Delay Models
in Data Networks

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 mostly 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 between the time the last bit is transmitted at the head node of the link and the time the last bit is received at the tail node. This is proportional to the physical distance between transmitter and receiver; it can be relatively substantial, particularly for a satellite link or a very high speed 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 developed extensively 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 on both 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, maintain
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 \( L/C \) seconds 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 were allocated to a single channel. The transmission time of a packet that is \( L \) bits long is \( Lm/C \), or \( m \) times larger 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 a 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 and simplifying packet processing at the nodes. Note also that when overhead is negligible, one can afford to make packets very small, thereby reducing delay through pipelining (cf. Fig. 2.37).
3.2 QUEUEING MODELS—LITTLE’S THEOREM

We consider queueing systems where customers arrive at random times to obtain service. 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 emphasize very much), customers represent ongoing conversations (or virtual circuits) between points in a network and service time corresponds to the duration of a conversation. In a related context, customers represent active calls in a telephone or circuit switched network and again service time corresponds to the duration of the call.

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).

These quantities will be estimated in terms of known information such as:

1. The customer arrival rate (i.e., the “typical” number of customers entering the system per unit time)
2. The customer service rate (i.e., the “typical” number of customers the system serves per unit time when it is constantly busy)

In many cases the customer arrival and service rates are not sufficient to determine the delay characteristics of the system. For example, if customers tend to arrive in groups, the average customer delay will tend to be larger than when their arrival times are regularly spaced apart. Thus to predict average delay, we will typically need more detailed (statistical) information about the customer interarrival and service times. In this section, however, we will largely ignore the availability of such information and see how far we can go without it.

3.2.1 Little’s Theorem

We proceed to clarify the meaning of the terms “average” and “typical” that we used somewhat liberally above in connection with the number of customers in the system, the customer delay, and so on. In doing so we will derive an important result known as Little’s Theorem.

Suppose that we observe a sample history of the system from time \( t = 0 \) to the indefinite future and we record the values of various quantities of interest as time
 progresses. In particular, let

\[ N(t) = \text{Number of customers in the system at time } t \]
\[ \alpha(t) = \text{Number of customers who arrived in the interval } [0, t] \]
\[ T_i = \text{Time spent in the system by the } i^{th} \text{ arriving customer} \]

Our intuitive notion of the "typical" number of customers in the system observed up to time \( t \) is

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

which we call the time average of \( N(\tau) \) up to time \( t \). Naturally, \( N_t \) changes with the time \( t \), but in many systems of interest, \( N_t \) tends to a steady-state \( N \) as \( t \) increases, that is,

\[ N = \lim_{t \to \infty} N_t \]

In this case, we call \( N \) the steady-state time average (or simply time average) of \( N(\tau) \). It is also natural to view

\[ \lambda_t = \frac{\alpha(t)}{t} \]

as the time average arrival rate over the interval \([0, t]\). The steady-state arrival rate is defined as

\[ \lambda = \lim_{t \to \infty} \lambda_t \]

(assuming that the limit exists). The time average of the customer delay up to time \( t \) is similarly defined as

\[ T_t = \frac{\sum_{i=0}^{n} T_i \alpha(t)}{\alpha(t)} \]  \hspace{1cm} (3.1)

that is, the average time spent in the system per customer up to time \( t \). The steady-state time average customer delay is defined as

\[ T = \lim_{t \to \infty} T_t \]

(assuming that the limit exists).

It turns out that the quantities \( N, \lambda, \) and \( T \) above 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

\[ N = \lambda T \]

Little's Theorem expresses the natural idea that crowded systems (large \( N \)) are associated with long customer delays (large \( T \)) and conversely. 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.
The theorem is really an accounting identity and its derivation is very simple. We will give a graphical proof under some simplifying assumptions. Suppose that the system is initially empty \([N(0) = 0]\) and that customers depart from the system in the order they arrive. Then the number of arrivals \(\alpha(t)\) and departures \(\beta(t)\) up to time \(t\) form a staircase graph as shown in Fig. 3.1. The difference \(\alpha(t) - \beta(t)\) is the number in the system \(N(t)\) at time \(t\). The shaded area between the graphs of \(\alpha(\tau)\) and \(\beta(\tau)\) can be expressed as

\[
\int_0^t N(\tau) \, d\tau
\]

and if \(t\) is any time for which the system is empty \([N(t) = 0]\), the shaded area is also equal to

\[
\sum_{i=1}^{\alpha(t)} T_i
\]

Dividing both expressions above with \(t\), we obtain

\[
\frac{1}{t} \int_0^t N(\tau) \, d\tau = \frac{1}{t} \sum_{i=1}^{\alpha(t)} T_i = \frac{\alpha(t)}{t} \sum_{i=1}^{\alpha(t)} \frac{T_i}{\alpha(t)}
\]

or equivalently,

\[
N_t = \lambda_t T_t \tag{3.2}
\]

Little's Theorem is obtained assuming that

\[
N_t \to N, \lambda_t \to \lambda, T_t \to T
\]

and that the system becomes empty infinitely often at arbitrarily large times. With a minor modification in the preceding argument, the latter assumption becomes unnecessary. To see this, note that the shaded area in Fig. 3.1 lies between \(\sum_{i=1}^{\alpha(t)} T_i\) and \(\sum_{i=1}^{\beta(t)} T_i\), so we obtain

\[
\frac{\beta(t)}{t} \sum_{i=1}^{\beta(t)} T_i \leq N_t \leq \lambda_t T_t
\]

Assuming that \(N_t \to N, \lambda_t \to \lambda, T_t \to T\), and that the departure rate \(\beta(t)/t\) up to time \(t\) tends to the steady-state arrival rate \(\lambda\), we obtain Little’s Theorem.

The simplifying assumptions used in the preceding graphical proof can be relaxed considerably, and one can construct an analytical proof that requires only that the limits \(\lambda = \lim_{t \to \infty} \alpha(t)/t, \delta = \lim_{t \to \infty} \beta(t)/t\), and \(T = \lim_{t \to \infty} T_t\) exist, and that \(\lambda = \delta\). In particular, it is not necessary that customers are served in the order they arrive, and that the system is initially empty (see Problem 3.41). Figure 3.2 explains why the order of customer service is not essential for the validity of Little’s Theorem.

### 3.2.2 Probabilistic Form of Little’s Theorem

Little’s Theorem admits also a probabilistic interpretation provided that we can replace time averages with statistical or ensemble averages, as we now discuss. Our preceding
Figure 3.1 Proof of Little's Theorem. If the system is empty at time $t$ [$N(t) = 0$], the shaded area can be expressed both as $\int_0^t N(\tau) \, d\tau$ and as $\sum_{i=1}^{\alpha(t)} T_i$. Dividing both expressions by $t$, equating them, and taking the limit as $t \to \infty$ gives Little’s Theorem.

If $N(t) > 0$, we have

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

and assuming that the departure rate $\beta(t)/t$ up to time $t$ tends to the steady-state arrival rate $\lambda$, the same argument applies.

analysis deals with a single sample function; now we will look at the probabilities of many sample functions and other events.

We first need to clarify the meaning of an ensemble average. Let us denote

$\rho_n(t) = \text{Probability of } n \text{ customers in the system at time } t$

(waing in queue or under service)

In a typical situation we are given the initial probabilities $\rho_n(0)$ at time 0, together with enough statistical information to determine, at least in principle, the probabilities $\rho_n(t)$ for all times $t$. For example, the probability distribution of the time between two successive arrivals (the interarrival time), and the probability distribution of the customers’ service time at various parts of the queueing system may be given. Then the average number in the system at time $t$ is given by

$$\bar{N}(t) = \sum_{n=0}^{\infty} n \rho_n(t)$$

Note that both $\bar{N}(t)$ and $\rho_n(t)$ 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 a steady-state in the sense that for some \( p_n \) (independent of the initial distribution), we have

\[ \lim_{t \to \infty} p_n(t) = p_n, \quad n = 0, 1, \ldots \]

The average number in the system at steady-state is given by

\[ \bar{N} = \sum_{n=0}^{\infty} np_n \]

and we typically have

\[ \bar{N} = \lim_{t \to \infty} \bar{N}(t) \]

Regarding average delay per customer, we are typically given enough statistical information 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 \( \bar{T}_k \), typically converges as \( k \to \infty \) to a steady-state value

\[ \bar{T} = \lim_{k \to \infty} \bar{T}_k \]

To make the connection with time averages, we note that almost every system of interest to us is ergodic in the sense that the time average, \( \bar{N} = \lim_{t \to \infty} N_t \), of a sample
function is, with probability 1, equal to the steady-state average $\overline{N} = \lim_{t \to \infty} N(t)$, that is,

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

Similarly, for the systems of interest to us, the time average of customer delay $T$ is also equal (with probability 1) to the steady-state average delay $\overline{T}$, that is,

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

Under these circumstances, Little's formula, $N = \lambda T$, holds with $N$ and $T$ being stochastic averages and with $\lambda$ given by

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

The equality of long term time and ensemble averages of various stochastic processes will often be accepted in this chapter on intuitive grounds. This equality can often be shown by appealing to general results from the theory of Markov chains (see Appendix A, at the end of this chapter, which states these results without proof). In other cases, this equality, though highly plausible, requires a specialized mathematical proof. Such a proof is typically straightforward for an expert in stochastic processes but requires background that is beyond what is assumed in this book. In what follows we will generally use the time average notation $T$ and $N$ in place of the ensemble average notation $\overline{T}$ and $\overline{N}$, respectively, implicitly assuming the equality of the corresponding time and ensemble averages.
3.2.3 Applications of Little's Theorem

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

Example 3.1

If $\lambda$ is the arrival rate in a transmission line, $\bar{N}_Q$ 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 $\bar{X}$ is the average transmission time, then Little's Theorem gives the average number of packets under transmission as

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

Since at most one packet can be under transmission, $\rho$ is also the line's utilization factor, (i.e., the proportion of time that the line is busy transmitting a packet).
Example 3.2

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 3.3

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 $\alpha K$ 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

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

Here the number in the system $N(t)$ is a deterministic function of time. Its form is shown in Fig. 3.3 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 holds with $N$ viewed as a time average.