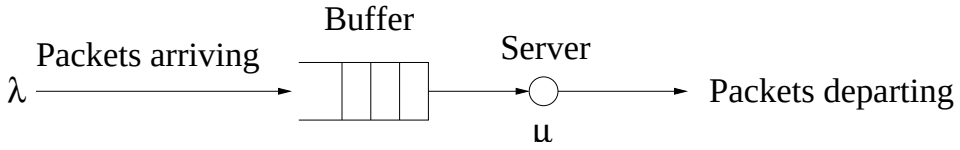
1 Queueing Delay

Consider a buffer where packets arrive at the buffer with (arrival) rate λ in packets per second, are queued and then transmitted at a rate of C bits per second (transmission rate of the link). Assuming that the average packet length is equal to $\bar{L}$, the service rate μ (in packets per second) of the buffer is equal to

$$\frac{1}{\mu} = \frac{\bar{L}}{C}.$$



In the following, we will develop a mathematical model that will allow us to study the queueing delay at such a buffer. To do this, we will make simplifying assumptions as more realistic assumptions make a meaningful analysis extremely difficult. Nevertheless, these models often provide a basis for an adequate delay approximation, and provide valuable qualitative results and worthwhile insights.

Naturally, the simpler the model the fewer conclusions we can get from it. We will start with a very simple model: the fluid flow model. This model will allow us to obtain insights regarding the stability of a queue.

1.1 Simple Model: Fluid Flow Model

In the fluid flow model, packets are assumed to be infinitesimal small and the packet arrival process is model as a (fluid) flow of rate λ. The buffer drains traffic at a rate μ. This situation is similar to a kitchen sink where the water pours into the sink via a faucet, and the sink drains water. The sink is stable (water does not spill over) as long as the rate at which water pours into the sink is equal or smaller to the rate at which the sink is able to drain water. When the rate λ is larger than μ, water will eventually will spill over (don't try this at home).

The fluid flow model provides insight regarding the stability of a buffer. As long as the packet rate λ is strictly smaller than μ, i.e. we have

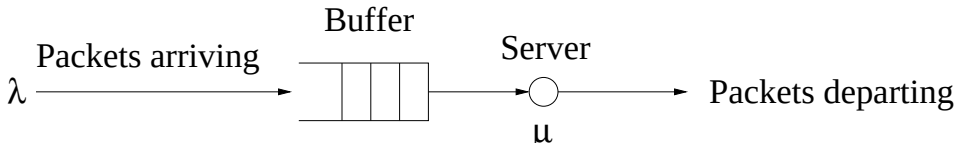
$$\lambda < \mu,$$

the buffer is stable (we will make this notion more precise in the following). However, the fluid flow model does not allow us to study the queueing delay of a packet. For this, we introduce a more sophisticated model using queueing theory.

2 Queueing Theory

Queueing systems model processes in which customers arrive, wait for their turn for service, and then depart. Supermarkets, ticket booths and checkout stands are examples of queueing systems. Here, we think of a queueing system as a buffer where packets arrive and wait until they are transmitted.

The queueing delay is the time the packets (customers) is assigned to a queue for transmission and the time it starts transmitting. During this time, the packet waits in a buffer to be serviced while other packet in the transmission queue are transmitted. This basic queueing model is illustrated by the figure below.



Unless stated otherwise, we will assume that the buffer can hold an infinite number of packets. This is a simplifying assumption capturing the case where the buffer space is large. As we will see when we consider the case of a finite buffer, this assumption can be relaxed without too much difficulties.

For the above queueing system, we are interested in determining quantities such as:

1. N, the average number of packets in the system (either in service or waiting in the queue).
2. T, the average delay of a packet (consisting of the queueing and transmission delay).

These quantities will depend on

1. the arrival rate λ and
2. the service rate μ.

2.1 Little's Theorem

Little's theorem provides a relation between the average number of packets in the system, the arrival rate, and the average delay, given by

$$N = \lambda T.$$

The beauty of Little's theorem is that it is very general. In addition, its derivation is very simple for the situation that we are interested in.

We use the following notation. Let,

- $A(t)$: number of packets that arrived in $[0, t]$
- $B(t)$: number of packets that departed in $[0, t]$
- $N(t) = A(t) - B(t)$: number of packets in the system (in queue and in service) at time t.
- T_i: Time spent in the system by the ith arriving packet.
- $N_t = \frac{1}{t}\int_0^t N(\tau)d\tau$: time average number of packets in the system up to time t.

  N_T changes with time t, but N_T tends to a steady-state N as t increases, that is,

- $N = \lim_{t\to\infty} N_t$: (time) average number of packets in the system.
- $\lambda_t = \frac{A(t)}{t}$: time average arrival rate over the interval $(0, t)$.

The steady-state arrival rate is defined as,

- $\lambda = \lim_{t\to\infty} \lambda_t$

Similarly, the time average of the customer delay up to time t is given by:

- $T_t = \frac{\sum_{i=0}^{A(t)} T_i}{A(t)}$

The steady-state time average packet delay is defined by,

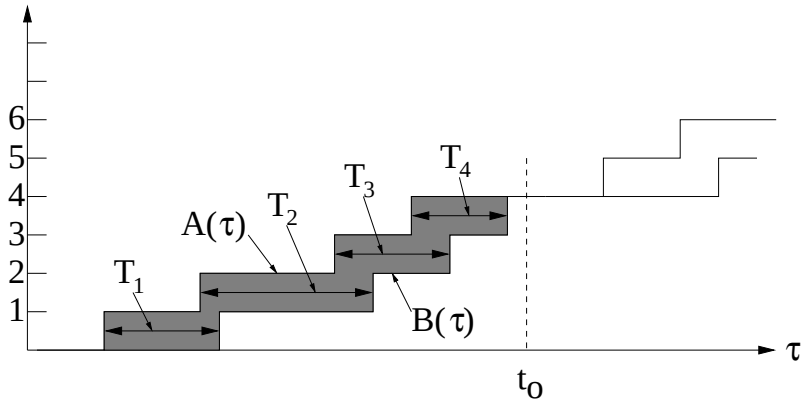
- $T = \lim_{t\to\infty} T_t$:

It turns out that the quantities N, λ and T above are related by a simple formula. This formula, known as Little's Theorem, has the form

$$N = \lambda T$$

This theorem expresses the idea that crowded systems (large N) are associated with long customer delays (large T) and vice versa. For example, on a rainy day, the traffic at rush hour moves slowly than the average (large T), while streets are more crowded (large N). Similarly, a fast-food restaurant (small T)needs a smaller waiting room (small N) than a regular restaurant for the same customer arrival rate.

Little's theorem can be proved using a graphical argument. Consider the figure below which indicates the number of arrivals $A(t)$, and the number of departures $B(t)$, up to time t as a staircase function. We assume that the system is empty at time $t = 0$ and at time t_0.



The difference $A(t) - B(t)$ is the number of packets in the system at time t, i.e. we have

$$N(t) = A(t) - B(t), \qquad t \geq 0.$$

The area between the function $A(t)$ and $B(t)$ is given by

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

At time t_0 (when the system is empty), this area is indicated by the shaded area. The shaded area is also equal to

$$\sum_{i=1}^{A(t_0)} T_i,$$

and we have

$$\int_0^{t_0} N(\tau)d\tau = \sum_{i=1}^{A(t_0)} T_i.$$

Dividing both sides by t_0, we obtain

$$\frac{1}{t_0}\int_0^{t_0} N(\tau)d\tau = \frac{1}{t_0}\sum_{i=1}^{A(t_0)} T_i. \tag{1}$$

Note that

$$\frac{1}{t_0}\int_0^{t_0} N(\tau)d\tau = N_{t_0},$$

and the left-handside in Equation 1 corresponds to the left-handside in Little's Theorem. Consequently, we would like the the right-handside in Equation 1 to correspond to the right-handside in Little's Theorem. Indeed, we have that

$$\frac{1}{t_0}\sum_{i=1}^{A(t_0)} T_i = \frac{A(t_0)}{t_0}\frac{\sum_{i=1}^{A(t)} T_i}{A(t_0)} = \lambda_{t_0} T_{t_0}.$$

Combining these two equations, we obtain that

$$N_{t_0} = \lambda_{t_0} T_{t_0}.$$

Assuming that the system becomes empty infinitely often at arbitrarily large times, and that the following limits exist,

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

then we obtain Little's theorem

$$N = \lambda T.$$

Little's theorem is valid under more general assumptions than we considered here. For example, it is not necessary that packets are served in the order that they arrive, and that the system is initially empty.

2.2 Probabilistic Formulation of Little's Theorem

Little's Theorem also admits a probabilistic interpretation provided that we replace time averages with statistical or ensemble averages. Before we go any further, we need to clarify the meaning of ensemble averages. So far we have been dealing with time-averages. This means that the system was observed for a long, long period of time. Statistical or ensemble averages refers to the system being observed at time t. For this let us denote,

- $p_n(t)$: Probability that n packets are in the system at time t (either waiting in the queue or under service).

  In a typical situation, we are given the initial probabilities $p_n(0)$ at time 0, together with enough statistical information to determine, the probabilities $p_n(t)$ for all times t. For example, the probability that how many people are present in the cinema. Then the average number of people in the cinema at time t is given by

- $\bar{N}(t) = \sum_{n=0}^{\infty} np_n(t)$: expected number of packets in the system at time t

  Note that both $\bar{N}(t)$ and $p_n(t)$ depends on t as well as the initial probability distribution $p_0(0), p_1(0), \ldots$. However, the queueing system will typically reach a steady-state in the sense that for some p_n (independent of some initial distribution), we have

- $p_n = \lim_{t\to\infty} p_n(t)$: steady-state probability that n packets are in the system

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

- $\bar{N} = \sum_{n=0}^{\infty} np_n$

  The average delay of the kth customer, denoted by $\bar{T}_k$, typically converges as $k \to \infty$ to a steady-state value

- $\bar{T} = \lim_{k\to\infty} \bar{T}_k$:

Every system that is of interest to us is ergodic in the sense that the time average, $N = \lim_{t\to\infty} N_t$, with probability 1, is equal to the steady-state average $\bar{N} = \lim_{k\to\infty} \bar{N}(t)$, that is,

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

Similarly, the time average of customer delay T, with probability 1, is also equal to the steady-state average delay $\bar{T}$, that is,

$$T = \lim_{k\to\infty} 1/k \sum_{n=0}^{\infty} T_i = \lim_{k\to\infty} \bar{T}_k = \bar{T}$$

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

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

3 Queueing Systems

Queueing systems model the situation where customers arrive at random times and queue to be served. We will concentrate mostly on infinite-buffer, single server queueing systems using first come, first out (FIFO) as a service discipline. The notation $A/B/m$ is widely used in the queueing literature for these systems where A is the interarrival-time probability density, B is the service-time probability density, and m the number of servers. The probability densities A and B are chosen from the set
M - exponential probability density (M stands for Markov)
D - all customers have the same value (D is for deterministic)
G - general (i.e. arbitrary probability density)

We will focus on $M/M/m$ systems where the first M indicates that customers arrive to a Poisson Process with rate λ, the second M indicates that the service rate is exponentially distributed with mean $1/\mu$, and m indicates that m servers are available. When $m > 1$, then a packet at the head of the buffer is routed to any server that is currently not busy, or to the first server that becomes available.

4 M/M/1 Queue

In this section, we analyze the $M/M/1$ queueing system. This system consists of a single buffer with a single server. Packets arrive according to a Poisson process with rate λ, and service times are independently and exponentially distributed with mean $1/\mu$ sec. We use the following notation.

- $A(t)$: number of packets that arrived in $[0, t]$
- $B(t)$: number of packets that departed in $[0, t]$
- $N(t) = A(t) - B(t)$: number of packets in the system (in queue and in service) at time t.
- λ : Arrival rate of the packets
- μ : Service rate of the server

4.1 Arrival Process $A(t)$

Consider the situation where time is divided into slots of length δ, where δ is assumed to be very small ($\delta << 1$). As packets arrive according to a Poisson process with rate λ, we have

$$\begin{aligned} P\{A(t+\delta) - A(t) = 0\} &= 1 - \lambda\delta + o(t), \\ P\{A(t+\delta) - A(t) = 1\} &= \lambda\delta + o(t), \\ P\{A(t+\delta) - A(t) \geq 2\} &= o(t), \end{aligned}$$

where $o(\delta)$ is a function of δ such that

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

4.2 Departure Process $B(t)$

Service times (transmission delays) are independently and exponentially distributed with parameter μ. When the system in not idle at time t, then we have

$$\begin{aligned} P\{B(t+\delta) - B(t) = 0\} &= 1 - \mu\delta + o(t), \\ P\{B(t+\delta) - B(t) = 1\} &= \mu\delta + o(t), \\ P\{B(t+\delta) - B(t) \geq 2\} &= o(t). \end{aligned}$$

To see this, note that for the interval $[t, t+\delta]$ we have

$$\begin{aligned} P\{B(t+\delta) - B(t) = 0\} &= e^{-\mu\delta} \\ &= 1 + \delta(-\mu e^{-\mu\delta} \mid_{\delta=0} + o(\delta) \\ &= 1 + \delta(-\mu) + o(\delta) \\ &= 1 - \mu\delta + o(\delta)), \end{aligned}$$

where we used the Taylor expansion for a function $f(x) : \Re \to \Re$ given by

$$f(x) = f(0) + xf'(0) + o(\delta)$$

4.3 System Dynamics $N(t)$

Let $N(t)$ be the number of packets in the system (either waiting in the buffer and being currently served) at time t. Using the above equations and assuming that $N(t) > 0$, we obtain the following transition probabilities.

$$\begin{aligned} P\{N(t+\delta) = n+1 \mid N(t) = n\} &= \Big(\lambda\delta + o(\delta)\Big)\Big(1 - \mu\delta + o(\delta)\Big) \\ &= \lambda\delta + \lambda\mu\delta^2 + o(\delta)\Big(1 - \mu\delta + o(\delta)\Big) \\ &= \lambda\delta + o(\delta), \end{aligned}$$

where we used that

$$\begin{aligned} &\lim_{\delta\to 0} \frac{\lambda\mu\delta^2}{\delta} = \delta = 0, \\ &\lim_{\delta\to 0} \frac{o(\delta)\mu\delta}{\delta} = o(\delta)\mu = 0, \\ &\lim_{\delta\to 0} o(\delta)^2 = 0. \end{aligned}$$

Similarly, the probability of 0 arrival and 1 departure in the interval $[t, t+\delta]$, is given by

$$P\{N(t+\delta) = n-1 \mid N(t) = n\} = \mu\delta + o(\delta).$$

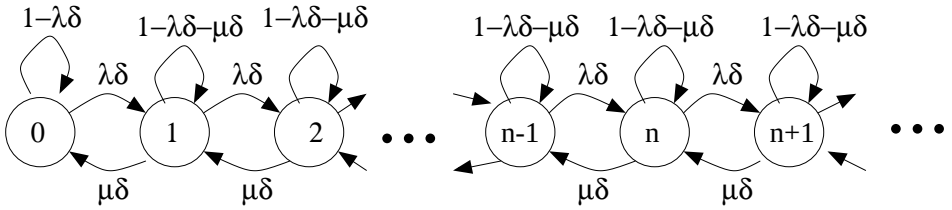
The probability of 0 arrival and 0 departure is given by

$$P\{N(t+\delta) = n \mid N(t) = n\} = 1 - \lambda\delta - \mu\delta + o(\delta)$$

To summarize, assuming that $N(t) > 0$, the system dynamics are given as follows,

$$\begin{aligned} P\{N(t+\delta) = n+1 \mid N(t) = n\} &= \lambda\delta + o(\delta), \\ P\{N(t+\delta) = n-1 \mid N(t) = n\} &= \mu\delta + o(\delta), \\ P\{N(t+\delta) = n \mid N(t) = n\} &= 1 - \lambda\delta - \mu\delta + o(\delta), \end{aligned}$$

and all other transition probabilities are of the order $o(\delta)$. Similarly, the transition probabilities an be derived for the case where $N(t) = 0$ (this is left as an exercise to the reader). Accordingly, we obtain the state transition diagram below. The state n represents that n packets are in the system. The transition probabilities shown are correct up to an $o(\delta)$ term.



4.4 Derivation of the Stationary Distribution

For the above state transition diagram, we can easily derive the steady-state probabilities p_n, $n = 0, 1, 2, ...$, that n packets are in the system, i.e.

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

We obtain the following system of equations for the steady-state probabilities,

$$\begin{aligned} p_0 &= (1 - \lambda\delta)p_0 + \mu\delta p_1, \\ p_n &= \lambda\delta p_{n-1} + (1 - \lambda\delta - \mu\delta)p_n + \mu\delta p_{n+1}, \qquad n = 1, 2, ... \end{aligned}$$

We can rewrite the first equation as

$$\mu p_1 = \lambda p_0,$$

or

$$p_1 = \rho p_0,$$

where $\rho = \lambda/\mu$. For $n = 1$, we have

$$p_1 = \lambda\delta p_0 + (1 - \lambda\delta - \mu\delta)p_1 + \mu\delta p_2,$$

or

$$\mu p_2 = (\lambda + \mu)p_1 - \lambda p_0 = (\lambda + \mu)\rho p_0 - \lambda p_0 = \rho^2 p_0.$$

Using this result recursively, we obtain

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, ...$$

There is another approach to obtain the above relation. In steady-state, the probability that the system is in state n and makes a transition to state $n + 1$ in the next transition interval has to be equal to the probability that the system is in state $n+1$ and makes a transition to state n, that is,

$$p_n \lambda\delta = p_{n+1}\mu\delta,$$

or

$$p_{n+1} = \rho p_n, \qquad n = 0, 1, 2, \ldots$$

It follows that,

$$p_n = \rho^n p_0, \qquad n = 0, 1, 2, \ldots.$$

When $\rho < 1$, that is the service rate exceeds the arrival rate, the probabilities p_n are all positive and add up to unity, so,

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

Combining the last two equations, we get

$$p_n = \rho^n (1 - \rho) \qquad n = 0, 1, 2, \ldots$$

where,

$$\rho = \frac{\lambda}{\mu} < 1$$

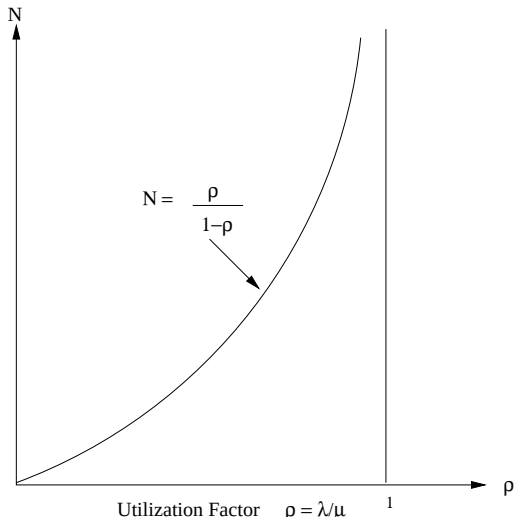
We can now calculate the expected number of packets in the system.

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

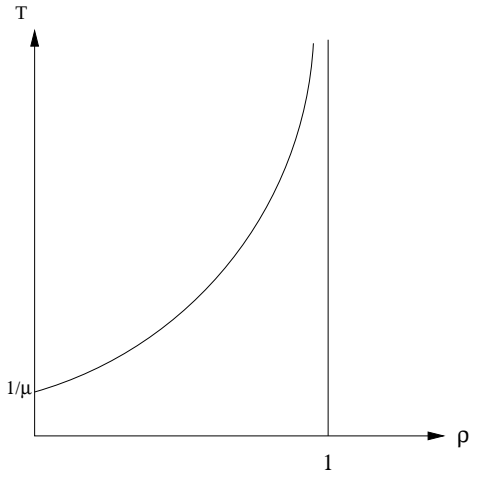
Using Little's Theorem, the average delay is given by

$$T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}.$$

The above equation illustrates that the system becomes unstable when ρ approaches 1 (system "blows up"). The graph of the average number of packets in the system as a function of ρ is shown below.



Similarly, the next figure shows the graph of the average delay T as a function of ρ. Note that the average delay T is equal to from $1/\mu$ when $\lambda = 0$. Why is this?



4.5 Utilization and Throughput

Next we compute the (system) utilization and the throughput for the above $M/M/1$ queue.

The utilization is equal to the probability P_b that the server is busy. Note that

$$P_b = 1 - p_0$$

where p_0 is the steady-state probability that the systems is idle (no packets are in service), and we obtain that

$$P_b = 1 - (1 - \rho) = \rho = \frac{\lambda}{\mu}.$$

This illustrates that the quantity $\rho = \frac{\lambda}{\mu}$ indeed has the interpretation of a utilization factor.

The throughput is equal to the (average) rate at which packets leave the system. When the server is busy, then the departure rate is equal to μ; when the server is idle then the departure rate is equal to 0. Therefore, the throughput is equal to

$$(1 - p_0)\mu = P_b\mu = \rho\mu = \lambda,$$

This result illustrates that (in a stable system, that is $\rho < 1$) all packets that enter the system will eventually leave.

4.6 Occupancy Distribution Seen by Arrivals

We are also interested in the stead-state probabilities a_n, $n = 0, 1, 2, ...$, that an arriving packets finds n packets in the system. It is important to be aware that for some arrival processes the times of packet arrivals are such the steady-state probabilities a_n are not equal to the steady-state probabilities p_n of having n packets in the system. However, for the $M/M/1$ queue we have that

$$a_n = p_n, \qquad n = 0, 1, 2, ...$$

Indeed, this property holds under very general conditions for queueing systems where packet (customers) arrive according to a Poisson process with a fixed rate λ. In particular, it holds for all systems that we consider in this course.

4.7 Scaling the Arrival/Service Rate

In the following, we study how scaling the arrival rate, and the service rate, affect the average number of packets in the system and the average delay.

4.7.1 Scaling the Arrival Rate

Assume that the arrival rate is increased from λ to $k\lambda$, where $k > 1$ is some scalar factor. Then the new link utilization ρ' is given by

$$\rho' = \frac{k\lambda}{\mu} = k\rho.$$

For this system, the new average number of packets in the system is

$$N' = \frac{\rho'}{1-\rho'} = \frac{k\rho}{1-k\rho} \geq \frac{k\rho}{1-\rho} = kN, \qquad k\rho < 1.$$

and the new average delay is

$$T' = \frac{1}{\mu - k\lambda} \geq \frac{1}{\mu - \lambda} = T, \qquad k\lambda < \mu$$

Thus, speeding up the arrival rate will increase both the average number of packets and the average delay. When the arrival rate is increased by a factor k, then the average number of packets is increased by at least a factor k.

4.7.2 Scaling the Service Rate

Next, assume that we increase the service rate from μ to $k\mu$, $k > 1$. Then the new link utilization ρ' is then equal to

$$\rho' = \frac{\lambda}{k\mu} = \frac{1}{k}\rho.$$

For this system, the average number of packets in the system is

$$N' = \frac{\rho'}{1-\rho'} = \frac{\frac{1}{k}\rho}{1-\frac{1}{k}\rho} = \frac{\rho}{k-\rho} \leq \frac{\rho}{k-k\rho} = \frac{1}{k}\frac{\rho}{1-\rho} = \frac{1}{k}N, \qquad \rho < 1$$

and the average delay is

$$T' = \frac{1}{k\mu - \lambda} \leq \frac{1}{k\mu - k\lambda} = \frac{1}{k}\frac{1}{\mu - \lambda} = \frac{1}{k}T, \qquad \lambda < \mu$$

Thus, speeding up the arrival rate by a factor k will decrease both the average number of packets, and the average delay, by at least a factor of $\frac{1}{k}$.

4.7.3 Scaling both the Arrival Rate and the Service Rate

Assume that arrival rate is increased from λ to $k\lambda$, and the service rate is increased from μ to $k\mu$, $k > 1$. Then the new link utilization is equal to

$$\rho' = \frac{k\lambda}{k\mu} = \frac{\lambda}{\mu}.$$

The average number of packets in the system is then

$$N' = \frac{\rho'}{1-\rho'} = \frac{\rho}{1-\rho} = N$$

and the average delay per packet is given by

$$T' = \frac{1}{k\mu - k\lambda} = \frac{1}{k}\frac{1}{\mu - \lambda} = \frac{1}{k}T.$$

Thus, speeding up the arrival rate, and the service rate, by a factor k will increase leave the link utilization and the average number of packets in the system unchanged, but will decrease the average delay by a factor of k. As the link utilization is unchanged, a new packet will see on average the same number of packets in the system, but as the packets are served k time as fast, the waiting time is decreased by a factor $\frac{1}{k}$.

4.8 Packet-Switching versus Circuit Switching

Assume that n independent Poisson packet streams with arrival rate $\frac{\lambda}{n}$ are multiplexed through a buffer with service rate μ (see figure below) The total arrival rate is then equal to λ; the average

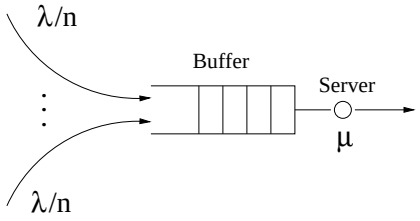


Figure 1: Packet-Switching

number of packets in the system is equal to

$$N = \frac{\rho}{1-\rho},$$

and the average delay is

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

Instead of multiplexing the packet streams (packet switching), assume that each streams is allocated a separate buffer which is served at a rate $\frac{\mu}{n}$ (circuit-switching). This situation is illustrated in the figure below. In this case, each buffer can be modeled as a $M/M/1$ queue with arrival rate

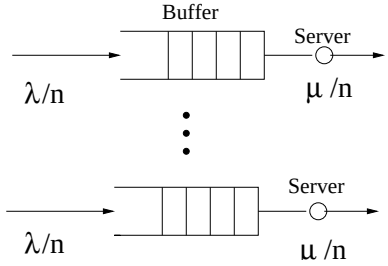


Figure 2: Circuit-Switching

$\frac{\lambda}{n}$ and service rate $\frac{\mu}{n}$. Using the results above, we obtain that the average number of packets in each of the n system is given by

$$N' = \frac{\rho}{1-\rho} = N,$$

and the average delay is

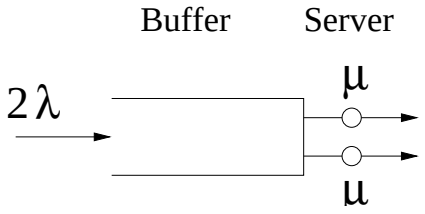
$$T' = \frac{n}{\mu - \lambda} = nT.$$

This illustrates, that using circuit switching performs very poorly (both in terms of average delay and buffer requirements) compared with packet switching when the number of separate streams becomes large.

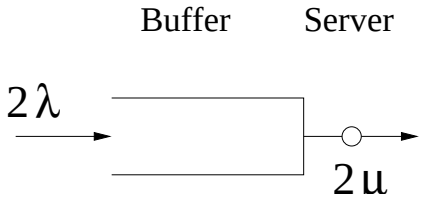
4.9 Updating versus Replacing a Server

Consider a system where packets are served by a single server at rate μ. We anticipate that the arrival rate will double from λ to 2λ, and consider two options.

(a) Updating: Add a second server to the system (see figure below).



(b) Replacing: Replace the old server with a server that is twice as fast, i.e. has a service rate 2μ.



Which option is leads to a better system performance?

4.9.1 Replacing the Server

When we replace the server, then the system can be modeled as a $M/M/1$ queue with arrival rate 2λ and service rate 2μ. Using the results above, we obtain that the average number of packets in the system is given by

$$N = \frac{\rho}{1-\rho},$$

and the average delay is

$$T = \frac{1}{2}\frac{1}{\mu - \lambda}.$$

4.9.2 Adding a Server

When we add a second server, then the system can be modeled as a M/M/2 queue (that is a queue with 2 servers) with arrival rate λ and service rate μ per servers. For this system, that state-transition diagram is given by Figure 3.

Setting $\rho = \frac{2\lambda}{2\mu} = \frac{\lambda}{\mu}$, we have

$$p_0 2\lambda\delta = p_1\mu\delta,$$

or

$$p_1 = \frac{2\lambda}{\mu} = 2\rho p_0,$$

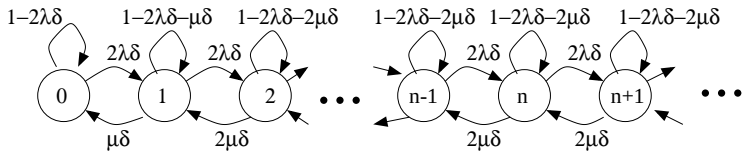


Figure 3: State diagram for $M/M/2$ queue

Similarly, for $n = 1, 2, ...$ we have

$$p_n 2\lambda\delta = p_{n+1} 2\mu\delta,$$

or

$$p_{n+1} = \frac{2\lambda}{2\mu} p_n = \frac{\lambda}{\mu} p_n = \rho p_n.$$

It follows that

$$p_n = 2\rho^n p_0, \qquad n = 1, 2, ...$$

Using the condition that

$$\sum_{n=0}^{\infty} p_n = 1,$$

we obtain that

$$\begin{aligned}
\sum_{n=0}^{\infty} p_n &= p_0 + \sum_{n=1}^{\infty} 2\rho^n p_0 \\
&= p_0 + 2p_0\rho \sum_{n=1}^{\infty} \rho^{n-1} \\
&= p_0 + 2p_0\rho \sum_{n=0}^{\infty} \rho^n \\
&= p_0 + 2p_0\rho \frac{1}{1-\rho} \\
&= p_0 \left(1 + \frac{2\rho}{1-\rho}\right) \\
&= p_0 \frac{1-\rho+2\rho}{1-\rho} \\
&= p_0 \frac{1+\rho}{1-\rho} \\
&= 1.
\end{aligned}$$

It follows that

$$\begin{aligned}
p_0 &= \frac{1-\rho}{1+\rho} \\
p_n &= 2\frac{1-\rho}{1+\rho}\rho^n = \frac{2}{1+\rho}(1-\rho)\rho^n, \qquad n = 1, 2, ...
\end{aligned}$$

Therefore, when we add a server, the average number of packet in the system N', and the average delay T', are given by

$$N' = \frac{2}{1+\rho} \cdot \frac{\rho}{1-\rho}$$

$$T' = \frac{1}{1+\rho} \cdot \frac{1}{\mu - \lambda}$$

4.10 Comparing the two Options

When λ is small (and therefore ρ is small), then we have that

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

and

$$N' \approx 2\frac{\rho}{1-\rho} = 2N.$$

Similarly, for the average delay we have that

$$T = \frac{1}{2}\frac{1}{\mu - \lambda}$$

and

$$T' \approx \frac{1}{\mu - \lambda} = 2T.$$

When λ approaches μ (and ρ approaches 1), then we have

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

and

$$N' \approx \frac{\rho}{1-\rho} = N,$$

and for the average delay we have

$$T = \frac{1}{2}\frac{1}{\mu - \lambda}$$

and

$$T' \approx \frac{1}{2}\frac{1}{\mu - \lambda} = T.$$

This means that when λ is very small and the system is lightly loaded, then replacing the server is the better option as its performance twice as good (both in terms of buffer requirement and average delay) compared with adding a second server of rate μ. When the system is heavily loaded (and ρ is close to 1), then the two options perform roughly identically. Why is that?