet transmission system whose arrival rate (in packets/sec) is increased from λ to $K\lambda$, where $K > 1$ is some scalar factor. The packet length distribution remains the same but the transmission capacity is increased by a factor of K, so the average packet transmission time is now $1/(K\mu)$ instead of $1/\mu$. It follows that the utilization factor ρ and, therefore, the average number of packets in the system remain the same

$$N = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

However, the average delay per packet is now $T = N/(K\lambda)$ and is therefore decreased by a factor of K. In other words, ***a transmission line K times as fast will accommodate K times as many packets/sec at K times smaller average delay per packet.*** This result is quite general, even applying to networks of queues. What is happening, as illustrated in Fig. 3.7, is that by increasing arrival rate and service rate by a factor K, the statistical characteristics of the queueing process are unaffected except for a change in time scale—the process is speeded up by a factor K. Thus, when a packet arrives, it will see ahead of it statistically the same number of packets as with a slower transmission line. However, the packets ahead of it will be moving K times faster.

Example 8: Statistical multiplexing compared with time- and frequency-division multiplexing

Assume that m statistically identical and independent Poisson packet streams each with an arrival rate of λ/m packets/sec are to be transmitted over a communication line. The packet lengths for all streams are independent and exponentially distributed. The average transmission time is $1/\mu$. If the streams are merged into a single Poisson stream, with rate λ, as in statistical multiplexing, the average delay per packet is

$$T = \frac{1}{\mu - \lambda}$$

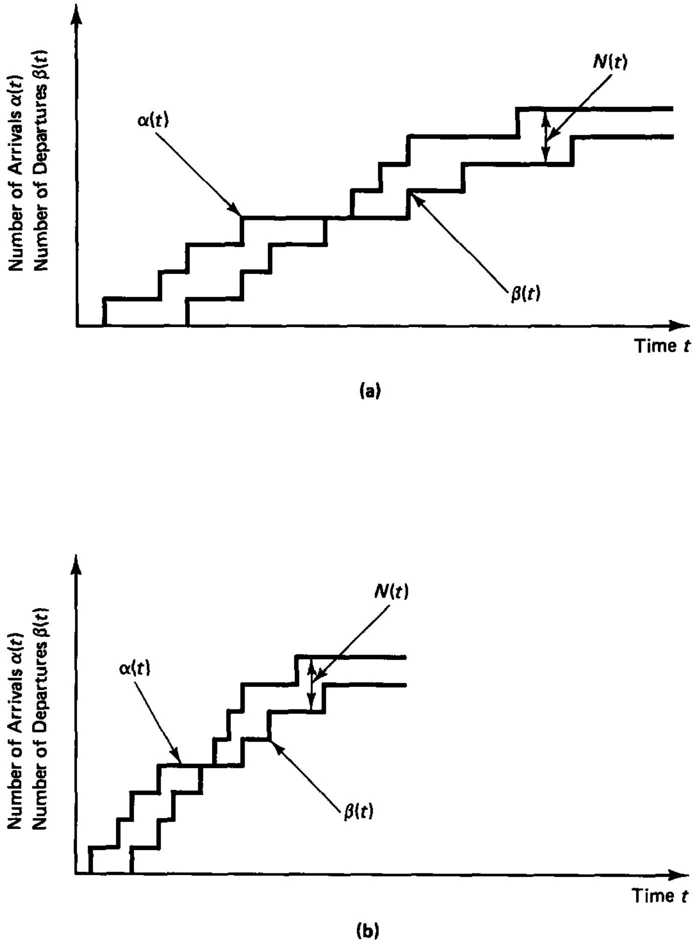


Figure 3.7 Increasing the arrival rate and the service rate by the same factor (see Ex. 7). (a) Sample paths of number of arrivals $\alpha(t)$ and departures $\beta(t)$ in the original system. (b) Corresponding sample paths of number of arrivals $\alpha(t)$ and departures $\beta(t)$ in the "speeded up" system, where the arrival rate and the service rate have been increased by a factor of two. The average number in the system is the same as before, but the average delay is reduced by a factor of two since customers are moving twice as fast.

If, instead, the transmission capacity is divided into m equal portions, one per packet stream as in time- and frequency-division multiplexing, each portion behaves like an $M/M/1$ queue with arrival rate λ/m and average service rate μ/m. Therefore, the average delay per packet is

$$T = \frac{m}{\mu - \lambda}$$

i.e., m times larger than for statistical multiplexing.

The preceding argument indicates that multiplexing a large number of traffic streams on separate channels in a transmission line performs very poorly in terms of delay. The performance is even poorer if the capacity of the channels is not allocated in direct proportion to the arrival rates of the corresponding streams—something that cannot be done (at least in the scheme considered here) if these arrival rates change over time. This is precisely why data networks that must contend with many low duty cycle traffic streams are organized on the basis of some form of statistical multiplexing. An argument in favor of time- and frequency-division multiplexing arises when each traffic stream is "regular" (as opposed to Poisson) in the sense that no packet arrives while another is transmitted, and thus there is no waiting in queue if that stream is transmitted on a dedicated transmission line. If several streams of this type are statistically multiplexed on a single transmission line, the average delay per packet will decrease, but the average waiting time in queue will become positive. For example in telephony each traffic stream is a voice conversation that is regular in the above sense, and time- and frequency-division multiplexing are still used widely.

3.3.2 Occupancy Distribution Upon Arrival

In our subsequent development, there are several situations where we will need a probabilistic characterization of a queueing system as seen by an arriving customer. In some systems, the times of customer arrivals are in some sense nontypical, so that the steady-state occupancy probabilities upon arrival

$$a_n = \lim_{t \to \infty} P\{N(t) = n | \text{ an arrival occurred just after time } t\} \tag{3.27}$$

need not be equal to the corresponding unconditional steady-state probabilities

$$p_n = \lim_{t \to \infty} P\{N(t) = n\} \tag{3.28}$$

It turns out, however, that for the $M/M/1$ system, we have

$$p_n = a_n, \quad n = 0, 1, \ldots \tag{3.29}$$

Indeed *this equality holds under very general conditions for queueing systems with Poisson arrivals regardless of the distribution of the service times.* The only additional requirement we need is that future arrivals are independent of the current number in the system. More precisely, we assume that for every time t and interval $\delta > 0$, the number of arrivals in the interval $(t, t+\delta)$ is independent of the number in the system at time t. Given the Poisson hypothesis, essentially this amounts to assuming that, at any time, the service times of previously arrived customers, and the future interarrival times are independent—something that is very reasonable for packet transmission systems. In particular, the assumption holds if the arrival process is Poisson and interarrival times and service times are independent.

The basic reason why $a_n = p_n$ is that the events $\{N(t) = n\}$ and {An arrival occurred just after t} are independent under our hypothesis. As a result, the conditional probability in Eq. (3.27) equals the unconditional probability in Eq. (3.28). For a more formal proof, let $A(t, t+\delta)$ be the event that an arrival occurs in the interval $(t, t+\delta)$. Let

$$p_n(t) = P\{N(t) = n\} \tag{3.30}$$

$$a_n(t) = P\{N(t) = n | \text{an arrival occurred just after time } t\} \tag{3.31}$$

We have, using Bayes' rule,

$$\begin{aligned} a_n(t) &= \lim_{\delta \to 0} P\{N(t) = n | A(t, t+\delta)\} \\ &= \lim_{\delta \to 0} \frac{P\{N(t) = n,\ A(t, t+\delta)\}}{P\{A(t, t+\delta)\}} \\ &= \lim_{\delta \to 0} \frac{P\{A(t, t+\delta) | N(t) = n\}\ P\{N(t) = n\}}{P\{A(t, t+\delta)\}} \end{aligned} \tag{3.32}$$

By assumption, the event $A(t, t+\delta)$ is independent of the number in the system at time t. Therefore,

$$P\{A(t, t+\delta) | N(t) = n\} = P\{A(t, t+\delta)\}$$

and we obtain from Eq. (3.32)

$$a_n(t) = P\{N(t) = n\} = p_n(t)$$

Taking the limit as $t \to \infty$, we obtain Eq. (3.29).

Thus, we have shown that the probability of an arrival finding n customers in the system equals the (unconditional) probability of n in the system. This is true at every time instant as well as in steady state regardless of the service time distribution. We can summarize this by saying that *when the arrival process is Poisson, an arriving customer finds the system in a "typical" state.*

As an example of what can happen if the arrival process is not Poisson, suppose that interarrival times are independent and uniformly distributed between two and four seconds, while customer service times are all equal to one second. Then an arriving customer always finds an empty system. On the other hand, the average number in the system as seen by an outside observer looking at a system at a random time is 1/3.

For a similar example where the arrival process is Poisson but the service times of customers in the system and the future arrival times are correlated, consider a packet transmission system where packets arrive according to a Poisson process. The transmission time of the n^{th} packet equals one half the interarrival time between packets n and $n+1$. A packet upon arrival finds the system empty. However, the average number in the system, as seen by an outside observer looking at the system is easily seen to be $1/2$.

3.3.3 Occupancy Distribution Upon Departure

Let us consider the distribution of the number of customers in the system just after a departure has occurred, *i.e.*, the probabilities

$$d_n(t) = P\{N(t) = n | \text{ a departure occurred just before time } t\}$$

The corresponding steady-state values are denoted

$$d_n = \lim_{t\to\infty} d_n(t), \quad n = 0, 1, \ldots$$

It turns out that

$$d_n = a_n, \quad n = 0, 1, \ldots$$

under very general assumptions—the only requirement essentially is that the system reaches a steady state with all n having positive steady-state probabilities, and that $N(t)$ changes in unit increments. (These assumptions certainly hold for a stable $M/M/1$ system ($\rho < 1$), but they also hold for most stable single-queue systems of interest.) For any sample path of the system and for every n, the number in the system will be n infinitely often (with probability one). This means that for each time the number in the system increases from n to $n+1$ due to an arrival, there will be a corresponding future decrease from $n+1$ to n due to a departure. Therefore, in the long run, the proportion of transitions from n to $n+1$ out of transitions from any k to $k+1$ equals the proportion of transitions from $n+1$ to n out of transitions from any $k+1$ to k which implies $d_n = a_n$. Therefore, ***in steady state, the system appears statistically identical to an arriving and a departing customer. When arrivals are Poisson,*** we saw earlier that $a_n = p_n$; so, in this case, ***both an arriving and a departing customer in steady state see a system that is statistically identical to the one seen by an observer looking at the system at a random time.***

3.4 THE M/M/m, M/M/∞, *AND* M/M/m/m *SYSTEMS*

We consider now a number of queueing systems that are similar to $M/M/1$ in that the arrival process is Poisson, the service times are independent, exponentially distributed, and independent of the interarrival times. Because of these assumptions, these systems can be modelled with continuous- or discrete-time Markov chains.

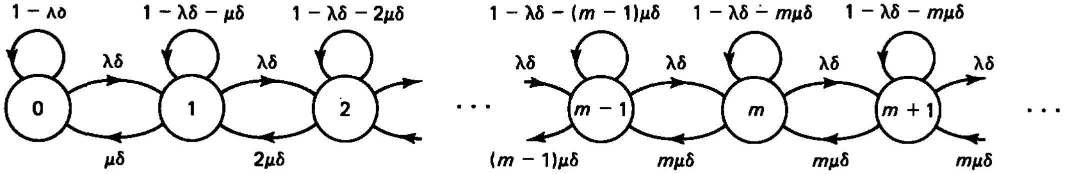


Figure 3.8 Discrete-time Markov chain for the $M/M/m$ system.

From the corresponding state transition diagram, we can derive a set of equations that can be solved for the steady-state occupancy probabilities. Application of Little's Theorem then yields the average delay per customer.

3.4.1 M/M/m: *The* m*-Server Case*

The $M/M/m$ queueing system is identical to the $M/M/1$ system except that there are m servers (or channels of a transmission line in a data communication context). A customer at the head of the queue is routed to any server that is available. The corresponding state transition diagram is shown in Fig. 3.8.

By writing down the equilibrium equations for the steady-state probabilities p_n and taking $\delta \rightarrow 0$, we obtain

$$\lambda p_{n-1} = n\mu p_n, \quad n \le m$$
$$\lambda p_{n-1} = m\mu p_n, \quad n > m$$

From these equations, we obtain

$$p_n = \begin{cases} p_0 \dfrac{(m\rho)^n}{n!}, & n \le m \\ \\ p_0 \dfrac{m^m \rho^n}{m!}, & n > m \end{cases} \tag{3.33}$$

where ρ is given by

$$\rho = \frac{\lambda}{m\mu} < 1 \tag{3.34}$$

We can calculate p_0 using Eq. (3.33) and the condition $\sum_{n=0}^{\infty} p_n = 1$. We obtain

$$p_0 = \left[1 + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} + \sum_{n=m}^{\infty} \frac{(m\rho)^n}{m!} \frac{1}{m^{n-m}}\right]^{-1}$$

and, finally,

$$p_0 = \left[\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1-\rho)}\right]^{-1} \tag{3.35}$$

The probability that an arrival will find all servers busy and will be forced to wait in queue is

$$P\{\text{Queueing}\} = \sum_{n=m}^{\infty} p_n$$
$$= \sum_{n=m}^{\infty} \frac{p_0 m^m \rho^n}{m!} = \frac{p_0(m\rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m}$$

and, finally,

$$P_Q \overset{\Delta}{=} P\{\text{Queueing}\} = \frac{p_0(m\rho)^m}{m!(1-\rho)} \tag{3.36}$$

where p_0 is given by Eq. (3.35). This equation is known as the *Erlang C formula* and is in wide use in telephony. (Denmark's A. K. Erlang is viewed as the foremost pioneer of queueing theory.)

The expected number of customers waiting in queue (not in service) is given by

$$N_Q = \sum_{n=0}^{\infty} n p_{m+n}$$

Using Eq. (3.33), we obtain

$$N_Q = \sum_{n=0}^{\infty} n p_0 \frac{m^m \rho^{m+n}}{m!} = \frac{p_0(m\rho)^m}{m!} \sum_{n=0}^{\infty} n\rho^n$$

Using Eq. (3.36) and the equation $(1-\rho)\sum_{n=0}^{\infty} n\rho^n = \rho/(1-\rho)$ encountered in the $M/M/1$ system analysis, we finally obtain

$$N_Q = P_Q \frac{\rho}{1-\rho} \tag{3.37}$$

Note that

$$\frac{N_Q}{P_Q} = \frac{\rho}{1-\rho}$$

represents the expected number found in queue by an arriving customer conditioned on the fact that he is forced to wait in queue, and is independent of the number of servers for a given $\rho = \lambda/m\mu$. This suggests in particular that, as long as there are customers waiting in queue, the queue size of the $M/M/m$ system behaves identically as in an $M/M/1$ system with service rate $m\mu$—the aggregate rate of the m servers. Some thought shows that indeed this is true in view of the memoryless property of the exponential distribution.

Using Little's Theorem and Eq. (3.37), we obtain the average time W a customer has to wait in queue:

$$W = \frac{N_Q}{\lambda} = \frac{\rho P_Q}{\lambda(1-\rho)} \tag{3.38}$$

The average delay per customer is, therefore,

$$T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{\rho P_Q}{\lambda(1-\rho)}$$

and, using $\rho = \lambda/m\mu$, we finally obtain

$$T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{P_Q}{m\mu - \lambda} \tag{3.39}$$

Using Little's Theorem again, the average number of customers in the system is

$$N = \lambda T = \frac{\lambda}{\mu} + \frac{\lambda P_Q}{m\mu - \lambda}$$

and, using $\rho = \lambda/m\mu$, we obtain

$$N = m\rho + \frac{\rho P_Q}{1-\rho}$$

Example 9: Using one vs. using multiple channels in statistical multiplexing

Consider a communication link serving m independent Poisson traffic streams with rate λ/m each. Suppose that the link is divided into m separate channels with one channel assigned to each traffic stream. However, if a traffic stream has no packet awaiting transmission, its corresponding channel is used to transmit a packet of another traffic stream. The transmission times of packets on each of the channels are exponentially distributed with mean $1/\mu$. The system can be modeled by the same Markov chain as the $M/M/m$ queue. Let us compare the average delays per packet of this system, and an $M/M/1$ system with the same arrival rate λ and service rate $m\mu$ (statistical multiplexing with one channel having m times larger capacity). In the former case, the average delay per packet is given by Eq. (3.39)

$$T = \frac{1}{\mu} + \frac{P_Q}{m\mu - \lambda}$$

while in the latter case, the average delay per packet is

$$\hat{T} = \frac{1}{m\mu} + \frac{\hat{P}_Q}{m\mu - \lambda}$$

where P_Q and $\hat{P}_Q$ denote the queueing probability in each case. When $\rho \ll 1$ (lightly loaded system) we have $P_Q \cong 0$, $\hat{P}_Q \cong 0$ and

$$\frac{T}{\hat{T}} \cong m$$

When ρ is only slightly less than 1, we have $P_Q \cong 1$, $\hat{P}_Q \cong 1$, $1/\mu \ll 1/(m\mu - \lambda)$ and

$$\frac{T}{\hat{T}} \cong 1$$

Therefore, for a light load, statistical multiplexing with m channels produces a delay almost m times larger than the delay of statistical multiplexing with the m channels combined in one (about the same as time- and frequency-division multiplexing). For a heavy load, the ratio of the two delays is close to one.

3.4.2 M/M/∞: *Infinite-Server Case*

In the limiting case where $m = \infty$ in the $M/M/m$ system, we obtain from Fig. 3.8

$$\lambda p_{n-1} = n\mu p_n\,, \quad n = 1, 2, \ldots$$

so

$$p_n = p_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}, \quad n = 1, 2, \ldots$$

From the condition $\sum_{n=0}^{\infty} p_n = 1$, we obtain

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1}$$
$$= e^{-\lambda/\mu}$$

so, finally,

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \frac{e^{-\lambda/\mu}}{n!}, \quad n = 0, 1, \ldots$$

Therefore, in steady state, *the number in the system is Poisson distributed with parameter* λ/μ. The average number in the system is

$$N = \frac{\lambda}{\mu}$$

By Little's Theorem, the average delay is N/λ or

$$T = \frac{1}{\mu}$$

This last equation can also be obtained by simply arguing that in an $M/M/\infty$ system, there is no waiting in queue, so T equals the average service time $1/\mu$. It can be shown that the number in the system is Poisson distributed even if the service time distribution is not exponential, *i.e.*, in the $M/G/\infty$ system (see Prob. 3.37).

Example 10: The quasistatic assumption

It is often convenient to assume that the external packet traffic entering a subnet node and destined for some other subnet node can be modeled by a stationary stochastic process that has a constant bit arrival rate (average bits/sec). This approximates a situation where the arrival rate changes slowly with time and constitutes what we refer to as the quasistatic assumption.

When there are only a few active sessions (*i.e.*, user pairs) for the given origin-destination pair, this assumption is seriously violated since the addition or termination of a single session can change the total bit arrival rate by a substantial factor. When, however, there are many active sessions, each with a bit arrival rate that is small relative to the total, it seems plausible that the quasistatic assumption is approximately valid. The reason is that session additions are statistically counterbalanced by session terminations, with variations in the total rate being relatively small. As analytical substantiation, let us assume that sessions are generated according to a Poisson process with rate λ, and terminate after a time which is exponentially distributed with mean $1/\mu$. Then the number of active sessions n evolves like the number of customers in an $M/M/\infty$ system, *i.e.*, is Poisson distributed with parameter λ/μ in steady-state. In particular, the mean and standard deviation of n are

$$N = E\{n\} = \lambda/\mu$$

$$\sigma_n = \left[E\left\{(n-N)^2\right\}\right]^{1/2} = (\lambda/\mu)^{1/2}$$

Suppose the i^{th} active session generates traffic according to a stationary stochastic process having a bit arrival rate γ_i bits/sec. Assume that the rates γ_i are independent random variables with common mean $E\{\gamma_i\} = \Gamma$, and second moment $s_\gamma^2 = E\{\gamma_i^2\}$. Then the total bit arrival rate for n active sessions is the random variable $f = \sum_{i=1}^n \gamma_i$, which has mean

$$F = E\{f\} = (\lambda/\mu)\Gamma$$

The standard deviation of f, denoted σ_f, can be obtained by writing

$$\sigma_f^2 = E\left\{ \left(\sum_{i=1}^n \gamma_i\right)^2 \right\} - F^2$$

and carrying out the corresponding calculations (Prob. 3.21). The result is

$$\sigma_f = (\lambda/\mu)^{\frac{1}{2}} s_\gamma$$

Therefore, we have

$$\sigma_f/F = (s_\gamma/\Gamma)(\mu/\lambda)^{1/2} \tag{3.40}$$

Suppose now that the average bit rate Γ of a session is small relative to the total F, *i.e.*, a "many-small-sessions assumption" holds. Then, since $\Gamma/F = \mu/\lambda$, we have that μ/λ is small. If we reasonably assume that s_γ/Γ has a moderate value, it follows from Eq. (3.40) that σ_f/F is small. Therefore, the total arrival rate f is approximately constant thereby justifying the quasistatic assumption.

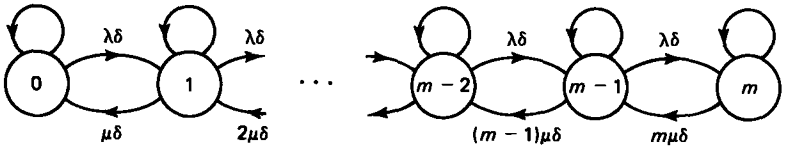


Figure 3.9 Discrete-time Markov chain for the $M/M/m/m$ system.

3.4.3 M/M/m/m: *The m-Server Loss System*

This system is identical to the $M/M/m$ system except that if an arrival finds all m servers busy, it does not enter the system and is lost—a model that is in wide use in telephony. (The last m in the $M/M/m/m$ notation indicates the limit on the number of customers in the system.) In data networks, it can be used as a model where arrivals correspond to requests for virtual circuit connections between two nodes and the number of virtual circuits allowed is m. The average service time $1/\mu$ is then the average duration of a virtual circuit conversation.

The corresponding state transition diagram is shown in Fig. 3.9. We have

$$\lambda p_{n-1} = n\mu p_n, \quad n = 1, 2, \ldots, m$$

so

$$p_n = p_0 \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}, \quad n = 1, 2, \ldots, m$$

Solving for p_0 in the equation $\sum_{n=0}^{m} p_n = 1$, we obtain

$$p_0 = \left[\sum_{n=0}^{m} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1}$$

The probability that an arrival will find all m servers busy and will therefore be lost is

$$p_m = \frac{(\lambda/\mu)^m/m!}{\sum_{n=0}^{m}(\lambda/\mu)^n/n!}$$

This equation is known as the *Erlang B formula.* It can be shown to hold even if the service time probability distribution is arbitrary, *i.e.*, for an $M/G/m/m$ system (see [Ros83], p. 170).

3.5 *THE* M/G/1 *SYSTEM*

Consider a single-server queueing system where customers arrive according to a Poisson process with rate λ, but the customer service times have a general distribution—not necessarily exponential as in the $M/M/1$ system. Suppose that customers are served in the order they arrive and that X_i is the service time of the i^{th} arrival. We assume that the random variables $(X_1, X_2, \ldots)$ are identically distributed, mutually independent, and independent of the interarrival times.

Let

$$\overline{X} = E\{X\} = \frac{1}{\mu} = \text{ Average service time}$$

$$\overline{X^2} = E\{X^2\} = \text{ Second moment of service time}$$

Our objective is to derive and understand the *Pollaczek-Khinchin (P-K) formula:*

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} \tag{3.41}$$

where W is the expected customer waiting time in queue and $\rho = \lambda/\mu = \lambda\overline{X}$. Given Eq. (3.41), the total waiting time, in queue and in service, is

$$T = \overline{X} + \frac{\lambda \overline{X^2}}{2(1-\rho)} \tag{3.42}$$

Applying Little's formula to W and T, we get the expected number of customers in the queue N_Q and the expected number in the system N

$$N_Q = \frac{\lambda^2 \overline{X^2}}{2(1-\rho)} \tag{3.43}$$

$$N = \rho + \frac{\lambda^2 \overline{X^2}}{2(1-\rho)} \tag{3.44}$$

For example, when service times are exponentially distributed, as in the $M/M/1$ system, we have $\overline{X^2} = 2/\mu^2$, and Eq. (3.41) reduces to the formula (see subsection 3.3.2)

$$W = \frac{\rho}{\mu(1-\rho)} \qquad (M/M/1\ )$$

When service times are identical for all customers (the $M/D/1$ system, where D means deterministic), we have $\overline{X^2} = 1/\mu^2$, and

$$W = \frac{\rho}{2\mu(1-\rho)} \qquad (M/D/1) \tag{3.45}$$

Since the $M/D/1$ case yields the minimum possible value of $\overline{X^2}$ for given μ, it follows that the values of W, T, N_Q, and N for an $M/D/1$ queue are lower bounds to the corresponding quantities for an $M/G/1$ queue of the same λ and μ. It is interesting to note that W and N_Q for the $M/D/1$ queue are exactly one half their values for the $M/M/1$ queue of the same λ and μ. The values of T and N for $M/D/1$, on the other hand, range from the same as $M/M/1$ for small ρ to one half of $M/M/1$ as ρ approaches 1. The reason is that the expected service time is the same in the two cases, and, for ρ small, most of the waiting occurs in service whereas, for ρ large, most of the waiting occurs in the queue.

We provide a proof of the Pollaczek-Khinchin formula based on the concept of the *mean residual service time.* This same concept will prove useful in a number of subsequent developments. For example, $M/G/1$ queues with priorities and reservation systems are analyzed later; then part of the service time is occupied with sending packets (*i.e.*, serving customers), and part with sending control information or making reservations for sending the packets.

Denote

W_i : The waiting time in queue of the i^{th} customer.

R_i : The residual service time seen by the i^{th} customer. By this we mean that if customer j is already being served when i arrives, R_i is the remaining time until customer j's service time is complete. If no customer is in service (*i.e.*, the system is empty when i arrives), then R_i is zero.

X_i : The service time of the i^{th} customer.

N_i : The number of customers found waiting in queue by the i^{th} customer upon arrival.

We have

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j$$

By taking expectations and using the fact that the random variables N_i and $X_{i-1}, \ldots, X_{i-N_i}$ are independent we have

$$E\{W_i\} = E\{R_i\} + E\left\{\textstyle\sum_{j=i-N_i}^{i-1} E\{X_j|N_i\}\right\} = E\{R_i\} + \overline{X}E\{N_i\}$$

Taking the limit as $i \to \infty$ we obtain

$$W = R + \frac{1}{\mu}N_Q \tag{3.46}$$

where

R : Mean residual time, defined as $R = \lim_{i\to\infty} E\{R_i\}$.

In Eq. (3.46) (and throughout this section) all long-term average quantities should be viewed as limits when time or customer index increase to infinity. Thus, W, R, and N_Q are limits (as $i \to \infty$) of the average waiting time, residual time, and number found in queue, respectively, corresponding to the i^{th} customer. We assume that these limits exist, and this is true of almost all systems of interest to us provided $\lambda < \mu$. Note that in Eq. (3.46) the average number in queue N_Q and the mean residual time R as seen by an arriving customer are also equal to the average number in queue and mean residual time seen by an outside observer at a random time. This is due to the Poisson character of the arrival process, which implies that the occupancy distribution upon arrival is typical (see subsection 3.3.2).

By Little's Theorem, we have

$$N_Q = \lambda W$$

and by substitution in Eq. (3.46), we obtain

$$W = R + \rho W \tag{3.47}$$

where $\rho = \lambda/\mu$ is the utilization factor; so, finally,

$$W = \frac{R}{1-\rho} \tag{3.48}$$

We can calculate R by a graphical argument. In Fig. 3.10 the residual service time $r(\tau)$, (*i.e.*, the remaining time for completion of the customer in service at time τ) is plotted as a function of τ. Note that when a new service of duration X begins, $r(\tau)$ starts at X and decays linearly for X time units. Consider a time t for which $r(t) = 0$. The time average of $r(\tau)$ in the interval $[0, t]$ is

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{t}\sum_{i=1}^{M(t)} \frac{1}{2}X_i^2 \tag{3.49}$$

where $M(t)$ is the number of service completions within $[0, t]$, and X_i is the service time of the i^{th} customer. We can also write this equation as

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} \tag{3.50}$$

and, assuming the limits below exist, we obtain

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{2}\lim_{t\to\infty}\frac{M(t)}{t}\cdot\lim_{t\to\infty}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} \tag{3.51}$$

The two limits on the right are the time averages of the departure rate (which equals the arrival rate) and the second moment of the service time, respectively, while the limit on the left is the time average of the residual time. Assuming that time averages can be replaced by ensemble averages, we obtain

$$R = \frac{1}{2}\lambda\overline{X^2} \tag{3.52}$$

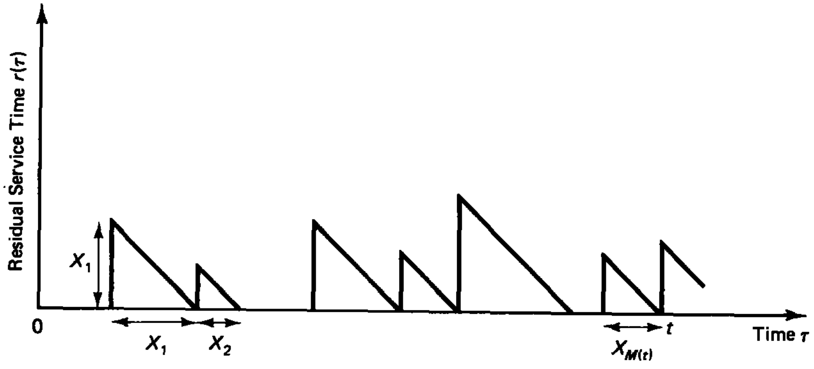


Figure 3.10 Derivation of the mean residual service time. During period $[0, t]$, the time average of the residual service time $r(\tau)$ is

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{t}\sum_{i=1}^{M(t)} \frac{1}{2}X_i^2 = \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)}$$

where X_i is the service time of the i^{th} customer, and $M(t)$ is the number of service completions in $[0, t]$. Taking the limit as $t \to \infty$ and equating time and ensemble averages we obtain the mean residual time $R = (1/2)\lambda\overline{X^2}$

The Pollaczek-Khinchin formula,

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} \tag{3.53}$$

now follows by combining Eqs. (3.48) and (3.52). Our derivation was based on two assumptions: (a) the existence of the steady-state averages W, R, and N_Q; and (b) the equality (with probability one) of the long-term time averages appearing in Eq. (3.51) with the corresponding ensemble averages. These assumptions can be justified by careful applications of the law of large numbers, but the details are beyond the scope of this text. However, these are natural assumptions for the systems of interest to us, and we will base similar derivations on graphical arguments and interchange of time averages with ensemble averages without further discussion.

One curious feature of Eq. (3.53) is that an $M/G/1$ queue can have $\rho < 1$ but infinite W if the second moment of the service distribution is ∞. What is happening in this case is that a small fraction of customers have incredibly long service times. When one of these customers is served, an incredible number of arrivals are queued and delayed by a significant fraction of that service time. Thus, the contribution to W is proportional to the square of the service time, leading to an infinite W if $\overline{X^2}$ is infinite.

The above derivation of the P-K formula assumed that customers were served in order of arrival, *i.e.*, that the number of customers served between the i^{th} arrival and service is just the number in queue at the i^{th} arrival. It turns out, however, that this formula is valid for any order of servicing customers as long as the order is determined independently of the required service time. To see this, suppose that the i^{th} and j^{th} customers are both in the queue and that they exchange places. The expected queueing time of customer i (over the service times of the customers in queue) will then be exchanged with that for customer j, but the average, over all customers, is unchanged. Since any service order can be considered as a sequence of reversals in queue position, the P-K formula remains valid (see also Problem 3.16).

To see why the P-K formula is invalid if the service order can depend on service time, consider a queue with two customers requiring 10 and 2 units of service time respectively. Assuming that the server becomes available at time 0, serving the first customer first results in one customer starting service at time 0 and the other at time 10. Serving the second customer first results in one customer starting at time 0 and the other at time 2. Thus, the average queueing time over the two customers is 5 in the first case and 1 in the second case. Clearly, queueing time is reduced by serving customers with small service time first. For this situation, the derivation of the P-K formula breaks down at Eq. (3.46) since the packets that will be transmitted before a newly arriving packet no longer have a mean service time equal to $1/\mu$.

Example 11: Delay Analysis of an ARQ System

Consider a goback n ARQ system such as the one discussed in section 2.4. Assume packets are transmitted in frames that are one time unit long, and there is a maximum wait for an acknowledgement of $n-1$ frames before a packet is retransmitted (see Fig. 3.11). In this system packets are retransmitted for two reasons:

1. A given packet transmitted in frame i might be rejected at the receiver due to errors, in which case the transmitter will transmit packets in frames $i+1, i+2, \ldots, i+n-1$, (if any are available), and then go back to retransmit the packet in frame $i+n$.
2. A packet transmitted in frame i might be accepted at the receiver but the corresponding acknowledgement (in the form of the receive number) might not have arrived at the transmitter by the time the transmission of packet $i+n-1$ is completed. This can happen due to errors in the return channel, large propagation delays, long return frames relative to the size of the goback number n, or a combination thereof.

We will assume (somewhat unrealistically) that retransmissions occur only due to #1 above, and that a packet is rejected at the receiver with probability p independently of other packets.

Consider the case where packets arrive at the transmitter according to a Poisson process with rate λ. It follows that the time interval between start of the first

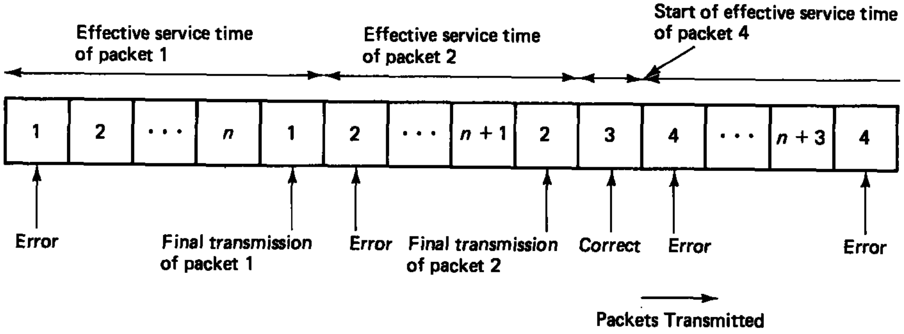

Effective service time of packet 1
Effective service time of packet 2
Start of effective service time of packet 4
1
2
• • •
n
1
2
• • •
n + 1
2
3
4
• • •
n + 3
4
Error
Final transmission of packet 1
Error
Final transmission of packet 2
Correct
Error
Error
Packets Transmitted

Figure 3.11 Illustration of the effective service times of packets in the ARQ system of Ex. 11. For example, packet 2 has an effective service time of $n+1$ because there was an error in the first attempt to transmit it following the last transmission of packet 1, but no error in the second attempt.

transmission of a given packet after the last transmission of the previous packet, and end of the last transmission of the given packet is $1+kn$ time units with probability $(1-p)p^k$ (this corresponds to k retransmissions following the last transmission of the previous packet—see Fig. 3.11). Thus, the transmitter's queue behaves like an $M/G/1$ queue with service time distribution given by

$$P\{X = 1 + kn\} = (1-p)p^k, \quad k = 0, 1, \ldots$$

The first two moments of the service time are

$$\overline{X} = \sum_{k=0}^{\infty}(1+kn)(1-p)p^k = (1-p)\left(\sum_{k=0}^{\infty}p^k + n\sum_{k=0}^{\infty}kp^k\right)$$

$$\overline{X^2} = \sum_{k=0}^{\infty}(1+kn)^2(1-p)p^k$$

$$= (1-p)\left(\sum_{k=0}^{\infty}p^k + 2n\sum_{k=0}^{\infty}kp^k + n^2\sum_{k=0}^{\infty}k^2p^k\right)$$

We now note that $\sum_{k=0}^{\infty}p^k = 1/(1-p)$, $\sum_{k=0}^{\infty}kp^k = p/(1-p)^2$, and $\sum_{k=0}^{\infty}k^2p^k = (p+p^2)/(1-p)^3$. (The first sum is the usual geometric series sum, while the other two sums are obtained by differentiating the first sum twice.) Using these formulas in the equations for $\overline{X}$ and $\overline{X^2}$ above, we obtain

$$\overline{X} = 1 + \frac{np}{1-p}$$

$$\overline{X^2} = 1 + \frac{2np}{1-p} + \frac{n^2(p+p^2)}{(1-p)^2}$$

The P-K formula gives the average packet time in queue and in the system (up to the end of the last transmission):

$$W = \frac{\lambda \overline{X^2}}{2(1-\lambda\overline{X})}$$

$$T = \overline{X} + W$$

3.5.1 M/G/1 *Queues with Vacations*

Suppose that at the end of each busy period, the server goes on "vacation" for some random interval of time. Thus, a new arrival to an idle system, rather than going into service immediately, waits for the end of the vacation period (see Fig. 3.12). If the system is still idle at the completion of a vacation, a new vacation starts immediately. For data networks, vacations correspond to the transmission of various kinds of control and record-keeping packets when there is a lull in the data traffic; other applications will become apparent later.

Let $V_1, V_2, \ldots$ be the durations of the successive vacations taken by the server. We assume that $V_1, V_2, \ldots$ are independent and identically distributed (IID) random variables, also independent of the customer interarrival times and service times. As before, the arrivals are Poisson and the service times are IID with a general distribution. A new arrival to the system has to wait in the queue for the completion of the current service or vacation and then for the service of all the customers waiting before it. Thus, Eq. (3.48) is still valid (*i.e.*, $W = R/(1-\rho)$), where now R is the mean residual time for completion of the service *or* vacation in process when the i^{th} customer arrives.

The analysis of this new system is the same as that of the P-K formula except that vacations must be included in the graph of residual service times $r(\tau)$ (see Fig. 3.13). Let $M(t)$ be the number of services completed by time t and $L(t)$ be the number of vacations completed by time t. Then (as in Eq. (3.49)), for any t where a service or vacation is just completed, we have

$$\frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{t}\sum_{i=1}^{M(t)} \frac{1}{2}X_i^2 + \frac{1}{t}\sum_{i=1}^{L(t)} \frac{1}{2}V_i^2 \tag{3.54}$$

As before, assuming that a steady state exists, $M(t)/t$ approaches λ with increasing t, and the first term on the right side of Eq. (3.54) approaches $\lambda\overline{X^2}/2$

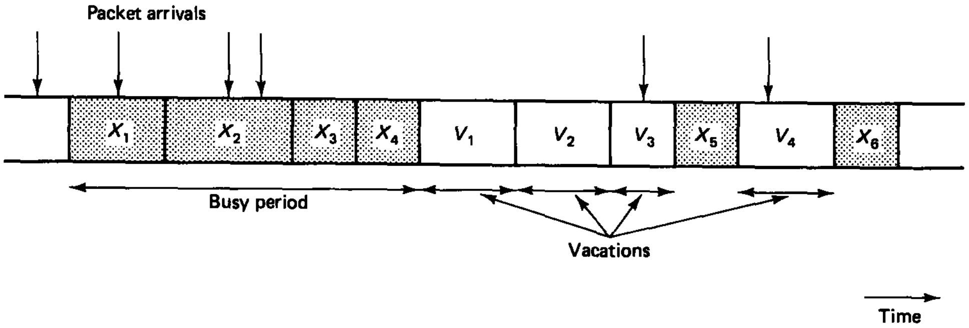


Figure 3.12 An $M/G/1$ system with vacations. At the end of a busy period, the server goes on vacation for time V with first and second moments $\overline{V}$ and $\overline{V^2}$, respectively. If the system is empty at the end of a vacation, the server takes a new vacation. An arriving customer to an empty system must wait until the end of the current vacation to get service.

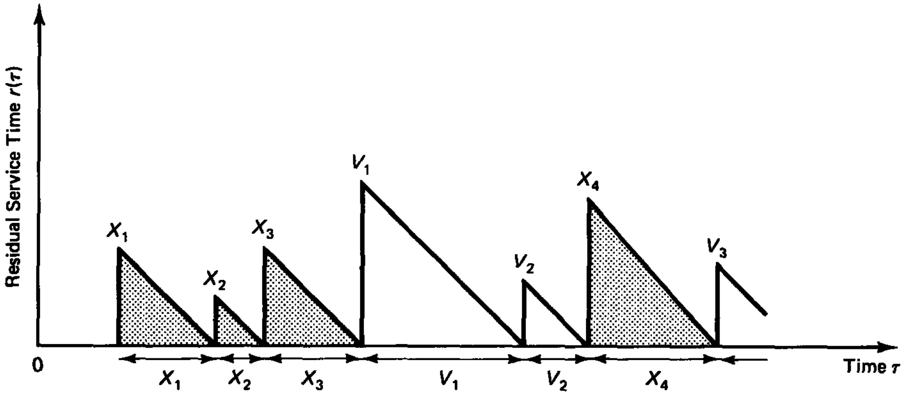


Figure 3.13 Residual service times for an $M/G/1$ system with vacations. Busy periods alternate with vacation periods.

as in the derivation of the P-K formula (cf. Eq. (3.52)). For the second term, note that as $t \to \infty$, the fraction of time spent serving customers approaches ρ, and thus the fraction of time occupied with vacations is $1 - \rho$. Assuming time averages can be replaced by ensemble averages we have $t(1-\rho)/L(t) \to \overline{V}$ with increasing t, and thus the second term in Eq. (3.54) approaches $(1-\rho)\overline{V^2}/(2\overline{V})$, where $\overline{V}$and $\overline{V^2}$ are the first and second moments of the vacation interval, respectively. Combining this with $W = R/(1-\rho)$, and assuming equality of the time and ensemble averages of R, we get

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}} \tag{3.55}$$

as the expected waiting time in queue for an $M/G/1$ system with vacations.

If we look carefully at the derivation of Eq. (3.55), we see that the mutual independence of the vacation intervals is not required (although the time and ensemble averages of the vacation intervals must still be equal) and the length of a vacation interval starting at time t need not be independent of the already completed service times or arrival times. Naturally, with this kind of dependence, it becomes more difficult to calculate $\overline{V}$ and $\overline{V^2}$, as these quantities might be functions of the underlying $M/G/1$ process.

Example 12: Frequency- and time-division multiplexing on a slot basis

We have m traffic streams of equal length packets arriving according to a Poisson process with rate λ/m each. If the traffic streams are frequency-division multiplexed on m subchannels of an available channel, the transmission time of each packet is m time units. Then, each subchannel can be represented by an $M/D/1$ queueing system and, by Eq. (3.45), with $\rho = \lambda$, $\mu = 1/m$, the average queueing delay per packet is

$$W_{\text{FDM}} = \frac{\lambda m}{2(1-\lambda)} \tag{3.56}$$

Consider the same FDM scheme with the difference that packet transmissions can start only at times m, $2m$, $3m, \ldots$, *i.e.*, at the beginning of a slot of m time units. We call this scheme *slotted frequency-division multiplexing* (SFDM), and note that it can be viewed as an $M/D/1$ queue with vacations. When there are no packets in the queue for a given stream at the beginning of a slot, the server takes a vacation for one slot, or m time units. Thus, $\overline{V} = m$, $\overline{V^2} = m^2$, and Eq. (3.55) becomes

$$W_{\text{SFDM}} = W_{\text{FDM}} + \frac{m}{2} \tag{3.57}$$

Finally, consider the case where the m traffic streams are time-division multiplexed in a scheme whereby the time axis is divided in m-slot frames with one slot dedicated to each traffic stream (see Fig. 3.14). Each slot is one time unit long and can carry a single packet. Then, if we compare this TDM scheme with the SFDM scheme, we see that the queue for a given stream in TDM is precisely the same as the queue for SFDM, and

$$W_{\text{TDM}} = W_{\text{SFDM}} = W_{\text{FDM}} + \frac{m}{2} = \frac{m}{2(1-\lambda)} \tag{3.58}$$

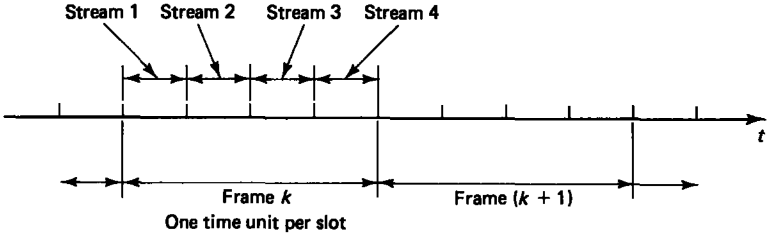


Figure 3.14 TDM with $m = 4$ traffic streams.

If we now look at the total delay for TDM, we get a different picture, since the service time is 1 unit of time rather than m units as in SFDM. By adding the service times to the queueing delays, we obtain

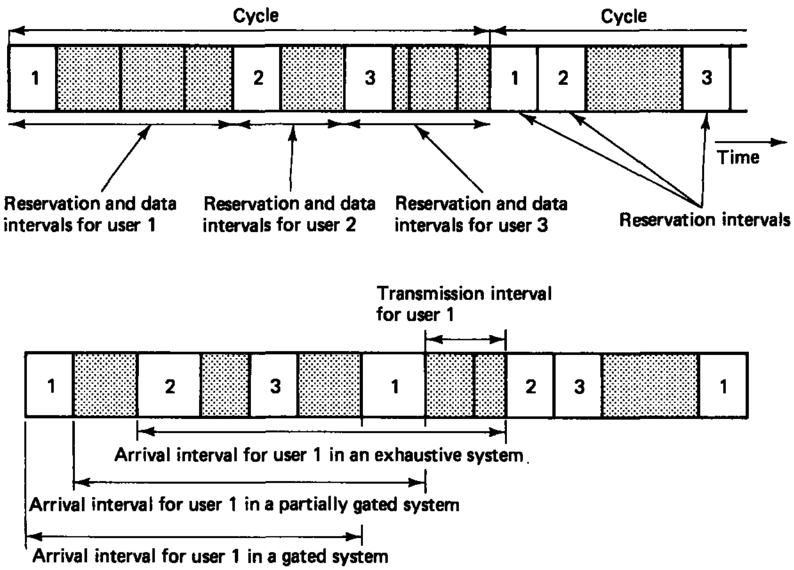
$$\begin{aligned}
T_{\mathrm{FDM}} &= m + \frac{\lambda m}{2(1-\lambda)} \\
T_{\mathrm{SFDM}} &= T_{\mathrm{FDM}} + \frac{m}{2} \\
T_{\mathrm{TDM}} &= 1 + \frac{m}{2(1-\lambda)} = T_{\mathrm{FDM}} - \left(\frac{m}{2} - 1\right)
\end{aligned} \tag{3.59}$$

Thus, the customer's average total delay is more favorable in TDM than in FDM (assuming $m > 2$). The longer average waiting time in queue for TDM is more than compensated by the faster service time. Note that the analysis above is generalized in Problem 3.22.

3.5.2 Reservations and Polling

Organizing transmissions from several packet streams into a statistical multiplexing system requires some form of scheduling. In some cases, this scheduling is naturally and easily accomplished; in other cases, however, some form of reservation or polling system is required.

A typical situation is one whereby there is a communication channel that can be accessed by several spatially separated users; however, only one user can transmit successfully on the channel at any one time. (This is called a *multiaccess channel*, and will be treated extensively in Chapter 4.) The communication resource of the channel can be divided over time into a portion used for packet transmissions and another portion used for reservation or polling messages that coordinate the packet transmissions. In other words, the time axis is divided into *data intervals*, where actual data is transmitted, and *reservation intervals*, used for scheduling future data. For uniform presentation, we use the term "reservation" even though "polling" may be more appropriate to the practical situation.


Cycle
Cycle
1
2
3
1
2
3
Time
Reservation and data intervals for user 1
Reservation and data intervals for user 2
Reservation and data intervals for user 3
Reservation intervals
Transmission interval for user 1
1
2
3
1
2
3
1
Arrival interval for user 1 in an exhaustive system
Arrival interval for user 1 in a partially gated system
Arrival interval for user 1 in a gated system

Packets arriving in the arrival interval shown are transmitted in the transmission interval shown

Figure 3.15 A reservation or polling system with three users. In the exhaustive version, a packet of a user that arrives during the user's reservation or data interval is transmitted in the same data interval. In the partially gated version, a packet of a user arriving during the user's data interval must wait for an entire cycle and be transmitted during the next data interval of the user. In the fully gated version, packets arriving during the user's reservation interval must also wait for an entire cycle. The figure shows, for the three systems, the association between the interval in which a packet arrives and the interval in which the packet is transmitted.

We will consider m traffic streams (also called users), and assume that each data interval contains packets of a *single* user. Reservations for these packets are made in the immediately preceding reservation interval. All users are taken up in cyclic order (see Fig. 3.15). There are several versions of this system differing in the rule for deciding which packets are transmitted during the data interval of each user. In the *gated* system, the rule is that only those packets that arrived prior to the user's preceding reservation interval are transmitted. By contrast, in the *exhaustive* system, the rule is that all available packets of a user are transmitted during the corresponding data interval, including those that arrived in this data interval or the preceding reservation interval. An intermediate version, which we call the *partially gated* system, results when the packets transmitted in a user's data interval are those that arrived up to the time this data interval began (and the corresponding reservation interval ended). A typical example of such reservation systems is one of the most common local area networks, the token ring. The users are connected by cable in a unidirectional loop. Each user transmits the current packet backlog, then gives the opportunity to a neighbor to transmit, and the process is repeated. (A more detailed description of the token ring is given in Chapter 4.)

We assume that the arrival processes of all users are independent Poisson with rate λ/m, and that the first and second moments of the packet transmission times are $\overline{X} = 1/\mu$ and $\overline{X^2}$, respectively. The utilization factor is $\rho = \lambda/\mu$. Interarrival times and transmission times are, as usual, assumed independent. While we assume that all users have identical arrival and service statistics, we allow the reservation intervals of different users to have different statistics.

Single-User System

Our general line of analysis of reservation systems can be better understood in terms of the special case where $m = 1$; so, we consider this case first. We may also view this as a system where all users share reservation and data intervals. Let V_ℓ be the duration of the ℓ^{th} reservation interval and assume that successive reservation intervals are independent and identically distributed random variables with first and second moments $\overline{V}$ and $\overline{V^2}$, respectively. We consider a gated system and assume that the reservation intervals are statistically independent of the arrival times and service durations. Finally, for convenience of exposition, we assume that packets are transmitted in the order of their arrival. As in our derivation of the P-K formula, expected delays and queue lengths are independent of service order as long as service order is independent of service requirement (*i.e.*, packet length).

Consider the i^{th} data packet arriving at the system. This packet must wait in queue for the residual time R_i until the end of the current packet transmission or reservation interval. It must also wait for the transmission of the N_i packets currently in the queue (this includes both packets for which reservations were already made in the last reservation interval and earlier arrivals waiting to make a reservation). Finally, the packet must wait during the next reservation interval $V_{\ell(i)}$, say, in which its reservation will be made (see Fig. 3.16). Thus, the expected queueing delay for the i^{th} packet is given by

$$E\{W_i\} = E\{R_i\} + E\{N_i\}/\mu + E\{V_{\ell(i)}\} \tag{3.60}$$

The similarity of this reservation system to the $M/G/1$ queue with vacations should be noted. The only difference is that in the gated reservation system, a reservation interval starts when all packets have been served from the previous interval, whereas in the vacation system, a vacation interval starts when *all* previous arrivals have been served. (In fact, the exhaustive version of this reservation system is equivalent to the vacation system.) The time-average mean residual time for the two systems is clearly the same (see Fig. 3.13), and is given by $\lambda\overline{X^2}/2+(1-\rho)\overline{V^2}/2\overline{V}$.

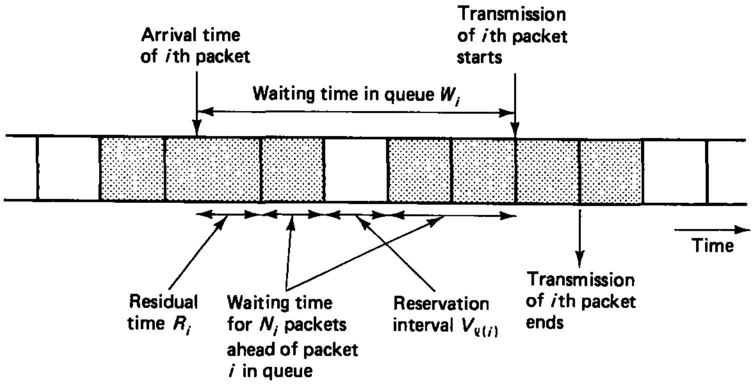

Arrival time of *i*th packet
Transmission of *i*th packet starts
Waiting time in queue W_i
Time
Residual time R_i
Waiting time for N_i packets ahead of packet *i* in queue
Reservation interval $V_{\ell(i)}$
Transmission of *i*th packet ends

Figure 3.16 Calculation of the average waiting time in the single-user gated system. The expected waiting time $E\{W_i\}$ of the i^{th} packet is

$$E\{W_i\} = E\{R_i\} + E\{N_i\}/\mu + E\{V_{\ell(i)}\}$$

The value of $\lim_{i\to\infty} E\{N_i\}/\mu$ is ρW in both systems, and finally the value of $\lim_{i\to\infty} E\{V_{\ell(i)}\}$ is just $\overline{V}$. Thus, from Eq. (3.60) the expected time in queue for the single-user reservation system is

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}} + \frac{\overline{V}}{1-\rho} \quad \text{(single user, gated)} \tag{3.61}$$

In the common situation where the reservation interval is a constant A, this simplifies to

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{3-\rho}{1-\rho}\right) \tag{3.62}$$

There is an interesting paradox associated with Eq. (3.61). We have seen that a fraction $1-\rho$ of time is used on reservations. Since there is one reservation interval of mean duration $\overline{V}$ per cycle, we can conclude that the expected cycle length must be $\overline{V}/(1-\rho)$ (Prob. 3.23 develops this result more carefully). The mean queueing delay in Eq. (3.61) can be an arbitrarily large multiple of this mean cycle length, which seems paradoxical since each packet is transmitted on the cycle following its arrival. The explanation of this is that more packets tend to arrive in long cycles than in short cycles, and thus mean cycle length is not representative of the cycle lengths seen by arriving packets; this is the same phenomenon that makes the mean residual service time used in the P-K formula derivation larger than one might think (see also Problem 3.15).

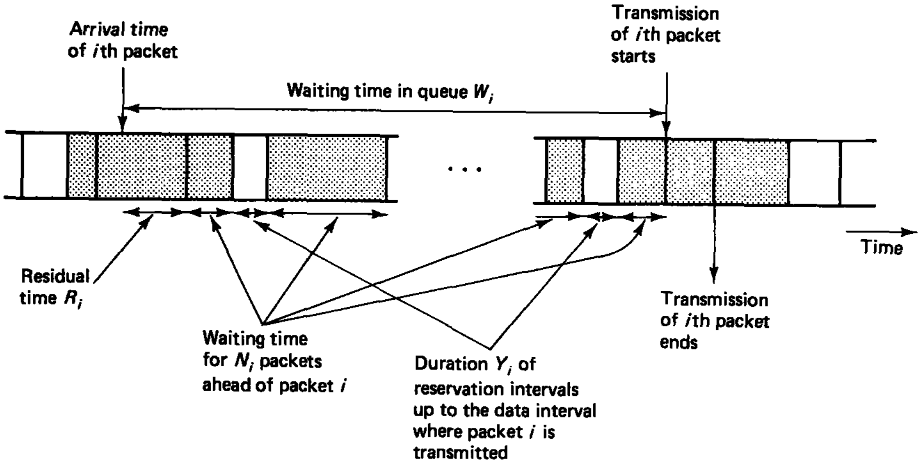

Arrival time of *i*th packet
Waiting time in queue W_i
Transmission of *i*th packet starts
Residual time R_i
Waiting time for N_i packets ahead of packet *i*
Duration Y_i of reservation intervals up to the data interval where packet *i* is transmitted
Transmission of *i*th packet ends
Time

Figure 3.17 Calculation of the average waiting time in the multiuser system. The expected waiting time $E\{W_i\}$ of the i^{th} packet is

$$E\{W_i\} = E\{R_i\} + E\{N_i\}/\mu + E\{Y_i\}$$

Multiuser System

Suppose that the system has m users, each with independent Poisson arrivals of rate λ/m. Again $\overline{X}$ and $\overline{X^2}$ are the first two moments of the service time for each user's packets. We denote by $\overline{V_i}$ and $\overline{V_i^2}$, respectively, the first two moments of the reservation intervals of user i. The service times and reservation intervals are all independent. We number the users $0, 1, \ldots, m-1$ and assume that the ℓ^{th} reservation interval is used to make reservations for user $\ell \text{mod}(m)$ and the subsequent (ℓ^{th}) data interval is used to send the packets corresponding to those reservations.

Consider the i^{th} packet arrival into the system (counting packets in order of arrival, regardless of user). As before, the expected delay for this packet consists of three terms: first, the mean residual time for the packet or reservation in progress; second, the expected time to transmit the number N_i of packets that must be transmitted before packet i; and third, the expected duration of reservation intervals (see Fig. 3.17). Thus,

$$E\{W_i\} = E\{R_i\} + E\{N_i\}/\mu + E\{Y_i\} \tag{3.63}$$

where Y_i is the duration of all the whole reservation intervals during which packet

i must wait before being transmitted. The time average mean residual time is calculated similarly as before, and is given by

$$R = \frac{\lambda \overline{X^2}}{2} + \frac{(1-\rho)\sum_{\ell=0}^{m-1}\overline{V_\ell^2}}{2\sum_{\ell=0}^{m-1}\overline{V}_\ell} \tag{3.64}$$

The number of packets N_i that i must wait for is not equal to the number already in queue, but the order of serving packets is independent of packet service time; thus, each packet served before i still has a mean transmission time $1/\mu$ as indicated in Eq. (3.63) and by Little's formula, the value of $\lim_{i\to\infty} E\{N_i\}/\mu$ is ρW as before. Letting $Y = \lim_{i\to\infty} E\{Y_i\}$, we can thus write the steady state version of Eq. (3.63)

$$W = R + \rho W + Y$$

or, equivalently,

$$W = \frac{R+Y}{1-\rho} \tag{3.65}$$

We first calculate Y for an exhaustive system. Denote

$$\alpha_{\ell j} = E\{Y_i |\text{ packet } i \text{ arrives in user } \ell\text{'s reservation or data interval and belongs to user } (\ell + j)\mathrm{mod}(m)\}$$

We have

$$\alpha_{\ell j} = \begin{cases} 0, & j = 0 \\ \overline{V}_{(\ell+1)\mathrm{mod}(m)} + \cdots + \overline{V}_{(\ell+j)\mathrm{mod}(m)}, & j > 0 \end{cases}$$

Since packet i belongs to any user with equal probability $1/m$, we have

$$E\{Y_i |\text{ packet } i \text{ arrives in user } \ell\text{'s reservation or data interval}\}$$
$$= \frac{1}{m}\sum_{j=1}^{m-1}\alpha_{\ell j} = \sum_{j=1}^{m-1}\frac{m-j}{m}\overline{V}_{(\ell+j)\mathrm{mod}(m)} \tag{3.66}$$

Since all users have equal data rate, the data intervals of all users have equal average length in steady state. Therefore, in steady state, a packet will arrive during user ℓ's data interval with probability ρ/m, and during user ℓ's reservation interval with probability $(1-\rho)\overline{V}_\ell \Big/ \left(\sum_{k=0}^{m-1}\overline{V}_k\right)$. Using this fact in Eq. (3.66) we obtain the following equation for $Y = \lim_{i\to\infty} E\{Y_i\}$

$$Y = \sum_{\ell=0}^{m-1}\left(\frac{\rho}{m} + \frac{(1-\rho)\overline{V}_\ell}{\sum_{k=0}^{m-1}\overline{V}_k}\right)\sum_{j=1}^{m-1}\frac{m-j}{m}\overline{V}_{(\ell+j)\mathrm{mod}(m)}$$
$$= \frac{\rho}{m}\sum_{j=1}^{m-1}\frac{m-j}{m}\left(\sum_{\ell=0}^{m-1}\overline{V}_\ell\right) + \frac{(1-\rho)}{\sum_{k=0}^{m-1}\overline{V}_k}\sum_{\ell=0}^{m-1}\sum_{j=1}^{m-1}\frac{m-j}{m}\overline{V}_\ell\overline{V}_{(\ell+j)\mathrm{mod}(m)} \tag{3.67}$$

The last sum above can be written

$$\sum_{\ell=0}^{m-1}\sum_{j=1}^{m-1}\frac{m-j}{m}\overline{V}_\ell\overline{V}_{(\ell+j)\mathrm{mod}(m)}=\frac{1}{2}\left[\left(\sum_{\ell=0}^{m-1}\overline{V}_\ell\right)^2-\sum_{\ell=0}^{m-1}\overline{V}_\ell^2\right]$$

(To see this, note that the right side above is the sum of all possible products $\overline{V}_\ell\overline{V}_{\ell'}$ for $\ell \neq \ell'$. The left side is the sum of all possible terms $(j/m)\overline{V}_\ell\overline{V}_{\ell'}$ and $[(m-j)/m]\overline{V}_\ell\overline{V}_{\ell'}$, where $j = |\ell - \ell'|$ and $\ell \neq \ell'$.) Using this expression, and denoting

$$\overline{V}=\frac{1}{m}\sum_{\ell=0}^{m-1}\overline{V}_\ell$$

as the reservation interval averaged over all users, we can write Eq. (3.67) as

$$\begin{aligned}Y&=\frac{\rho\overline{V}(m-1)}{2}+\frac{(1-\rho)m\overline{V}}{2}-\frac{(1-\rho)\sum_{\ell=0}^{m-1}\overline{V}_\ell^2}{2m\overline{V}}\\&=\frac{(m-\rho)\overline{V}}{2}-\frac{(1-\rho)\sum_{\ell=0}^{m-1}\overline{V}_\ell^2}{2m\overline{V}}.\end{aligned}\tag{3.68}$$

Combining Eq. (3.64), (3.65), and (3.68), we obtain

$$W=\frac{\lambda\overline{X^2}}{2(1-\rho)}+\frac{(m-\rho)\overline{V}}{2(1-\rho)}+\frac{\sum_{\ell=0}^{m-1}\left(\overline{V_\ell^2}-\overline{V}_\ell^2\right)}{2m\overline{V}}$$

Denoting

$$\sigma_V^2=\frac{\sum_{\ell=0}^{m-1}\left(\overline{V_\ell^2}-\overline{V}_\ell^2\right)}{m}$$

as the variance of the reservation intervals averaged over all users, we finally obtain

$$W=\frac{\lambda\overline{X^2}}{2(1-\rho)}+\frac{(m-\rho)\overline{V}}{2(1-\rho)}+\frac{\sigma_V^2}{2\overline{V}}\qquad\text{(exhaustive)}\tag{3.69}$$

The partially gated system is the same as the exhaustive except that if a packet of a user arrives during a user's own data interval (an event of probability ρ/m in steady state), it is delayed by an additional $m\overline{V}$, the average sum of reservation intervals in a cycle. Thus, Y is increased by $\rho\overline{V}$ in the preceding calculation, and we obtain

$$W=\frac{\lambda\overline{X^2}}{2(1-\rho)}+\frac{(m+\rho)\overline{V}}{2(1-\rho)}+\frac{\sigma_V^2}{2\overline{V}}\qquad\text{(partially gated)}\tag{3.70}$$

Consider finally the fully gated system. This is the same as the partially gated system except that if a packet of a user arrives during a user's own reservation interval (an event of probability $(1-\rho)/m$ in steady state), it is delayed by an additional $m\overline{V}$. This increases Y by an additional $(1-\rho)\overline{V}$ and results in the equation

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{(m+2-\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \quad \text{(gated)} \qquad (3.71)$$

In comparing these results with the single user system, consider the case where the reservation interval is a constant A/m. Thus, A is the overhead or reservation time for an entire cycle of reservations for each user, which is usually the appropriate parameter to compare with A in Eq. (3.62). We then have ($\overline{V} = A/m$, $\sigma_V^2 = 0$)

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{1-\rho/m}{1-\rho}\right) \quad \text{(exhaustive)} \qquad (3.72)$$

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{1+\rho/m}{1-\rho}\right) \quad \text{(partially gated)} \qquad (3.73)$$

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2}\left(\frac{1+(2-\rho)/m}{1-\rho}\right) \quad \text{(gated)} \qquad (3.74)$$

It can be seen that delay is somewhat reduced in the multiuser case; essentially, packets are delayed by roughly the same amount until the reservation time in all cases but delay is quite small after the reservation in the multiuser case.

Limited Service Systems

We now consider a variation of the multiuser system whereby, in each user's data interval, *only the first* packet of the user waiting in queue (if any) is transmitted (rather than *all* waiting packets). We concentrate on the gated and partially gated versions of this system, since an exhaustive version does not make sense. As before, we have

$$E\{W_i\} = E\{R_i\} + E\{N_i\}/\mu + E\{Y_i\}$$

and by taking the limit as $i \to \infty$, we obtain

$$W = R + \rho W + Y \qquad (3.75)$$

Here R is given by Eq. (3.64) as before. To calculate the new formula for Y for the partially gated system, we argue as follows. A packet arriving during user ℓ's data or reservation interval will belong to any one of the users with equal probability $1/m$. Therefore, in steady state, the expected number of packets waiting in the individual queue of the user that owns the arriving packet, averaged over all users, is $\lim_{i\to\infty} E\{N_i\}/m = \lambda W/m$. Each of these packets causes an extra cycle of reservations $m\overline{V}$, so Y is increased by an amount $\lambda W\overline{V}$. Using this fact in Eq. (3.75), we see that

$$W = \frac{R+\tilde{Y}}{1-\rho-\lambda\overline{V}}$$

where $\tilde{Y}$ is the value of Y obtained earlier for the partially gated system without the single-packet-per-data-interval restriction. Equivalently, we see from Eq. (3.65), that *the single-packet-per-data-interval restriction results in an increase of the average waiting time for the partially gated system by a factor*

$$\frac{1-\rho}{1-\rho-\lambda\overline{V}}$$

Using this fact in Eq. (3.70), we obtain

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho-\lambda\overline{V})} + \frac{(m+\rho)\overline{V}}{2(1-\rho-\lambda\overline{V})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\lambda\overline{V})}$$

(limited service, partially gated) (3.76)

Consider now the gated version. Y_i is the same as for the partially gated system except for an additional cycle of reservation intervals of average length $m\overline{V}$ associated with the event where packet i arrives during the reservation interval of its owner, and the subsequent data interval is empty. It is easily verified (Prob. 3.24) that the latter event occurs with steady-state probability $(1-\rho-\lambda\overline{V})/m$. Therefore, for the gated system Y equals the corresponding value for the partially gated system plus $(1-\rho-\lambda\overline{V})\overline{V}$. This adds $\overline{V}$ to the value of W for the partially gated system, and the average waiting time now is

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho-\lambda\overline{V})} + \frac{(m+2-\rho-2\lambda\overline{V})\overline{V}}{2(1-\rho-\lambda\overline{V})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\lambda\overline{V})}$$

(limited service, gated) (3.77)

Note that it is not enough that $\rho = \lambda/\mu < 1$ for W to be bounded; rather, $\rho + \lambda\overline{V} < 1$ is required or, equivalently,

$$\lambda\left(\frac{1}{\mu}+\overline{V}\right) < 1$$

This is due to the fact that each packet requires a separate reservation interval of average length $\overline{V}$, thereby effectively increasing the average transmission time from $1/\mu$ to $1/\mu+\overline{V}$.

As a final remark, consider the case of a very large number of users m and a very small average reservation interval $\overline{V}$. An examination of the equation given for the average waiting time W of every system considered so far shows that as $m \to \infty$, $\overline{V} \to 0$, $\sigma_V^2/\overline{V} \to 0$, and $m\overline{V} \to A$, where A is a constant, we have

$$W \to \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{A}{2(1-\rho)}$$

It can be shown (Prob. 3.23) that $A/(1-\rho)$ is the average length of a cycle (m successive reservation and data intervals). Thus, W approaches the $M/G/1$ average waiting time plus one half the average cycle length.

3.5.3 Priority Queueing

Consider the $M/G/1$ system with the difference that arriving customers are divided into n different priority classes. Class 1 has the highest priority, class 2 has the second highest, and so on. The arrival rate and the first two moments of service time of each class k are denoted λ_k, $\overline{X}_k = 1/\mu_k$, and $\overline{X_k^2}$, respectively. The arrival processes of all classes are assumed independent, Poisson, and independent of the service times.

Nonpreemptive Priority

We first consider the nonpreemptive priority rule whereby a customer undergoing service is allowed to complete service without interruption even if a customer of higher priority arrives in the meantime. A separate queue is maintained for each priority class. When the server becomes free, the first customer of the highest nonempty priority queue enters service. This priority rule is one of the most appropriate for modeling packet transmission systems.

We will develop an equation for average delay for each priority class, which is similar to the P-K formula and admits a similar derivation. Denote

N_Q^k : Average number in queue for priority k

W_k : Average queueing time for priority k

$\rho_k = \lambda_k/\mu_k$: System utilization for priority k

R : Mean residual service time

We assume that the overall system utilization is less than one, *i.e.*,

$$\rho_1 + \rho_2 + \cdots + \rho_n < 1$$

When this assumption is not satisfied, there will be some priority class k such that the average delay of customers of priority k and lower will be infinite while the average delay of customers of priority higher than k will be finite. Problem 3.16 takes a closer look at this situation.

Similarly, as in the derivation of the P-K formula given earlier, we have for the highest priority class

$$W_1 = R + \frac{1}{\mu_1} N_Q^1$$

Eliminating N_Q^1 from this equation using Little's Theorem

$$N_Q^1 = \lambda_1 W_1$$

we obtain

$$W_1 = R + \rho_1 W_1$$

and, finally,

$$W_1 = \frac{R}{1-\rho_1} \tag{3.78}$$

For the second priority class, we have a similar expression for the queueing delay W_2 except that we have to count the additional queueing delay due to customers of higher priority that arrive while a customer is waiting in queue. This is the meaning of the last term in the formula

$$W_2 = R + \frac{1}{\mu_1}N_Q^1 + \frac{1}{\mu_2}N_Q^2 + \frac{1}{\mu_1}\lambda_1 W_2$$

Using Little's Theorem ($N_Q^k = \lambda_k W_k$) we obtain

$$W_2 = R + \rho_1 W_1 + \rho_2 W_2 + \rho_1 W_2$$

which yields

$$W_2 = \frac{R + \rho_1 W_1}{1 - \rho_1 - \rho_2}$$

Using the expression $W_1 = R/(1-\rho_1)$ obtained earlier, we finally have

$$W_2 = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

The derivation is similar for all priority classes $k > 1$. The formula for the waiting time in queue is

$$W_k = \frac{R}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)} \tag{3.79}$$

The average delay per customer of class k is

$$T_k = \frac{1}{\mu_k} + W_k \tag{3.80}$$

The mean residual service time R must now be derived. As in the earlier derivation of the P-K formula (compare with Fig. 3.10) we have

$$R = \frac{1}{2}\left(\sum_{i=1}^{n}\lambda_i\right)\overline{X^2} \tag{3.81}$$

where $\overline{X^2}$ denotes the second moment of service time averaged over all priority classes. In particular,

$$\overline{X^2} = \frac{\lambda_1}{\sum_{i=1}^{n}\lambda_i}\overline{X_1^2} + \cdots + \frac{\lambda_n}{\sum_{i=1}^{n}\lambda_i}\overline{X_n^2}$$

Substitution in Eq. (3.81) yields

$$R = \frac{1}{2}\sum_{i=1}^{n} \lambda_i \overline{X_i^2} \tag{3.82}$$

The average waiting time in queue and the average delay per customer for each class is obtained from Eqs. (3.79), (3.80), and (3.82)

$$W_k = \frac{\sum_{i=1}^{n} \lambda_i \overline{X_i^2}}{2(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)} \tag{3.83}$$

$$T_k = \frac{1}{\mu_k} + W_k$$

Note that it is possible to affect the average delay per customer by choosing the priority classes appropriately. It is generally true that average delay tends to be reduced when customers with short service times are given higher priority. (For an example from common experience, consider the supermarket practice of having special checkout counters for customers with few items. A similar situation can be seen in copying machine waiting lines, where people often give priority to others that need to make just a few copies.) An analytical substantiation can be obtained by considering a nonpreemptive system and two customer classes A and B with respective arrival and service rates λ_A, μ_A, and λ_B, μ_B. A straightforward calculation using the formulas above shows that if $\mu_A > \mu_B$, then the average delay per customer (averaged over both classes)

$$T = \frac{\lambda_A T_A + \lambda_B T_B}{\lambda_A + \lambda_B}$$

is smaller when A is given priority over B than when B is given priority over A. For related results, see Prob. 3.19.

The analysis given above does not extend easily to the case of multiple servers, primarily because there is no simple formula for the mean residual time. If, however, the service times of all priority classes are identically and exponentially distributed, there is a convenient characterization of R. Equation (3.79) then yields a closed-form expression for the average waiting times W_k (see Prob. 3.17).

Preemptive Resume Priority

One of the features of the nonpreemptive priority rule is that the average delay of a priority class depends on the arrival rate of lower priority classes. This is evident from Eq. (3.83) and is due to the fact that a high priority customer must wait for a lower priority customer already in service. This dependence is not present in the ***preemptive resume priority discipline***, whereby service of a customer is interrupted when a higher priority customer arrives, and is resumed from the point of interruption once all customers of higher priority have been served.

As we consider the calculation of T_k, the average time in the system of priority k customers, we should keep in mind that the presence of customers of priorities $k+1$ through n does not affect this calculation. Therefore, we can treat each priority class as if it were the lowest in the system.

The system time T_k consists of three terms. The first is the customer's average service time $1/\mu_k$. The second is the average time required, upon arrival of a priority k customer, to service customers of priority 1 to k already in the system, *i.e.*, the average unfinished work corresponding to priorities 1 through k. This term is equal to the average waiting time in the corresponding, ordinary $M/G/1$ system (without priorities), where the customers of priorities $k+1$ through n are neglected, *i.e.*, (cf. Eq. (3.48))

$$\frac{R_k}{1-\rho_1-\cdots-\rho_k}$$

where R_k is the mean residual time

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i \overline{X_i^2}}{2} \tag{3.84}$$

This follows since, at all times, the unfinished work (sum of remaining service times of all customers in the system) of an $M/G/1$-type system is independent of the priority discipline of the system. This is true of all queueing systems that are conservative in the sense that the server is always busy when the system is nonempty, and customers leave the system only after receiving their required service. The third term in the expression for T_k is the average waiting time for customers of priorities 1 through $k-1$ who arrive while the customer of class k is in the system. This term is

$$\sum_{i=1}^{k-1} \frac{1}{\mu_i} \lambda_i T_k = \sum_{i=1}^{k-1} \rho_i T_k$$

for $k > 1$, and is zero for $k = 1$. Collecting these terms, we obtain the equation

$$T_k = \frac{1}{\mu_k} + \frac{R_k}{1-\rho_1-\cdots-\rho_k} + \left(\sum_{i=1}^{k-1} \rho_i\right) T_k \tag{3.85}$$

The final result is, for $k = 1$,

$$T_1 = \frac{(1/\mu_1)(1-\rho_1)+R_1}{1-\rho_1} \tag{3.86}$$

and, for $k > 1$,

$$T_k = \frac{(1/\mu_k)(1-\rho_1-\cdots-\rho_k)+R_k}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)} \tag{3.87}$$

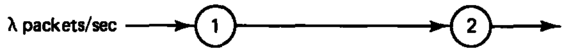


Figure 3.18 Two equal capacity transmission lines in tandem. If all packets have equal length, there is no queueing delay in the second queue.

where R_k is given by Eq. (3.84). As for the nonpreemptive system, there is no easy extension of this formula to the case of multiple servers unless the service times of all priority classes are identically and exponentially distributed (see Prob. 3.17).

3.6 NETWORKS OF TRANSMISSION LINES

In a data network, there are many transmission queues that interact in the sense that a traffic stream departing from one queue enters one or more other queues, perhaps after merging with portions of other traffic streams departing from yet other queues. Analytically, this has the unfortunate effect of complicating the character of the arrival processes at downstream queues. The difficulty is that the packet interarrival times become strongly correlated with packet lengths once packets have traveled beyond the first queue at their entry point in the network. As a result it is impossible to carry out a precise and effective analysis comparable to the one provided for queueing systems such as $M/M/1$, $M/G/1$, etc.

As an illustration of the phenomena that complicate the analysis, consider two transmission lines of equal capacity in tandem, as shown in Fig. 3.18. Assume that Poisson arrivals of rate λ packets/sec enter the first queue, and that all packets have *equal* length. Therefore, the first queue is $M/D/1$ and the average packet delay there is given by the Pollaczek-Khinchin formula. However, at the second queue, the interarrival times must be greater than or equal to $1/\mu$ (the packet transmission time). Furthermore, because the packet transmission times are equal at both queues, each packet arriving at the second queue will complete transmission at or before the time the next packet arrives, so there is *no waiting at the second queue.* Therefore, a delay model based on Poisson assumptions is totally inappropriate for the second queue.

Consider next the case of the two tandem transmission lines where packet lengths are exponentially distributed, and are independent of each other as well as of the interarrival times at the first queue. Then the first queue is $M/M/1$. The second queue, however, *cannot* be modeled as $M/M/1$. The reason is, again, that *the interarrival times at the second queue are strongly correlated with the packet lengths.* To see this, consider a busy period at the first queue where several packets are transmitted one after the other. The interarrival time at the second queue

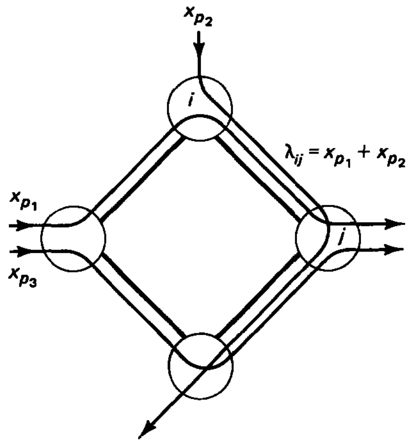


Figure 3.19 A network of transmission lines. The total arrival rate λ_{ij} at a link (i,j) is equal to the sum of arrival rates x_p of all packet streams p traversing the link.

between two such packets equals the transmission time of the second packet. As a result, long packets will typically wait less time at the second queue than short packets, since their transmission at the first queue takes longer, thereby giving the second queue more time to empty out. For a traffic analogy, consider a slow truck traveling on a busy narrow street together with several faster cars. The truck will typically see empty space ahead of it while being closely followed by the faster cars.

As an indication of the difficulty of analyzing queueing network problems involving dependent interarrival and service times, no analytical solution is known for even the simple tandem queueing problem of Fig. 3.18 involving Poisson arrivals and exponentially distributed service times. In the real situation where packet lengths and interarrival times are correlated, a simulation has shown that under heavy traffic conditions, average delay per packet is smaller than in the idealized situation where there is no such correlation. The reverse is true under light traffic conditions. It is not known whether and in what form this result can be extended to more general networks.

Consider now a network of communication links as shown in Fig. 3.19. Assume that there are several packet streams each following a path p that consists of a sequence of links through the network. Let x_p, in packets/sec, be the arrival rate of the packet stream associated with the path p. Then, the total arrival rate at link

(i,j) is

$$\lambda_{ij} = \sum_{\substack{\text{all } p \\ \text{traversing} \\ \text{link } (i,j)}} x_p \tag{3.88}$$

We have seen from the special case of two tandem queues that even if the packet streams are Poisson with independent packet lengths at their point of entry into the network, this property is lost after the first transmission line. To resolve the dilemma, it was suggested by Kleinrock [Kle64] that merging several packet streams on a transmission line has an effect akin to restoring the independence of interarrival times and packet lengths. It was concluded that it is often appropriate to adopt an $M/M/1$ queueing model for each communication link regardless of the interaction of traffic on this link with traffic on other links (see also the discussion preceding Jackson's theorem in section 3.8). This is known as the *Kleinrock independence approximation* and seems to be a reasonably good approximation for systems involving Poisson stream arrivals at the entry points, packet lengths that are nearly exponentially distributed, a densely connected network, and moderate to heavy traffic loads. Based on this $M/M/1$ model, the average number of packets in queue or service at (i,j) is

$$N_{ij} = \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{3.89}$$

where $1/\mu_{ij}$ is the average packet transmission time on link (i,j). The average number of packets summed over all queues is

$$N = \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{3.90}$$

so by Little's Theorem, the average delay per packet (neglecting processing and propagation delays) is

$$T = \frac{1}{\gamma} \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{3.91}$$

where $\gamma = \sum_p x_p$ is the total arrival rate in the system. If the average processing and propagation delay d_{ij} at link (i,j) is not negligible, this formula should be adjusted to

$$T = \frac{1}{\gamma} \sum_{(i,j)} \left(\frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} + \lambda_{ij} d_{ij} \right) \tag{3.92}$$

Finally, the average delay per packet of a traffic stream traversing a path p is given by

$$T_p = \sum_{\substack{\text{all } (i,j) \\ \text{on path } p}} \left(\frac{\lambda_{ij}}{\mu_{ij}(\mu_{ij} - \lambda_{ij})} + \frac{1}{\mu_{ij}} + d_{ij} \right) \tag{3.93}$$

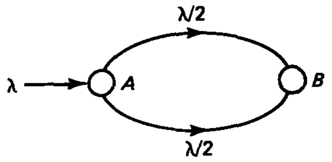


Figure 3.20 A Poisson process with rate λ divided among two links. If division is done by randomization, each link behaves like an $M/M/1$ queue. If division is done by metering, the whole system behaves like an $M/M/2$ queueing system.

where the three terms in the sum above represent average waiting time in queue, average transmission time, and processing and propagation delay, respectively.

In many networks, the assumption of exponentially distributed packet lengths is not appropriate. Given a distribution function on packet lengths, one may keep the approximation of independence between queues but replace (3.89) with the P-K formula for average number in the system. Equations (3.90) to (3.93) would then be modified in an obvious way.

It should be mentioned that the main approximation involved in Eq. (3.90) is due to the correlation of the packet lengths and the packet interarrival times at the various queues in the network. If somehow this correlation was not present, *e.g.*, if a packet upon departure from a transmission line was assigned a new length drawn from an exponential distribution, then the average number of packets in the system would be given indeed by the formula

$$N = \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}}$$

This fact (by no means obvious) is a consequence of Jackson's Theorem, which will be discussed in section 3.8.

In datagram networks that involve multiple path routing for some origin-destination pairs, the accuracy of the $M/M/1$ approximation deteriorates for another reason which is best illustrated by an example.

Example 13

Suppose node A sends traffic to node B along two equal capacity links in the simple network of Fig. 3.20. Packets arrive at A according to a Poisson process with rate λ packets/sec. Packet transmission times are exponentially distributed and independent of interarrival times similarly as in the $M/M/1$ system. Assume that the arriving traffic is to be divided equally among the two links. However, how should this division be implemented? Consider the following possibilities.

1. **Randomization**: Here each packet is assigned upon arrival at A to one of the two links based on the outcome of a fair coin flip. It is then possible to show that the arrival process on each of the two queues is

Poisson and independent of the packet lengths (see Prob. 3.11). Therefore, each of the two queues behaves like an $M/M/1$ queue with arrival rate $\lambda/2$ and average delay per packet

$$T_R = \frac{1}{\mu - \lambda/2} = \frac{2}{2\mu - \lambda} \tag{3.94}$$

which is consistent with the Kleinrock independence approximation.

2. **Metering**: Here each arriving packet is assigned to the queue that currently has the smallest total backlog in bits (equivalently to the queue that will empty its current backlog first). This scheme works like an $M/M/2$ system with arrival rate λ and with each link playing the role of a server. Using the result of subsection 3.4.1, the average delay per packet can be calculated to be

$$T_M = \frac{2}{(2\mu - \lambda)(1+\rho)} \tag{3.95}$$

where $\rho = \lambda/2\mu$.

Comparing Eqs. (3.94) and (3.95), we see that metering performs better than randomization in terms of delay by a factor $1/(1+\rho)$. This is basically the same advantage that statistical multiplexing with multiple channels holds over time-division multiplexing as discussed in the Ex. 9 of subsection 3.4.1. Generally, it is preferable to use some form of metering rather than randomization when dividing traffic among alternate routes. However, in contrast with randomization, metering destroys the Poisson character of the arrival process at the point of division. In our example, when metering is used, the interarrival times at each link are neither exponentially distributed nor independent of preceding packet lengths. Therefore, the use of metering (which is recommended for performance reasons) tends to degrade the accuracy of the $M/M/1$ approximation.

3.7 *TIME REVERSIBILITY—BURKE'S THEOREM*

The analysis of the $M/M/1$, $M/M/m$, $M/M/\infty$, and $M/M/m/m$ systems was based on the relation that for any state j, the steady-state probability of j times the transition probability from j to $j+1$ is equal to the steady state probability of state $j+1$ times the transition probability from $j+1$ to j. These relations, called *detailed balance equations*, are valid for any Markov chain with integer states in which transitions can only occur between neighboring states, *i.e.*, from j to $j-1$, j, or $j+1$; these Markov chains are called *birth-death* processes. The detailed balance equations lead to an important property called time reversibility as we now explain.

Consider an irreducible, aperiodic, discrete-time Markov chain $X_n, X_{n+1}, \ldots$ having transition probabilities P_{ij} and stationary distribution $\{p_j | j \geq 0\}$. Suppose that the chain is in steady state, *i.e.*,

$$P\{X_n = j\} = p_j\,, \quad \text{for all } n$$

(this occurs if the initial state is chosen according to the stationary distribution, and is equivalent to imagining that the process began at time $-\infty$).

Suppose we trace the sequence of states going backwards in time. That is, starting at some n, consider the sequence of states $X_n, X_{n-1}, \ldots$. This sequence is itself a Markov chain as seen by the following calculation

$$
\begin{aligned}
&P\{X_m = j | X_{m+1} = i,\ X_{m+2} = i_2, \ldots, X_{m+k} = i_k\} \\
&= \frac{P\{X_m = j, X_{m+1} = i,\ X_{m+2} = i_2, \ldots, X_{m+k} = i_k\}}{P\{X_{m+1} = i,\ X_{m+2} = i_2, \ldots, X_{m+k} = i_k\}} \\
&= \frac{P\{X_m = j\}\, P\{X_{m+1} = i | X_m = j\} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_m = j, X_{m+1} = i\}}{P\{X_{m+1} = i\} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_{m+1} = i\}} \\
&= \frac{p_j P_{ji} P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_{m+1} = i\}}{p_i P\{X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_{m+1} = i\}} = \frac{p_j P_{ji}}{p_i}
\end{aligned}
$$

Thus, conditional on the state at time $m+1$, the state at time m is independent of that at times $m+2, m+3\ldots$ The backward transition probabilities are given by

$$P_{ij}^* = P\{X_m = j | X_{m+1} = i\} = \frac{p_j P_{ji}}{p_i}, \qquad i, j \geq 0 \tag{3.96}$$

If $P_{ij}^* = P_{ij}$ for all i, j (*i.e.*, the transition probabilities of the forward and reversed chain are identical), we say that the chain is *time reversible.*

We list some properties of the reversed chain:

1. The reversed chain is irreducible, aperiodic, and has the same stationary distribution as the forward chain. (This property can be shown either by elementary reasoning using the definition of the reversed chain, or by verifying the equality $p_j = \sum_{i=0}^{\infty} p_i P_{ij}^*$ using Eq. (3.96).) The intuitive idea here is that the reversed chain corresponds to the same process, looked at in the reversed time direction. Thus, if the steady state probabilities correspond to time averages (as they must for the concept of steady state to be meaningful), then the steady state is the same in both directions. Note that, in view of the equality of the stationary distribution of the forward and reverse chains, Eq. (3.96) can be intuitively explained. It expresses the fact that (with probability one) the proportion of transitions from j to i out of all transitions in the forward chain (which is $p_j P_{ji}$) equals the proportion of transitions from i to j out of all transitions in the reversed chain (which is $p_i P_{ij}^*$).
2. If we can find nonnegative numbers p_i, $i \geq 0$, summing to unity and find a transition probability matrix $P^* = [P_{ij}^*]$ such that

$$p_i P_{ij}^* = p_j P_{ji}, \quad i, j \geq 0 \tag{3.97}$$

then $\{p_i | i \geq 0\}$ is the stationary distribution and P_{ij}^* are the transition probabilities of the reversed chain. (To see this, add Eq. (3.97) over all j to obtain

$$\sum_{j=0}^{\infty} p_j P_{ji} = p_i \sum_{j=0}^{\infty} P_{ij}^* = p_i \tag{3.98}$$

and conclude that $\{p_i | i \geq 0\}$ is the stationary distribution.) This property, which holds regardless of whether the chain is time reversible, is useful if we can guess at the nature of the reversed chain and verify Eq. (3.97), thereby obtaining both the p_j and P^*_{ij}; see section 3.8.

3. A chain is time reversible if and only if the detailed balance equations hold:

$$p_i P_{ij} = p_j P_{ji}, \quad i, j \geq 0$$

This follows from Eq. (3.96), and the definition of time reversibility. Otherwise explained, a chain is time reversible if, for all i and j, the proportion of transitions from i to j out of all transitions equals the proportion of transitions from j to i. In particular, the chains corresponding to the queueing systems $M/M/1$, $M/M/m$, $M/M/\infty$, and $M/M/m/m$ discussed in sections 3.3 and 3.4 are time reversible (in the limit as $\delta \to 0$). More generally chains corresponding to birth-death processes ($P_{ij} = 0$ if $|i - j| > 1$) are time reversible.

The idea of time reversibility extends in a straightforward manner to continuous-time Markov chains. The corresponding analysis can be carried out either directly, or by discretizing time in intervals of length δ, considering the corresponding discrete-time chain, and passing back to the continuous chain by taking the limit as $\delta \to 0$. All results regarding the reversed chain carry over almost verbatim from their discrete-time counterparts by replacing transition probabilities with transition rates. In particular if the continuous-time chain has transition rates q_{ij}, is irreducible, and has a stationary distribution $\{p_j | j \geq 0\}$, then:

1. The reversed chain is a continuous-time Markov chain with the same stationary distribution as the forward chain and with transition rates

$$q^*_{ij} = \frac{p_j q_{ji}}{p_i}, \quad i, j \geq 0 \tag{3.99}$$

2. If a probability distribution $\{p_j | j \geq 0\}$ and nonnegative numbers q^*_{ij} (for $i, j = 0, 1, \ldots$) can be found such that

$$p_i q^*_{ij} = p_j q_{ji}, \quad i, j \geq 0 \tag{3.100}$$

and for all $i \geq 0$

$$\sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} q^*_{ij} \tag{3.101}$$

then $\{p_j | j \geq 0\}$ is the stationary distribution of both the forward and the reversed chain, and q^*_{ij} are the transition rates of the reversed chain.

3. The forward chain is time reversible if and only if its stationary distribution and transition rates satisfy the detailed balanced equations

$$p_i q_{ij} = p_j q_{ji}, \quad i, j \geq 0$$

Consider now the $M/M/1$, $M/M/m$, and $M/M/\infty$ queueing systems. We assume that the initial state is chosen according to the stationary distribution so that the queueing systems are in steady state at all times. The reversed process can be represented by another queueing system where departures correspond to arrivals of the original system and arrivals correspond to departures of the original system (see Fig. 3.21). Because time reversibity holds for all these systems as discussed above, the forward and reversed systems are statistically indistinguishable in steady state. In particular by using the fact that the departure process of the forward system corresponds to the arrival process of the reversed system, we obtain the following result:

Burke's Theorem. Consider an $M/M/1$, $M/M/m$, or $M/M/\infty$ system with arrival rate λ. Suppose the system starts in steady state. Then the following hold true:

(a) The departure process is Poisson with rate λ.

(b) At each time t, the number of customers in the system is independent of the sequence of departure times prior to t.

(c) If customers are served in the order they arrive, then, given that a customer departs at time t, the arrival time of that customer is independent of the departure process prior to t.

Proof. (a) This follows from the fact that the forward and reversed systems are statistically indistinguishable in steady state, and the departure process in the forward system is the arrival process in the reversed system. (b) As shown in Fig. 3.22, for a fixed time t, the departures prior to t in the forward process are also the arrivals after t in the reversed process. The arrival process in the reversed system is independent Poisson, so the future arrival process does not depend on the current number in the system, which in forward system terms means that the past departure process does not depend on the current number in the system. (c) Consider a customer arriving at time t_1 and departing at time t_2 (see Fig. 3.23). In reversed system terms, the arrival process is independent Poisson, so the arrival process to the left of t_2 is independent of the times spent in the system of customers that arrived at or to the right of t_2. In particular $t_2 - t_1$ is independent of the (reversed system) arrival process to the left of t_2. In forward system terms, this means that $t_2 - t_1$ is independent of the departure process to the left of t_2. **QED**

Note that both parts (b) and (c) of Burke's Theorem are quite counterintuitive. One would expect that a recent stream of closely spaced departures suggests a busy system with an atypically large number of customers in queue. Yet Burke's Theorem shows that this is not so. Note carefully, however, that Burke's Theorem says nothing about the state of the system *before* a stream of closely spaced departures. Such a state would tend to have abnormally many customers in queue in

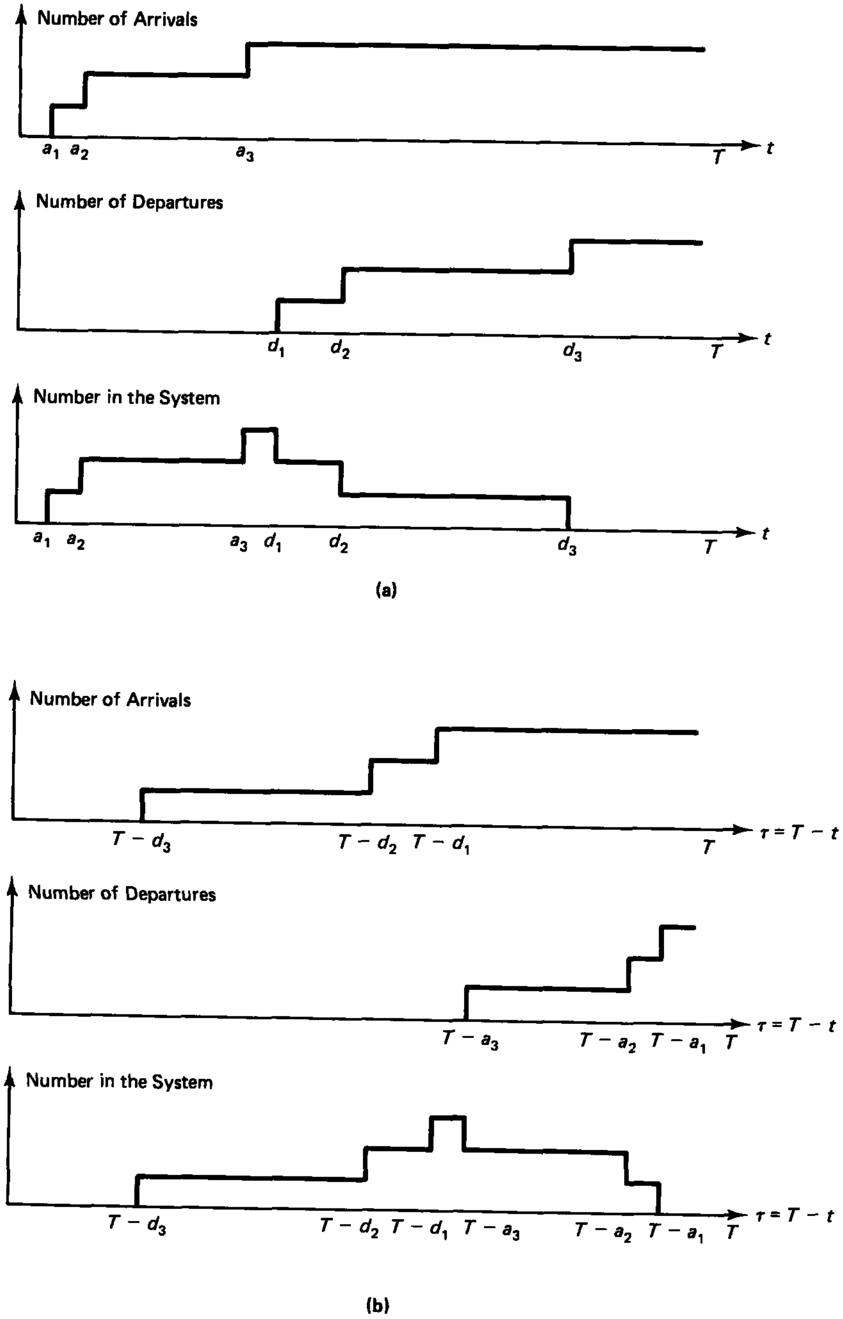


Figure 3.21 (a) Forward system number of arrivals, number of departures, and occupancy during $[0, T]$. (b) Reversed system number of arrivals, number of departures, and occupancy during $[0, T]$.

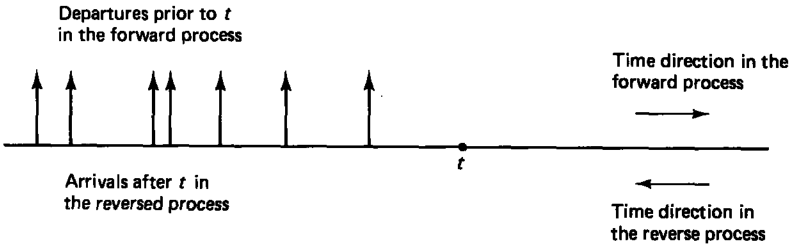

Departures prior to t in the forward process
Time direction in the forward process
t
Arrivals after t in the reversed process
Time direction in the reverse process

Figure 3.22 Customer departures *prior* to time t in the forward system, become customer arrivals *after* time t in the reversed system.

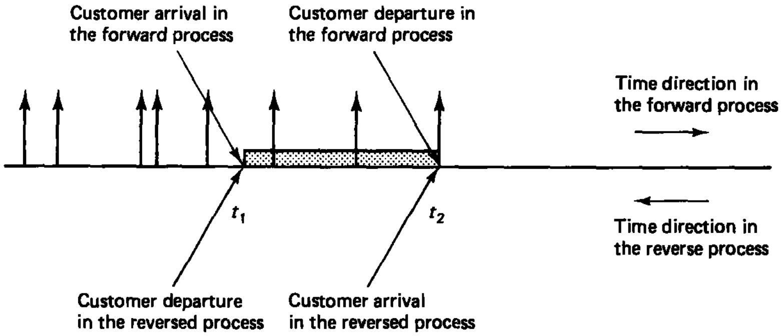

Customer arrival in the forward process
Customer departure in the forward process
Time direction in the forward process
t_1
t_2
Time direction in the reverse process
Customer departure in the reversed process
Customer arrival in the reversed process

Figure 3.23 Proof of part (c) of Burke's Theorem. In reversed system terms, customer arrivals after t_2 do not affect the time spent in the system of the customer that arrived at time t_2. In forward system terms, the time spent in the system by the customer $(t_2 - t_1)$ is independent of the departure process prior to the customer's departure.

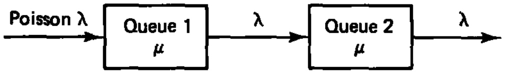


Figure 3.24 Two queues in tandem. The service times at the two queues are exponentially distributed and mutually independent. Using Burke's Theorem, we can show that the number of customers in queues 1 and 2 are independent at a given time and

$$P\{n \text{ at queue } 1,\ m \text{ at queue } 2\} = \rho_1^n(1-\rho_1)\rho_2^m(1-\rho_2),$$

i.e., the two queues behave as if they are independent $M/M/1$ queues in isolation.

accordance with intuition.

As an application of the theorem we analyze the simple queueing network involving Poisson arrivals and two queues in tandem with exponential service times (see Fig. 3.24). There is a major difference between this system and the one discussed in the previous section in that here we assume that the service times of a customer at the first and second queue are mutually independent as well as independent of the arrival process. This is what we called the Kleinrock independence approximation in the previous section. As a result of this assumption we will see that the occupancy distribution in the two queues is the same as if they were independent $M/M/1$ queues in isolation. This fact will also be shown in a more general context in the next section.

Let the rate of the Poisson arrival process be λ, and let the mean service times at queues 1 and 2 be $1/\mu_1$ and $1/\mu_2$ respectively. Let $\rho_1 = \lambda/\mu_1$ and $\rho_2 = \lambda/\mu_2$ be the corresponding utilization factors, and assume that $\rho_1 < 1$ and $\rho_2 < 1$. We will show that under steady-state conditions the following hold true:

(a) The number of customers presently at queue 1 and at queue 2 are mutually independent, and independent of the sequence of past departure times from queue 2. Furthermore,

$$P\{n \text{ at queue } 1,\, m \text{ at queue } 2\} = \rho_1^n(1-\rho_1)\rho_2^m(1-\rho_2) \tag{3.102}$$

(b) Assuming that customers are served at each queue in the order they arrive, the times (including service) spent by a customer in queue 1 and in queue 2 are mutually independent, and independent of the departure process from queue 2 prior to the customer's departure from the system.

To prove (a) above we first note that queue 1 is an $M/M/1$ queue so, by part (a) of Burke's Theorem, the departure process from queue 1 is Poisson and independent of the service times at queue 2. Therefore, queue 2, viewed in isolation, is an $M/M/1$ queue. Thus, from the results of section 3.1,

$$\begin{aligned} P\{n \text{ at queue } 1\} &= \rho_1^n(1-\rho_1), \\ P\{m \text{ at queue } 2\} &= \rho_2^m(1-\rho_2) \end{aligned} \tag{3.103}$$

From part (b) of Burke's theorem it follows that the number of customers presently in queue 1 is independent of the sequence of earlier arrivals at queue 2 and therefore also of the number of customers presently in queue 2. This implies that

$$P\{n \text{ at queue } 1,\, m \text{ at queue } 2\} = P\{n \text{ at queue } 1\} \cdot P\{m \text{ at queue } 2\}$$

and using Eq. (3.103) we obtain the desired product form (3.102).

To prove (b) above note that, by part (c) of Burke's Theorem, the time spent by a customer in queue 1 is independent of the sequence of arrival times at queue 2 prior to the customer's arrival at queue 2. However, these arrival times (together with the corresponding independent service times) determine the time the customer spends at queue 2 as well as the departure process from queue 2 prior to the customer's departure from the system. This proves the statement made in (b) above.

The assertion (b) above on the independence of the times spent by the same customer at queues 1 and 2 is quite counterintuitive, since one expects that a large number of customers found at queue 1 is likely to be reencountered at queue 2. As an illustration of how delicate this independence result is and how careful one should be about accepting such results without much thought, we note that the times a customer spends waiting *before entering service* at the two queues are *not* independent. To see this suppose $\mu_1 = \mu_2 = \mu$ and that λ is very small relative to μ. Then, almost all customers have zero waiting time at both queues. However, given that the wait of a customer at queue 1 is positive, the wait of the same customer at queue 2 will be positive with probability at least 1/2 (the probability that the customer will have a larger service time at queue 1 than the service time of the customer immediately ahead at queue 2). Therefore, for the same customer, the times spent waiting in queues 1 and 2 are not independent—they become independent when the corresponding service times are added.

Note that, by part (a) of Burke's Theorem, the arrival and the departure processes at both queues are Poisson. This fact together with facts (a) and (b) above can be similarly shown for a much broader class of queueing networks with Poisson arrivals and independent, exponentially distributed service times. We call such networks *acyclic* and define them as follows. We say that queue j is a *downstream neighbor* of queue i if there is a positive probability that a departing customer from queue i will next enter queue j. We say that queue j *lies downstream* of queue i if there is a sequence of queues starting from i and ending at j such that each queue after i in the sequence is a downstream neighbor of its predecessor. A queueing network is called acyclic if it is impossible to find two queues i and j such that j lies downstream of i, and i lies downstream of j. Having an acyclic network is essential for the Poisson character of the arrival and departure processes at each queue to be maintained (see the next section). However, the product form (3.102) of the occupancy distribution generalizes in a natural way to networks that are not acyclic as we show in the next section.

3.8 NETWORKS OF QUEUES—JACKSON'S THEOREM

As discussed in section 3.6, the main difficulty with analysis of networks of transmission lines is that the packet interarrival times after traversing the first queue are correlated with their lengths. It turns out that if somehow this correlation were eliminated (which is the premise of the Kleinrock independence approximation) and randomization is used to divide traffic among different routes, then the average number of packets in the system can be derived as if each queue in the network was $M/M/1$. This is an important result known as Jackson's Theorem. In this section we will derive a simple version of this theorem.

Consider a network of K single server queues in which customers arrive from outside the network at each queue i in accordance with independent Poisson processes at rate r_i. Once a customer is served at queue i, it proceeds to join each queue j with probability P_{ij} or to exit the network with probability $1 - \sum_{j=1}^{K} P_{ij}$. The total customer arrival rate at queue j, denoted λ_j, satisfies

$$\lambda_j = r_j + \sum_{i=1}^{K} \lambda_i P_{ij}\,, \quad j = 1, \ldots, K \tag{3.104}$$

These equations represent a linear system in which the total rates λ_j, $j = 1, \ldots, K$, constitute a set of K unknowns. We assume that they can be solved uniquely to yield λ_j, $j = 1, \ldots, K$ in terms of r_j, P_{ij}, i, $j = 1, \ldots, K$. It can be shown that this is guaranteed under very general assumptions—for instance, if all the departure probabilities $\left(1 - \sum_{j=1}^{K} P_{ij}\right)$ are positive, $i = 1, \ldots, K$, or, more generally, if for every queue i_1, there is a queue i with $\left(1 - \sum_{j=1}^{K} P_{ij}\right) > 0$ and a sequence $i_1, i_2, \ldots, i_k, i$ such that $P_{i_1 i_2} > 0, \ldots, P_{i_k i} > 0$.

The service times of customers at the i^{th} queue are assumed exponentially distributed with mean $1/\mu_i$, and are assumed mutually independent and independent of the arrival process at the queue. The utilization factor of each queue is denoted

$$\rho_i = \frac{\lambda_i}{\mu_i}\,, \quad i = 1, \ldots, K\,, \tag{3.105}$$

and we assume $\rho_i < 1$ for all i.

In order to model a packet network such as the one considered in section 3.6 within the framework described above, it is necessary to accept several simplifying conditions in addition to assuming Poisson arrivals and exponentially distributed packet lengths. The first is the independence of packet lengths and interarrival times discussed earlier. The second is relevant to datagram networks, and has to do with the assumption that bifurcation of traffic at a network node can be modeled reasonably well by a randomization process whereby each departing packet from queue i joins queue j with probability P_{ij}—this need not be true, as discussed in section 3.6. Still a packet network differs from the model of this section because it involves several traffic streams which may have different routing probabilities at each node, and which maintain their identity as they travel along different routes. This difficulty can be partially addressed by using an extension of Jackson's Theorem that applies to a network with multiple classes of customers. Within this more general framework, we can model traffic streams corresponding to different origin-destination pairs as different classes of customers. If all traffic streams have the same average packet length, it turns out that Jackson's Theorem as stated below is valid assuming the simplifying conditions mentioned earlier.

Turning now to analysis, we view the system as a continuous-time Markov chain with states $n_1, n_2, \ldots, n_K$ where n_i denotes the number of customers at queue i. Let $P(n_1, \ldots, n_K)$ denote the stationary distribution of the chain. We have:

Jackson's Theorem. Assuming $\rho_i < 1$, $i = 1, \ldots, K$, we have for all $n_1, \ldots, n_K \geq 0$

$$P(n_1, \ldots, n_K) = P_1(n_1)P_2(n_2) \cdots P_K(n_K) \tag{3.106}$$

where

$$P_i(n) = \rho_i^n(1 - \rho_i), \quad n \geq 0 \tag{3.107}$$

Proof. We use a technique outlined in the previous section whereby we guess at the transition rates of the reversed process and verify that, together with the stationary probabilities given by Eqs. (3.106) and (3.107), they satisfy the conditions for such rates given by Eqs. (3.100) and (3.101) of the previous section. (The Markov chain is not time reversible here. Nonetheless, the use of the reversed process is both analytically convenient and conceptually useful.)

More specifically denote state vectors as

$$n = (n_1, n_2, \ldots, n_K)$$

and denote (cf. Eq. (3.106))

$$P(n) = P_1(n_1)P_2(n_2) \cdots P_K(n_K) \tag{3.108}$$

For any two state vectors n and n', let $q_{nn'}$ be the corresponding transition rate. Jackson's Theorem will be proved if we can find rates $q^*_{nn'}$ such that for all n, n'

$$P(n)q_{nn'} = P(n')q^*_{n'n} \tag{3.109}$$

and

$$\sum_m q_{nm} = \sum_m q^*_{nm} \tag{3.110}$$

For state vectors n and n' of the form

$$n = (n_1, \ldots, n_i, \ldots, n_K)$$
$$n' = (n_1, \ldots, n_i + 1, \ldots, n_K)$$

we have

$$q_{nn'} = r_i \tag{3.111}$$

$$q_{n'n} = \mu_i \left(1 - \textstyle\sum_j P_{ij}\right) \tag{3.112}$$

If we define

$$q^*_{nn'} = \lambda_i \left(1 - \textstyle\sum_j P_{ij}\right) \tag{3.113}$$

$$q^*_{n'n} = \frac{\mu_i r_i}{\lambda_i}, \tag{3.114}$$

we see that Eq. (3.109) is satisfied.

Next consider state vectors n and n' of the form

$$n = (n_1, \ldots, n_i, \ldots, n_j, \ldots, n_K)$$
$$n' = (n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_K)$$

We have

$$q_{nn'} = \mu_j P_{ji} \tag{3.115}$$

If we define

$$q^*_{n'n} = \frac{\mu_i \lambda_j P_{ji}}{\lambda_i} \tag{3.116}$$

we see again that Eq. (3.109) is satisfied.

Since for all other types of pairs of state vectors n, n', we have

$$q_{nn'} = 0 \tag{3.117}$$

we can define

$$q^*_{n'n} = 0 \tag{3.118}$$

and be assured that Eq. (3.109) holds for all n and n'. Finally, a straightforward calculation using Eqs. (3.111) through (3.118) and Eq. (3.104) verifies that Eq. (3.110) holds. **QED**

Note that the transition rates $q^*_{nn'}$ defined by Eqs. (3.113), (3.114), (3.116), and (3.118), are those of the reversed process. It can be seen that the reversed process corresponds to a network of queues where traffic arrives at queue i from outside the network according to a Poisson process with rate $\lambda_i \left(1 - \sum_j P_{ij}\right)$ (cf. Eq. (3.113)). The routing probability from queue i to queue j in the reversed process is $\lambda_j P_{ji} / \left(r_i + \sum_k \lambda_k P_{ki}\right)$ (cf. Eq. (3.114) and (3.116)).

There are several extensions of Jackson's theorem. For example, the product form (3.106) holds if each queue i has multiple servers, say m_i, rather than a single server. In that case, the formula corresponding to (3.107) is identical to the one of the $M/M/m_i$ system. Other extensions involve closed networks where there is a fixed number of customers circulating inside the network with no external arrivals or departures allowed (see Prob. 3.28).

We now turn to interpretation of Jackson's Theorem. First, from Eq. (3.106) we see that the numbers of customers at distinct queues at a given time are *independent.* The equation for the distribution of customers at each queue i is identical

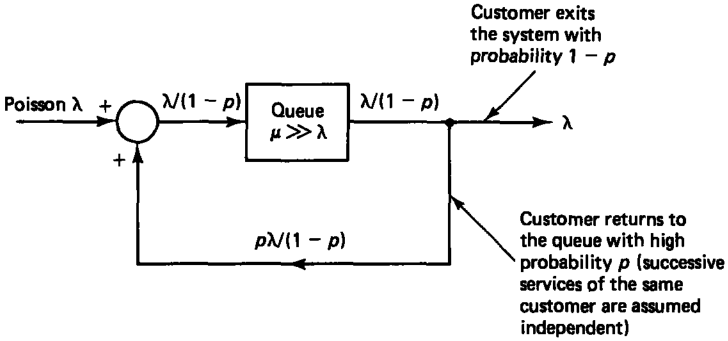

Customer exits the system with probability 1 − p
Poisson λ
+
+
λ/(1 − p)
Queue
μ ≫ λ
λ/(1 − p)
λ
pλ/(1 − p)
Customer returns to the queue with high probability p (successive services of the same customer are assumed independent)

Figure 3.25 Example of a queue within a network where the external arrival process is Poisson but the total arrival process at the queue is not Poisson. An external arrival is typically processed fast (since μ is much larger than λ), and with high probability returns to the queue through the feedback loop. As a result, the total queue arrival process typically consists of bursts of arrivals with each burst triggered by the arrival of a single customer from the outside.

as for an $M/M/1$ queue (compare Eq. (3.107) and the corresponding equations in section 3.3). This is a remarkable result particularly since one can show by example that the arrival process at each queue need *not* be a Poisson process. In fact, if there is a possibility that a customer may visit a queue more than once (a situation called feedback), the arrival process will not be Poisson. As an example (see Fig. 3.25), suppose that there is a single queue with a service rate which is very large relative to the arrival rate from the outside. Suppose also that with probability p near unity, a customer upon completion of service is fed back into the queue. Hence, when an arrival occurs at the queue, there is a large probability of another arrival at the queue in a short time (namely, the feedback arrival), whereas at an arbitrary time point, there will be only a very slight chance of an arrival occurring shortly since λ is small. In other words, queue arrivals tend to occur in bursts triggered by the arrival of a single customer from the outside. Hence, the queue arrival process does not have independent interarrival times and cannot be Poisson.

Unfortunately, our proof of Jackson's theorem is based on algebraic manipulation, and gives little insight as to why this remarkable result holds. For this reason we provide a heuristic explanation for the case of the feedback network of Fig. 3.25. This explanation can be generalized and made rigorous albeit at the expense of a great deal of technical complications (see [Wal83]).

Suppose that we introduce a delay Δ in the feedback loop of the single-queue

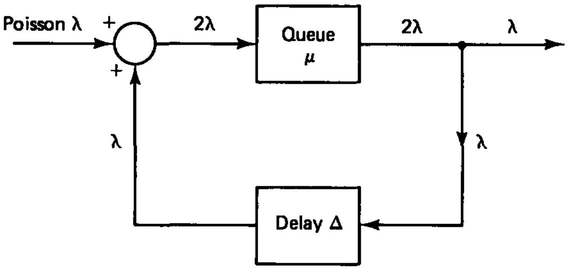


Figure 3.26 Heuristic explanation of Jackson's Theorem. Consider the introduction of an arbitrarily small positive delay Δ in the feedback loop of the network of Figure 3.25. An occupancy distribution of the queue that equals the $M/M/1$ equilibrium, and a content of the delay line that is an independent Δ segment of a Poisson process form an equilibrium distribution of the overall system. Therefore, the $M/M/1$ equilibrium distribution is an equilibrium for the queue as suggested by Jackson's Theorem even though the total arrival process to the queue is not Poisson.

network discussed above (see Fig. 3.26, where for convenience, we have chosen $p = 1/2$). Let us denote by $n(t)$ the number in the queue at time t, and by $f_\Delta(t)$ the content of the delay line at time t. The interpretation here is that $f_\Delta(t)$ is a function of time that specifies the customer output of the delay line in the subsequent Δ interval $(t, t+\Delta]$. Suppose that the initial distribution $n(0)$ of the queue state at time 0, is equal to the steady state distribution of an $M/M/1$ queue, *i.e.*,

$$P\{n(0) = n\} = \rho^n(1-\rho) \tag{3.119}$$

where $\rho = 2\lambda/\mu$ is the utilization factor. Suppose also that $f_\Delta(0)$ is a portion of a Poisson arrival process with rate λ. The customers in $f_\Delta(0)$ have service times that are independent, exponentially distributed with parameter μ. We assume that $n(0)$ and $f_\Delta(0)$ are independent. Then, the input to the queue over the interval $[0, \Delta)$ will be the sum of two independent Poisson streams which are independent of the number in queue at time 0. It follows that the queue will behave in the interval $[0, \Delta)$ like an $M/M/1$ queue in equilibrium. Therefore, $n(\Delta)$ will be distributed according to the $M/M/1$ steady-state distribution of Eq. (3.119), and by part (b) of Burke's theorem, $n(\Delta)$ will be independent of the departure process from the queue in the interval $[0, \Delta)$, or, equivalently, of $f_\Delta(\Delta)$—the delay line content at time Δ. Furthermore, by part (a) of Burke's Theorem, $f_\Delta(\Delta)$ will be Poisson. Thus, to summarize, we started out with independent initial conditions $n(0)$ and $f_\Delta(0)$ which had the equilibrium distribution of an $M/M/1$ queue and the statistics of a Poisson process, respectively, and Δ seconds later we obtained corresponding quantities $n(\Delta)$ and $f_\Delta(\Delta)$ with the same properties. Using the same reasoning, we can show that for all t, $n(t)$ and $f_\Delta(t)$ have the same properties. It follows that the $M/M/1$ steady-state distribution of Eq. (3.119) is an equilibrium distribution for the queueing system for an arbitrary positive value of the feedback delay Δ, and this strongly suggests the validity of Jackson's Theorem. Note that this argument does not suggest that the feedback process and, therefore, also the total arrival process to the queue are Poisson. Indeed, it can be seen that successive Δ portions of the feedback arrival stream are correlated since, with probability 0.5, a departing customer from the queue appears as an arrival Δ seconds later. Therefore, over the interval $[0, \infty)$, the feedback process is not Poisson. This is consistent with our earlier observations regarding the example of Fig. 3.25.

3.9 SUMMARY

Queueing models provide qualitative insights on the performance of data networks, and quantitative predictions of average packet delay. An example of the former is the comparison of time-division and statistical multiplexing, while an example of the latter is the delay analysis of reservation systems.

To obtain tractable queueing models for data networks, it is frequently necessary to make simplifying assumptions. A prime example is the Kleinrock independence approximation discussed in section 3.6. Delay predictions based on this approximation are adequate for many uses. A more accurate alternative is simulation which, however, can be slow, expensive, and lacking in insight.

Little's Theorem is a simple but extremely useful result since it holds under very general conditions. To proceed beyond this theorem we assumed Poisson arrivals and independent interarrival and service times. This led to the $M/G/1$ system, and its extensions in reservation and priority queueing systems. We analyzed a surprisingly large number of important delay models using simple graphical arguments. An alternative analysis was based on the use of Markov chain models and led to the derivation of the occupancy probability distribution of the $M/M/1$, $M/M/m$, and related systems.

Reversibility is an important notion that helps to prove and understand Jackson's Theorem, and provides a taste of advanced queueing topics.

3.10 NOTES, SOURCES, AND SUGGESTED READING

Section 3.2. Little's Theorem was formalized in [Lit61]. Rigorous proofs under various assumptions are given in [Sti72] and [Sti74]. Several applications in finding performance bounds of computer systems are described in [StA85].

Section 3.3. For a general background on the Poisson process, Markov chains, and related topics, see [Ros80], [Ros83], and [KaT75]. Standard texts on queueing theory include [Coo81], [GrH85], [HeS82], and [Kle75]. A reference for the fact that Poisson arrivals see a typical occupancy distribution (subsection 3.3.2) is [Wol82a].

Section 3.4. Queueing systems that admit analysis via Markov chain theory include those where the service times have an Erlang distribution; see [Kle76], Chap. 4. For extensions to more general models and computational methods, see [Kei79], [Neu81], [Haj82], and [Twe82].

Section 3.5. The P-K formula is often derived by using z-transforms; see [Kle75]. This derivation is not very insightful, but gives the probability distribution of the system occupancy (not just the mean that we obtained via our much simpler analysis). For more on delay analysis of ARQ systems see [AnP86] and [ToW79].

The results on polling and reservation systems are of recent origin; see [Coo70], [Eis79], [FeA85], [FuC85], [IEE86], and [Kue79]. The original references that are closest to our analysis are [Has72] for unlimited service systems, [NoT78] for limited service systems, and [Hum78] for nonsymmetric polling systems. Reference [Tak86] is a monograph devoted to polling. There are two main reservation and polling systems considered in the literature: the symmetric case, where all users have identical arrival and reservation interval statistics, and the nonsymmetric case, where these statistics are user dependent. The former case admits simple expressions for the mean waiting times while the latter does not. We have considered the partially symmetric case, where all users have identical arrival statistics but different reservation interval statistics. The fact that simple expressions hold for this case has not been known earlier and, in this respect, our formulas are original. Our treatment in terms of simple graphical arguments is also original. The result of Prob. 3.25 on limited service systems with shared reservation and data intervals is new.

An extensive treatment of priority queueing systems is [Jai68]. A simpler, less comprehensive exposition is given in [Kle75].

Section 3.6. Delay analysis for data networks in terms of $M/M/1$ approximations was introduced in [Kle64]. References [Wol82b] and [PiW82] study via analysis and simulation the behavior of two queues in tandem when the service times of a customer at the two queues are dependent. The special issues [IEE86] provides a view of recent work on the subject.

Section 3.7. The notion of reversibility in queueing networks is explored in depth in [Kel79].

Section 3.8. There is an extensive literature on product form solutions of queueing networks following Jackson's original paper [Jac57]. The survey [DiK85] lists 314 references. Two books on the subject are [Kel79] and [BrB80]. The heuristic explanation of Jackson's theorem is due to [Wal83].

PROBLEMS

3.1 Customers arrive at a fast-food restaurant at a rate of five per minute, and wait to receive their order for an average of five minutes. Customers eat in the restaurant with probability 0.5, and carry out their order without eating with probability 0.5. A meal requires an average of 20 minutes. What is the average number of customers in the restaurant?

3.2 An absent-minded professor schedules two student appointments for the same time. The appointment durations are independent and exponentially distributed with mean thirty minutes. The first student arrives on time, but the second student arrives five minutes late. What is the expected time between the arrival of the first student and the departure of the second student?

3.3 A person enters a bank and finds all of the four clerks busy serving customers. There are no other customers in the bank, so the person will start service as soon as one of the customers in service leaves. Customers have independent, identical, exponential distribution of service time.

(a) What is the probability that the person will be the last to leave the bank assuming no other customers arrive?

(b) If the average service time is one minute, what is the average time the person will spend in the bank?

(c) Will the answer in (a) change if there are some additional customers waiting in a common queue, and customers begin service in the order of their arrival?

3.4 Consider a packet stream whereby packets arrive according to a Poisson process with rate 10 packets/sec. If the interarrival time between any two successive packets is less than the transmission time of the first, the two packets are said to collide. (This notion will be made more meaningful in Chapter 4 when we talk about multiaccess schemes.) Find the probability that a packet collides with either its predecessor or its successor assuming:

(a) All packets have a transmission time of 20 msec.

(b) Packets have independent, exponentially distributed transmission times with mean 20 msec.

3.5 A communication line capable of transmitting at a rate of 50 Kbits/sec will be used to accommodate 10 sessions each generating Poisson traffic at a rate 150 packets/min. Packet lengths are exponentially distributed with mean 1000 bits.

(a) For each session, find the average number of packets in queue, the average number in the system, and the average delay per packet when the line is allocated to the sessions by using:

(1) 10 equal capacity time-division multiplexed channels

(2) statistical multiplexing.

(b) Repeat (a) for the case where 5 of the sessions transmit at a rate of 250 packets/min while the other 5 at a rate 50 packets/min.

3.6 Repeat part (a) of Prob. 3.5 for the case where packet lengths are not exponentially distributed, but 10% of the packets are 100 bits long and the rest are 1500 bits long. Repeat the problem for the case where the short packets are given nonpreemptive priority over the long packets.

3.7 A communication line is divided in two identical channels each of which will serve a packet traffic stream where all packets have equal transmission time T, and equal interarrival time $R > T$. Consider, alternatively, statistical multiplexing of the two traffic streams by combining the two channels into a single channel with transmission time $T/2$ for each packet. Show that the average system time of a packet will be decreased from T to something between $T/2$ and $3T/4$, while the variance of waiting time in queue will be increased from 0 to as much as $T^2/16$.

3.8 Persons arrive at a taxi stand with room for five taxis according to a Poisson process with rate one per minute. A person boards a taxi upon arrival if one is available and otherwise waits in a line. Taxis arrive at the stand according to a Poisson process with rate two per minute. An arriving taxi that finds the stand full departs immediately; otherwise, it picks up a customer if at least one is waiting, or else joins the queue of waiting taxis. What is the steady-state probability distribution of the taxi queue size?

3.9 A communication node A receives Poisson packet traffic from two other nodes 1 and 2 at rates λ_1 and λ_2, respectively, and transmits it on a link with capacity C bits/sec. The two input streams are assumed independent and their packet lengths are identically and exponentially distributed with mean L bits. A packet from node 1 is always accepted by A. A packet from node 2 is accepted only if the number of packets in A (in queue or under transmission) is less than a given number $K > 0$; otherwise, it is assumed lost.

(a) What is the range of values of λ_1 and λ_2 for which the expected number of packets in A will stay bounded as time increases?

(b) For λ_1 and λ_2 in the range of part (a) find the steady-state probability of having n packets in $A\,(0 \leq n < \infty)$. Find the average time needed by a packet from source 1 to clear A once it enters A, and the average number of packets in A from source 1. Repeat for packets from source 2.

3.10 (a) Derive Eqs. (3.11) to (3.14).

(b) Show that if the arrivals in two disjoint time intervals are independent and Poisson distributed with parameters $\lambda\tau_1$, $\lambda\tau_2$, then the number of arrivals in the union of the intervals is Poisson distributed with parameter $\lambda(\tau_1 + \tau_2)$.

(c) Show that if k independent Poisson processes $A_1, \ldots, A_k$ are combined into a single process $A = A_1 + A_2 + \cdots + A_k$, then A is Poisson with rate λ equal to the sum of the rates $\lambda_1, \ldots \lambda_k$ of $A_1, \ldots, A_k$. Show also that the probability that the first arrival of the combined process comes from A_1 is λ_1/λ independently of the time of arrival.

(d) Suppose we know that in an interval $[t_1, t_2]$ only one arrival of a Poisson process has occurred. Show that, conditional on this knowledge, the time of this arrival is uniformly distributed in $[t_1, t_2]$.

3.11 Packets arrive at a transmission facility according to a Poisson process with rate λ. Each packet is independently routed with probability p to one of two transmission lines and with probability $(1 - p)$ to the other. Show that the arrival processes at the two transmission lines are Poisson with rates λp and $\lambda(1 - p)$ respectively. Furthermore the two processes are independent.

3.12 Consider a system that is identical to $M/M/1$ except that when the system empties out, service does not begin again until k customers are present in the system (k is given). Once service begins it proceeds normally until the system becomes empty again. Find the steady state probabilities of the number in the system, the average number in the system, and the average delay per customer.

3.13 A telephone company establishes a direct connection between two cities expecting Poisson traffic with rate 30 calls/min. The durations of calls are independent and exponentially distributed with mean three minutes. Interarrival times are independent of call durations. How many circuits should the company provide to ensure that an attempted call is blocked (because all circuits are busy) with probability less than 0.01? It is assumed that blocked calls are lost, *i.e.*, a blocked call is not attempted again.

3.14 Consider an $M/M/\infty$ queue with servers numbered $1, 2, \ldots$ There is an additional restriction that upon arrival a customer will choose the lowest numbered server that is idle at the time. Find the fraction of time that each server is busy. Will the answer change if the number of servers is finite? *Hint*: Argue that in steady-state the probability that all of the first m servers are busy is given by the Erlang B formula of the $M/M/m/m$ system. Find the total arrival rate to servers $(m + 1)$ and higher, and from this the arrival rate to each server.

3.15 In the $M/G/1$ system, show that

$$P\{\text{the system is empty}\} = 1 - \lambda\overline{X}$$
$$\text{Average length of time between busy periods } = 1/\lambda$$
$$\text{Average length of busy period } = \frac{\overline{X}}{1 - \lambda\overline{X}}$$
$$\text{Average number of customers served in a busy period } = \frac{1}{1 - \lambda\overline{X}}$$

Consider the following argument: When a customer arrives the probability that another customer is being served is $\lambda\overline{X}$. Since the served customer has mean service time $\overline{X}$, the average time to complete the service is $\overline{X}/2$. Therefore the mean residual service time is $\lambda\overline{X}^2/2$. What is wrong with this argument?

3.16 *$M/G/1$ System with Arbitrary Order of Service.* Consider the $M/G/1$ system with the difference that customers are not served in the order they arrive. Instead, upon completion of a customer's service, one of the waiting customers in queue is chosen according to some rule, and is served next. Show that the P-K formula for the average waiting time in queue W remains valid provided the relative order of arrival of the customer chosen is independent of the service times of the customers waiting in queue. *Hint*: Argue that the independence hypothesis above implies that, at any time t, the number $N_Q(t)$ of customers waiting in queue is independent of the service times of these customers. Show that this in turn implies that $U = R + \rho W$ where R is the mean residual time, and U is the average steady-state unfinished work in the system (total remaining service time of the customers in the system). Argue that U and R are independent of the order of customer service.

3.17 *Priority Systems with Multiple Servers.* Consider the systems of subsection 3.5.3 where all priority classes have exponentially distributed service times with common mean $1/\mu$. Assume that there are m servers.

(a) Consider the nonpreemptive system. Show that Eq. (3.79) yields the average queueing times with the mean residual time R given by

$$R = \frac{P_Q}{m\mu}$$

where P_Q is the steady-state probability of queueing given by the Erlang C formula of Eq. (3.36). (Here $\rho_i = \lambda_i/(m\mu)$ and $\rho = \sum_{i=1}^{n} \rho_i$.)

(b) Consider the preemptive resume system. Write a formula for $W_{(k)}$—the average time in queue averaged over the first k priority classes. Use Little's Theorem to show that the average time in queue of a k^{th} priority class customer can be obtained recursively from

$$W_1 = W_{(1)}$$

$$W_k = \frac{1}{\lambda_k}\left[W_{(k)}\sum_{i=1}^{k}\lambda_i - W_{(k-1)}\sum_{i=1}^{k-1}\lambda_i\right], \quad k = 2, 3, \ldots, n$$

3.18 Consider the nonpreemptive priority queueing system of subsection 3.5.3 for the case where the available capacity is sufficient to handle the highest priority traffic, but cannot handle the traffic of all priorities, *i.e.*,

$$\rho_1 < 1 < \rho_1 + \rho_2 + \cdots + \rho_n$$

Find the average delay per customer of each priority class. *Hint*: Determine the departure rate of the highest priority class that will experience infinite average delay, and the mean residual service time.

3.19 Consider an n-class, nonpreemptive priority system:

(a) Show that the sum $\sum_{k=1}^{n}\rho_k W_k$ is independent of the priority order of classes, and in fact

$$\sum_{k=1}^{n}\rho_k W_k = \frac{R\rho}{1-\rho}$$

where $\rho = \rho_1 + \rho_2 + \cdots + \rho_n$. (This is known as the $M/G/1$ conservation law [Kle64].) *Hint*: Use Eq. (3.79). Alternatively argue that $U = R + \sum_{k=1}^{n}\rho_k W_k$, where U is the average steady-state unfinished work in the system (total remaining service time of customers in the system), and U and R are independent of the priority order of the classes.

(b) Suppose there is a cost c_k per unit time for each class k customer that waits in queue. Show that cost is minimized when classes are ordered so that

$$\frac{\overline{X_1}}{c_1} \le \frac{\overline{X_2}}{c_2} \le \cdots \le \frac{\overline{X_n}}{c_n}$$

Hint: Express the cost as $\sum_{k=1}^{n}(c_k/\overline{X}_k)(\rho_k W_k)$ and use part (a). Use also the fact that interchanging the order of any two adjacent classes leaves the waiting time of all other classes unchanged.

3.20 *$M/M/1$ Shared Service System.* Consider a system which is the same as $M/M/1$ except that whenever there are n customers in the system they are all served simultaneously at an equal rate $1/n$ per unit time. Argue that the steady-state occupancy distribution is the same as for the $M/M/1$ system. *Note*: It can be shown that the steady-state occupancy distribution is the same as for $M/M/1$ even if the service time distribution is not exponential, *i.e.*, for an $M/G/1$ type of system ([Ros83], p. 171).

3.21 In Ex. 10 of subsection 3.4.2, verify the formula $\sigma_f = (\lambda/\mu)^{1/2} s_\gamma$. *Hint*: Write

$$E\{f^2\} = E\{(\textstyle\sum_{i=1}^{n}\gamma_i)^2\} = E\{E\{(\textstyle\sum_{i=1}^{n}\gamma_i)^2\,|n\}\},$$

and use the fact that n is Poisson distributed.

3.22 Show that Eq. (3.59) for the average delay of time-division multiplexing on a slot basis can be obtained as a special case of the results for the limited service reservation system. Generalize the expression (3.59) for the case where slot lengths are random and independent, and the traffic streams do not have equal rates and identical slot length distributions. *Hint*: Consider the gated system with zero packet length.

3.23 Show that the expected cycle lengths in the single-user and multiuser reservation systems are $\overline{V}/(1-\rho)$ and $m\overline{V}/(1-\rho)$, respectively. *Hint*: If L_k is the length of the k^{th} cycle show that $E\{L_{k+1}|L_k\} = \overline{V} + \rho L_k$ for the single-user case.

3.24 Consider the limited service reservation system. Show that for both the gated and the partially gated versions:

(a) The steady-state probability of arrival of a packet during a reservation interval is $1-\rho$.

(b) The steady-state probability of a reservation interval being followed by an empty data interval is $(1-\rho-\lambda\overline{V})/(1-\rho)$. *Hint*: If p is the required probability, argue that the ratio of the times used for data intervals and for reservation intervals is $(1-p)\overline{X}/\overline{V}$.

3.25 *Limited Service Reservation System with Shared Reservation and Data Intervals.* Consider the gated version of the limited service reservation system with the difference that the m users share reservation and data intervals, *i.e.*, all users make reservations in the same interval and transmit at most one packet each in the subsequent data interval. Show that

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho-\lambda\overline{V}/m)} + \frac{(1-\rho)\overline{V^2}}{2(1-\rho-\lambda\overline{V}/m)\overline{V}} + \frac{(1-\rho\alpha-\lambda\overline{V}/m)\overline{V}}{1-\rho-\lambda\overline{V}/m}$$

where $\overline{V}$ and $\overline{V^2}$ are the first two moments of the reservation interval, and α satisfies

$$\frac{\overline{K}+(\hat{K}-1)(2\overline{K}-\hat{K})}{2m\overline{K}} - \frac{1}{2m} \le \alpha \le \frac{1}{2} - \frac{1}{2m}$$

where

$$\overline{K} = \frac{\lambda \overline{V}}{1-\rho}$$

is the average number of packets per data interval, and $\hat{K}$ is the smallest integer which is larger than $\overline{K}$. Verify that the formula for W becomes exact as $\rho \to 0$ (light load), and as $\rho \to 1 - \lambda\overline{V}/m$ (heavy load). *Hint:* Verify that

$$W = R + \lambda W + \left(1 + \frac{\lambda W}{m} - S\right)\overline{V}$$

where $S = \lim_{i\to\infty} E\{S_i\}$ and S_i is the number (0 or 1) of packets of the owner of packet i that will start transmission between the time of arrival of packet i and the end of the cycle in which packet i arrives. Try to obtain bounds for S by considering separately the cases where packet i arrives in a reservation and in a data interval.

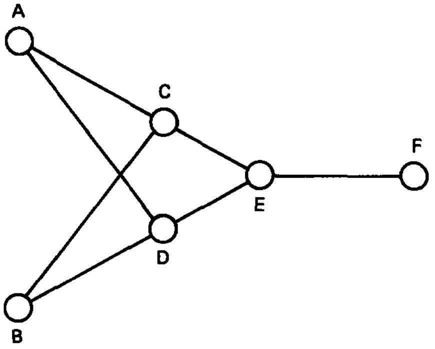


Figure 3.27

3.26 Consider the network in Fig. 3.27. There are four sessions: ACE, ADE, BCEF, and BDEF sending Poisson traffic at rates 100, 200, 500, and 600 packets/min., respectively. Packet lengths are exponentially distributed with mean 1000 bits. All transmission lines have capacity 50 kbits/sec, and there is a propagation delay of 2 msec on each line. Using the Kleinrock independence approximation, find the average number of packets in the system, the average delay per packet (regardless of session), and the average delay per packet of each session.

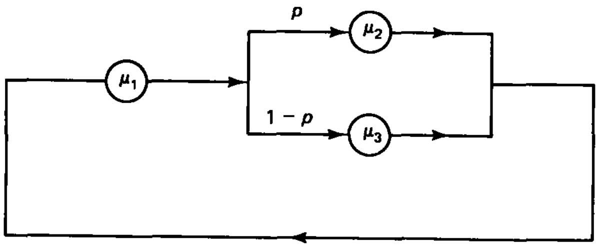


Figure 3.28

3.27 *Bounds on the Throughput of a Closed Queueing Network.* Packets enter the network of transmission lines shown in Fig. 3.28 at point A and exit at point B. A packet is first transmitted on one of the lines $L_1, \ldots, L_K$, where it requires on the average a transmission time $\overline{X}$, and is then transmitted in line L_{K+1}, where it requires on the average a transmission time $\overline{Y}$. To effect flow control, a maximum of $N \geq K$ packets are admitted into the system. Each time a packet exits the system at point B, an acknowledgement is sent back and reaches point A after a fixed time $\overline{Z}$. At that time, a new packet is allowed to enter the system. Use Little's Theorem to find upper and lower bounds for the system throughput under two circumstances:

(a) The method of routing a packet to one of the lines $L_1, \ldots, L_K$ is unspecified.

(b) The routing method is such that whenever one of the lines $L_1, \ldots, L_K$ is idle, there is no packet waiting at any of the other lines.

3.28 *Analysis of a Closed Queueing Network.* Certain systems are best modelled by *closed* queueing networks, where the number of customers in the system is fixed. Consider a closed queueing network with 3 customers shown in Fig. 3.29 together with the probabilities that a departing customer from a queue enters another.

(a) How would you define the state of this system? Draw the state transition diagram and label the states and transition rates on your figure.

(b) Show that the system is reversible and that occupancy probabilities have the product form

$$P(n_1, n_2, n_3) = \alpha \left(\frac{1}{\mu_1}\right)^{n_1} \left(\frac{p}{\mu_2}\right)^{n_2} \left(\frac{1-p}{\mu_3}\right)^{n_3}$$

where $n_1 + n_2 + n_3 = 3$, and α is a normalizing constant.

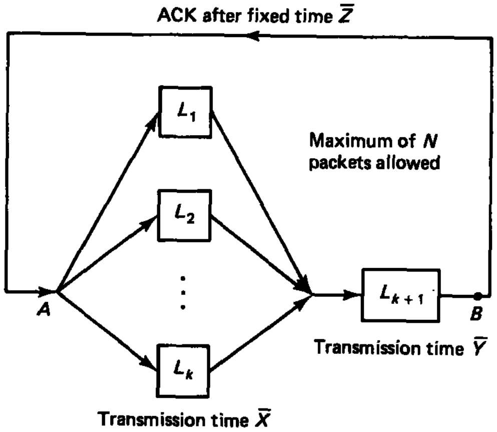

ACK after fixed time $\bar{Z}$
L_1
L_2
L_k
Maximum of N packets allowed
L_{k+1}
A
B
Transmission time $\bar{Y}$
Transmission time $\bar{X}$

Figure 3.29

3.29 Consider two queues with independent Poisson arrivals and independent exponentially distributed service times. The arrival and service rates are denoted λ_i, μ_i, for $i = 1, 2$, respectively. The two queues share a waiting room with finite capacity B (including customers in service). Arriving customers that find the waiting room full are lost. Show that for $m + n \leq B$, the steady-state probabilities are

$$P\{m \text{ in queue } 1, \ n \text{ in queue } 2\} = c\rho_1^m \rho_2^n$$

where $\rho_i = \lambda_i/\mu_i$, $i = 1, 2$, and c is a normalizing constant.

3.30 Consider the $M/M/1/m$ system which is the same as $M/M/1$ except that there can be no more than m customers in the system and customers arriving when the system is full are lost.

(a) Derive the steady-state occupancy probabilities.

(b) Show that the system is reversible, and characterize the departure process.

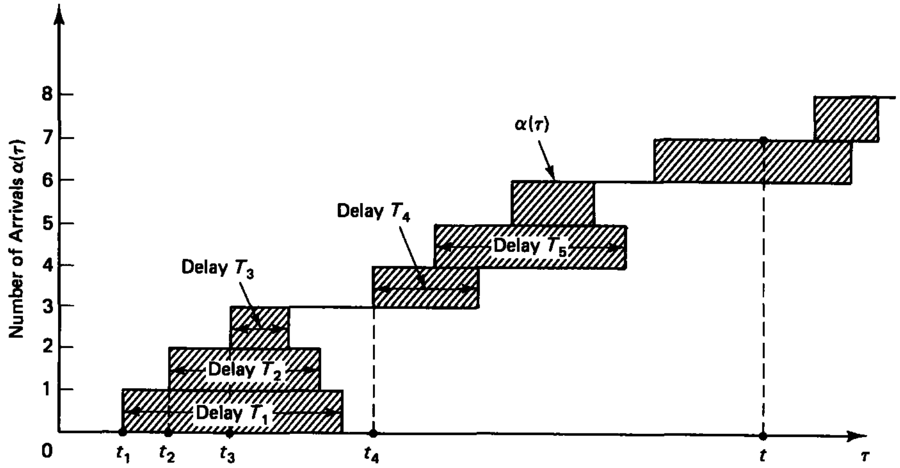


Figure 3.30

3.31 *Little's Theorem for Arbitrary Order of Service.* Use Fig. 3.30 to derive Little's Theorem for systems where the order of customer service is arbitrary, including cases where servers are shared by several customers, and customer service can be interrupted to serve customers of higher priority. In this figure t_i is the arrival time of the i^{th} customer, and T_i is the customer's system time, *i.e.*, the time between the customer's arrival and departure from the system. *Hint*: Calculate the shaded area up to time t in two different ways.

3.32 *Little's Theorem for Arbitrary Order of Service; Analytical Proof.* Consider the analysis of Little's theorem in section 3.2 and the notation introduced there. Assume that the time average arrival and departure rates exist and are equal

$$\lambda = \lim_{k\to\infty} \alpha(t)/t = \lim_{t\to\infty} \beta(t)/t$$

and that the limit defining the time average system time

$$T = \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} T_i$$

exists. Show that, regardless of the order customers are served, Little's Theorem ($N = \lambda T$) holds with

$$N = \lim_{t\to\infty} \frac{1}{t}\int_0^t N(\tau)d\tau$$

Hint: Show that for all t, we have

$$\sum_{i=1}^{\beta(t)} T_i \leq \int_0^t N(\tau)d\tau \leq \sum_{i=1}^{\alpha(t)} T_i$$

3.33 *A Generalization of Little's Theorem.* Consider an arrival/departure system with arrival rate λ, where entering customers are forced to pay money to the system according to some rule.

(a) Argue that the following identity holds:

$$\text{Average rate at which the system earns} = \lambda \cdot (\text{Average amount a customer pays})$$

(b) Show that Little's Theorem is a special case.

(c) Consider the $M/G/1$ system and the following cost rule: Each customer pays at a rate of y per unit time when its remaining service time is y, whether in queue or in service. Show that the formula in (a) can be written as

$$W = \lambda \left(\overline{X}W + \frac{\overline{X^2}}{2} \right)$$

which is the Pollaczek–Khinchin formula.

3.34 *$M/G/1$ Queue with Random Sized Batch Arrivals.* Consider the $M/G/1$ system with the difference that customers are arriving in batches according to a Poisson process with rate λ. Each batch has n customers, where n has a given distribution and is independent of customer service times. Adapt the proof of section 3.5 to show that the waiting time in queue is given by

$$W = \frac{\lambda \overline{n}\overline{X^2}}{2(1-\rho)} + \frac{\overline{X}(\overline{n^2} - \overline{n})}{2\overline{n}(1-\rho)}$$

Hint: Use the equation $W = R + \rho W + W_B$ where W_B is the average waiting time of a customer for other customers that arrived in the same batch.

3.35 *$M/G/1$ Queue with Overhead for Each Busy Period.* Consider the $M/G/1$ queue with the difference that the service of the first customer in each busy period requires an increment Δ over the ordinary service time of the customer. We assume that Δ has a given distribution and is independent of all other random variables in the model. Let $\rho = \lambda\overline{X}$ be the utilization factor. Show that

(a) $P_0 = P\{$ the system is empty$\} = (1-\rho)/(1+\lambda\overline{\Delta})$.

(b) Average length of busy period= $(\overline{X}+\overline{\Delta})/(1-\rho)$

(c) The average waiting time in queue is

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{\lambda[\overline{(X+\Delta)^2} - \overline{X^2}]}{2(1+\lambda\overline{\Delta})}$$

3.36 *Single Vacation $M/G/1$ System.* Consider the $M/G/1$ system with the difference that each busy period is followed by a single vacation interval. Once this vacation is over, an arriving customer to an empty system starts service immediately. Assume that vacation intervals are independent, indentically distributed, and independent of the customer interarrival and service times. Derive the average waiting time in queue.

3.37 *The $M/G/\infty$ System.* Consider a queueing system with Poisson arrivals at rate λ. There are an infinite number of servers, so that each arrival starts service at an idle server immediately on arrival. Each server has a general service time distribution and $F_X(x) = P\{X \leq x\}$ denotes the probability that a service starting at any given time τ is completed by time $\tau + x$ ($F_X(x) = 0$ for $x \leq 0$). The servers have independent and identical service time distributions.

(a) For x and δ $(0 < \delta < x)$ very small, find the probability that there was an arrival in the interval $[\tau - x,\ \tau - x + \delta]$ *and* that this arrival is still being served at time τ.

(b) Show that the mean service time for any arrival is given by

$$\overline{X} = \int_0^\infty [1 - F_X(x)]\, dx$$

Hint: Use a graphical argument or integration by parts.

(c) Use (a) and (b) to verify that the number in the system is Poisson distributed with mean $\lambda\overline{X}$.

APPENDIX A: *Review of Markov Chain Theory*

The purpose of this appendix is to provide a brief summary of the results we need from discrete- and continuous-time Markov chain theory. We refer the reader to books on stochastic processes for detailed accounts.

3A.1 Discrete-Time Markov Chains

Consider a discrete-time stochastic process $\{X_n | n = 0, 1, 2, \ldots\}$ that takes values from the set of nonnegative integers, so the states that the process can be in are $i = 0, 1 \ldots$. The process is said to be a *Markov chain* if whenever it is in state i, there is a fixed probability P_{ij} that it will next be in state j regardless of the process history prior to arriving at i. That is, for all $n > 0$, $i_{n-1}, \ldots, i_0, i, j$

$$\begin{aligned} P_{ij} &= P\{X_{n+1} = j | X_n = i,\ X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} \\ &= P\{X_{n+1} = j | X_n = i\} \end{aligned}$$

We refer to P_{ij} as the *transition probabilities*. They must satisfy

$$P_{ij} \geq 0\,, \ \sum_{j=0}^{\infty} P_{ij} = 1\,, \quad i = 0, 1, \ldots$$

The corresponding transition probability matrix is denoted

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \cdot & \cdot & \cdot & \cdot \quad \cdot \quad \cdot \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \cdot & \cdot & \cdot & \cdot \quad \cdot \quad \cdot \end{bmatrix}$$

Consider the n-step transition probabilities

$$P_{ij}^n = P\{X_{n+m} = j | X_m = i\}\,, \quad n \geq 0, i, j \geq 0\,.$$

The *Chapman-Kolmogorov equations* provide a method for calculating P_{ij}^n. They are given by

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m\,, \quad n, m \geq 0, i, j \geq 0$$

From these equations, we see that P_{ij}^n are the elements of the matrix P^n (the transition probability matrix P raised to the n^{th} power).

We say that two states i and j *communicate* if for some n and n', we have $P_{ij}^n > 0$, $P_{ji}^{n'} > 0$. If all states communicate, we say that the Markov chain is *irreducible*. We say that the Markov chain is *aperiodic* if for each state i there is no integer $d \geq 2$ such that $P_{ii}^n = 0$ except when n is a multiple of d. A probability distribution $\{p_j \mid j \geq 0\}$ is said to be a *stationary distribution* for the Markov chain if

$$p_j = \sum_{i=0}^{\infty} p_i P_{ij}, \quad j \geq 0. \tag{3A.1}$$

We will restrict attention to irreducible and aperiodic Markov chains, since this is the only type we will encounter. For such a chain, denote

$$p_j = \lim_{n \to \infty} P_{jj}^n, \quad j \geq 0$$

It can be shown that the limit above exists and when $p_j > 0$, then $1/p_j$ equals the *mean recurrence time of* j, *i.e.*, the expected number of transitions between two successive visits to state j. If $p_j = 0$, the mean recurrence time is infinite. Another interpretation is that p_j represents the proportion of time the process visits j on the average. The following result will be of primary interest:

Theorem. In an irreducible, aperiodic Markov chain, there are two possibilities:

1. $p_j = 0$ for all $j \geq 0$ in which case the chain has no stationary distribution.
2. $p_j > 0$ for all $j \geq 0$ in which case $\{p_j \mid j \geq 0\}$ is the unique stationary distribution of the chain.

A typical example of case 1 above is an $M/M/1$ queueing system where the arrival rate λ exceeds the service rate μ.

In case 2, there arises the issue of characterizing the stationary distribution $\{p_j \mid j \geq 0\}$. For queueing systems, the following technique is often useful. Multiplying the equation $P_{jj} + \sum_{\substack{i=0 \\ i \neq j}}^{\infty} P_{ji} = 1$ by p_j and using Eq. (3A.1), we have

$$p_j \sum_{\substack{i=0 \\ i \neq j}}^{\infty} P_{ji} = \sum_{\substack{i=0 \\ i \neq j}}^{\infty} p_i P_{ij} \tag{3A.2}$$

These equations are known as the *global balance equations*. They state that, at equilibrium, the probability of a transition out of j (left side of Eq. (3A.2)) equals the probability of a transition into j (right side of Eq. (3A.2)).

The global balance equations can be generalized to apply to an entire set of states. Consider a subset of states S. By adding Eq. (3A.2) over all $j \epsilon S$, we obtain

$$\sum_{j \epsilon S} p_j \sum_{i \notin S} P_{ji} = \sum_{i \notin S} p_i \sum_{j \epsilon S} P_{ij} \tag{3A.3}$$

which means that *the probability of a transition out of the set of states S equals the probability of a transition into S.*

An intuitive explanation of these equations is based on the fact that when the Markov chain is irreducible, the state (with probability one) will return to the set S infinitely many times. Therefore, for each transition out of S there must be (with probability one) a reverse transition into S at some later time. As a result, the proportion of transitions out of S (over all transitions) equals the proportion of transitions into S. This is precisely the meaning of the global balance equations (3A.3).

3A.2 Detailed Balance Equations

As an application of the global balance equations, consider a Markov chain typical of queueing systems and, more generally, birth-death systems where two successive states can only differ by unity as in Fig. 3A.1. We assume that $P_{i,i+1} > 0$ and $P_{i+1,i} > 0$ for all i. This is a necessary and sufficient condition for the chain to be irreducible. Consider the sets of states

$$S = \{0, 1, \ldots, n\}$$

Application of Eq. (3A.3) yields

$$p_n P_{n,n+1} = p_{n+1} P_{n+1,n}\,, \quad n = 0, 1, \ldots \tag{3A.4}$$

i.e., in steady state, the probability of a transition from n to $n+1$ equals the probability of a transition from $n+1$ to n. These equations can be very useful in computing the stationary distribution $\{p_j | j \geq 0\}$ (see sections 3.3 and 3.4).

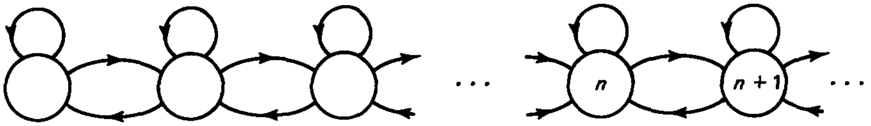


Figure 3A.1 Transition probability diagram for a birth-death process.

Equation (3A.4) is a special case of the equations

$$p_j P_{ji} = p_i P_{ij}\,, \quad i, j \geq 0 \tag{3A.5}$$

known as the ***detailed balance equations***. These equations need not hold in any given Markov chain. However, in many important special cases, they do hold and greatly simplify the calculation of the stationary distribution. A common method of verifying the validity of the detailed balance equations for a given irreducible, aperiodic Markov chain is to hypothesize their validity and try to solve them for the steady-state probabilities p_j, $j \geq 0$. There are two possibilities; either the system (3A.5) together with $\sum_j p_j = 1$ is inconsistent or else a distribution $\{p_j | j \geq 0\}$ satisfying Eq. (3A.5) will be found. In the latter case, this distribution will clearly also satisfy the global balance equations (3A.2). These equations are equivalent to the condition

$$p_j = \sum_{i=0}^{\infty} p_i P_{ij}, \quad j = 0, 1, \ldots$$

so, by the theorem given earlier, $\{p_j | j \geq 0\}$ is the unique stationary distribution.

3A.3 Partial Balance Equations

Some Markov chains have the property that their stationary distribution $\{p_j | j \geq 0\}$ satisfies a set of equations which is intermediate between the global and the detailed balance equations. For every node j, consider a partition $S_j^1, \ldots, S_j^k$ of the complementary set of nodes $\{i | i \geq 0, i \neq j\}$ and the equations

$$p_j \sum_{i \epsilon S_j^m} P_{ji} = \sum_{i \epsilon S_j^m} p_i P_{ij}, \quad m = 1, 2, \ldots, k \tag{3A.6}$$

Equations of the form above are known as a set of *partial balance equations*. If a distribution $\{p_j | j \geq 0\}$ solves a set of partial balance equations, then it will also solve the global balance equations so it will be the unique stationary distribution of the chain. A technique that often proves useful is to guess the right set of partial balance equations satisfied by the stationary distribution and then proceed to solve them.

3A.4 Continuous-Time Markov Chains

A continuous-time Markov chain is a process $\{X(t) | t \geq 0\}$ taking values from the set of states $i = 0, 1, \ldots$ that has the property that each time it enters state i:

1. The time it spends in state i is exponentially distributed with parameter ν_i. We may view ν_i as the average rate (in transitions/sec) at which the process makes a transition when at state i.
2. When the process leaves state i, it will enter state j with probability P_{ij}, where $\sum_j P_{ij} = 1$.

   We will be interested in chains for which:

   1. The number of transitions in any finite length of time is finite with probability one (such chains are called *regular*).
   2. The discrete-time Markov chain with transition probabilities P_{ij} (called the *imbedded chain*) is irreducible.

Under the preceding conditions, it can be shown that the limit

$$p_j = \lim_{t \to \infty} P\{X(t) = j | X(0) = i\} \tag{3A.7}$$

exists and is independent of the initial state i. Furthermore if the imbedded chain has a stationary distribution $\{\pi_j | j \geq 0\}$, the steady-state probabilities p_j of the continuous chain are all positive and satisfy

$$p_j = \frac{\pi_j/\nu_j}{\sum_{i=0}^{\infty} \pi_i/\nu_i}, \quad j = 0, 1, \ldots \tag{3A.8}$$

The interpretation here is that π_j represents the proportion of visits to state j, while p_j represents the proportion of time spent in state j in a typical system run.

For every i and j, denote

$$q_{ij} = \nu_i P_{ij} \tag{3A.9}$$

Since ν_i is the rate at which the process leaves i and P_{ij} is the probability that it then goes to j, it follows that q_{ij} is the rate at which the process makes a transition to j when at state i. Consequently, q_{ij} is called the ***transition rate*** from i to j.

Since we will often analyze continuous-time Markov chains in terms of their time-discretized versions, we describe the general method for doing this.

Consider any $\delta > 0$, and the discrete-time Markov chain $\{X_n | n \geq 0\}$, where

$$X_n = X(n\delta), \qquad n = 0, 1, \ldots$$

The stationary distribution of $\{X_n\}$ is clearly $\{p_j | j \geq 0\}$, the stationary distribution of the continuous chain (cf. Eq. (3A.7)). The transition probabilities of $\{X_n | n \geq 0\}$ are

$$\overline{P_{ij}} = \delta q_{ij} + o(\delta), \quad i \neq j$$
$$\overline{P_{ii}} = 1 - \delta \sum_{j \neq i} q_{ij} + o(\delta)$$

Using these expressions in the global balance equations for the discrete chain (cf. Eq. (3A.2)) and taking the limit as $\delta \to 0$, we obtain

$$p_j \sum_{\substack{i=0 \\ i \neq j}}^{\infty} q_{ji} = \sum_{\substack{i=0 \\ i \neq j}}^{\infty} p_i q_{ij}, \quad j = 0, 1, \ldots \tag{3A.10}$$

These are the global balance equations for the continuous chain. Similarly, the detailed balance equations take the form

$$p_j q_{ji} = p_i q_{ij}, \quad i, j = 0, 1, \ldots, \tag{3A.11}$$

One can also write a set of partial balance equations and attempt to solve them for the distribution $\{p_j | j \geq 0\}$. If a solution is found, it provides the stationary distribution of the continuous chain.

APPENDIX B: *Summary of Results*

Notation

p_n: Steady-state probability of having n customers in the system

λ: Arrival rate (inverse of average interarrival time)

μ: Service rate (inverse of average service time)

N: Average number of customers in the system

N_Q: Average number of customers waiting in queue

T: Average customer time in the system

W: Average customer waiting time in queue (does not include service time)

$\overline{X}$: Average service time

$\overline{X^2}$: Second moment of service time

Little's Theorem

$$N = \lambda T$$

$$N_Q = \lambda W$$

Poisson distribution with parameter m

$$p_n = \frac{e^{-m} m^n}{n!}, \quad n = 0, 1, \ldots$$

$$\text{Mean } = \text{Variance } = m$$

Exponential distribution with parameter λ

$$P\{\tau \le s\} = 1 - e^{-\lambda s}, \quad s \ge 0.$$

$$\text{Density: } p(\tau) = \lambda e^{-\lambda \tau}$$

$$\text{Mean } = 1/\lambda$$

$$\text{Variance } = 1/\lambda^2$$

Summary of $M/M/1$ System Results

(a) Utilization factor (proportion of time the server is busy)

$$\rho = \frac{\lambda}{\mu}$$

(b) Probability of n customers in the system

$$p_n = \rho^n (1 - \rho), \quad n = 0, 1, \ldots$$

(c) Average number of customers in the system

$$N = \frac{\rho}{1-\rho}$$

(d) Average customer time in the system

$$T = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}$$

(e) Average number of customers in queue

$$N_Q = \frac{\rho^2}{1-\rho}$$

(f) Average waiting time in queue of a customer

$$W = \frac{\rho}{\mu-\lambda}$$

Summary of $M/M/m$ System Results

(a) Ratio of arrival rate to maximal system service rate

$$\rho = \frac{\lambda}{m\mu}$$

(b) Probability of n customers in the system

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}\right]^{-1}, \quad n = 0$$

$$p_n = \begin{cases} p_0 \dfrac{(m\rho)^n}{n!}, & n \le m \\ p_0 \dfrac{m^m \rho^n}{m!}, & n > m \end{cases}$$

(c) Probability that an arriving customer has to wait in queue (m customers or more in the system)

$$P_Q = \frac{p_0(m\rho)^m}{m!(1-\rho)} \qquad \text{(Erlang C Formula)}$$

(d) Average waiting time in queue of a customer

$$W = \frac{\rho P_Q}{\lambda(1-\rho)}$$

(e) Average number of customers in queue

$$N_Q = \frac{\rho P_Q}{1-\rho}$$

(f) Average customer time in the system

$$T = \frac{1}{\mu} + W$$

(g) Average number of customers in the system

$$N = m\rho + \frac{\rho P_Q}{1-\rho}$$

Summary of $M/M/m/m$ System Results

(a) Probability of m customers in the system

$$p_0 = \left[\sum_{n=0}^{m}\left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1}$$

$$p_n = p_0\left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}, \quad n = 1, 2, \ldots, m$$

(b) Probability that an arriving customer is lost

$$p_m = \frac{(\lambda/\mu)^m/m!}{\sum_{n=0}^{m}(\lambda/\mu)^n/n!} \qquad \text{(Erlang B Formula)}$$

Summary of $M/G/1$ System Results

(a) Utilization factor

$$\rho = \frac{\lambda}{\mu}$$

(b) Mean residual service time

$$R = \frac{\lambda\overline{X^2}}{2}$$

(c) Pollaczek-Khinchin formula

$$W = \frac{R}{1-\rho} = \frac{\lambda\overline{X^2}}{2(1-\rho)}$$

$$T = \frac{1}{\mu} + W$$

$$N_Q = \frac{\lambda^2\overline{X^2}}{2(1-\rho)}$$

$$N = \rho + \frac{\lambda^2\overline{X^2}}{2(1-\rho)}$$

(d) Pollaczek-Khinchin formula for $M/G/1$ queue with vacations

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}}$$

$$T = \frac{1}{\mu} + W$$

where $\overline{V}$ and $\overline{V^2}$ are the first two moments of the vacation interval.

Summary of Reservation/Polling Results

(a) Average waiting time (m-user system, unlimited service)

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m-\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(exhaustive)}$$

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m+\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(partially gated)}$$

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{(m+2-\rho)\overline{V}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} \qquad \text{(gated)}$$

where $\rho = \lambda/\mu$, and $\overline{V}$ and σ_V^2 are the mean and variance of the reservation intervals, respectively, averaged over all users

$$\overline{V} = \frac{1}{m}\sum_{\ell=0}^{m-1} \overline{V}_\ell$$

$$\sigma_V^2 = \frac{1}{m}\sum_{\ell=0}^{m-1} \left(\overline{V_\ell^2} - \overline{V}_\ell^2\right)$$

(b) Average waiting time (m-user system, limited service)

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho-\lambda\overline{V})} + \frac{(m+\rho)\overline{V}}{2(1-\rho-\lambda\overline{V})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\lambda\overline{V})} \qquad \text{(partially gated)}$$

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho-\lambda\overline{V})} + \frac{(m+2-\rho-2\lambda\overline{V})\overline{V}}{2(1-\rho-\lambda\overline{V})} + \frac{\sigma_V^2(1-\rho)}{2\overline{V}(1-\rho-\lambda\overline{V})} \qquad \text{(gated)}$$

(c) Average time in the system

$$T = \frac{1}{\mu} + W$$

Summary of Priority Queueing Results

(a) *Nonpreemptive Priority.* Average waiting time in queue for class k customers

$$W_k = \frac{\sum_{i=1}^{n} \lambda_i \overline{X_i^2}}{2(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}$$

(b) *Nonpreemptive Priority.* Average time in the system for class k customers

$$T_k = \frac{1}{\mu_k} + W_k$$

(c) *Preemptive Resume Priority.* Average time in the system for class k customers

$$T_k = \frac{(1/\mu_k)(1 - \rho_1 - \cdots - \rho_k) + R_k}{(1 - \rho_1 - \cdots - \rho_{k-1})(1 - \rho_1 - \cdots - \rho_k)}$$

where

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i \overline{X_i^2}}{2}$$