CHAPTER 17

Queuing theory for packet networks

17.1 Introduction

In packet switching, secure bundles of information are assembled, addressed and transmitted through a network without the need for dedicated end-to-end connection paths to be established. The packets are individually transported and delivered by the network to the required destination. The network additionally ensures that packets are output in the correct order at the receiver.

Packet switched networks have existed for many years. Early examples were Euronet which links nine EC countries and Switzerland, running the Direct Information Access Network-Europe (DIANE) service. In 1981, British Telecom opened its first national public network, known as the Packet Switched Service (PSS). This system is controlled by a network management centre, based on duplicated minicomputers. PSS uses the X.25 protocol (see Section 18.6.1). The most well known network today is the Internet which supports the information retrieval service known as the World Wide Web (WWW).

The Joint Academic Network (JANET), Figure 18.2, and its successor SuperJANET are funded by the UK research councils. JANET has a node at every university and runs on leased lines. It provides a service for the communication of data, such as computer files and electronic mail, between sites and onward via the Internet (see section 18.7.6). All these networks introduce queuing, with consequent delays and possible loss of traffic data.

Queuing theory [Nussbaumer] can be used to model these or other networks where customers or data packets arrive, wait their turn for handling or service, are subsequently serviced, and are then transmitted through the network. (Supermarket checkouts, ticket booths, and doctors' waiting rooms are all commonly encountered examples of queuing systems.) Queuing theory was developed originally to model analogue teletraffic but is now widely applied to digital packet traffic [Tanenbaum]. A queuing system can be characterised by the following five attributes:

- The interarrival-time probability density function.
- The service-time probability density function.
- The number of servers, or server processes.
- The queuing discipline.
- The amount of buffer, or waiting, space in the queues.

The interarrival-time probability density function (pdf) describes the interval between consecutive arrivals. After a sufficiently long sampling time, the arrival times can be grouped to obtain the pdf which characterises the arrival process.

Each customer requires a certain amount of the server's time which varies from customer to customer. To analyse a queuing system, the service-time pdf, like the interarrival-time pdf, must be known.

The number of servers speaks for itself. Many banks, for example, have one queue for all customers. Whenever a teller is free, the customer at the front of the queue goes directly to that teller. Such a system is a multiserver queuing system. In other banks, each teller has his, or her,' own private queue. This corresponds to a collection of independent single-server queues.

The queuing discipline describes the order in which customers are taken from the queue. Supermarkets use first come, first served. Hospital emergency rooms often use sickest attended to first. In friendly office environments, shortest job first often prevails at the photocopy machine. Not all queuing systems have an infinite amount of buffer space. When too many customers are queued up in a finite number of available slots, some customers can get lost or rejected.

This chapter concentrates predominantly on infinite-buffer, single-server systems using a first come first served queuing discipline. The Kendall notation A/B/m [Kleinrock] is widely used in the queuing literature for these systems. A represents the interarrival-time pdf, B the service-time pdf, and m the number of servers employed. The probability densities A and B are usually chosen from the following set:

M – Markov (implying an exponential pdf);

D – deterministic (all customers have the same constant value implying an impulsive pdf);

G – general (i.e. some arbitrary pdf);

E_k – Erlang distributed.

The state of the art ranges from the M/M/1 system, about which everything is known, to the G/G/m system, for which no exact analytical solution is yet available. We will concentrate on the M/M/1 model. Figure 17.1 shows the queue for such a single server process. We now need to develop a mathematical analysis for queuing systems which can be used to show what limits or restricts the practical performance of these systems. This is achieved by first examining arrivals only, before progressing to model the combined arrival and service processes within the queuing buffer memory.

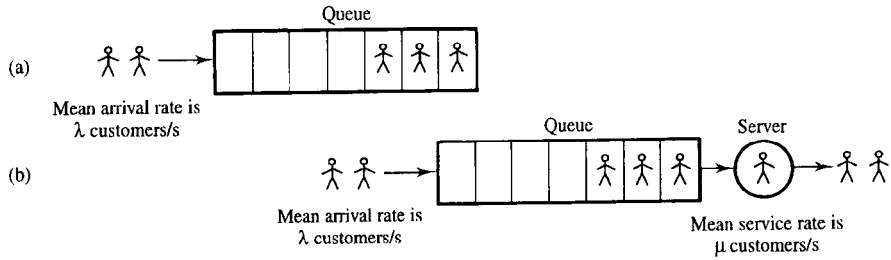

Figure 17.1 *Single server queue model and outcomes for a counting process: (a) arrivals only; (b) arrivals and departures.*

17.2 The arrival process

We make the following assumptions:

- The arrival process is memoryless in the sense that any arrival is statistically independent of all other arrivals;
- The arrival process is statistically stationary. (This implies that the probability of an arrival occuring in any small time interval depends only on the interval's width and not its location in time.)

Queuing systems in which the only transitions are to adjacent states are known as birth–death systems, which is a mathematician's terminology for a counting process with both arrivals and departures. We are considering, at present, *only* arrivals.

We have to model the evolution of the system from one state to another [Gelenbe and Pujolle] where the *state* is synonymous with the number of customers waiting for service. The approach via the Markov probability chain [Chung] is inappropriate, since the probability of transition between any two states at any given point in time, t, is zero. While we cannot characterise the *probability* of transition, we *can* characterise the *rate* of transitions between two states. Suppose that for two particular states the rate of transitions between them is a constant λ. What we mean by this is that in a time δt we can expect an average of $\lambda\ \delta t$ transitions. If δt is very small then $\lambda\ \delta t$ is a number much smaller than unity, and the probability of more than one transition in time δt is vanishingly small. Under these conditions, we can think of $\lambda\ \delta t$ as the probability of one transition (P_1) in time δt, and $(1 - \lambda\ \delta t)$ as the probability of no transition (P_0) in this time.

This leads us to a transition diagram and associated set of differential equations. The transition diagram, Figure 17.2, associates a node with each state. Within node j we denote the probability of being in that state at time t as $P_j(t)$.

17.2.1 pdf for j arrivals in t seconds

Assume arrivals (A) are governed by a randomly distributed pure birth process as shown in Figure 17.3 where the arrival rate does not depend on the state of the system. The

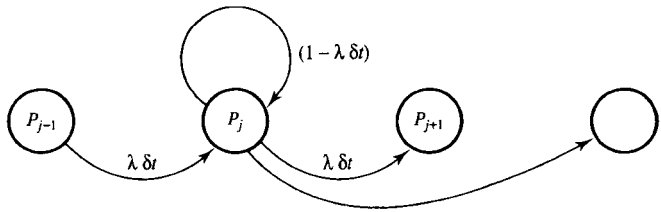

Figure 17.2 *Markov model for queue states corresponding to $j-1$, j, $j+1$ packet arrivals.*

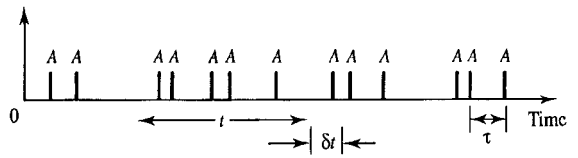

Figure 17.3 *Example of randomly distributed arrivals (A) with interarrival time τ.*

only arrivals then the probability of being in state j after δt seconds is dependent on P_j and P_{j-1}:

$$P_j(t+\delta t) = P_j(t)\,(1-\lambda\,\delta t) + P_{j-1}(t)\,\lambda\,\delta t \tag{17.1}$$

hence:

$$\frac{P_j(t+\delta t) - P_j(t)}{\delta t} = \lambda(P_{j-1}(t) - P_j(t))$$

Now let $\delta t \to 0$:

$$\frac{dP_j(t)}{dt} = \lambda(P_{j-1}(t) - P_j(t)) \tag{17.2}$$

For $j = 0$, $P_{j-1} = 0$, as we cannot have less than zero arrivals. Therefore:

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \tag{17.3}$$

The solution to equation (17.3) is:

$$P_0(t) = e^{-\lambda t} \tag{17.4}$$

which represents the probability of no arrivals in time t. This is plotted in Figure 17.4, starting with a probability of 1 at $t = 0$, decaying asymptotically to zero with an exponential time constant of λ. Thus at $t = 1/\lambda$, $P_0(1/\lambda) = 0.37$. This is the start of the derivation of the Poisson distribution for the number of arrivals j in t seconds for an exponential interarrival time distribution, with a mean arrival rate of λ. It can further be shown that:

$$P_j(t) = \frac{(\lambda t)^j}{j!}\,e^{-\lambda t} \tag{17.5(a)}$$

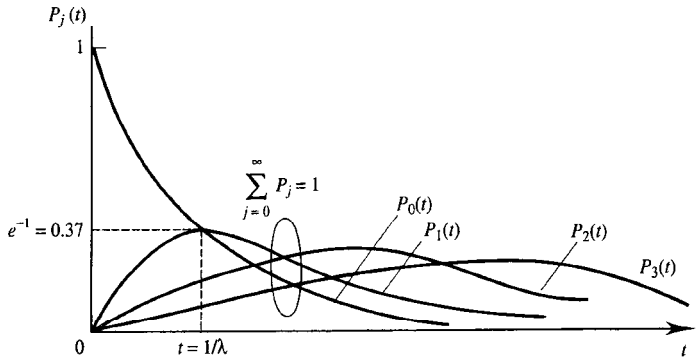

Figure 17.4 *Probability of j arrivals plotted against time for j = 0, 1, 2, 3.*

i.e.:

$$P_1(t) = \lambda t\, e^{-\lambda t} \qquad (17.5(b))$$

$$P_2(t) = \tfrac{1}{2}(\lambda t)^2\, e^{-\lambda t} \quad etc \qquad (17.5(c))$$

Figure 17.4 shows that the probability of only one arrival, $P_1(t)$, peaks at the time interval $t = 1/\lambda$. If Figure 17.4 is plotted with the horizontal axis as offered traffic, λt, then the peak value of $P_1(t)$ occurs at an offered traffic value of unity. For longer time intervals, it becomes increasingly likely that there will be more than one arrival. The probabilities of there being two or three arrival events, $P_2(t)$ and $P_3(t)$, peak at later time intervals and also have progressively smaller probability peaks, so that, for any specified time interval, all the probabilities sum to unity as required, i.e.:

$$\sum_{j=0}^{\infty} P_j(t) = 1 \qquad (17.6)$$

17.2.2 CD and pdf for the time between arrivals

If the time between successive arrivals is τ, as in Figure 17.3, the probability that τ is less than, or equal to, some value of time t, $P(\tau \leq t)$, is given by:

$$P(\tau \leq t) = 1 - P(\tau > t) \qquad (17.7(a))$$

But $P(\tau > t)$ is the probability of no arrivals in time t. Thus, for the Poisson process:

$$P(\tau > t) = P_0(t) = e^{-\lambda t} \qquad (17.7(b))$$

hence:

$$P(\tau \leq t) = 1 - e^{-\lambda t} \qquad (17.7(c))$$

Equation (17.7(c)) is the cumulative distribution (CD) function, equations (3.10) and

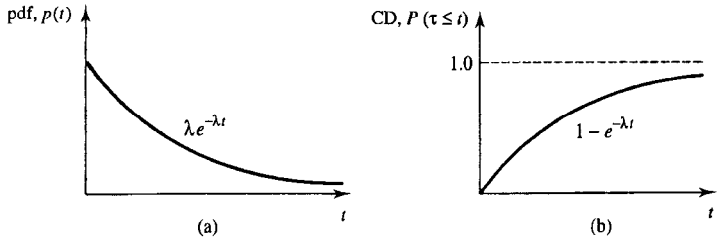

Figure 17.5 *(a) pdf; and (b) CD for time between successive arrivals.*

(3.13(b)), for the time between arrivals, Figure 17.5. This implies an exponential interarrival probability density function (pdf). As shown in section 3.2.4, the pdf is obtained by differentiating the CD, i.e. equation (17.7(c)):

$$P(t \leq \tau < t + dt) = p(t) = \frac{dP(\tau \leq t)}{dt} = \lambda e^{-\lambda t} \qquad (17.7(d))$$

where the mean or average value of t is $1/\lambda$ and the variance (or second central moment) is $1/\lambda^2$ (see section 3.2.5).

17.2.3 Other arrival patterns

These can be specified to be:

- Unpunctual – which occur at $t = a + E_1, 2a + E_2, \cdots, ma + E_m$, where E_m is a random variable;
- Discrete-time arrivals – these can only occur at a discrete set of allowed instants;
- Non-stationary – where the probabilities vary with time;
- Correlated – where the arrival rate may be affected by the state of the system.

17.3 The service process

The service or transmission mechanism is described by the service time distribution, which defines the capacity, or number of servers, which must be deployed to handle the given traffic.

17.3 1 Service time distributions

As for arrival times there are two extremes. We can have constant (deterministic) service times or alternatively we can have 'completely random' (stastically independent) service times, the latter again leading to an exponential pdf given by:

$$P(t) = \mu\, e^{-\mu t} \qquad (17.8)$$

for a service rate of μ customers/s, Figure 17.1. As before, the mean value, or average

service time, is $1/\mu$ and the variance is $1/\mu^2$. The statistical independence of succesive service times (resulting in an exponential distribution of service time pdf) means that service time, like arrival interval, has been modelled as a Poisson process. It should be noted, however, that:

- service times may be discrete (word or packet multiples);
- service time may be non-stationary.

The queuing discipline determines how customers are selected from the queue and allocated to servers.

17.3.2 Single server queues

These typically use one of the following queuing disciplines:

- First-in-first-out (FIFO) – the simple queue;
- Last-in-first-out (LIFO or last come first served, LCFS);
- First-in-random-out (FIRO);
- Priority queuing.

17.3.3 Multiserver queues

Here service is allocated according to rules such as:

- Rotation – customers assigned in strict rotation to each queue;
- Random selection – customers themselves decide which queue to join;
- Single queue – customer at the head of queue goes to the next available server.

17.4 The simple single server queue

It is instructive to find the distribution of queue lengths and waiting times in an M/M/1 system, where total waiting time equals queuing time plus service time.

17.4.1 Simple queue analysis

We define the *state* of the system as the number of customers waiting (i.e. state m implies m customers waiting or a queue of length m), and denote the probability of being in a state m at time t as $P_m(t)$. Assume, when the system is in state m, customers arrive randomly at an average rate λ_m, and are randomly serviced at rate μ_m (i.e. the average service time is $1/\mu_m$).

What is the probability of such a system, Figure 17.6, being in state m at time $t + \delta t$? This is obtained by extending equation (17.1) to include departures as well as arrivals:

$$\begin{aligned} P_m(t+\delta t) &= P_{m-1}(t)\lambda_{m-1}\delta t + P_{m+1}(t)\mu_{m+1}\delta t + P_m(t)(1-\mu_m\delta t)(1-\lambda_m\delta t) \\ &= P_{m-1}(t)\lambda_{m-1}\delta t + P_{m+1}(t)\mu_{m+1}\delta t + P_m(t)[1-(\mu_m+\lambda_m)\delta t] \end{aligned} \quad (17.9)$$

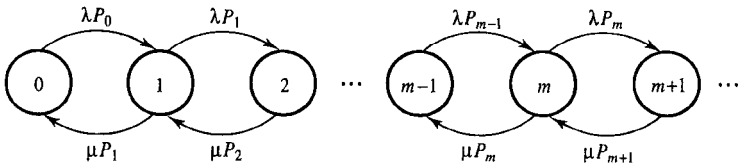

Figure 17.6 *State transition-rate diagram for a simple queue.*

(for small δt). Therefore:

$$\frac{P_m(t+\delta t) - P_m(t)}{\delta t} = \lambda_{m-1} P_{m-1}(t) + \mu_{m+1} P_{m+1}(t) - (\mu_m + \lambda_m) P_m(t) \quad (17.10)$$

In the limit as $\delta t \to 0$:

$$\frac{dP_m(t)}{dt} = \lambda_{m-1} P_{m-1}(t) + \mu_{m+1} P_{m+1}(t) - (\mu_m + \lambda_m)\, P_m(t) \quad (17.11)$$

(for $m \geq 0$ where $P_{-1}(t) = 0$). This is the full Chapman-Kolmogorov equation [Gelenbe and Pujolle] for arrivals and departures. It states that the rate of increase of probability with time for state m is equal to the rate at which transitions into that state, from states $m-1$ and $m+1$, are occurring (multiplied by the current probability of those states) minus the rate at which transitions out of state m are occurring (multiplied by the current probability of state m). To solve equation (17.11) we must specify the initial conditions. The process starts in state zero, as there are no arrivals before time t_0, i.e.:

$$P_m(t_0) = \begin{cases} 1, & m = 0 \\ 0, & m > 0 \end{cases}$$

and assuming a stationary solution so that $dP_m(t)/dt = 0$ then:

$$0 = \lambda_{m-1} P_{m-1} + \mu_{m+1} P_{m+1} - (\mu_m + \lambda_m)\, P_m \quad (17.12(a))$$

where:

$$P_{-1} = P_{-2} = \cdots = 0 \quad (17.12(b))$$

$$\lambda_{-1} = \lambda_{-2} = \cdots = 0 \quad (17.12(c))$$

$$\mu_0 = \mu_{-1} = \cdots = 0 \quad (17.12(d))$$

and:

$$P_0 + P_1 + P_2 + \cdots = 1 \quad (17.13)$$

A typical queue result is shown in Figure 17.7. Here the vertical steps in the solid staircase occur with new customers arriving while the vertical steps in the dashed staircase imply service has been completed for these customers. The horizontal displacement measures the queue plus service time for each customer or unique arrival.

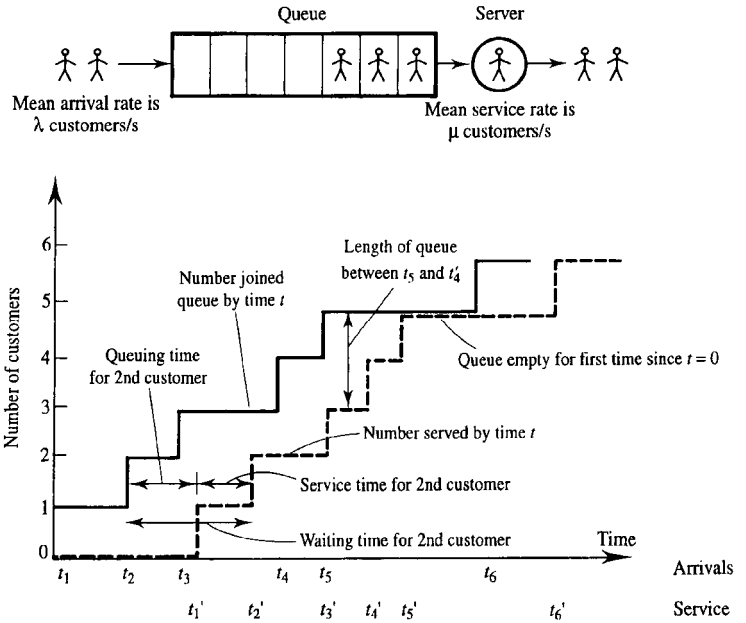

Figure 17.7 *Typical queue performance showing customer (packet) arrivals and departures (after service).*

17.4.2 Queue parameters

The total delay in a time t, $\gamma(t)$, is the sum of the waiting times. If the total number of customers who have arrived in a time t is denoted by $\alpha(t) = \lambda t$ then:

$$\text{average delay,} \qquad T = \frac{\gamma(t)}{\alpha(t)} \tag{17.14(a)}$$

$$\text{average queue length,} \qquad N = \frac{\gamma(t)}{t} \tag{17.14(b)}$$

$$\text{average arrival rate,} \qquad \lambda = \frac{\alpha(t)}{t} \tag{17.14(c)}$$

Now we can rewrite the average queue length as $N = \gamma(t)/t = (\gamma(t)/\alpha(t)) \times (\alpha(t)/t)$, where N is given by the sum, over all m, of the product of queue length, m, and its probability, P_m, to obtain Little's result:

$$N = \sum_{m=0}^{\infty} mP_m = T\lambda \tag{17.15}$$

The performance of a queuing system is controlled by its utilisation factor, defined as:

$$\rho = \frac{\text{demand for service}}{\text{maximum rate of supply}} \tag{17.16}$$

where the demand for service = arrival rate × mean service time = λ/μ. This is also a measure of the traffic intensity in erlangs [Dunlop and Smith], named after the Danish pioneer of teletraffic theory. (In telephony systems the total applied traffic in erlangs is equal to $\lambda\tau$ where λ is the arrival rate for call connections and τ is the average call duration. A circuit carrying one call continuously then carries one erlang of traffic.)

The service rate is often controlled directly by the output transmission rate and packet length. For example a 2 Mbit/s link (Chapter 19) with 500 byte packets and 8-bit bytes, has a service rate, $\mu = (2 \times 10^6)/(500 \times 8) = 500$ packet/s. For single server queues, maximum capacity or rate of supply is 1 s of service/s, i.e. the maximum rate of supply = 1. In this case:

$$\rho = \lambda/\mu \tag{17.17}$$

and $\rho < 1$ is required to prevent server overload.

17.4.3 Classical queue with single server

Assume a Poisson arrival process and exponentially distributed service times so that the arrival and service rates are independent of the state of the system. Thus λ_m simplifies to λ and μ_m simplifies to μ. Further assume that infinite queuing space is available and that customers are served on FIFO basis.

From equation (17.5) and Figure 17.6, by applying the conditions of equilibrium or detailed balancing, the input and output transitions to and from a state must occur at the same rate. Thus starting with state zero, $\lambda P_0 = \mu P_1$, these balances can be written as:

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0 \\ P_2 &= \frac{\lambda^2}{\mu^2} P_0 \\ P_m &= \left(\frac{\lambda}{\mu}\right)^m P_0 = \rho^m P_0 \end{aligned} \tag{17.18}$$

But $\Sigma_{m=0}^{\infty} P_m = 1$, as shown previously (Figure 17.4 and equation (17.6)). Now $\Sigma_{m=0}^{\infty} P_m = \Sigma_{m=0}^{\infty} \rho^m P_0 = P_0 \Sigma_{m=0}^{\infty} \rho^m = 1$ and the sum of the geometric series $\Sigma_{m=0}^{\infty} \rho^m = 1/(1-\rho)$. Therefore:

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu} \tag{17.19}$$

Figure 17.8 (curve (a)) shows, as ρ increases from 0 to 1, how the probability of an empty queue, P_0, decreases. Also by applying equation (17.18):

$$P_m = \rho^m P_0 = \rho^m (1 - \rho) \tag{17.20}$$

implying a geometric distribution for P_m. The probability of the single server being busy is $1 - P_0 = \rho$. $\rho < 1$ ensures that the server has more capacity than is required. Figure 17.9 shows a typical queue length pdf, for $\rho = 0.5$.

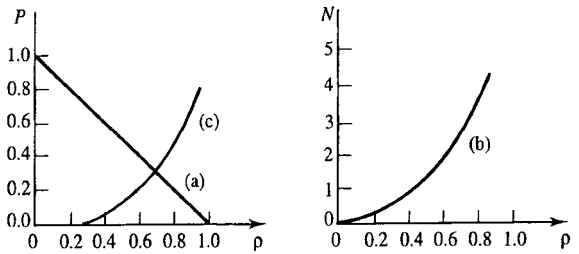

Figure 17.8 *Average queue size or length: (a) probability of an empty queue; (b) mean queue length, N, in packets; (c) probability that queue length exceeds 4.*

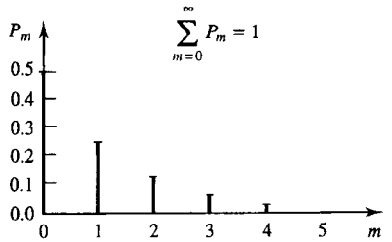

Figure 17.9 *Queue length pdf for* $\rho = \frac{1}{2}$.

EXAMPLE 17.1

For a single server queuing system with Poisson distributed arrivals of average rate 1 message/s and Poisson distributed service of capacity 3 messages/s calculate the probability of receiving no messages in a 5 s period. Also find the probabilities of queue lengths of 0, 1, 2, 3. If the queue length is limited to 4 what percentage of messages will be lost?

From equation (17.5(a)):

$$P_0(t) = e^{-\lambda t}$$

and $\lambda = 1$ and $t = 5$ for a 5 s period. Thus the probability of no arrivals in a 5 s period is given by:

$$P_0(5) = e^{-5} = 0.00674$$

Now from equation (17.18):

$$P_m = \left(\frac{\lambda}{\mu}\right)^m P_0$$

and from equation (17.19):

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{1}{3} = \frac{2}{3}$$

Thus the various queue lengths can be calculated as:

$$P_1 = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}$$

$$P_2 = \left(\frac{1}{3}\right)^2 \times \frac{2}{3} = \frac{2}{27}$$

$$P_3 = \left(\frac{1}{3}\right)^3 \times \frac{2}{3} = \frac{2}{81}$$

This differs slightly from Figure 17.9 in that the P_0 value is larger and, in consequence, the subsequent magnitudes fall off more rapidly. Finally for a queue length limited to 4 the probability of exceeding this restricted length is given by:

$$P(m > 4) = 1 - P(m \leq 4) = 1 - (P_0 + P_1 + P_2 + P_3 + P_4)$$

$$= 1 - \left(\frac{2}{3} + \frac{2}{9} + \frac{2}{27} + \frac{2}{81} + \frac{2}{243}\right) = 0.0041 = 0.41\%$$

17.4.4 Queue length and waiting times

The average queue length, equation (17.15), $N = \Sigma_{m=0}^{\infty} m P_m = (1-\rho)\Sigma_{m=0}^{\infty} m\rho^m$. Further, as the sum of the geometric series $\Sigma_{m=0}^{\infty} m\rho^m = \rho/(1-\rho)^2$ the mean queue length is given by:

$$N = \sum_{m=k+1}^{\infty} P_m = \frac{\rho}{1-\rho} \tag{17.21}$$

Figure 17.8 (curve (b)) shows how N increases with increasing ρ. At $\rho = ½$, $N = 1$ and for $\rho > ½$, N increases above unity. As traffic intensity increases and ρ approaches 1 the queue length becomes infinite. If the queue is restricted to some finite value k then the probability of exceeding this value is given by:

$$P(m > k) = \sum_{m=k+1}^{\infty} P_m$$

$$= (1-\rho) \sum_{m=k+1}^{\infty} \rho^m$$

$$= (1-\rho)\left[\sum_{m=0}^{\infty} \rho^m - \sum_{m=0}^{k} \rho^m\right]$$

$$= (1-\rho)\left[\frac{1}{1-\rho} - \frac{1-\rho^{k+1}}{1-\rho}\right]$$

$$= \rho^{k+1} \tag{17.22}$$

This is shown in Figure 17.8 (curve (c)) for various values of ρ. Clearly when ρ exceeds about 0.8 there is a problem. To find average delay or waiting time, T, we use Little's

result, equation (17.15) and equation (17.21):

$$T = \frac{\text{average queue length}}{\text{average arrival rate}} = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} \qquad (17.23(a))$$

or, using equation (17.17):

$$T = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda} \qquad (17.23(b))$$

Average delay (normalised by μ) is plotted in Figure 17.10, against ρ in the range $0 \leq \rho < 1$, and it is this key result which forms the basis of network delay analysis. When constrained by a finite queue length of k, then $\Sigma_{m=0}^{k} P_m = 1 = P_0 \Sigma_{m=0}^{k} \rho^m$. For this case $P_0 = (1-\rho)/(1-\rho^{k+1})$.

For packets transmitted over a link, at a bit rate of R_b bit/s with packet size K bits, the mean packet delay is, from equation (17.23(b)):

$$T = \frac{1}{R_b/K - \lambda} \qquad (17.24)$$

where R_b/K represents the packet transmission or service rate, μ, in packet/s and λ is the packet arrival rate.

EXAMPLE 17.2

Consider a switch at which packets arrive according to a Poisson distribution. The mean arrival rate is 3 packet/s. The service time is exponentially distributed with a mean value of 100 ms. Assume the packet comprises 70 8-bit bytes and the output transmission rate is 5.6 kbit/s. How long does a packet have to wait in the queue?

The mean service rate is $\mu = 5600/(8 \times 70) = 10$ packet/s. From equation (17.23(b)):

$$T = 1/(\mu(1-\rho)) = 1/(\mu-\lambda)$$

and we find that the mean packet delay is $T = 0.143$ s for queuing plus service. Since the mean service time is 100 ms, the mean queuing time is therefore 143 – 100 = 43 ms.

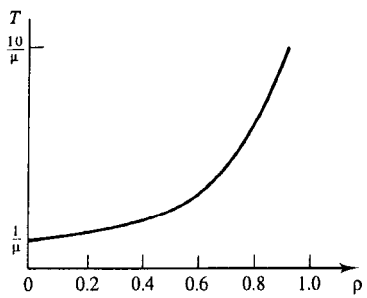

Figure 17.10 *Average time delay (T) for queue against ρ.*

Calculations, such as those shown above, allow us to model the delays on packet data links. The mean overall network delay is given by the sum of the transmitted packets times their delay divided by the total number of transmitted packets. (This type of queue analysis is also used for determining the performance of the Banyan switch networks in Chapter 18 in order to find their blocking performance.)

17.5 Packet speech transmission

Early packet networks for communication (see Chapter 18) were:

- store-and-forward, hard-wired, long-haul networks, e.g. the American ARPANET (with a channel capacity of 50 kbit/s), and the British SuperJANET (Chapter 18);
- satellite networks, e.g. the American Atlantic SATNET (with a channel capacity of 64 kbit/s);
- radio networks, e.g. the American PRNET (with a channel capacity of 100 to 400 kbit/s, for local data distribution).

The most general distinction is that, unlike the circuit switched TDMA telephone system (Figure 17.11), in a packet switched network the individual data (D) and voice (V) packets are switched (Figure 17.12) in accordance with the header information within each packet. The interface between the user and a packet network can be a data terminal, a host computer, or a packet voice terminal which comprises a telephone handset with a full range of control and signalling capabilities.

17.5.1 The components of packet speech

Analogue speech must first be converted into a digital sequence by a coder using waveform processing, e.g. delta modulation, Figure 5.28, or other coding techniques such as LPC, section 9.7. The vocoder is favoured when there is limited channel capacity and speech must be compressed to lower data rates than PCM. LPC produces fewer packet/s (1.7 to 7.4) than delta modulation which generates 9.4 to 38.5 packet/s at a bit rate of 16 kbit/s.

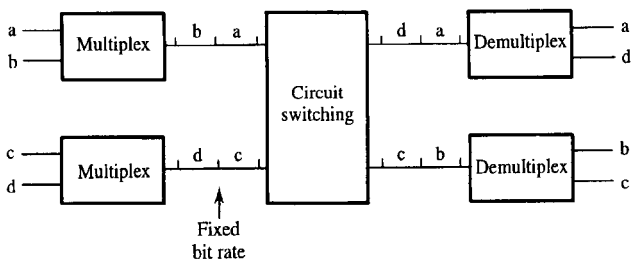

Figure 17.11 *Conventional circuit switched network (e.g. TDMA digital telephony transmission example, as described in section 6.5).*

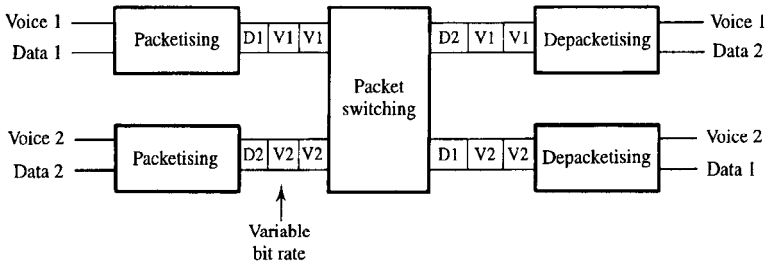

Figure 17.12 *Packet switched network with variable rate traffic where voice rate exceeds the input data rate.*

The digital bit stream is next partitioned into segments and some control (header) information added to each segment to form a packet. The control information consists of a time stamp and a sequence number, Figure 17.13, to assist reconstruction at the receiver. To reduce the bandwidth required for speech transmission, silences between bursts of speech are not packetised or transmitted. The time stamp enables the receiver to generate the appropriate duration of silence before processing the subsequent speech samples. Time stamps also allow reordering of packets which are received out of order. Since the time stamp cannot differentiate silence from packet loss, a sequence number is included to allow detection of lost packets.

It is important that a continuous stream of bits is provided at the receiver in order to produce smooth speech. This is achieved by delaying the arriving packets at the receiver queueing buffer, Figure 17.14. The size of this buffer must be sufficiently long to avoid packet loss due to overflow. Normally the delay will be chosen so that, statistically, a large proportion (e.g. 95%) of packets will be expected to arrive in the time allocated to the buffer delay, which is usually in the range 100 to 170 ms. The delay must not be so long that it dominates the performance of the speech transmission system, however.

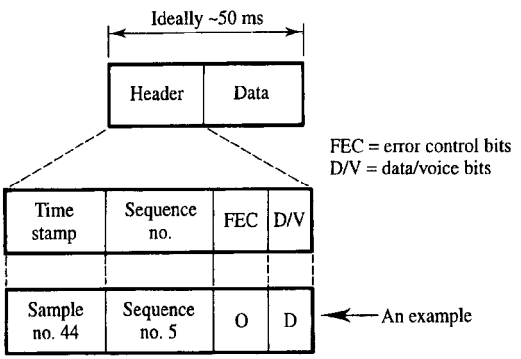

Figure 17.13 *Details of the header information carried within a data packet.*

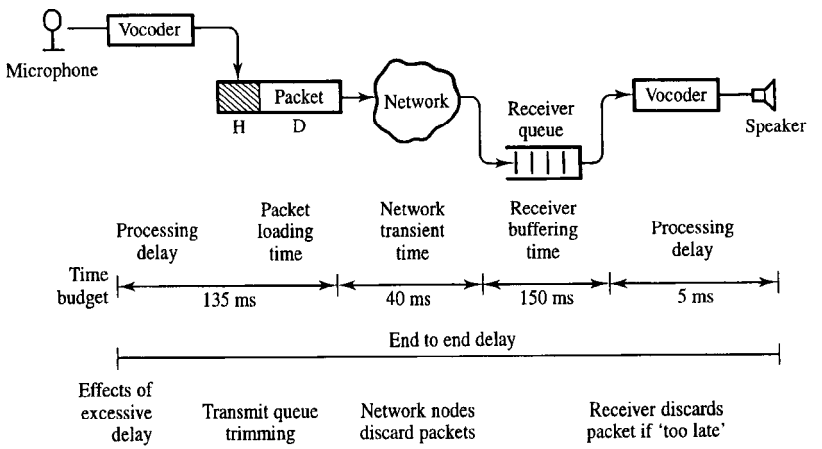

Figure 17.14 *Packet speech transmission system with inherent transmission delays.*

EXAMPLE 17.3

An X.25 packet switch has a single outgoing transmission link at 2 Mbit/s. The average length of each packet is 960 bytes. If the average packet delay through the switch, assuming an M/M/1 queue, is to be less than 15 ms, determine: (i) the maximum gross input packet rate to the switch; (ii) the average length of the queue; (iii) the utilisation factor through the switch if each packet in the input is converted into ATM cells having 48 bytes of data and a 5 byte overhead (see section 19.7.2 for an explanation of ATM).

In the X.25 switch the outgoing link is 2 Mbit/s. An average packet is 960 bytes or 7680 bits. If service time is equal to packet length, then: $\mu = 2 \times 10^6/7680 = 260.4$ packet/s

(i) Now if $T \leq 15$ ms, from equation (17.23(b)):

$$T = \frac{1}{\mu(1-\rho)} = \frac{1}{260.4(1-\rho)} \leq 15 \times 10^{-3} \quad \text{and} \quad \rho_{\min} = 0.744$$

Maximum gross input rate is $\lambda = \rho\mu = 0.744 \times 260.4 = 193.7$ packet/s.

(ii) Average length of queue from equation (17.21) is:

$$N = \rho/(1-\rho) = 2.91 \text{ packets}$$

(iii) ATM switch

Each packet is 960 bytes, and corresponds to 960/48 = 20 ATM cells.

The cell input rate is therefore $20 \times 193.7 = 3{,}874$ cell/s.

Now the output rate is $2 \times 10^6/(48+5) \times 8 = 4{,}717$ cell/s. Therefore $\rho = 3874/4717 = 0.821$.

17.5.2 Speech service performance

Three key parameters that describe the performance of a packet speech service are end-to-end delay, throughput and reliability. Delay is the time between speech being presented to the system, to the packet carrying that speech being played at the loudspeaker. Experimental tests show that few people notice any quality degradation if delay is kept below 0.3 s while a delay above 1.5 s is intolerable. Throughput is limited by the processing capability of the nodes. Reliability is defined as the proportion of packets that arrive at the destination in time to be used to reconstruct the speech.

Appropriate choice of packet size and rate can minimise delay and allow high throughput. In particular we must control the number of bits in the message header. Since the header is constant for every packet, regardless of size, to maintain high channel utilisation the number of speech bits per packet should be maximised. Large packets are also more desirable from the point of view of network throughput. However, to minimise the effects of lost packets and delay at the transmitter, packets should be short and, ideally, a packet should contain no more than 50 ms of speech. The trade-off is particularly difficult for narrowband speech, e.g. using LPC, because 50 ms of 2.4 kbit/s speech comprises only 120 data bits. Typical packet size for speech transmission across the Internet is about 300 bits comprising 100 to 170 ms speech segments. Channel loading information should be provided to the voice terminal so that packet rate and size can be varied according to the network load. In cases where the network is lightly loaded, it is capable of supporting a higher packet rate, hence smaller packets with less delay are used while, if the network loading is high, packets are made larger and packet rate is reduced.

17.6 Summary

Packets are groups of data bits to which have been added (as headers and/or trailers) addressing and other control information to facilitate routing through a digital data network. Queuing theory can be used to model packet behaviour at the switching nodes of a network and, in particular, can be used to predict average packet delay, average queue length and probability of packet loss at a node, given the packet interarrival-time pdf, the packet service-time pdf, the number of servers, the queuing discipline and the queuing storage space. For real-time applications, such as packet speech, resources can be saved by not transmitting empty or 'silent' packets. This necessitates packets being time stamped, however, for the receiver to regenerate the appropriate speech gaps.

In current Ethernet based networks data rates are in the range 10 to 100 Mbit/s and propagation delays (typically less than 5 μs) are small compared with the time required to transmit a 1 kbit packet. With the trend towards Gbit/s optical fibre links, operating over long distances, propagation delay becomes much longer than the packet duration, which will significantly alter the analysis of these systems.

17.7 Problems

17.1. The number of messages arriving at a particular node in a message switched computer network may be assumed to be Poisson distributed. Given that the average arrival rate is 5 messages per minute, calculate the following:
(a) Probability of receiving no messages in an interval of 2 minutes. [4.5×10^{-5}]
(b) Probability of receiving just 1 message in the next 30 s. [0.2]
(c) Probability of receiving 10 messages in any 30 s period. [2.16×10^{-4}]

17.2. In Problem 17.1: (a) What is the probability of having a gap between two successive messages of greater than 20 s? (b) What is the probability of having gaps between messages in the range 20 to 30 s inclusive? [0.189, 0.106]

17.3. Assuming a classical queue with a single server, determine the probabilities of having: (a) an empty queue, (b) a queue of 4 or more 'customers'. You should assume that the utilisation factor for the service is 0.6. [0.4, 0.13]

17.4. If the queue length in Problem 17.3 is limited to 10, what percentage of customers are lost to the service? [0.36%]

17.5. A packet data network links London, Northampton and Southend for credit card transaction data. It is realised with a 2-way link between London and Northampton and, separately, between London and Southend. Both links operate with primary rate access at 2 Mbit/s in each direction. If the Northampton and Southend nodes send 200 packets/s to each other and the London node sends 50 packets/s to Northampton what is the mean packet delay on the network when the packets have a mean size of 2 kbit? [1.275 ms]