QUEUING ANALYSIS

Some examples of queues are:

- The number of patients in a doctor's waiting room.
- The number of customers in a store checkout line.
- The number of packets stored in a router's buffer.
- The number of print jobs present in a printer's queue.
- The number of workstations requesting access to the LAN.

Some history

- 1962, Leonard Kleinrock finished his PhD in MIT describing the underlying principles of packet-switching networks -- queuing theory. He predicted that packet switching offers the most promising model for communications between computers.
- 1968, Kleinrock's team at UCLA prepares the network measurement system, which became the first node of the ARPANET (=> Internet)
- 1969, Vint Cerf, Jon Postel, and Steve Crocken worked with Kleinrock on first "Request for comments (RFC)" documents, the design of the APRANET.
- 1973, Bob Kahn and Vint Cerf gave their first paper on TCP. (later it became two protocols TCP and IP)

Kendall's notation

This notation is represented as A/B/c/n/p, where:

A: Arrival statistics
B: Service or departure statistics
c: Number of servers
n: Buffer size
p: Customer population size

The final two fields are optional and are assumed infinite if they are omitted.

Kendall's notation

$A/B/c/n/p$

The letters A and B denoting arrival and server statistics are given the following notations:

D: **Deterministic**, process has fixed arrival or service rates
M: **Markovian**, process is Poisson or binomial
G **General**

Queues to be studied

The queues we shall deal with here will be one of the following types:

1. Single arrival, single departure infinite-size queues in which the transition matrix $\mathbf{P}$ is tridiagonal. Such a queue will be denoted by the symbols $M/M/1$.
2. Single arrival, single departure finite-size queues in which $\mathbf{P}$ is tridiagonal. Such a queue will be denoted by the symbols $M/M/1/B$.

Tridiagonal:
A **square matrix** with **nonzero** elements only on the diagonal and slots horizontally or vertically adjacent the diagonal (i.e., along the subdiagonal and superdiagonal.)

Throughput (*Th*)

Most often we are interested in estimating the rate of customers leaving the queue; which is expressed as customers per time step or customers per second. We call this rate the **average output traffic** $N_a(out)$, or **throughput** (Th) of a queue. The throughput is given by

$$Th = \text{ output data rate } = N_a(out)$$

The units of Th in the above expression are packets/time step. Notice that this definition implies that Th could could never be negative.

Efficiency or Access Probability

The **efficiency** (η) of the queue or its **access probability** (p_a) essentially measures the effectiveness of the queue at processing data present at the input.

We define the **access probability** (p_a) or **efficiency** η as the ratio of the average output traffic relative to the average input traffic.

$$p_a = \eta = \frac{N_a(out)}{N_a(in)}$$

Efficiency or Access Probability

This can be expressed in terms of the throughput

$$p_a = \eta = \frac{N_a(out)}{N_a(in)} = \frac{Th}{N_a(in)} \qquad \leq 1$$

Notice that the access probability or efficiency could never be negative and could never be more than one.

Traffic Conservation

We can write the traffic conservation as

$$N_a(in) = N_a(out) + N_a(lost)$$

where $N_a(lost)$ is the average number of lost traffic or customers per unit time.

Dividing by $N_a(in)$ to normalize we get

Traffic Conservation

Dividing the previous equation by $N_a(in)$ to normalize we get

$$\eta + L = 1$$

where L is the customer, or traffic loss probability

$$L = 1 - \eta$$

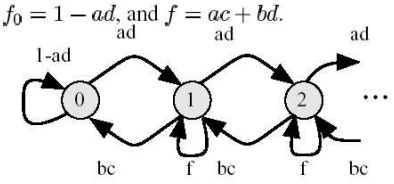
Systems that have high efficiency will have low loss probability and vice versa.

M/M/1 Queue

Assume that when a packet arrives, it could be serviced at the same time step.

$$\mathbf{P} = \begin{bmatrix} f_0 & bc & 0 & 0 & \cdots \\ ad & f & bc & 0 & \cdots \\ 0 & ad & f & bc & \cdots \\ 0 & 0 & ad & f & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where $b = 1 - a$, $d = 1 - c$, $f_0 = 1 - ad$, and $f = ac + bd$.



Analysis of M/M/1 Queue

The difference equations for steady-state distribution vector are

$$\mathbf{P}\,\mathbf{s} = \mathbf{s}$$

which produces the following difference equations

$$\begin{aligned} ad\,s_0 - bc\,s_1 &= 0 \\ ad\,s_0 - g\,s_1 + bc\,s_2 &= 0 \\ ad\,s_{i-1} - g\,s_i + bc\,s_{i+1} &= 0 \qquad i > 0 \end{aligned}$$

where $g = 1 - f$ and s_i is the probability that the system is in state i.

Analysis of M/M/1 Queue

The solution to the above equations is given as

$$s_1 = \left(\frac{a\,d}{b\,c}\right) s_0$$

$$s_2 = \left(\frac{a\,d}{b\,c}\right)^2 s_0$$

$$s_3 = \left(\frac{a\,d}{b\,c}\right)^3 s_0$$

Analysis of M/M/1 Queue

and in general

$$s_i = \left(\frac{a\,d}{b\,c}\right)^i s_0 \qquad i \geq 0$$

It is more convenient to write s_i in the form

$$s_i = \rho^i\, s_0 \qquad i \geq 0$$

Analysis of M/M/1 Queue

where ρ is the **distribution index**

$$\rho = \frac{a\,d}{b\,c} < 1$$

Analysis of M/M/1 Queue

The normalizing condition helps us find a solution to the distribution vector.

$$\sum_{i=0}^{\infty} s_i = 1$$

The solution to each component of **s** is

$$s_i = (1-\rho)\rho^i \qquad i \geq 0$$

M/M/1 Queue Performance

The average input traffic is given by

$$N_a(in) = 1 \times a + 0 \times b = a$$

M/M/1 Queue Performance

The average output traffi c is

$$N_a(out) = ac\, s_0 + \sum_{i=1}^{\infty} c\, s_i$$

M/M/1 Queue Performance

Simplifying we get

$$\begin{aligned} N_a(out) &= a\, c\, s_0 + c\,(1 - s_0) \\ &= c - bc\, s_0 \\ &= a \end{aligned}$$

M/M/1 Queue Performance

The throughput for the $M/M/1$ queue is given by

$$Th = N_a(out) = a$$

M/M/1 Queue Performance

The effi ciency of the $M/M/1$ queue is given by

$$\eta = \frac{N_a(out)}{N_a(in)} = \mathbf{1}$$

M/M/1 Queue Performance

The average queue size is given by the equation

$$Q_a = \sum_{i=0}^{\infty} i\, s_i = \frac{\rho}{1 - \rho}$$

M/M/1 Queue Performance

We can invoke Little's result to estimate the **wait time**, which is the average number of time steps a packet spends in the queue before it is routed

$$Q_a = W \times Th$$

where W is the average number of time steps that a packet spends in the queue.

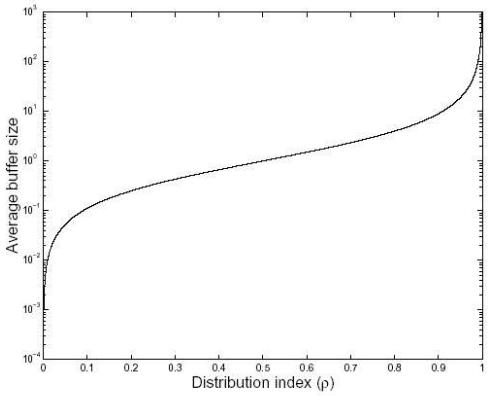
Little's Law: The average number of things in the system (queue) is the product of the average rate at which things leave the system and the average time each one spends in the system.

M/M/1 Queue Performance

Thus W is given by

$$W = \frac{\rho}{a(1-\rho)}$$

M/M/1 Queue Performance



Average queue size versus the distribution index ρ for the $M/M/1$ queue.

Example

Consider the $M/M/1$ queue with the following parameters $a = 0.6$ and $c = 0.8$. Find the equilibrium distribution vector and the queue performance.

Performance: system throughput, efficiency, average number of packets in the system, and average waiting time of each packet

Solution:

The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0.88 & 0.32 & 0 & 0 & 0 & 0 & \cdots \\ 0.12 & 0.56 & 0.32 & 0 & 0 & 0 & \cdots \\ 0 & 0.12 & 0.56 & 0.32 & 0 & 0 & \cdots \\ 0 & 0 & 0.12 & 0.56 & 0.32 & 0 & \cdots \\ 0 & 0 & 0 & 0.12 & 0.56 & 0.32 & \cdots \\ 0 & 0 & 0 & 0 & 0.12 & 0.56 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The steady-state distribution vector is

$$\mathbf{s} = \begin{bmatrix} 0.6250 & 0.2344 & 0.0879 & 0.0330 & 0.0124 & 0.0046 & \cdots \end{bmatrix}^t$$

The probability of being in state i decreases exponentially as i increases.

The queue performance is as follows:

Th	$= 0.6$	packets/time step
η	$= 1$	
Q_a	$= 0.6$	packets
W	$= 1$	time steps

Example

Investigate the queue in the previous example when the arrival probability is very close to the departure probability.

Solution:

For the queue to remain stable, we must have $a < c$. Let us try $a = 0.6$ and $c = a + 0.01$. The steady-state distribution vector is

$$\mathbf{s} = \begin{bmatrix} 0.0410 & 0.0393 & 0.0377 & 0.0361 & 0.0347 & 0.0332 & \cdots \end{bmatrix}^t$$

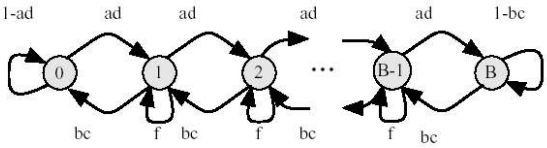
The queue performance is as follows:

Th	$= 0.6$	packets/time step
η	$= 1$	
Q_a	$= 23.4$	packets
W	$= 39$	time steps

The average queue length and the queuing delay are substantially increased.

M/M/1/B Queue

This queue is similar to the discrete-time $M/M/1$ queue except that the queue has finite size B. The state transition diagram is shown below.



State transition diagram for the discrete-time $M/M/1/B$ queue.

M/M/1/B Queue

This results in the transition matrix is given by

$$\mathbf{P} = \begin{bmatrix} f_0 & bc & 0 & \cdots & 0 & 0 & 0 \\ ad & f & bc & \cdots & 0 & 0 & 0 \\ 0 & ad & f & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & f & bc & 0 \\ 0 & 0 & 0 & \cdots & ad & f & bc \\ 0 & 0 & 0 & \cdots & 0 & ad & 1-bc \end{bmatrix}$$

where $f_0 = 1 - ad$ and $f = ac + bd$.

Analysis of M/M/1/B Queue

The difference equations for the steady-state distribution vector are obtained from the equation

$$\mathbf{P}\,\mathbf{s} = \mathbf{s}$$

Analysis of M/M/1/B Queue

The equation produces the following difference equations

$$\begin{aligned} ad\, s_0 - bc\, s_1 &= 0 \\ ad\, s_0 - g\, s_1 + bc\, s_2 &= 0 \\ ad\, s_{i-1} - g\, s_i + bc\, s_{i+1} &= 0 \qquad 0 < i < B \end{aligned}$$

where $g = ad + bc$ and s_i is the component of the distribution vector corresponding to state i.

Analysis of M/M/1/B Queue

The solution to the above equations is given as

$$\begin{aligned} s_1 &= \left(\frac{a\,d}{b\,c}\right) s_0 \\ s_2 &= \left(\frac{a\,d}{b\,c}\right)^2 s_0 \\ s_3 &= \left(\frac{a\,d}{b\,c}\right)^3 s_0 \end{aligned}$$

Analysis of M/M/1/B Queue

and in general

$$s_i = \rho^i\, s_0 \qquad 0 \le i \le B$$

where ρ is the distribution index for the $M/M/1/B$ queue

$$\rho = \frac{a\,d}{b\,c}$$

Analysis of M/M/1/B Queue

The complete solution is obtained from the above equations plus the condition

$$\sum_{i=0}^{B} s_i = 1$$

Analysis of M/M/1/B Queue

$$s_i = \frac{(1-\rho)\rho^i}{1-\rho^{B+1}} \qquad 0 \le i \le E$$

Analysis of M/M/1/B Queue

Note that ρ for the finite-sized queue *can* be more than one. In that case the queue will not be stable in the following sense:

$$s_0 < s_1 < s_2 \cdots < s_B$$

M/M/1/B Queue Performance

Throughput or output traffic for the $M/M/1/B$ queue is given by

$$\begin{aligned} Th &= N_a(out) \\ &= ac\, s_0 + \sum_{i=1}^{B} c\, s_i \\ &= ac\, s_0 + c\,(1 - s_0) \\ &= c\ (1 - b\, s_0) \end{aligned}$$

This throughput is measured in units of packets/time step.

M/M/1/B Queue Performance

The input traffic is given by

$$N_a(in) = 1 \times a + 0 \times b = a$$

Input traffic is measured in units of packets/time step.

M/M/1/B Queue Performance

The efficiency of the $M/M/1/B$ queue is given by

$$\begin{aligned} \eta &= \frac{N_a(out)}{N_a(in)} \\ &= \frac{Th}{a} \\ &= \frac{c\,(1 - b\, s_0)}{a} \end{aligned}$$

M/M/1/B Queue Performance

Data is lost in the $M/M/1/B$ queue when The above equation is simply the probability that a packet is lost which equals the probability that the queue is full, and a packet arrives, and no packets can leave.

The average lost traffic $N_a(lost)$ is given by

$$N_a(lost) = s_B\ a\ d$$

Lost traffic is measured in units of packets/time step.

M/M/1/B Queue Performance

The packet loss probability L is the ratio of lost traffic relative to the input traffic

$$L = \frac{N_a(lost)}{N_a(in)} = s_B\ d$$

M/M/1/B Queue Performance

The average queue size is given by the equation

$$\begin{aligned} Q_a &= \sum_{i=0}^{B} i\ s_i \\ &= \frac{\rho \times \left[1 - (B+1)\rho^B + B\rho^{B+1}\right]}{(1-\rho) \times (1 - \rho^{B+1})} \end{aligned}$$

Queue size is measured in units of packets.

M/M/1/B Queue Performance

Little's result is is used to estimate the average number of time steps a packet spends in the queue before it is routed.

$$Q_a = W \times Th$$

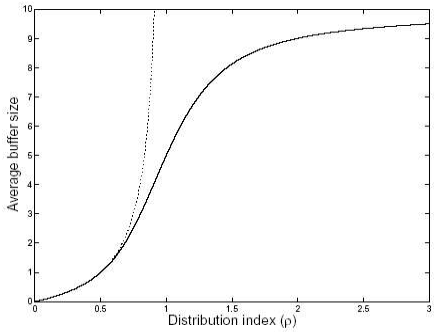
M/M/1/B Queue Performance

The the wait time is simply

$$W = \frac{Q_a}{Th}$$

Wait time is measured in units of time steps.

M/M/1/B Queue Performance



Average queue size versus the distribution index ρ for the $M/M/1/B$ queue when $B = 10$ (solid line). The dotted line is average queue size for an infinite-size $M/M/1$ queue.

Example

Consider the $M/M/1/B$ queue with the following parameters $a = 0.6$ $c = 0.8$ and $B = 4$. Find the equilibrium distribution vector and the queue performance.

Solution:

The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0.88 & 0.32 & 0 & 0 & 0 \\ 0.12 & 0.56 & 0.32 & 0 & 0 \\ 0 & 0.12 & 0.56 & 0.32 & 0 \\ 0 & 0 & 0.12 & 0.56 & 0.32 \\ 0 & 0 & 0 & 0.12 & 0.68 \end{bmatrix}$$

The steady-state distribution vector is

$$\mathbf{s} = \begin{bmatrix} 0.6297 & 0.2361 & 0.0885 & 0.0332 & 0.0125 \end{bmatrix}^t$$

The queue performance is as follows:

$N_a(out) = Th =$	0.5985	packets/time step
$\eta =$	0.9975	
$N_a(lost) =$	1.5×10^{-3}	packets/time step
$L =$	0.0025	
$Q_a =$	0.5626	packets
$W =$	0.9401	time steps

As expected we have

$$N_a(out) + N_a(lost) = N_a(in)$$

Example

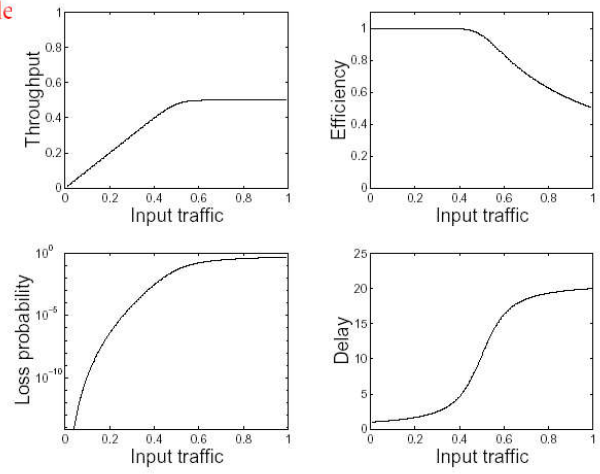
Find the performance of the queue in the previous example when the queue size becomes $B = 20$.

The queue performance is as follows.

$N_a(out) = Th =$	0.6	packets/time step
$\eta =$	1	
$N_a(lost) =$	2.2682×10^{-10}	packets/time step
$L =$	3.7804×10^{-10}	
$Q_a =$	0.6	packets
$W =$	1	time steps

We see that increasing the queue size exponentially decreases the loss probability. The throughput is not changed by much but the wait time is slightly increased due to the increased average queue size.

Example



$M/M/1/B$ throughput, efficiency, loss probability, and delay to plot versus input traffic when $B = 10$ and $c = 0.5$.

The important things to note from this example are:

1. The throughput of the queue could not exceed the maximum value for the average output traffi c.
2. The effi ciency of the queue is very close to 100% until the input traffi c approaches the maximum output traffi c c.
3. Packet loss probability is always present but starts to increase when the input traffi c approaches the packet maximum output traffi c c.
4. Packet delay increases sharply when the input traffi c approaches the packet maximum output traffi c c.
5. Congestion conditions occur as soon as the input traffi c exceeds the maximum output traffi c c. Congestion is characterized by decreased effi ciency, increased packet loss and increased delay.

Performance Bounds on M/M/1/B Queue

Under full load conditions, the $M/M/1/B$ become full and we can assume

$$\begin{aligned} a &\rightarrow 1 \\ b &\rightarrow 0 \\ s_0 &\rightarrow 0 \\ s_B &\rightarrow 1 \\ Q_a &\rightarrow B \end{aligned}$$

Performance Bounds on M/M/1/B Queue

The maximum throughput is given) by

$$\begin{aligned} Th(\max) &= N_a(out)_{\max} \\ &= c \end{aligned}$$

The departure probability is most important for determining the maximum throughput of the queue.

Performance Bounds on M/M/1/B Queue

The minimum effi ciency of the queue is given by

$$\eta(\min) = c$$

The departure probability is most important for determining the effi ciency of the queue.

Performance Bounds on M/M/1/B Queue

The maximum lost traffi c is given by

$$N_a(lost)_{\max} = d = 1 - c$$

Performance Bounds on M/M/1/B Queue

The maximum packet loss probability is given by

$$L(\max) = 1 - c$$

Performance Bounds on M/M/1/B Queue

Waiting time can be approximated by W= B/c
Larger buffer size results in longer queueing delay as expected.