Example 3.8 Increasing the Arrival and Transmission Rates by the Same Factor

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

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

However, the average delay per packet is now $T = N/(K\lambda)$ and is therefore decreased by a factor of K. In other words, *a transmission line K times as fast will accommodate K times as many packets/sec at K times smaller average delay per packet.* This result is quite general, even applying to networks of queues. What is happening, as illustrated in Fig. 3.8, is that by increasing arrival rate and service rate by a factor K, the statistical characteristics of the queueing process are unaffected except for a change in time scale—the process is speeded up by a factor K. Thus, when a packet arrives, it will see ahead of it statistically the same number of packets as with a slower transmission line. However, the packets ahead of it will be moving K times faster.

Example 3.9 Statistical Multiplexing Compared with Time- and Frequency-Division Multiplexing

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

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

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

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

that is, m times larger than for statistical multiplexing.

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

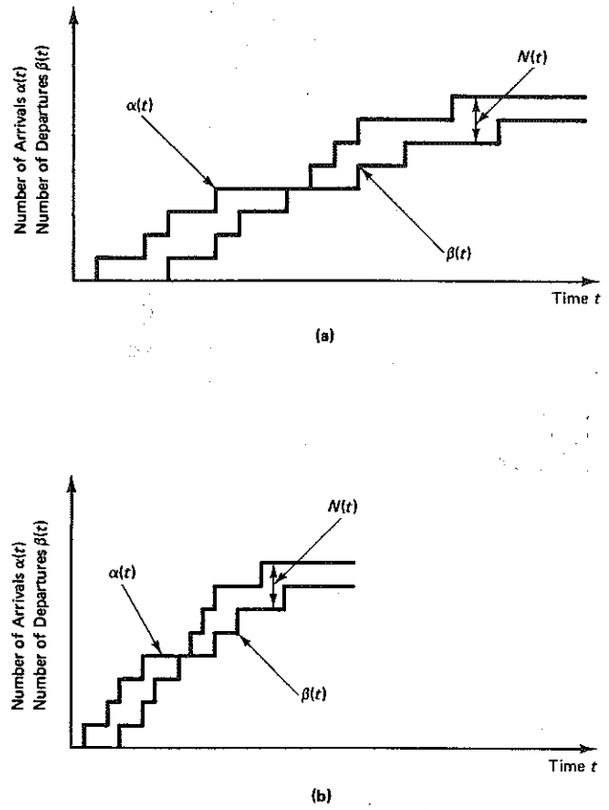

Figure 3.8 Increasing the arrival rate and the service rate by the same factor (see Example 3.8). (a) Sample paths of number of arrivals $\alpha(t)$ and departures $\beta(t)$ in the original system. (b) Corresponding sample paths of number of arrivals $\alpha(t)$ and departures $\beta(t)$ in the "speeded up" system, where the arrival rate and the service rate have been increased by a factor of 2. The average number in the system is the same as before, but the average delay is reduced by a factor of 2 since customers are moving twice as fast.

3.3.2 Occupancy Distribution upon Arrival

In our subsequent development, there are several situations where we will need a probabilistic characterization of a queueing system as seen by an arriving customer. It is

possible that the times of customer arrivals are in some sense nontypical, so that the steady-state occupancy probabilities upon arrival,

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

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

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

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

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

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

For a formal proof of the equality $a_n = p_n$ under the preceding assumption, let $A(t, t+\delta)$ be the event that an arrival occurs in the interval $(t, t+\delta)$. Let

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

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

We have, using Bayes' rule,

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

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

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

and we obtain from Eq. (3.32)

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

Taking the limit as $t \to \infty$, we obtain $a_n = p_n$.

As an example of what can happen if the arrival process is not Poisson, suppose that interarrival times are independent and uniformly distributed between 2 and 4 sec,

while customer service times are all equal to 1 sec. Then an arriving customer always finds an empty system. On the other hand, the average number in the system as seen by an outside observer looking at a system at a random time is $1/3$. (The time in the system of each customer is 1 sec, so by Little's Theorem, N is equal to the arrival rate λ, which is $1/3$ since the expected time between arrivals is 3.)

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

3.3.3 Occupancy Distribution upon Departure

Let us consider the distribution of the number of customers in the system just after a departure has occurred, that is, the probabilities

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

The corresponding steady-state values are denoted

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

It turns out that

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

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

3.4 THE $M/M/m$, $M/M/\infty$, $M/M/m/m$, AND OTHER MARKOV SYSTEMS

We consider now a number of queueing systems that are similar to $M/M/1$ in that the arrival process is Poisson and the service times are independent, exponentially dis-