A Brief Intro to QUEUEING THEORY

The "Memoryless" arrival process indicates a Poisson arrival process, in which the interarrival times have an exponential distribution.

Likewise, the "Memoryless" service process indicates that the service times have an exponential distribution.

Kendall's Notation

- \( M \): Memoryless (Markovian)
- \( E_k \): Erlang-k
- \( D \): Deterministic
- \( GI \): General Interarrival times (but i.i.d.)
- \( G \): General service times (but i.i.d.)

LITTLE's Queueing Formula

\[ L = \lambda W \]

Intuitive argument:
Suppose that you join a queue and spend \( W \) minutes before you have been served and leave.
During those \( W \) minutes, customers have been arriving and joining the queue behind you at the average rate of \( \lambda \) per minute. Thus, when you are ready to leave, you should expect to see \( \lambda W \) customers remaining in the system behind you.

You enter queue
\[ \rightarrow \quad \text{entered queue behind you} \quad \text{time } W \]

M/M/1
Interarrival times and service times both have exponential distributions, with parameters \( \lambda \) & \( \mu \), respectively.
That is, the "customers" arrive at the rate of \( \lambda \) per unit time, and are served at the rate \( \mu \) per unit of time.
It is assumed that the queue has infinite capacity, and that \( \mu > \lambda \) (so that the queue length does not tend to increase indefinitely.)
In this case, it is possible to derive the probability distribution of the number of customers in the queueing system.

\[ \pi_i = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^i \]

M/M/c

Most theoretical results have been obtained for the case in which both inter-arrival times and service times are memoryless (have exponential dist'n):

- \( \mathbb{E} \): M/M/1
- \( \mathbb{E} \): M/M/c (c > 1)
- \( \mathbb{E} \): M/M/1/N
- \( \mathbb{E} \): M/M/1/N/N

A case in which service time is not memoryless:

- \( \mathbb{E} \): M/G/1

Exercises
Using this probability distribution, we can then derive the average number of customers in the system:

\[ L = \sum_{i=0}^{\infty} i \pi_i = \sum_{i=0}^{\infty} i \left(1 - \frac{\lambda}{\mu}\right)^i \left(\frac{\lambda}{\mu}\right)^i \]

\[ L = \frac{\lambda}{\mu} \left(1 - \frac{\lambda}{\mu}\right) = \frac{\rho}{1 - \rho} \]

where \( \rho = \frac{\lambda}{\mu} < 1 \)

For the M/M/1 queueing system, Little's formula implies that:

\[ W = \frac{L}{\lambda} = \frac{\rho}{\lambda (1 - \rho)} \]

\[ W = \frac{1}{\mu - \lambda} \]

For the M/M/1 queueing system, then:

\[ W_q = W - \frac{1}{\mu} \quad \Rightarrow \quad W_q = \frac{\lambda}{\mu (\mu - \lambda)} \]

\[ L_q = \lambda W_q \quad \Rightarrow \quad L_q = \frac{\lambda^2}{\mu (\mu - \lambda)} \]

**Example**

An average of 24 trucks per 8-hour day arrive to be unloaded &/or loaded, which requires an average of 15 minutes.

The loading dock can handle only a single truck at a time.

Assume that the arrival process is Poisson, and that the service times have exponential distribution.

This loading dock is modeled as an M/M/1 queue.

**M/M/1**

\( \lambda = \text{arrival rate} = 3/\text{hour} \)

\( \mu = \text{service rate} = 4/\text{hour} \)

\( \rho = \frac{\lambda}{\mu} = 0.75 \)

Utilization of the server

Average number of trucks in system

\( L = \frac{\rho}{1 - \rho} = 0.75 \times 1.75 = 3 \)

Average time in system per truck

\( W = \frac{L}{\mu} = \frac{3}{3/\text{hr}} = 1 \text{ hr.} \)

Steady-state Behavior

**M/M/1**

\( \lambda = \text{arrival rate} = 3/\text{hour} \)

\( \mu = \text{service rate} = 4/\text{hour} \)

Average time in the queue

\( W_q = W - \frac{1}{\mu} = 1 \text{ hr.} - \frac{1}{4/\text{hr}} = 0.75 \text{ hr.} \)

Average length of the queue

\( L_q = \lambda W_q = (3/\text{hr})(0.75 \text{ hr}) = 2.25 \)

**M/M/c**

- Arrival & Service processes are memoryless, i.e., interarrival times have Exponential distribution with mean 1/\( \lambda \).
- Service times have Exponential distribution with mean 1/\( \mu \).
- Number of servers is c.
- Capacity of queueing system is infinite.

If the arrival rate \( \lambda \) is less than the combined rate \( c \mu \) at which the servers can work, then the system will have a steady-state distribution, given by:

\[ \pi_0 = \frac{1}{\sum_{n=0}^{c} \frac{(c \mu)^n}{n!}} \]

\[ \pi_j = \frac{(c \mu)^j}{j!} \pi_0 \quad j = 1, 2, \ldots, c \]

\[ \pi_j = \frac{(c \mu)^{c+j}}{c!} \pi_0 \quad j = c, c+1, \ldots \]

where \( \rho = \frac{\lambda}{c \mu} < 1 \)
Probability that all servers are busy:

\[
\sum_{j=0}^{\infty} \pi_j = \frac{(c \rho)^j}{c! (1 - \rho)} \pi_0 \quad \text{where} \quad \rho = \frac{\lambda}{c \mu} < 1
\]

This, then, is the probability that an arriving customer will be required to wait for service!

\[\rho = \frac{\lambda}{c \mu}\]

Average Length of Queue (not including those being served)

\[L_q = \frac{\rho (c \rho)^j}{c! c^j} \pi_0 \left(1 - \frac{1}{1 - \rho}\right)^2\]

Once \(L_q\) is computed, then we can compute (using Little's formula)

\[W_q = \frac{L_q}{\lambda}, \quad W = W_q + \frac{1}{\mu}, \quad L = \lambda W\]

Example: Pooled vs. Separate Servers

Compare two queuing systems:

\[\lambda = 4/\text{hr} \quad \rightarrow \quad \mu = 5/\text{hr}\]

\[\lambda = 8/\text{hr} \quad \rightarrow \quad \mu = 5/\text{hr}\]

Separate queue per server

Pooled servers

Average waiting time:

\[W_q = \frac{1}{\mu (\mu - \lambda)}\]

\[\frac{4/\text{hr}}{(5/\text{hr}) (5/\text{hr} - 4/\text{hr})} = 0.8 \text{ hr} = 48 \text{ minutes}\]

Single M/M/2 queue

\[\lambda = 8/\text{hr} \quad \rightarrow \quad \mu = 5/\text{hr}\]

Pooled servers

Rather than maintaining a separate queue for each server, customers enter a common queue.

\[\rho = \frac{\lambda}{2\mu} = \frac{8/\text{hr}}{2 \times 5/\text{hr}} = 0.8 < 1\]

which implies that a steady state exists!
Queues - a brief introduction

**M/M/1/N**

- Arrival & Service processes are **memoryless**, i.e., interarrival times have Exponential distribution with mean $1/\lambda$.
- Service times have Exponential distribution with mean $1/\mu$.
- Single server.
- **Capacity of queueing system is finite:** $N$ (including customer currently being served).
- Arriving customers **balk** when queue is full.

**Steady-state Distribution**

\[
\pi_0 = \frac{1 - \rho}{1 - \rho^{N+1}}
\]

\[
\pi_j = \rho^j \pi_0 = \rho^j \left( \frac{1 - \rho}{1 - \rho^{N+1}} \right)
\]

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

Note that $\rho$ is not restricted to be less than 1 for steady state to exist.

**Average Number of Customers in System**

\[
L = \sum_{j=0}^{N} j \pi_j
\]

\[
L = \frac{\rho \left[ 1 - (N+1)\rho^N + N\rho^{N+1} \right]}{(1 - \rho^{N+1})(1 - \rho)}
\]

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

**Average Time in System per Customer**

Little's Formula:

\[
W = \frac{L}{\lambda}
\]

\[
\rho = \frac{\lambda}{\mu} < 1
\]

Since arrival rate is zero when there are $N$ in system.

**Special Case:** $\lambda = \mu$, i.e., $\rho = \frac{\lambda}{\mu} = 1$.

Arrival rate = Service rate

All states are equally likely!

System is, on average, half-full!

**M/M/1/N/N**

- Single server.
- **Finite Source Population of size $N$**
- Arrival & Service processes are **memoryless**, i.e., service times have Exponential distribution with mean $1/\mu$.
- A departing customer returns to the queue after a time having an Exponential distribution with mean $1/\lambda$.
Each customer, after being served returns to the source population for a length of time having exponential distribution with mean \( \frac{1}{\lambda} \).

**Example**

An operator can be assigned to service (load, unload, adjust, etc.) several automatic machines in a factory.

- Running time of each machine before it must be serviced has exponential distribution, with mean 120 minutes.
- Service time has an exponential distribution with mean 12 minutes.

To achieve a desired utilization of 87.5% for the machines, how many machines should be assigned to the operator?

\[
\frac{1}{\pi_0} = 1 + 0.3 \cdot 0.06 + 0.006 = 1.366
\]

\[
\begin{align*}
\pi_0 &= \frac{1}{1.366} = 0.732965 \\
\pi_1 &= 0.3 \pi_0 = 0.2196 \\
\pi_2 &= 0.06 \pi_0 = 0.0439 \\
\pi_3 &= 0.006 \pi_0 = 0.0044
\end{align*}
\]

**M/G/1**

- Arrival process is Memoryless, i.e., interarrival times have Exponential distribution with mean \( \frac{1}{\lambda} \).
- Single server
- Service times are independent, identically distributed, but not necessarily exponential. Mean service time is \( \frac{1}{\mu} \) with variance \( \sigma^2 \).
- Queue capacity is infinite

**Steady-state Distribution**

First calculate the probability \( \pi_0 \) that the server is idle.

Other probabilities are then multiples of \( \pi_0 \)

\[
\begin{align*}
\pi_0 &= \frac{1}{N} \\
\pi_j &= \frac{N!}{(N-j)!} \rho^j \pi_0
\end{align*}
\]

where \( \rho = \frac{\lambda}{\mu} \)

This can be modeled as a M/M/1 queueing system with finite source population.

Machine operator = server

Machines = customers

\( \mu = \) 5/hour

\( \lambda = 0.5/\)hour

\[\begin{align*}
\pi_0 &= 0.732965 \\
\pi_1 &= 0.2196 \\
\pi_2 &= 0.0439 \\
\pi_3 &= 0.0044
\end{align*}\]

If 0 machines are in system, then 3 are busy processing jobs.
If 1 machine is in system, then 2 are busy processing jobs, etc.

Average utilization of the machines will be

\[
\frac{3 \pi_0 + 2 \pi_1 + \pi_2 + 0 \pi_3}{3} = 89.3\%
\]

**M/G/1**

A steady-state distribution exists if \( \rho = \frac{\lambda}{\mu} < 1 \)

i.e., if service rate exceeds the arrival rate.

\[
\begin{align*}
\pi_0 &= 1 - \rho = \text{probability that server is idle} \\
1 - \pi_0 &= \rho = \text{probability that server is busy}
\end{align*}
\]

i.e., utilization of server

There is no convenient formula for the probability of \( j \) customers in system when \( j > 0 \).
Queues - a brief intro

M/G/1

Steady-state Characteristics

\[
L_q = \frac{\lambda^2 \sigma^2 + \rho^2}{2(1 - \rho)} \quad \text{average number of customers waiting}
\]

After calculating \( L_q \), Little's Formula allows us to compute:

\[
W_q = \frac{L_q}{\lambda}, \quad W = W_q + \frac{1}{\mu}, \quad L = \lambda W = L_q + \rho
\]

For the M/M/1 queue, the standard deviation equals the mean service time, i.e., \( \sigma = \frac{1}{\mu} \) and the coefficient of variation equals 1.0

Using these formulae for the M/G/1 queueing system with \( \sigma^2 = \frac{1}{\mu^2} \) will give results consistent with the formulae for M/M/1.

\[
L_q = \frac{\rho^2}{2(1 - \rho)}
\]

The average number in the queue will be minimized when the service time is constant, i.e., \( \sigma^2 = 0 \).

In this case, the average number in the queue will be exactly half of that for the exponential distribution:

\[
L_q = \frac{\rho^2}{2(1 - \rho)}
\]

\[ \rho = \frac{\lambda}{\mu} < 1 \]

The UI Dept. of Public Safety has 5 patrol cars.

A patrol car breaks down and requires service once every 30 days.

The dept. has 2 mechanics, each of whom takes an average of 3 days to repair a car.

Time between breakdowns & repair times have exponential distribution.

What is...

the average # of patrol cars in good condition
the average down time for a car that needs repair

A small bank is trying to determine how many tellers to employ.

The total cost of employing a teller is $100/day.

A teller can serve an average of 60 customers per day (i.e., 8 minutes/customer).

An average of 50 customers per day visit the bank.

Arrivals form a Poisson process & service times have exponential distribution.

If delay cost per Poisson process & service time is $100/day (i.e., about 214/minute), how many tellers should be employed?

An average of 40 cars/hr. are tempted to use the drive-in window at the Hot Dog King.

If 5 cars (including the one at the window) are in line, no car will join the line.

It takes an average of 4 minutes to serve each car (with time having exponential dist’n)

What is...

...average # of cars waiting in line?
...# cars per hour served?
...average waiting time per car?

The mean number of customers in the system (including the one being served) is 4.41673
What fraction of the time will all 3 lanes be filled?

On the average, how many persons will be swimming?

How many lanes should be allocated to lap swimming to ensure that at most 5% of all prospective swimmers will be turned away?

- An average of 10 persons/hour arrive at the YMCA intending to swim laps.
- Each swimmer intends to swim an average of 30 minutes.
- The Y has 3 lanes open for lap swimming. Each lane can handle 2 swimmers.
- If all 3 lanes are occupied by 2 swimmers, a prospective swimmer becomes disgusted and goes running.

- The manager of an office must decide whether to rent a second copier.
- The cost of a machine is $40 per 8-hour day, whether used or not.
- An average of 4 workers/hour need to use the copier, and each uses it for an average of 10 minutes.
- Interarrival times & copying times are exponentially distributed.
- Employees are paid $8/hour, which is assumed to be the cost to the firm of a worker waiting in line for the copier. How many copiers should be rented?

### Steady State Distn

<table>
<thead>
<tr>
<th>servers - 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

Mean Queue Length (L) - 1.3333
Mean # Servers Busy - 0.66667
P(# idle servers > 1) = 0.3333

### Steady State Distn

<table>
<thead>
<tr>
<th>servers - 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Mean Queue Length (L) - 0.016667
Mean # Servers Busy - 0.4
P(# idle servers > 1) = 0.933333
Queues - a brief intro

### Queues - 3

<table>
<thead>
<tr>
<th># servers - 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Queue Length (L) = 0.0012685</td>
</tr>
<tr>
<td>Mean # Servers Busy = 0.4</td>
</tr>
<tr>
<td>P(# idle servers &gt; 1) = 0.99175</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>P</th>
<th>CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.400000</td>
<td>0.670103</td>
</tr>
<tr>
<td>1</td>
<td>0.400000</td>
<td>0.870103</td>
</tr>
<tr>
<td>2</td>
<td>0.400000</td>
<td>0.910103</td>
</tr>
<tr>
<td>3</td>
<td>0.400000</td>
<td>0.945103</td>
</tr>
<tr>
<td>4</td>
<td>0.400000</td>
<td>0.959103</td>
</tr>
<tr>
<td>5</td>
<td>0.400000</td>
<td>0.963103</td>
</tr>
<tr>
<td>6</td>
<td>0.400000</td>
<td>0.967103</td>
</tr>
<tr>
<td>7</td>
<td>0.400000</td>
<td>0.971103</td>
</tr>
<tr>
<td>8</td>
<td>0.400000</td>
<td>0.975103</td>
</tr>
<tr>
<td>9</td>
<td>0.400000</td>
<td>0.979103</td>
</tr>
<tr>
<td>10</td>
<td>0.400000</td>
<td>0.983103</td>
</tr>
<tr>
<td>11</td>
<td>0.400000</td>
<td>0.987103</td>
</tr>
<tr>
<td>12</td>
<td>0.400000</td>
<td>0.991103</td>
</tr>
<tr>
<td>13</td>
<td>0.400000</td>
<td>0.995103</td>
</tr>
<tr>
<td>14</td>
<td>0.400000</td>
<td>0.999103</td>
</tr>
<tr>
<td>15</td>
<td>0.400000</td>
<td>1.000103</td>
</tr>
</tbody>
</table>

#### An automated car wash will wash a car in 10 minutes.

- Arrivals occur an average of 15 minutes apart (exponentially distributed).
- On the average, how many cars are waiting in line for the car wash?

If the car wash could be speeded up, what wash time would reduce the average wait to 5 minutes?

- Each airline passenger & his/her luggage must be checked to prevent weapons carried onto the plane.
- At the local airport, 10 passengers/minute arrive at the checkpoint.
- A checkpoint can check 12 passengers/minute (with exponential distribution).

What is the probability that an arriving passenger must wait to be checked?
What is the average time that a passenger spends at the checkpoint?

Mean Queue Length (L) = 1.3333
Mean number of servers busy = 0.8888
Probability that at least one server is idle = 0.1111