APPENDIX A

Delay and Loss Performance

A key feature of communication networks is the sharing of resources such as transmission bandwidth, storage, and processing capacity. Because the demand for these resources is unscheduled, the situation can arise where resources are not available when a user places a request. This situation typically leads to a delay or loss in service. In this appendix we develop some simple but fundamental models to quantify the delay and loss performance. The appendix is organized as follows:

1. *Little's formula* relates the average occupancy in the system to the average time spent in the system. This formula is extremely powerful in obtaining average delay performance of complex systems.
2. *Basic queueing model* for a multiplexer allows us to account for arrival rate, message length, transmission capacity, buffer size, and performance measures such as delay and loss.
3. *M/M/1* model provides a simple, basic multiplexer model that allows us to explore trade-offs among the essential system parameters.
4. *M/G/1* model provides a more precise description of service times and message lengths.
5. *Erlang B blocking formula* quantifies blocking performance in loss systems.

A.1 DELAY ANALYSIS AND LITTLE'S FORMULA

Figure A.1 shows a basic model for a delay/loss system. Customers arrive to the system according to some arrival pattern. These customers can be connection requests, individual messages, packets, or cells. The system can be an individual transmission line, a multiplexer, a switch, or even an entire network. The custo-

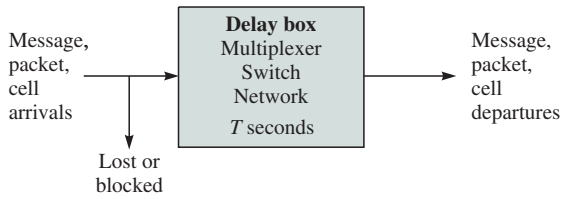


FIGURE A.1 Network delay analysis

mer spends some time T in the system. After this time the customer departs the system. It is possible that under certain conditions the system is in a blocking state, for example, due to lack of resources. Customers that arrive at the system when it is in this state are blocked or lost. We are interested in the following performance measures:

- Time spent in the system: T.
- Number of customers in the system: $N(t)$.
- Fraction of arriving customers that are lost or blocked: P_b.
- Average number of messages/second that pass through the system: throughput.

Customers generally arrive at the system in a random manner, and the time that they spend in the system is also random. In this section we use elementary probability to assess the preceding performance measures.

A.1.1 Arrival Rates and Traffic Load Definitions

We begin by introducing several key system variables and some of their averages. Let $A(t)$ be the number of arrivals at the system in the interval from time 0 to time t. Let $B(t)$ be the number of blocked customers and let $D(t)$ be the number of customer departures in the same interval. The number of customers in the system at time t is then given by

$$N(t) = A(t) - D(t) - B(t)$$

because the number that have entered the system up to time t is $A(t) - B(t)$ and because $D(t)$ of these customers have departed by time t. Note that we are assuming that the system was empty at $t = 0$. The long-term **arrival rate** at the system is given by

$$\lambda = \lim_{t\to\infty} \frac{A(t)}{t} \text{customers/second}$$

The **throughput** of the system is equal to the long-term departure rate, which is given by

$$\text{throughput} = \lim_{t\to\infty} \frac{D(t)}{t} \text{customers/second}$$

The **average number in the system** is given by

$$E[N] = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(t')dt' \text{ customers}$$

The **fraction of blocked customers** is then

$$P_b = \lim_{t\to\infty} \frac{B(t)}{A(t)}$$

Figure A.2 shows a typical sample function $A(t)$, the number of arrivals at the system. We assume that we begin counting customers at time $t = 0$. The first customer arrives at time τ_1, and so $A(t)$ goes from 0 to 1 at this time instant. The second arrival is τ_2 seconds later. Similarly the nth customer arrival is at time $\tau_1 + \tau_2 + \ldots + \tau_n$, where τ_i is the time between the arrival of the $i-1$ and the ith customer. The arrival rate up to the time when the nth customer arrives is then given by $n/(\tau_1 + \tau_2 + \ldots + \tau_n)$ customers/second. Therefore, the long-term arrival rate is given by

$$\lambda = \lim_{n\to\infty} \frac{n}{\tau_1 + \tau_2 + \ldots + \tau_n} = \lim_{n\to\infty} \frac{1}{(\tau_1 + \tau_2 + \ldots + \tau_n)/n} = \frac{1}{E[\tau]}$$

In the preceding expression we assume that all of the interarrival times are statistically independent and have the same probability distribution and that their average or expected value is given by $E[\tau]$. *Thus the average arrival rate is given by the reciprocal of the average interarrival time.*

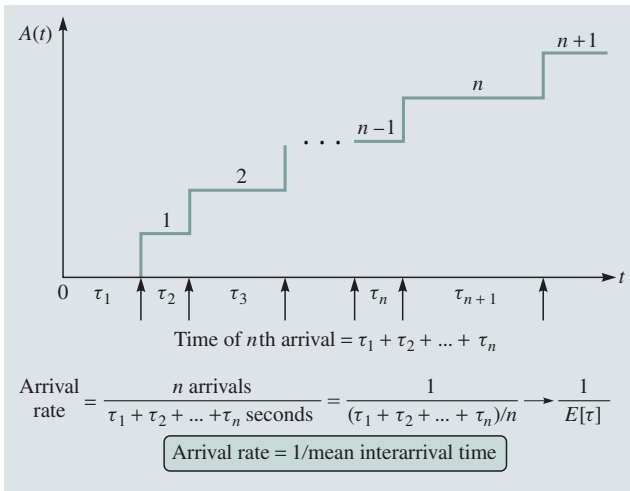


FIGURE A.2 Arrivals at a system as a sample function

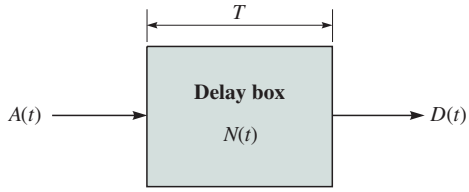


FIGURE A.3 Little's formula

A.1.2 Little's Formula

Next we will develop **Little's formula**, which relates the average time spent in the system $E[T]$ to the arrival rate λ and the average number of customers in the system $E[N]$ by the following formula:

$$E[N] = \lambda E[T]$$

We will assume, as shown in Figure A.3, that the system does not block any customers. The number in the system $N(t)$ varies according to $A(t) - D(t)$.

Suppose we plot $A(t)$ and $D(t)$ in the same graph as shown in Figure A.4. $A(t)$ increases by 1 each time a customer arrives, and $D(t)$ increases by 1 each time a customer departs. The number of customers in the system $N(t)$ is given by the difference between $A(t)$ and $D(t)$. The number of departures can never be greater than the number of arrivals, and so $D(t)$ lags behind $A(t)$ as shown in the figure. Assume that customers are served in first-in, first-out (FIFO) fashion. Then the time T_1 spent by the first customer in the system is the time that elapses between the instant when $A(t)$ goes from 0 to 1 to the instant when $D(t)$ goes from 0 to 1. Note that T_1 is also the area of the rectangle defined by these two time instants in the figure. A similar relationship holds for all subsequent times $T_2, T_3, \ldots$

Consider a time instant t_0 where $D(t)$ has caught up with $A(t)$; that is, $N(t_0) - A(t_0) - D(t_0) = 0$. Note that the area between $A(t)$ and $D(t)$ is given by

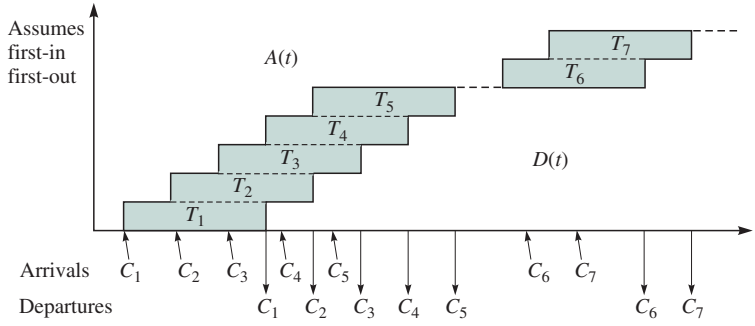


FIGURE A.4 Arrivals and departures in a FIFO system

the sum of the times T_0 spent in the system by the first $A(t_0)$ customers. The time average of the number of customers in the system up to time t_0 is then

$$\frac{1}{t_0}\int_0^{t_0} N(t')dt' = \frac{1}{t_0}\sum_{j=1}^{A(t_0)} T_j$$

If we multiply and divide the preceding expression by $A(t_0)$, we obtain

$$\frac{1}{t_0}\int_0^{t_0} N(t')dt' = \frac{A(t_0)}{t_0}\left\{\frac{1}{A(t_0)}\sum_{j=1}^{A(t_0)} T_j\right\}$$

This equation states that, up to time t_0, the average number of customers in the system is given by the product of the average arrival rate $A(t_0)/t_0$ and the arithmetic average of the times spent in the system by the first $A(t_0)$ customers. Little's formula follows if we assume that

$$E[T] = \lim_{A(t_0)\to\infty}\left\{\frac{1}{A(t_0)}\sum_{j=1}^{A(t_0)} T_j\right\}$$

It can be shown that Little's formula is valid even if customers are not served in order of arrival [Bertsekas 1987].

Now consider a system in which customers can be blocked. The above derivation then applies if we replace $A(t)$ by $A(t) - B(t)$, the actual number of customers who enter the system. The actual arrival rate into a system with blocking is $\lambda(1 - P_b)$, since P_b is the fraction of arrivals that are blocked. It then follows that Little's formula for a system with blocking is

$$E[N] = \lambda(1 - P_b)E[T]$$

In the preceding derivation we did not specify what constitutes a "system," so Little's formula can be applied in many different situations. Thus we can apply Little's formula to an individual transmission line, to a multiplexer, to a switch, or even to a network.

We now show the power of Little's formula by finding the average delay that is experienced by a packet in traversing a packet-switching network. Figure A.5 shows an entire packet-switching network that consists of interconnected packet switches. We assume that when a packet arrives at a packet switch the packet is routed instantaneously and placed in a multiplexer to await transmission on an outgoing line. Thus each packet switch can be viewed as consisting of a set of multiplexers. We begin by applying Little's formula to the network as a whole. Let N_{net} be the total number of packets in the network, let T_{net} be the time spent by the packet in the network, and let λ_{net} be the total packet arrival rate to the network, Little's formula then states that

$$E[N_{net}] = \lambda_{net}E[T_{net}]$$

This formula implies that the average delay experienced by packets in traversing the network is

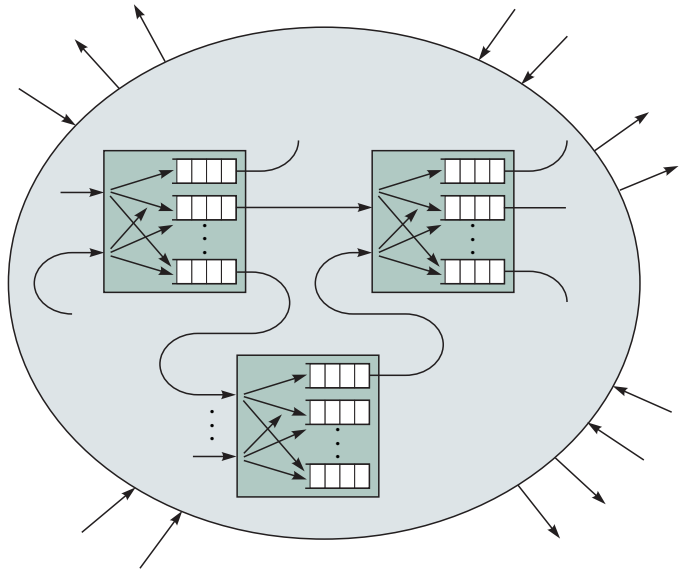


FIGURE A.5 Packet-switching network delay

$$E[T_{net}] = E[N_{net}]/\lambda_{net}$$

We can refine the preceding equation by applying Little's formula to each individual multiplexer. For the mth multiplexer Little's formula gives

$$E[N_m] = \lambda_m E[T_m]$$

where λ_m is the packet arrival rate at the multiplexer and $E[T_m]$ is the average time spent by a packet in the multiplexer. The total number of packets in the network N_{net} is equal to the sum of the packets in all the multiplexers:

$$E[N_{net}] = \sum_m E[N_m] = \sum_m \lambda_m E[T_m]$$

By combining the preceding three equations, we obtain an expression for the total delay experienced by a packet in traversing the entire network:

$$E[T_{net}] = E[N_{net}]/\lambda_{net} = \frac{1}{\lambda_{net}} \sum_m \lambda_m E[T_m]$$

Thus the network delay depends on the overall arrival rates in the network, the arrival rate to individual multiplexers, and the delay in each multiplexer. The arrival rate at each multiplexer is determined by the routing algorithm. The delay in a multiplexer depends on the arrival rate and on the rate at which the associated transmission line can transmit packets. Thus the preceding formula succinctly incorporates the effect of routing as well as the effect of the capacities of the transmission lines in the network. For this reason the preceding expression is frequently used in the design and management of packet-switching networks. To

obtain $E[T_m]$, it is necessary to analyze the delay performance of each multiplexer. This is our next topic.

A.2 BASIC QUEUEING MODELS

The pioneering work by Erlang on the traffic engineering of telephone systems led to the development of several fundamental models for the analysis of resource-sharing systems. In a typical application customers demand resources at random times and use the resources for variable durations. When all the resources are in use, arriving customers form a line or "queue" to wait for resources to become available. *Queueing theory* deals with the analysis of these types of systems.

A.2.1 Arrival Processes

Figure A.6 shows the basic elements of a queueing system. Customers arrive at the system with interarrival times $\tau_1, \tau_2, \ldots, \tau_n$. We will assume that the interarrival times are independent random variables with the same distribution. The results for the arrival process developed in Figure A.2 then hold. In particular, the *arrival rate to the system* is given by

$$\lambda = \frac{1}{E[\tau]} \text{customers/second}$$

Several special cases of arrival processes are of interest. We say that arrivals are *deterministic* when the interarrival times are all equal to the same constant value. We say that the arrival times are *exponential* if the interarrival times are exponential random variables with mean $E[\tau] = 1/\lambda$:

$$P[\tau > t] = e^{-t/E[\tau]} = e^{-\lambda t} \text{ for } t > 0$$

The case of exponential interarrival times is of particular interest because it leads to tractable analytical results. It can be shown that when the interarrival

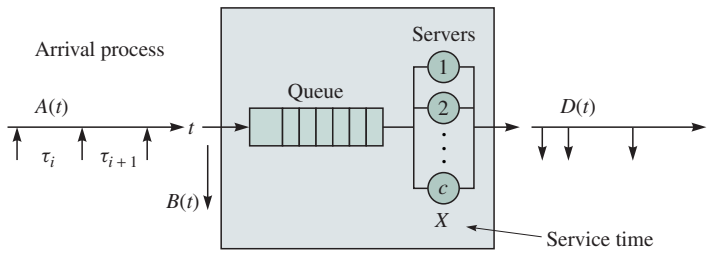


FIGURE A.6 Queueing model

times are exponential, then the number of arrivals $A(t)$ in an interval of length t is given by a Poisson random variable with mean $E[A(t)] = \lambda t$:

$$P[A(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \text{ for } k = 0, 1, \ldots$$

For this reason, the case of exponential interarrival times is also called the *Poisson arrival process*.

A.2.2 Service Times

Resources are denoted by "servers" because their function is to serve customer requests. The time required to service a customer is called the **service time** and is denoted by X. In our discussion the server is typically a transmission line and the service time can be the time required to transmit a message or the duration of a telephone call. The maximum rate at which a server can process customers is attained when the server is continuously busy. When this is the case, the average time between customer departures is equal to the average service time. The processing capacity of a single server is given by the maximum throughput or departure rate. From the discussion leading to the arrival rate formula, clearly the processing capacity is given by

$$\mu = \frac{1}{E[X]} \text{customers/second}$$

The processing capacity μ can be likened to the maximum flow that can be sustained over a pipe. The number of servers c in a queueing system can be greater than one. The total processing capacity of a queueing system is then given by $c\mu$ customers/second.

An ideal queueing system is one where customers arrive at equal intervals and in which they require a constant service time. As long as the service time is less than the interarrival time, each customer arrives at an available server and there is no waiting time. In general, however, the interarrival time and the service times are random. The combination of a long service time followed by a short interarrival time can then lead to a situation in which the server is not available for an arriving customer. For this reason, in many applications a queue is provided so that a customer can wait for an available server, as shown in Figure A.6. When a server becomes available the next customer to receive service is selected according to the service discipline. Possible service disciplines are FIFO; last-in, first-out (LIFO); service according to priority class; and random order of service. We usually assume FIFO service disciplines.

The *maximum* number of customers allowed in a queueing system is denoted by K. Note that K includes both the customers in queue and those in service. We denote the total number of customers in the system by $N(t)$, the number in queue by $N_q(t)$, and the number in service by $N_s(t)$. When the system is full, that is, $N(t) = K$, then new customers arrivals are blocked or lost.

A.2.3 Queueing System Classification

Queueing systems are classified by a notation that specifies the following characteristics:

- Customer arrival pattern.
- Service time distribution.
- Number of servers.
- Maximum number in the system.

For example, in Figure A.7 the queueing system M/M/1/K corresponds to a queueing system in which the interarrival times are exponentially distributed (M)[1]; the service times are exponentially distributed (M); there is a single server (1); and at most K customers are allowed in the system. The M/M/1/K model was used to illustrate the typical delay and loss performance of a data multiplexer in Chapter 5. If there is no maximum limit in the number of customers allowed in the system, the parameter K is left unspecified. Thus the M/M/1 system is identical to the above system except that it has no maximum limit on the number of customers allowed in the system.

The M/G/1 is another example of a queueing system where the arrivals are exponential, the service times have a *general* distribution, there is a single server, and there is no limit on the customers allowed in the system. Similarly, the M/D/1 system has constant, that is, *deterministic*, service times.

Figure A.8 shows the parameters that are used in analyzing a queueing system. The total time that a customer spends in the system is denoted by T, which consists of the time spent waiting in queue W plus the time spent in service X. When a system has blocking, P_b denotes the fraction of customers that are blocked. Therefore, the actual arrival rate into the system is given by $\lambda(1 - P_b)$. This value is the arrival rate that should be used when applying Little's formula.

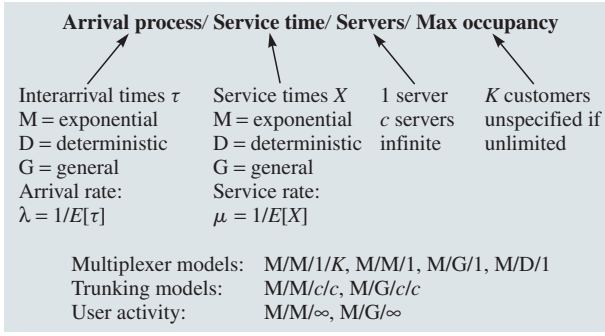


FIGURE A.7 Queueing model classification

[1]The notation M is used for the exponential distribution because it leads to a Markov process model.

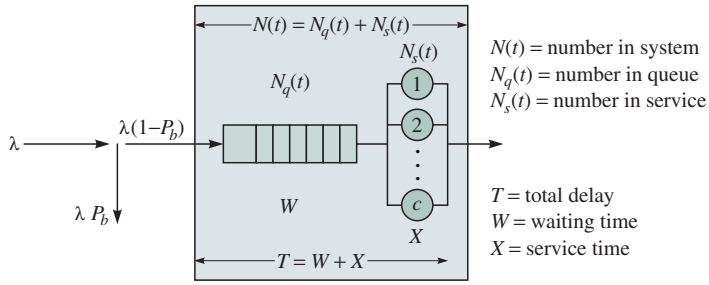


FIGURE A.8 Queueing system variables

Thus the average number in the system and the average delay in the system are related by

$$E[N] = \lambda(1 - P_b)E[T]$$

If we apply Little's formula where the "system" is just the queue, then the average number of customers in queue and the average waiting time are related by

$$E[N_q] = \lambda(1 - P_b)E[W]$$

Finally, if the system is defined as a set of servers, then the average number of customers in service and the average service time are related by

$$E[N_s] = \lambda(1 - P_b)E[X]$$

The preceding three equations are very useful in relating average occupancy and average delay performance. Typically it is relatively simple to obtain the averages associated with occupancies, that is, $N(t)$, $N_q(t)$ and $N_s(t)$.

Finally, we revisit some of the terms introduced earlier in the appendix. The *traffic load* or offered load is the rate at which "work" arrives at the system:

$$\begin{aligned} a &= \lambda \text{ customers/second} \times E[X] \text{ seconds/customer} \\ &= \lambda/\mu \text{ Erlangs.} \end{aligned}$$

The *carried load* is the average rate at which the system does work. It is given by the product of the average service time per customer, $E[X]$, and the actual rate at which customers enter the system, $\lambda(1 - P_b)$. Thus we see that the carried load is given by $a(1 - P_b)$.

The *utilization* ρ is defined as the average fraction of servers that are in use:

$$\rho = \frac{E[N_s]}{c} = \frac{\lambda}{c\mu}(1 - P_b)$$

Note that when the system has a single server, then the utilization ρ is also equal to the proportion of time that the server is in use.

A.3 M/M/1: A BASIC MULTIPLEXER MODEL

In this section we develop the M/M/1/K queueing system, shown in Figure A.9, as a basic model for a multiplexer. The interarrival times τ in this system have mean $E[\tau] = 1/\lambda$ as an exponential distribution. Let $A(t)$ be the number of arrivals in the interval 0 to t; then as indicated above $A(t)$ has a Poisson distribution.

The average packet length is $E[L]$ bits per packet, and the transmission line has a speed of R bits/second. So the average packet transmission time is $E[X] = E[L]/R$ seconds. This transmission line is modeled by a single server that can process packets at a maximum rate of $\mu = R/E[L]$ packets/second. We assume that the packet transmission time X has an exponential distribution:

$$P[X > t] = e^{-t/E[X]} = e^{-\mu t} \text{ for } t > 0$$

We also assume that the interarrival times and packet lengths are independent of each other. We will first assume that at most K packets are allowed in the system. We later consider the case were K is infinite.

In terms of long-term flows, packets arrive at this system at a rate of λ packets/second, and the maximum rate at which packets can depart is μ packets/second. If $\lambda > \mu$, then the system will necessarily lose packets because the system is incapable of handling the arrival rate λ. If $\lambda < \mu$, then on the average the system can handle the rate λ, but it will occasionally lose packets because of temporary surges in arrivals or long consecutive service times. We will now develop a model that allows us to quantify these effects.

Consider what events can happen in the next Δt seconds. In terms of arrivals there can be 0, 1, or >1 arrivals. Similarly there can be 0, 1, or >1 departures. It can be shown that if the interarrival times are exponential, then

$$P[1 \text{ arrival in } \Delta t] = \lambda \Delta t + o(\Delta t)$$

where $o(\Delta t)$ denotes terms that are negligible relative to Δt, as $\Delta t \to 0$.[2] Thus the probability of a single arrival is proportion to λ. Similarly, it can also be shown that probability of no arrivals in Δt seconds is given by

$$P[0 \text{ arrival in } \Delta t] = 1 - \lambda \Delta t + o(\Delta t)$$

The preceding two equations imply that only two events are possible as Δt becomes very small: one arrival or no arrival. Since the service times also have an

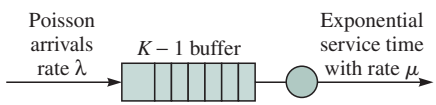


FIGURE A.9 M/M/1/K queue

[2] In particular, a function $g(x)$ is $o(x)$ if $g(x)/x \to 0$ as $x \to 0$; that is, $g(x)$ goes to 0 faster than x does.

exponential distribution, it can be shown that a customer in service will depart in the next Δt seconds with probability

$$P[1 \text{ departure in } \Delta t] = \mu \Delta t + o(\Delta t)$$

and that the probability that the customer will continue its service after an additional Δt seconds is

$$P[0 \text{ departure in } \Delta t] = 1 - \mu \Delta t + o(\Delta t)$$

A.3.1 M/M/1 Steady State Probabilities and the Notion of Stability

We can determine the probability of changes in the number of customers in the system by considering the various possible combinations of arrivals and departures:

$$\begin{aligned} P[0 \text{ arrival } \& \text{ } 0 \text{ departure in } \Delta t] &= \{1 - \mu \Delta t + o(\Delta t)\}\{1 - \lambda \Delta t + o(\Delta t)\} \\ &= 1 - (\lambda + \mu)\Delta t + o(\Delta t) \end{aligned}$$

The preceding equation gives the probability that the number in the system is still $n > 0$ after Δt seconds.

$$\begin{aligned} P[1 \text{ arrival } \& \text{ } 0 \text{ departure in } \Delta t] &= \{\lambda \Delta t + o(\Delta t)\}\{1 - \mu \Delta t + o(\Delta t)\} \\ &= \lambda \Delta t + o(\Delta t) \end{aligned}$$

The preceding equation gives the probability that the number in the system increases by 1 in Δt seconds.

$$\begin{aligned} P[\text{no arrival } \& \text{ } 1 \text{ departure in } \Delta t] &= \{1 - \lambda \Delta t + o(\Delta t)\}\{\mu \Delta t + o(\Delta t)\} \\ &= \mu \Delta t + o(\Delta t). \end{aligned}$$

The preceding equation gives the probability that the number in the system decreases by 1 in Δt seconds. Note that the preceding equations imply that $N(t)$ always changes by single arrivals or single departures.

Figure A.10 shows the state transition diagram for $N(t)$, the number in the system. $N(t)$ increases by 1 in the next Δt seconds with probability $\lambda \Delta t$ and decreases by 1 in the next Δt seconds with probability $\mu \Delta t$. Note that every transition n to $n + 1$ cannot recur until the reverse transition $n + 1$ to n occurs.

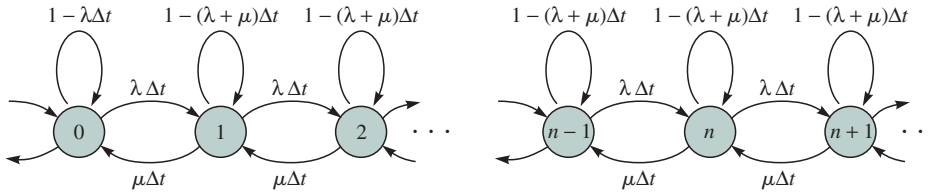


FIGURE A.10 State transition diagram

Therefore, if the system is stable, that is, if it does not grow steadily to infinity, then the long-term transition rate from n to $n+1$ must equal the long-term transition rate from $n+1$ to n.

Let p_n be the probability that n customers are in the system; then p_n is also the proportion of time that the system is in state n. Therefore, $p_n\lambda\Delta t$ is the transition from state n to state $n+1$. Similarly, $p_{n+1}\mu\Delta t$ is the transition rate from $n+1$ to n. This discussion implies that the two transition rates must be equal. Therefore

$$p_{n+1}\mu\Delta t = p_n\lambda\Delta t$$

This implies that

$$p_{n+1} = (\lambda/\mu)p_n \quad n = 0, 1, \ldots, K$$

Repeated applications of the preceding recursion imply that

$$p_{n+1} = (\lambda/\mu)^{n+1}p_0 = p^n p_0 \quad n = 0, 1, \ldots, K$$

To find p_0, we use the fact that the probabilities must add up to 1:

$$1 = p_0 + p_1 + p_2 + \ldots + p_K = p_0\{1 + \rho + \rho^2 + \rho^3 + \ldots \rho_K\}$$
$$= p_0\frac{1-\rho^{K+1}}{1-\rho}$$

which implies that

$$p_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

Finally we obtain the probabilities for the number of customers in the system:

$$P[N(t) = n] = p_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, \text{ for } n = 0, 1, \ldots, K$$

The probability of blocking or loss in the M/M/1/K system is given by $P_{loss} = p_K$, which is the proportion of time that the system is full.

Consider what happens to the state probabilities as the load ρ is varied. For ρ less than 1, which corresponds to $\lambda < \mu$, the probabilities decrease exponentially as n increases; thus the number in the system tends to cluster around $n = 0$. In particular, adding more buffers is beneficial when $\lambda < \mu$, since the result is a reduction in loss probability. When $\rho = 1$, the normalization condition implies that all the states are equally probable; that is, $p_n = 1(K+1)$. Once ρ is greater than 1, the probabilities actually increase with n and tend to cluster toward $n = K$; that is, the system tends to be full, as expected. Note that adding buffers when $\lambda > \mu$ is counterproductive, since the system will fill up the additional buffers. This result illustrates a key point in networking: The arrival rate should not be allowed to exceed the maximum capacity of a system for extended periods of time. The role of *congestion control* procedures inside the network is to deal with this problem.

The average number of customers in the system $E[N]$ is given by

$$E[N] = \sum_{n=0}^{K} np_n = \sum_{n=0}^{K} n\frac{(1-\rho)\rho^n}{1-\rho^{K+1}} = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

The preceding equation is valid for ρ not equal to 1. When $\rho = 1$, $E[N] = K/2$. By applying Little's formula, we obtain the average delay in an M/M/1/K system:

$$E[T] = \frac{E[N]}{\lambda(1-P_K)}$$

Now consider the M/M/1 system that has $K = \infty$. The state transition diagram for this system is the same as in Figure A.10 except that the states can assume all nonzero integer values. The probabilities are still related by

$$p_{n+1} = (\lambda/\mu)^{n+1}p_0 = \rho^n p_0 \quad n = 0, 1, \ldots$$

where $\rho = \lambda/\mu$. The normalization condition is now

$$\begin{aligned} 1 &= p_0 + p_1 + p_2 + \ldots = p_0\{1 + \rho + \rho^2 + \rho^3 + \ldots\} \\ &= p_0 \frac{1}{1-\rho} \end{aligned}$$

Note that the preceding power series converges only if $\rho < 1$, which corresponds to $\lambda < \mu$. This result agrees with our intuition that an infinite-buffer system will be stable only if the arrival rate is less than the maximum departure rate. If not, the number in the system would grow without bound. We therefore find that the state probabilities are now

$$P\{N(t) = n\} = p_n = (1-\rho)\rho^n \quad n = 0, 1, \ldots \rho < 1$$

The average number in the system and the average delay are then given by

$$\begin{aligned} E[N] &= \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{(1-\rho)} \\ E[T] &= \frac{E[N]}{\lambda} = \frac{1/\mu}{(1-\rho)} \end{aligned}$$

The average waiting time is obtained as follows:

$$E[W] = E[T] - E[X] = \frac{(1/\mu)\rho}{(1-\rho)}.$$

These equations show that the average delay and the average waiting time grow without bound as ρ approaches 1. Thus we see that when arrivals or service times are random, perfect scheduling is not possible, so the system cannot be operated at $\lambda = \mu$.

A.3.2 Effect of Scale on Performance

The expressions for the M/M/1 system allow us to demonstrate the typical behavior of queueing systems as they are increased in scale. Consider a set of m separate M/M/1 systems, as shown in Figure A.11. Each system has an arrival rate of λ customers/second and a processing rate of μ customers/second. Now suppose that it is possible to combine the customer streams into a single stream with arrival rate $m\lambda$ customers/second. Also suppose that the processing capacities are combined into a single processor with rate $m\mu$ customers/second. The mean delay in the separate systems is given by

$$E[T_{separate}] = \frac{1/\mu}{(1-\rho)}$$

The combined system has an arrival rate $\lambda' = m\lambda$ and a processing rate $\mu' = m\mu$; therefore, its utilization is $\rho' = \lambda'/\mu' = \rho$. Therefore, the mean delay in the combined system is

$$E[T_{combined}] = \frac{1/\mu'}{(1-\rho)} = \frac{1/m\mu}{(1-\rho)} = \frac{1}{m}E[T_{separate}]$$

Thus we see that the combined system has a total delay of $1/m$ of the separate systems.

The improved performance of the combined system arises from improved global usage of the processors. In the separate systems some of the queues may be empty while others are not. Consequently, some processors can be idle, even though there is work to be done in the system. In the combined system the processor will stay busy as long as customers are waiting to be served.

A.3.3 Average Packet Delay in a Network

In the beginning of this section, we used Little's formula to obtain an expression for the average delay experienced by a packet traversing a network. As shown in Figure A.5 the packet-switching network was modeled as many interconnected multiplexers. The average delay experienced by a packet in traversing the network is

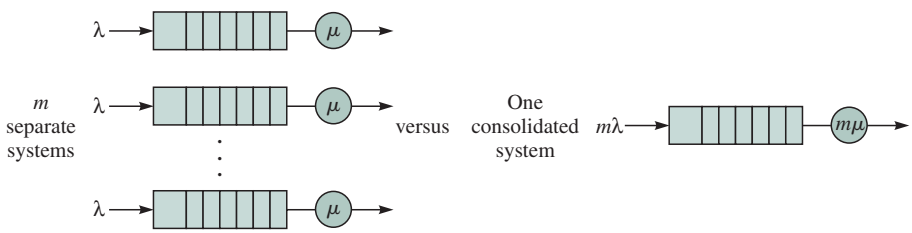


FIGURE A.11 Multiplexing gain and effect of scale

$$E[T_{net}] = \frac{1}{\lambda_{net}} \sum_m \lambda_m E[T_m]$$

where $E[T_m]$ is the average delay experienced in each multiplexer. To apply the results from our M/M/1 analysis, we need to make several assumptions. The most important assumption is that the service times that a packet experiences at different multiplexers are independent of each other. In fact, this assumption is untrue, since the service time is proportional to the packet length, and a packet has the same length as it traverses the network. Nevertheless, it has been found that this *independence assumption* can be used in larger networks.

If we model each multiplexer by an M/M/1 queue, we can then use the corresponding expression for the average delay:

$$E[T_m] = (1/\mu)(1 - \rho_m), \quad \text{where } \rho_m = \lambda_m/\mu$$

where λ_m is the packet arrival rate at the mth multiplexer. The average packet delay in the network is then

$$E[T_m] = \sum_m \frac{1}{\lambda_{net}} \left(\frac{\rho_m}{1 - \rho_m} \right)$$

This simple expression was first derived by [Kleinrock 1964]. It can be used as the basis for selecting the transmission speeds in a packet-switching network. It can also be used to synthesize routing algorithms that distribute the flows of packets over the network so as to keep the overall packet delay either minimum or within some range.

A.4 THE M/G/1 MODEL

The M/M/1 models derived in the previous sections are extremely useful in obtaining a quick insight into the trade-offs between the basic queueing systems parameters. The M/G/1 queueing model allows us to consider a more general class of resource sharing systems. In particular, the service time can have any distribution and is not restricted to be exponential. The derivation of the M/G/1 results is beyond the scope of our discussion; we simply present the results and apply them to certain multiplexer problems. The reader is referred to [Leon-Garcia 1994] for a more detailed discussion of this model.

As shown in Figure A.12, the M/G/1 model assumes Poisson arrivals with rate λ, service times X with a general distribution $F_x(x) = P[X \leq x]$ that has a mean $E[X]$ and a variance $VAR[X]$, a single server, and unlimited buffer space.

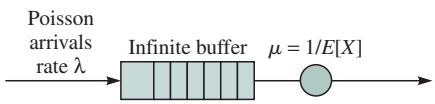


FIGURE A.12 M/G/1 queueing system

The mean waiting time in an M/G/1 queueing system is given by

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

Using the fact that $E[X^2] = VAR[X] + E[X]^2$, we obtain

$$E[W] = \frac{\lambda(VAR[X] + E[X]^2}{2(1-\rho)} = \frac{\rho(1+C_X^2)}{2(1-\rho)}E[X]$$

where the coefficient of variation of the service time is given by $C_X^2 = VAR[X]/E[X]^2$.

The mean delay of the system is obtained by adding the mean service time to $E[W]$:

$$E[T] = E[W] + E[X]$$

The mean number in the system $E[N]$ and the mean number in queue $E[N_q]$ can then be found from Little's formula.

A.4.1 Service Time Variability and Delay

The mean waiting time in an M/G/1 system increases with the coefficient of variation of the service time. Figure A.13 shows the coefficient of variation for several service time distributions. The exponential service time has a coefficient of variation of 1 and serves as a basis for comparison. The constant service time has zero variance and hence has a coefficient of variation of zero. Consequently, its mean waiting time is one-half that of an M/M/1 system. The greater randomness in the service times of the M/M/1 system results in a larger average delay in the M/M/1 system.

Figure A.13 also shows two other types of service time distributions. The Erlang distribution has a coefficient of variation between 0 and 1, and the hyperexponential distribution has a coefficient of variation greater than 1. These two distributions can be used to model various degrees of randomness in the service time relative to the M/M/1 system.

In Chapter 6 we used the M/G/1 formula in assessing various types of schemes for sharing broadcast channels.

	M/D/1	M/Er/1	M/M/1	M/H/1
Interarrivals	Constant	Erlang	Exponential	Hyperexponential
C_X^2	0	<1	1	>1
$E[W]/E[W_{M/M/1}]$	1/2	1/2<, <1	1	>1

FIGURE A.13 Comparison of mean waiting times in M/G/1 systems

A.4.2 Priority Queueing Systems

The M/G/1 model can be generalized to the case where customers can belong to one of K priority classes. When a customer arrives at the system, the customer joins the queue of its priority class. Each time a customer service is completed, the next customer to be served is selected from the head of the line of the highest priority nonempty queue. We assume that once a customer begins service, it cannot be preempted by subsequent arrivals of higher-priority customers.

We will assume that the arrival at each priority class is Poisson with rate λ_k and that the average service time of a class k customer is $E[X_k]$, so the load offered by class k is $\rho_k = \lambda_k E[X_k]$.

It can be shown that if the total load is less than 1; that is

$$\rho = \rho_1 + \rho_2 + \ldots + \rho_K < 1$$

then the average waiting time for a type k customer is given by

$$E[W_k] = \frac{\lambda E[X^2]}{(1 - \rho_1 - \rho_2 - \ldots - \rho_{k-1})(1 - \rho_1 - \rho_2 - \ldots - \rho_k)}$$

where

$$E[X^2] = \frac{\lambda_1}{\lambda} E[X_1^2] + \frac{\lambda_2}{\lambda} E[X_2^2] + \ldots + \frac{\lambda_K}{\lambda} E[X_K^2]$$

The mean delay is found by adding $E[X_k]$ to the corresponding waiting time. The interesting result in the preceding expression is in the terms in the denominator that indicate at what load a given class saturates. For example, the highest priority class has average waiting time

$$E[W_1] = \frac{\lambda E[X^2]}{(1 - \rho_1)}$$

so it saturates as ρ_1 approaches 1. Thus the saturation point of class 1 is determined only by its own load. On the other hand, the waiting time for class 2 is given by

$$E[W_2] = \frac{\lambda E[X^2]}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

The class 2 queue will saturate when $\rho_1 + \rho_2$ approaches 1. Thus the class 2 queue saturation point is affected by the class 1 load. Similarly, the class k queue saturation point depends on the sum of the loads of the classes of priority up to k. This result was used in Chapter 7 in the discussion of priority queueing disciplines in packet schedulers.

A.4.3 Vacation Models and Multiplexer Performance

The M/G/1 model with vacations arises in the following way. Consider an M/G/1 system in which the server goes on vacation (becomes unavailable) whenever it empties the queue. If upon returning from vacation, the server finds that the system is still empty, the server takes another vacation, and so on until it finds customers in the system. Suppose that vacation times are independent of each other and of the other variables in the system. If we let V be the vacation time, then the average waiting time in this system is

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

The M/G/1 vacation model is very useful in evaluating the performance of various multiplexing and medium access control systems. As a simple example consider an ATM multiplexer in which cells arrive according to a Poisson process and where cell transmission times are constrained to begin at integer multiples of the cell time. When the multiplexer empties the cell queue, then we can imagine that the multiplexer goes away on vacation for one cell time, just as modeled by the M/G/1 vacation model. If the cell transmission time is the constant X, then a vacation time is also $V = X$. Therefore, the average waiting time in this ATM multiplexer is

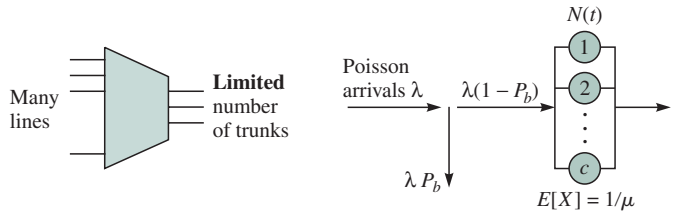
$$E[W] = \frac{\lambda X^2}{2(1-\rho)} + \frac{E[X^2]}{2E[X]} = \frac{\rho X}{2(1-\rho)} + \frac{X}{2}$$

The first term is the average waiting time in an ordinary M/G/1 system. The second term is the average time that elapses from the arrival instant of a random customer arrival to the beginning of the next cell transmission time.

A.5 ERLANG B FORMULA: M/M/c/c System

In Figure A.14 we show the M/M/c/c queueing model that can be used to model a system that handles trunk connection requests from many users. We assume trunk requests with exponential interarrival times with rate λ requests/second. Each trunk is viewed as a server, and the connection time is viewed as the service time X. Thus each trunk or server has a service rate $\mu = 1/E[X]$. We assume that connection requests are blocked if all the trunks are busy.

The state of the preceding system is given by $N(t)$, the number of trunks in use. Each new connection increases $N(t)$ by 1, and each connection release decreases $N(t)$ by 1. The state of the system then takes on the values $0, 1, \ldots, c$. The probability of a connection request in the next Δt seconds is given by $\lambda \Delta t$. The M/M/c/c queueing model differs from the M/M/1 queueing model in terms of the departure rate. If $N(t) = n$ servers are busy, then each server will complete its service in the next Δt seconds with probability $\mu \Delta t$. The



- Blocked calls are cleared from the system; no waiting allowed.
- Performance parameter: P_b = fraction of arrivals that are blocked.
- $P_b = P[N(t) = c] = B(c, a)$ where $a = \lambda/\mu$.
- $B(c, a)$ is the Erlang B formula, which is valid for ***any*** service time distribution.

$$B(c, a) = \frac{\frac{a^c}{c!}}{\sum_{j=0}^{c} \frac{a^j}{j!}}$$

FIGURE A.14 M/M/c/c and the Erlang B formula

probability that one of the connections completes its service and that the other $n - 1$ continue their service in the next Δt seconds is given by

$$n(\mu\Delta t)(1 - \mu\Delta t)^{n-1} \approx n\mu\Delta t$$

The probability that two connections will complete their service is proportional to $(\mu\Delta t)^2$, which is negligible relative to Δt. Therefore, we find that the departure rate when $N(t) = n$ is $n\mu$.

Figure A.15 shows the state transition diagram for the M/M/c/c system. Proceeding as in the M/M/1/K analysis, we have

$$p_{n+1}(n + 1)\mu\Delta t = p_n\lambda\Delta t$$

This result implies that

$$p_{n+1} = \frac{\lambda}{(n + 1)\mu} p_n \quad n = 0, 1, \ldots, c$$

Repeated applications of the preceding recursion imply that

$$p_{n+1} = \frac{(\lambda/\mu)^{n+1}}{(n + 1)!} p_0 \quad n = 0, 1, \ldots, c$$

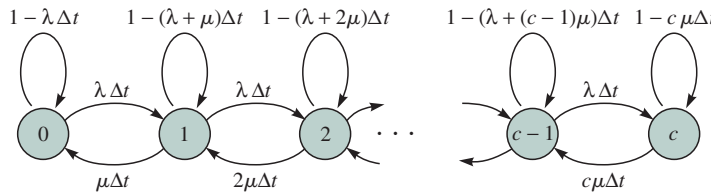


FIGURE A.15 State transition diagram for M/M/c/c

Let the offered load be denoted by $a = \lambda/\mu$. To find p_0, we use the fact that the probabilities must add up to 1:

$$1 = p_0 + p_1 + p_2 + \ldots + p_c = p_0 \sum_{n=0}^{c} \frac{a^n}{n!}$$

Finally, we obtain the probabilities for the number of customers in the system:

$$P[N(t) = n] = p_n = \frac{\frac{a^n}{n!}}{\sum_{k=0}^{c} \frac{a^k}{k!}}, \text{ for } n = 0, 1, \ldots, c$$

The probability of blocking in the M/M/c/c system is given by $P_b = p_c$, which is the proportion of time that the system is full. This result leads to the Erlang B formula $B(c, a)$ that was used in Chapter 4.

Finally, we note that the Erlang B formula also applies to the M/G/c/c system; that is, the service time distribution need not be exponential.

FURTHER READING

Bertsekas, D. and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

Kleinrock, L., *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964.

Leon-Garcia, A., *Probability and Random Process for Electrical Engineering*, Addison-Wesley, Reading, Massachusetts, 1994.