Queueing Theory

VK
Room: M1.30
knightva@cf.ac.uk
www.vincent-knight.com

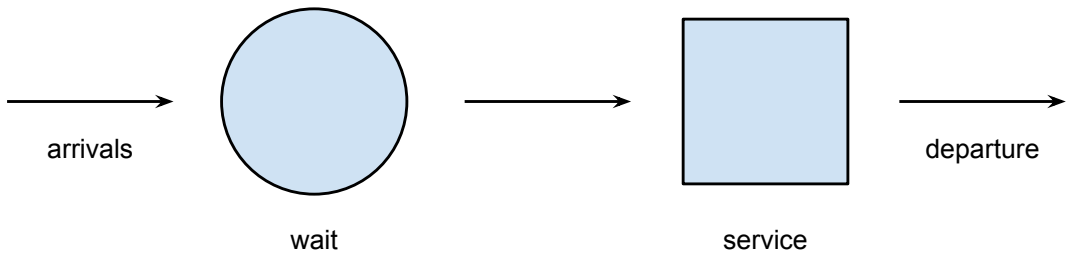
Last updated: October 17, 2013.

Overview

Description of Queueing Processes

Queueing Theory

A queueing system consists of:



- Arrival process
- Waiting regime
- Service process
- Departure process

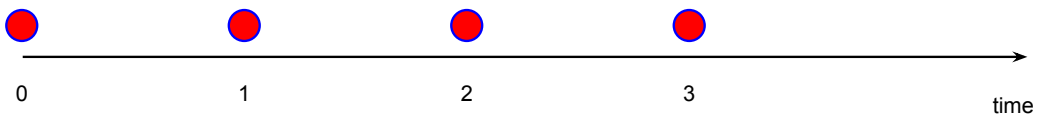
Examples

- Customers waiting to pay at the supermarket
- Patients at a medical clinic waiting to see a doctor
- Passengers waiting at the bus stop
- Aeroplanes circling an airport waiting to land
- Parts on a production line waiting for further processing

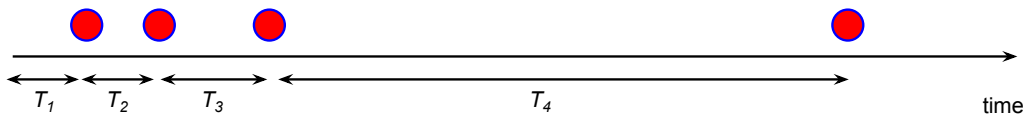
Arrival Process

The arrival process can be:

- Deterministic:



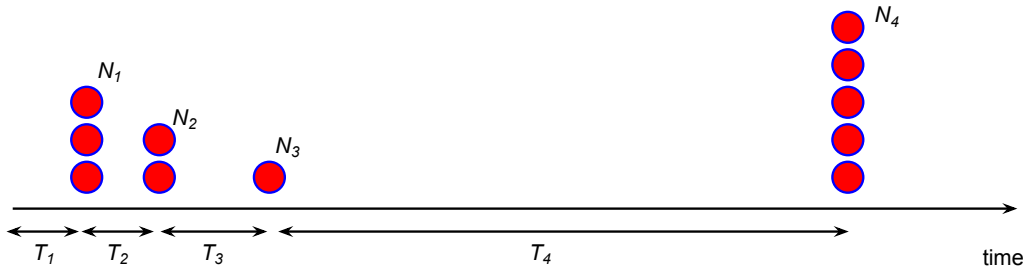
- Random:



Here $T_1, T_2, T_3, \ldots$ are random *interarrival* times.

Arrival Process

- Batched:



  The *batch sizes* N_i and *interarrival* times T_i can be deterministic or random.

- Typed arrivals: arrivals can be of different types, requiring different types of service.

We consider random arrivals on this course.

Waiting Regime

The waiting regime typically consists of a buffer size. This is just the maximum number of people/units that can wait in the queue to be served. People/units that arrive when the buffer is full are either lost to the system, or come back later.

Service Process

The service process typically consists of:

- Service times: can be deterministic, random, batched and/or depend on the type of customer. They can also depend on the queue size.
- Number of servers
- The service discipline:
  - FIFO: **F**irst **I**n **F**irst **O**out
  - LIFO: **L**ast **I**n **F**irst **O**out
  - SIRO: **S**ervice **I**n **R**andom **O**rder

Departure Process

This is the outcome of the arrival, waiting and service processes.

Examples

- Customer waiting to pay at the supermarket. Random arrivals. Multiple servers with random service times, number may depend on queue size. May have typed customers such as “8 items of less” or “pay by cash”,
- Patients at a medical clinic waiting to see a doctor. Deterministic arrivals (appointment times). Random service times.
- Passengers waiting at the bus stop. Random arrivals. Random service times with batched service.
- Aeroplanes circling an airport waiting to land. Random arrivals. Deterministic service times (approximately every 2 minutes at Heathrow).

Aims of queueing analysis

Aims of queueing analysis

In general we want to know things like:

- Average time a customer is in the system
- Average queue length
- Utilisation of servers (proportion of time busy)

These are examples of *performance measures* for the system. We may be designing or trying to improve a queueing system. We would like to be able to gauge the effect of:

- A change in arrival rate
- A change in service time
- A change in the number of servers
- A change in the service regime

Approaches to queueing analysis

There are a variety of approaches to the study of queueing systems:

- **Real World:** real-world systems provide the best information, but are expensive to experiment with.
- **Simulation:** simulation allows cheap analysis of the effects of change to a system and is the only practical way to deal with very complex queueing systems.
- **Theory:** theoretical analysis is only possible for relatively simple systems, but provides unrivalled insight into why queues behave as they do,

This course takes the theoretical approach to queueing.

Example

Suppose 3 customers arrive one just after the other with service requirements (in units of time):

$$10, 20, 30$$

In a FIFO regime the average time in the system will be:

$$\frac{10 + 30 + 60}{3} = \frac{100}{3}$$

In a LIFO regime the average time in the system will be:

$$\frac{30 + 50 + 60}{3} = \frac{140}{3}$$

The arrival process

The most important characteristic of the arrival process is:

$$\lambda = \text{ average number of arrivals per unit time}$$

Note that λ depends on the time unit you use, so that $\lambda = 2$ arrivals per **minute** is equivalent to $\lambda = 120$ arrivals per **hour**.

The arrival rate can change over time, in which case we use the notation $\lambda(t)$ for the arrival rate at time t.

We assume that arrivals follow a Poisson process for the rest of the course.

The arrival process

If an arrival process is a Poisson process, the number of arrivals occurring within any interval of time of length t, follows a **Poisson** distribution with parameter λt.

The probability that there are n arrivals in a time interval of length t is equal to:

$$p(n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \text{ for } n = 0, 1, 2, \ldots$$

and is independent of the number of units currently in the system or the history of arrivals prior to the start of the interval.

The average number of arrivals in an interval of length t is λt and the variance is also equal to λt (properties of the Poisson distribution).

Infinitesimal Arrival Rate

Consider a short interval of length γt and let $N(t, t+\gamma t)$ be the number of arrivals between t and $t+\gamma t$. From the expression for the Poisson distribution:

$$P(N(t, t+\gamma t) = 1) = (\lambda\gamma t)e^{-\lambda\gamma t} = (\lambda\gamma t)\left(1 - \lambda\gamma t + \frac{(\lambda\gamma t)^2}{2} - \dots\right)$$

where we have expanded $e^{-\lambda\gamma t}$ in a Taylor series. As $\gamma t \to 0$, $(\gamma t)^2$ and higher order terms become negligible and:

$$P(N(t, t+\gamma t) = 1) \approx \lambda\gamma t$$

$$P(N(t, t+\gamma t) = 0) \approx 1 - \lambda\gamma t$$

$$P(N(t, t+\gamma t) > 1) \approx 0$$

Inter-Arrival Times

For a poisson process, the time between events (arrivals) must follow a negative exponential distribution.

To show this, assume we start observing a Poisson process immediately after an event (arrival) and say this occurred at time 0. The probability that we have no events at time t is:

$$P(N(0,t) = 0) = e^{-\lambda t}$$

but this is equal to the probability that the time between two successive events (arrivals) is greater than t. Writing X as the time between two successive event (arrivals), this means that:

$$P(X > t) = e^{-\lambda t}$$

$$P(X \leq t) = 1 - e^{-\lambda t}$$

Inter-Arrival Times

We know that $P(X \leq t) = F(t)$, where $F(t)$ is the **cumulative density function** (cdf) of the distribution of time between events. The **probability density function** (pdf) is given by

$$f(t) = \frac{dF(t)}{dt}$$

$$f(t) = \lambda e^{-\lambda t}$$

but this is the pdf of a negative exponential distribution. Thus, inter arrival times $\sim NegExp(\lambda)$.

The memoryless property

The most important property of the negative exponential distribution is the memoryless property. This says that *if you have a NegExp(λ) inter arrival time, and have already waited time t for the next arrival, then the time remaining until the next arrival still have NegExp(λ) distribution.*

i.e. the amount of time you have waited tells you nothing about how long you still have to wait.

The exponential distribution is the **only** continuous distribution with this property. It can be used to describe e.g. the arrival of telephone calls at an exchange, the arrival of customer at a store ...

The memoryless property

The memoryless property is equivalent to the conditional probability statement:

$$P(T_i > s + t \mid T_i > s) = P(T_i > t)$$

The proof is a straight forward application of the definition of conditional probability, and the fact that $P(T_i > t) = e^{-\lambda t}$:

$$\begin{aligned} P(T_i > s + t \mid T_i > s) &= \frac{P(\ T_i > s \mid T_i > s+t\ )P(\ T_i < s+t\ )}{P(T_i > s)} \\ &= \frac{P(T_i > s+t)}{P(T_i > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T_i > t) \end{aligned}$$

Memoryless property paradox

Consider the following argument. Suppose we have a Poisson process of rate λ, and we turn up at some random time t to observe it. On average, we will arrive half way between two arrivals. Thus the expected time until the next arrival will be half the expected time between any two arrivals, that is $\frac{1}{2\lambda}$.

Memoryless property paradox

Consider the following argument. Suppose we have a Poisson process of rate λ, and we turn up at some random time t to observe it. On average, we will arrive half way between two arrivals. Thus the expected time until the next arrival will be half the expected time between any two arrivals, that is $\frac{1}{2\lambda}$.

But the memoryless property tells us that the time from our appearance to the next arrival should still be $NegExp(\lambda)$, with mean $\frac{1}{\lambda}$: a contradiction!

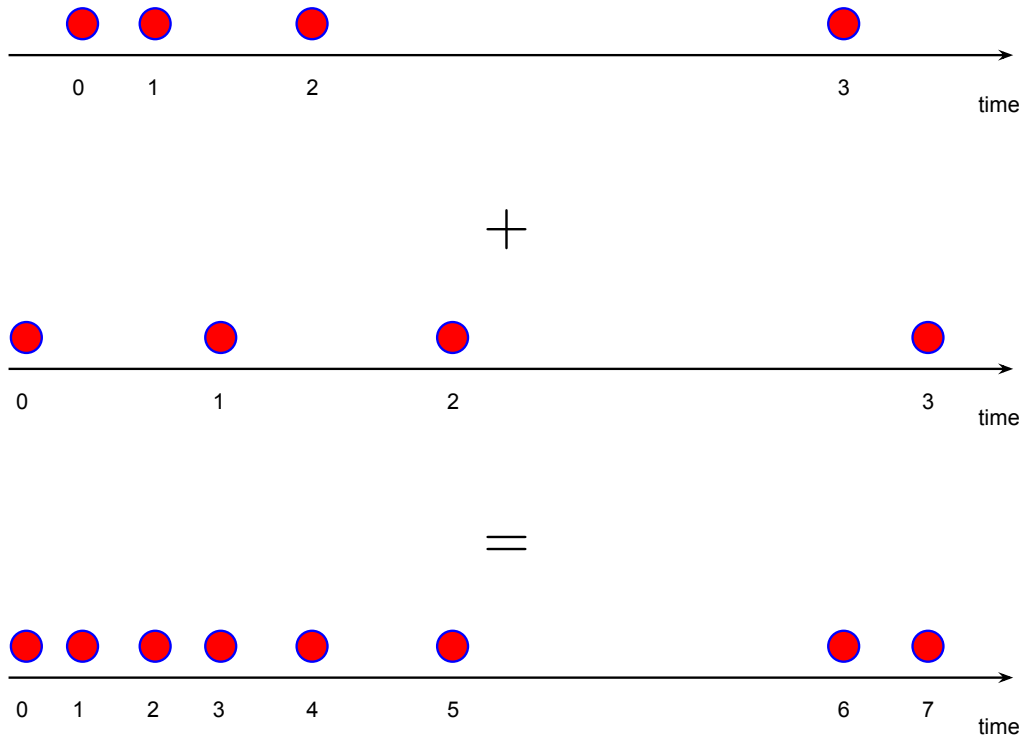
Memoryless property paradox

Of course there is a flaw in the previous argument.

If we turn up at a random time, then we are more likely to turn up between two widely spaced arrivals than between two closely spaced arrival. Thus, the inter arrival period we turn up in will on average be larger than the norm, and so its expected length will be larger than the norm (in fact, exactly twice the norm)

Merging and thinning

The poisson process has many useful properties. Two of these concern merging and thinning.

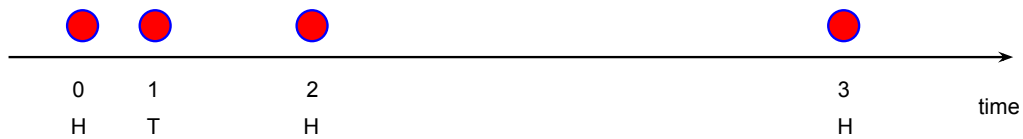


Merging and thinning

If we merge a Poisson process rate λ_1 with an independent Poisson process rate λ_2, then the result is a Poisson process rate $\lambda_1 + \lambda_2$. By merging we mean that we add all of the arrivals together.

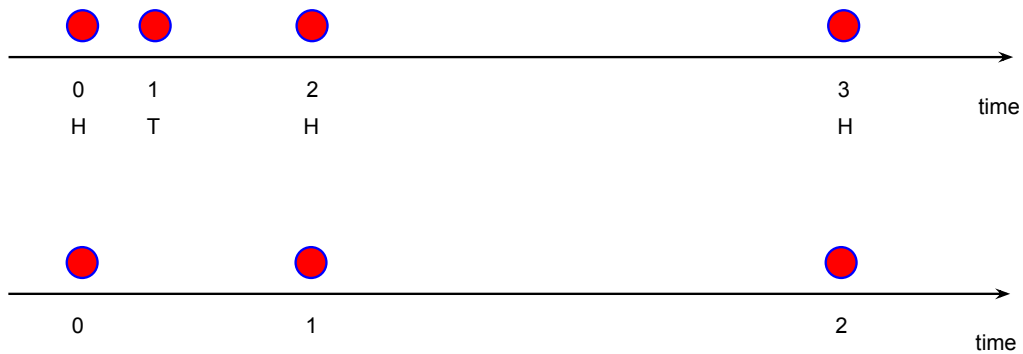
Merging and thinning

We thin a process by selecting arrivals at random. For example we could toss a coin for each arrival: heads we keep it, tails it is discarded.



Merging and thinning

We thin a process by selecting arrivals at random. For example we could toss a coin for each arrival: heads we keep it, tails it is discarded.



Merging and thinning

If we start with a Poisson process rate λ, and the probability of keeping an arrival is p, then the results is a Poisson process rate $p\lambda$.

(A proof of these results is not part of the course)

The service process

Most simple queueing models assume that that service times have a negative exponential distribution with parameter μ, so that the pdf is written as:

$$f(t) = \mu e^{-\mu t}$$

where $\frac{1}{\mu}$ is the average length of a service.

If this is true, the service process is also a Poisson process and service times will also obey the memoryless property. For example, if you arrive at a cash desk and find the server busy and the service process is a Poisson process, the expected time until the service finishes serving will be independent of how long the current service has been in progress.

The service process

What is the expected waiting time of a unit that joins the queue and finds n unites ahead of it($n-1$ in the queue and 1 being served)?

$$\text{Total expected time} = n \times \frac{1}{\mu} = \frac{n}{\mu}$$

$$\text{Variance} = n \times \frac{1}{\mu^2} = \frac{n}{\mu^2}$$

$$\text{Standard deviation} = \frac{\sqrt{n}}{\mu}$$

The distribution of the waiting times is the convolution of n negative exponential distributions. I.e. a gamma distribution with parameters (n, μ):

$$f(t) = \frac{\mu t^{n-1} e^{-\mu t}}{(n-1)!}, \ t \geq 0$$

Classification of queues

There is a classification scheme for commonly encountered queues (originally devised by David Kendall). A general queue is denoted:

$$A/B/m/n$$

where we make the following assumptions:

1. Inter-arrival times are independent and give by some distribution A.
2. Service times are independent and given by some distribution B.
3. There are m servers.
4. There is a buffer of size n.

Standard notation for distributions

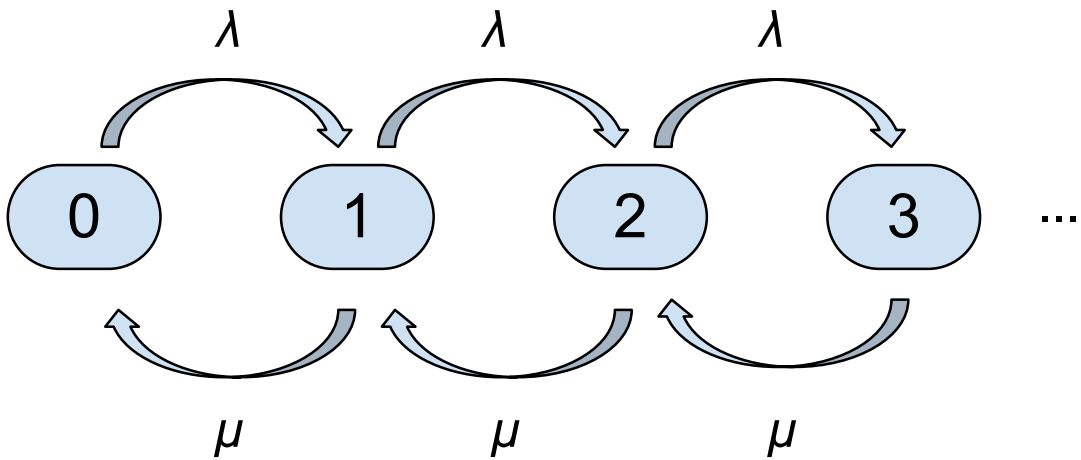
There is also some standard notation for the possible types of distribution A and B:

- M the exponential distribution (Markovian)
- D deterministic
- E_k the Erlang distribution with k stages
- H_k the hyper-exponential with k channels
- G a general distribution

The $M/M/1$ queue

The $M/M/1$ queue

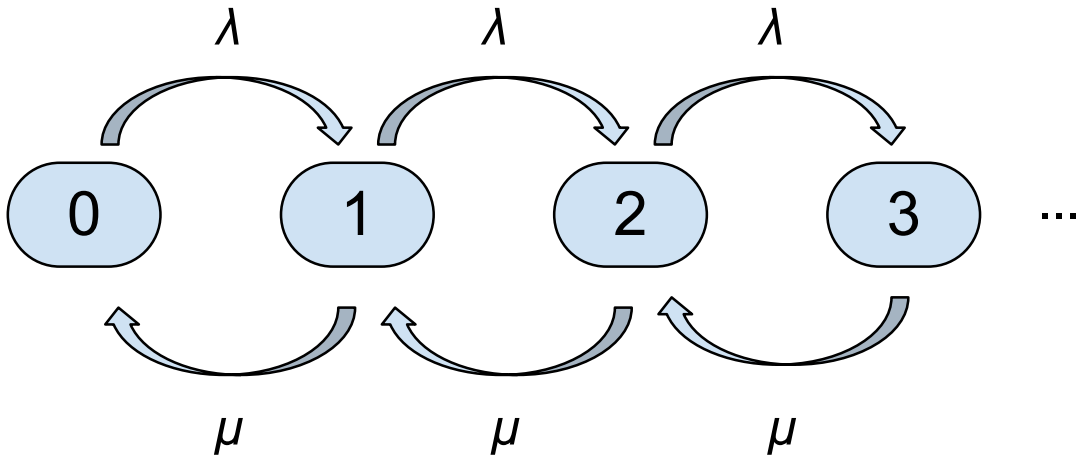
The $M/M/1$ queue has Markov arrival and service processes, one server and infinite waiting spaces. We can describe how a queueing system works using a *transition diagram*. The transition diagram for the $M/M/1$ queue is given below:



We study queues as continuous Markov chains.

Stability of queues

Recall:



This has rate matrix:

$$\begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \dots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Steady state of a single server queue

Consider a single server queue with infinite buffer, arrival rate λ and service rate μ, as shown on the previous slide.

Using the transition diagram on the previous slide, and balancing the probability flows into and out of states, the steady-state equations for this model are:

$$\pi_0\lambda = \pi_1\mu$$
$$\pi_1(\lambda+\mu) = \pi_0\lambda + \pi_2\mu$$
$$\pi_2(\lambda+\mu) = \pi_1\lambda + \pi_3\mu$$
$$\vdots$$
$$\pi_i(\lambda+\mu) = \pi_{i-1}\lambda + \pi_{i+1}\mu$$

We solve these iteratively.

Steady state of a single server queue

The first equation gives:

$$\pi_1 = \frac{\lambda}{\mu}\pi_0$$

Substituting this into the second equation gives:

$$\pi_2 = \frac{\lambda+\mu}{\mu}\frac{\lambda}{\mu}\pi_0 - \frac{\lambda}{\mu}\pi_0 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0$$

Similarly, the third equation gives: $\pi_3 = \left(\frac{\lambda}{\mu}\right)^3 \pi_0$. We postulate:

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0$$

this will be proved using an inductive process.

Proof by induction

We assume that $\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0$ is true for all $i \leq n$ and show that this implies that it must also hold for the $(n+1)^{th}$ term. We know:

$$\pi_i(\lambda + \mu) = \pi_{i-1}\lambda + \pi_{i+1}\mu$$

but $\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0$ and $\pi_{i-1} = \left(\frac{\lambda}{\mu}\right)^{i-1} \pi_0$, thus:

$$\pi_{i+1} = \frac{\lambda}{\mu}\pi_i = \left(\frac{\lambda}{\mu}\right)^{i+1} \pi_0$$

as required. We have shown that **if** the result is true for π_i and π_{i-1}, it is true for π_{i+1}. We have also shown that it is true for π_0 and π_1, therefore it **must** be true for all π_i.

Steady state of a single server queue

To find π_0, we use the additional equation $\sum_{i=0}^{\infty} \pi_i = 1$:

$$\sum_{i=0}^{\infty} \pi_i = \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i \pi_0 = \frac{\pi_0}{1 - \frac{\lambda}{\mu}} = 1$$

thus, $\pi_0 = 1 - \frac{\lambda}{\mu}$ and so $\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i$.

Note: this proof only holds for $\frac{\lambda}{\mu} < 1$ as otherwise $\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i$ does not converge. If $\lambda \geq \mu$ then the sum is infinite and there is no solution to the steady state equations. In this case the queue is **unstable**: its length grows indefinitely. (If $\lambda \geq \mu$ then customers are arriving faster than the server can deal with them.)

Steady state of a single server queue

The *traffic intensity* is given by $\rho = \frac{\lambda}{\mu}$. The $M/M/1$ queue is stable if and only if $\rho < 1$.

For a stable queue, we can use the steady state distribution to describe the behaviour of the queue:

- The proportion of time the server is busy is $1 - \pi_0 = \rho$. The proportion of time the system is idle is π_0.
- The average number of units in the system is:

$$\begin{aligned} L_c = \sum_{i=1}^{\infty} i\pi_i &= (1-\rho)\rho \sum_{i=1}^{\infty} i\rho^{i-1} \\ &= (1-\rho)\rho \frac{1}{(1-\rho)^2} \\ &= \frac{\rho}{(1-\rho)} \end{aligned}$$

Steady state of a single server queue

- The average number of units in the queue is:

$$\begin{aligned} L_q = \sum_{i=1}^{\infty}(i-1)\pi_i &= \sum_{i=1}^{\infty} i\pi_i - \sum_{i=1}^{\infty}\pi_i \\ &= L_c - (1-\pi_0) \\ &= \frac{\rho^2}{(1-\rho)} \end{aligned}$$

Steady state of a single server queue

- The average time in the queue W_q (assuming FIFO) is the average number of units in the system when the new unit arrives (L_c), multipled by the average time for one unit to be served $\left(\frac{1}{\mu}\right)$. Therefore:

$$W_q = \frac{\rho}{\mu(1-\rho)}$$

Steady state of a single server queue

- The average time spent in the system W_c (assuming FIFO) is the average spent in the queue (W_q) plus the average time it takes to service one unit. Therefore:

$$W_c = W_q + \frac{1}{\mu}$$
$$= \frac{1}{\mu(1-\rho)}$$

(note that we have used the memoryless property to tell us that when you arrive, the service time remaining for the person currently being served is still $NegExp(\lambda)$)

Little's queueing formulae

For any $G/G/m/n$ queue take:

$$\lambda = \text{arrivals per unit time}$$

$$L_c = \text{mean number in system} \quad W_c = \text{mean time in system}$$

$$L_q = \text{mean number in queue} \quad W_q = \text{mean time in queue}$$

$$L_s = \text{mean number in service} \quad W_s = \text{mean time in service}$$

then provided the queue has a long-term steady state:

$$L_c = \lambda W_c$$

$$L_q = \lambda W_q$$

$$L_s = \lambda W_s$$

Note that the units match: λ is measured in customer/time, L is measured in customers and W is measured in time.

Little's queueing formulae

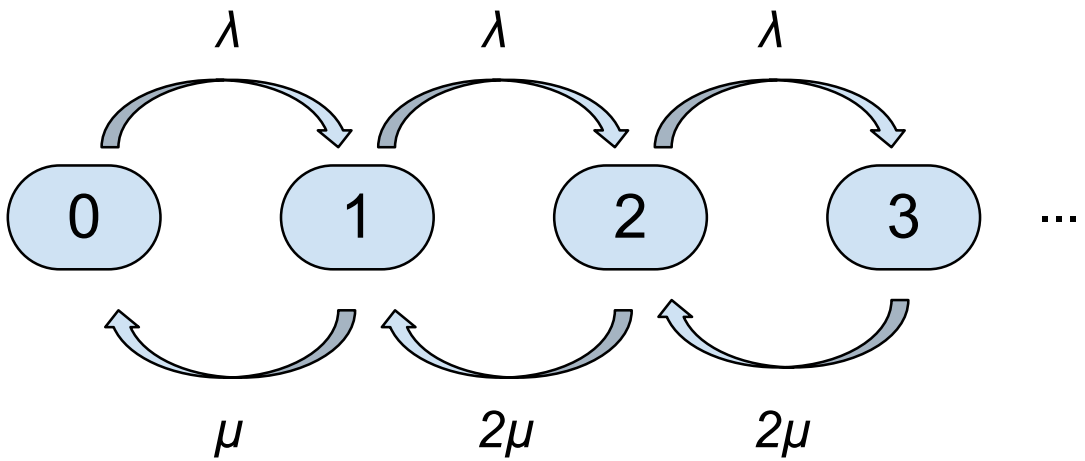
For any $M/M/1$ queue:

$$L_c = \frac{\rho}{1-\rho} \quad W_c = \frac{1}{\mu(1-\rho)}$$
$$L_q = \frac{\rho^2}{1-\rho} \quad W_q = \frac{\rho}{\mu(1-\rho)}$$
$$L_s = \rho \quad W_s = \frac{1}{\mu}$$

Multi-server queues

Multi-server queues

The $M/M/2$ queue.
Suppose we have arrivals at rate λ, and two servers who each serve at rate μ. The operational difference between two servers rate μ each, and a single server rate 2μ, is that if there is only one person in the system then only one server is active at rate μ.



Multi-server queues

The steady state equations are

$$\pi_0\lambda = \pi_1\mu$$
$$\pi_1(\lambda + \mu) = \pi_0\lambda + \pi_2 2\mu$$
$$\pi_2(\lambda + 2\mu) = \pi_1\lambda + \pi_3 2\mu$$
$$\vdots$$
$$\pi_i(\lambda + 2\mu) = \pi_{i-1}\lambda + \pi_{i+1}2\mu$$

Solving these iteratively we obtain:

$$\pi_i = \frac{\lambda^i}{2^{i-1}\mu^i}\pi_0 \text{ for all } i \geq 1$$

Multi-server queues

Setting $\sum_{i=0}^{\infty} \pi_i = 1$ gives:

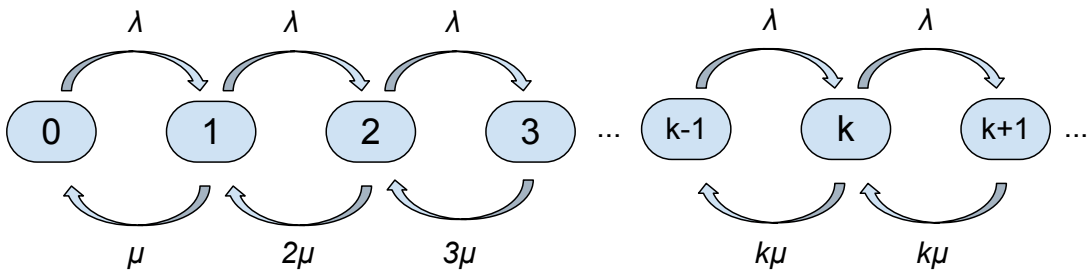
$$\begin{aligned}\sum_{i=0}^{\infty} \pi_i &= \pi_0 \left(1 + \frac{\lambda}{\mu} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{2^{i-1}\mu^{i-1}}\right) \\ &= \pi_0 \left(1 + \frac{\lambda}{\mu} \frac{1}{1 - \frac{\lambda}{2\mu}}\right) \\ &= 1\end{aligned}$$

This only works if $\lambda < 2\mu$, otherwise the sum does not converge and no solution exists. In this case we have $\rho = \frac{\lambda}{2\mu}$. So for $\rho < 1$ we have:

$$\pi_0 = \frac{1-\rho}{1+\rho} \qquad \pi_i = 2\frac{1-\rho}{1+\rho}\rho^i$$

Multi-server queues

We have arrivals at a rate λ and k servers who each serve at a rate μ.



Multi-server queues

The steady state equations are

$$\begin{aligned}
\pi_0\lambda &= \pi_1\mu \\
\pi_1(\lambda+\mu) &= \pi_0\lambda + \pi_2 2\mu \\
\pi_2(\lambda+2\mu) &= \pi_1\lambda + \pi_3 3\mu \\
&\vdots \\
\pi_k(\lambda+k\mu) &= \pi_{k-1}\lambda + \pi_{k+1}k\mu \\
&\vdots \\
\pi_i(\lambda+k\mu) &= \pi_{i-1}\lambda + \pi_{i+1}k\mu \text{ for } i > k
\end{aligned}$$

Multi-server queues

Solving these iteratively we obtain:

$$\pi_i = \begin{cases} \frac{\left(\frac{\lambda}{\mu}\right)^i}{i!}\pi_0 \text{ for } i < k \\ \frac{\left(\frac{\lambda}{\mu}\right)^i}{k!k^{i-k}}\pi_0 \text{ for } i \geq k \end{cases}$$

We now set $\sum_{i=0}^{\infty} \pi_i = 1$ to give:

$$1 = \left(\sum_{i=0}^{k-1} \frac{\left(\frac{\lambda}{\mu}\right)^i}{i!} + \sum_{i=k}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^i}{k!k^{i-k}} \right) \pi_0$$

Multi-server queues

Consider the second term in the sum:

$$\sum_{i=k}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^i}{k!k^{i-k}} = \frac{1}{k!} \sum_{i=k}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^i}{k^{i-k}}$$

Let $j = i - k$, then:

$$\begin{aligned}
\frac{1}{k!} \sum_{i=k}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^i}{k^{i-k}} &= \frac{1}{k!} \sum_{j=0}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^{j+k}}{k^j} \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} \sum_{j=0}^{\infty} \frac{\left(\frac{\lambda}{\mu}\right)^j}{k^j} \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} \sum_{j=0}^{\infty} \left(\frac{\lambda}{k\mu}\right)^j \\
&= \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!\left(1-\frac{\lambda}{k\mu}\right)}
\end{aligned}$$

Multi-server queues

Thus:

$$\pi_0 = \frac{1}{\left(\sum_{i=0}^{k-1} \frac{\left(\frac{\lambda}{\mu}\right)^i}{i!} + \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!\left(1-\left(\frac{\lambda}{k\mu}\right)\right)} \right)}$$

Multi-server queues

The probability an arrival will have to wait for service is the probability that all servers are busy:

$$P(\text{wait for service}) = \sum_{i=k}^{\infty} \pi_i = \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!\left(1 - \left(\frac{\lambda}{k\mu}\right)\right)}\pi_0$$

The expected number of servers busy is:

$$E(\text{channels busy}) = \sum_{i=1}^{k-1} i\pi_i + k\sum_{i=k}^{\infty} \pi_i = \frac{\lambda}{\mu}$$

The expected number in the queue is:

$$L_q = \sum_{i=k+1}^{\infty} (i-k)\pi_i = \frac{\left(\frac{\lambda}{\mu}\right)^{k+1}\pi_0}{kk!\left(1 - \frac{\lambda}{k\mu}\right)^2}$$

Multi-server queues

The expected number in the system is the expected number in the queue plus the expected number in service:

$$L_c = L_q + \frac{\lambda}{\mu}$$

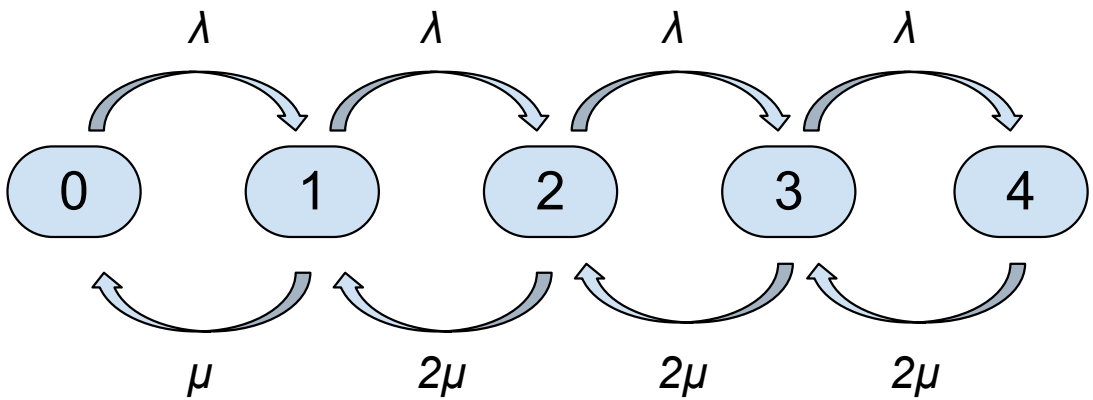
We can use Little's formula to find the expected time in the system W_c and the queue W_q:

$$W_c = \frac{L_c}{\lambda}$$
$$W_q = \frac{L_q}{\lambda}$$

Multi-server queue with finite buffer

Consider a queue with arrival rate λ, 2 servers each with service rate μ, and a finite buffer of size 2. That is, at most 2 units can wait in the queue. Arrivals when the buffer is full are assumed to be lost to the system.

Let the state be the number of customers in the system, which thus varies from 0 to 4. The transition diagram for this system is:



Multi-server queue with finite buffer

As we have a finite buffer there are a finite number of steady-state equations:

$$\pi_0\lambda = \pi_1\mu$$

$$\pi_1(\lambda + \mu) = \pi_0\lambda + \pi_2 2\mu$$

$$\pi_2(\lambda + 2\mu) = \pi_1\lambda + \pi_3 2\mu$$

$$\pi_3(\lambda + 2\mu) = \pi_2\lambda + \pi_4 2\mu$$

$$\pi_4 2\mu = \pi_3\lambda$$

Note that these equations are not independent: if you add all the equations together and cancel like terms you just get $0 = 0$. This is not a problem however, as we have the additional equation:

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

Multi-server queue with finite buffer

Putting $\rho = \frac{\lambda}{2\mu}$, the solution to this system is:

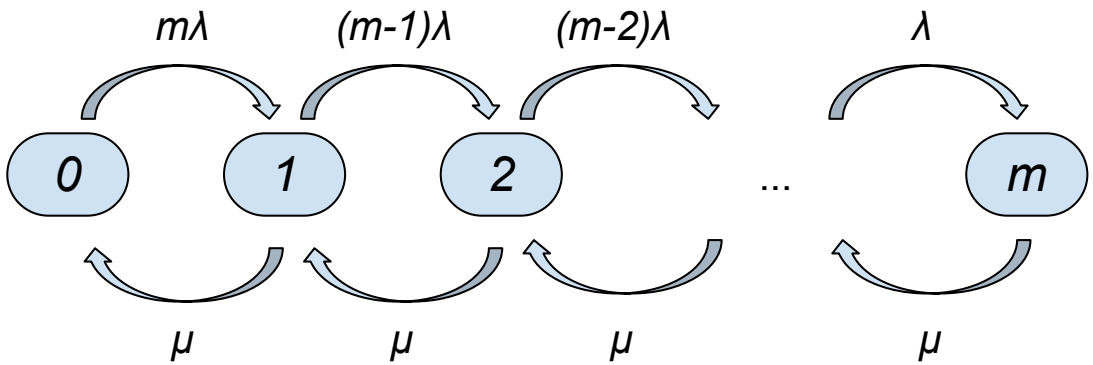
$$\pi_0 = \frac{1-\rho}{1+\rho-2\rho^5}$$
$$\pi_i = 2\rho^i \pi_0 \text{ for } i \geq 1$$

Note, in this case the steady-state equations have a solution even if $\rho \geq 1$. This is possible because the finite buffer prevents the length of the queue heading off to infinity.

Exercise: use the steady state distribution to calculate the server utility, average number in the system and average time spent waiting. What happens as $\rho \to \infty$.

Machine interference model

Consider a shop floor with m machines and a single operator, whose job is to reset/repair machines when they jam or break down. Suppose that each machine breaks down at a rate λ (average time between breakdowns is $\frac{1}{\lambda}$), and that the operator repairs machines at rate μ (average time to repair a machine is $\frac{1}{\mu}$).



Machine interference model

The steady state equations:

$$\begin{aligned}
\pi_0 m\lambda &= \pi_1 \mu \\
\pi_1((m-1)\lambda + \mu) &= \pi_0 m\lambda + \pi_2 \mu \\
&\vdots \\
\pi_{m-1}(\lambda + \mu) &= \pi_{m-2} 2\lambda + \pi_m \mu \\
\pi_m \mu &= \pi_{m-1} \lambda
\end{aligned}$$

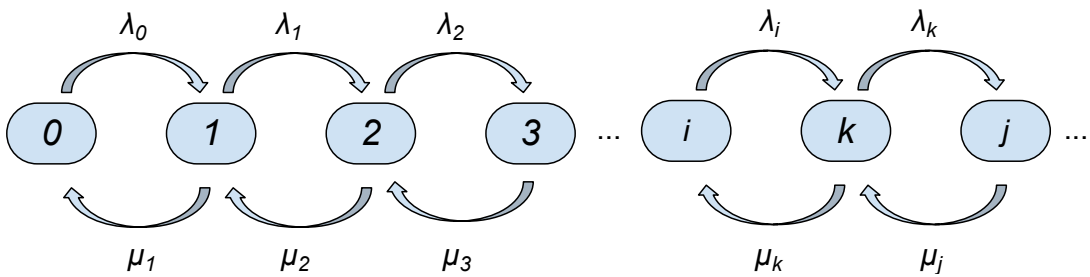
The solution to these equations is:

$$\pi_i = \frac{m!}{(m-i)!} \left(\frac{\lambda}{\mu}\right)^i \pi_0$$

General Birth-Death Process

All of the examples we have seen so far can be viewed as examples of a birth-death process. If we let the state i correspond to the size of population, then a transition $i \to i+1$ corresponds to a birth, and a transition $i \to i-1$ corresponds to a death. (Note that we allow transitions from $0 \to 1$, which correspond to immigration from a separate population.)

A general birth death process allows the birth rate and death rate to depend on the current state, i.e. we jump from $i \to i+1$ at rate λ_i, and from $i \to i-1$ at rate μ_i. The transition diagram is:



General Birth-Death Process

The steady state equations are:

$$\pi_0\lambda_0 = \pi_1\mu_1$$

$$\pi_1(\lambda_1 + \mu_1) = \pi_0\lambda_0 + \pi_2\mu_2$$

$$\vdots$$

$$\pi_i(\lambda_i + \mu_i) = \pi_{i-1}\lambda_{i-1} + \pi_{i+1}\mu_{i+1}$$

The solution to these equations is, for $i \geq 1$:

$$\pi_i = \frac{\lambda_0\lambda_1 \ldots \lambda_{i-1}}{\mu_1\mu_2 \ldots \mu_i}\pi_0$$

Thus, a steady-state distribution exists provided:

$$\sum_{i=1}^{\infty} \frac{\lambda_0\lambda_1 \ldots \lambda_{i-1}}{\mu_1\mu_2 \ldots \mu_i} < \infty$$