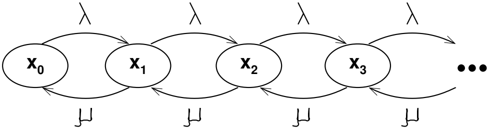
Performance Modelling — Lecture 6
Solving Queueing Models

Jane Hillston
School of Informatics
The University of Edinburgh
Scotland

2nd February 2017

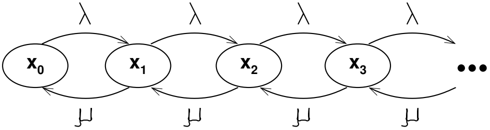
Single Queues: $M/M/1$

Consider a $M/M/1$ queue with infinite capacity:



Single Queues: $M/M/1$

Consider a $M/M/1$ queue with infinite capacity:



If we write the global balance equations for this system we can soon recognise a regular pattern emerging.

$$
\begin{aligned}
\lambda\pi_0 &= \mu\pi_1 \\
(\lambda+\mu)\pi_1 &= \lambda\pi_0 + \mu\pi_2 \\
(\lambda+\mu)\pi_2 &= \lambda\pi_1 + \mu\pi_3 \\
&\vdots
\end{aligned}
$$

Single Queues: $M/M/1$

$$
\begin{aligned}
\lambda \pi_0 &= \mu \pi_1 \\
(\lambda + \mu)\pi_1 &= \lambda \pi_0 + \mu \pi_2 \\
(\lambda + \mu)\pi_2 &= \lambda \pi_1 + \mu \pi_3 \\
&\vdots
\end{aligned}
$$

Single Queues: $M/M/1$

$$
\begin{aligned}
\lambda \pi_0 &= \mu \pi_1 \\
(\lambda + \mu)\pi_1 &= \lambda \pi_0 + \mu \pi_2 \\
(\lambda + \mu)\pi_2 &= \lambda \pi_1 + \mu \pi_3 \\
&\vdots
\end{aligned}
$$

Using simple algebra we can rewrite these:

$$
\begin{aligned}
\pi_1 &= \frac{\lambda}{\mu}\pi_0 \\
\pi_2 &= \frac{\lambda}{\mu}\pi_1 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0 \\
\pi_3 &= \frac{\lambda}{\mu}\pi_2 = \left(\frac{\lambda}{\mu}\right)^3 \pi_0 \\
&\vdots
\end{aligned}
$$

$M/M/1$: state probabilities

$$
\begin{aligned}
\pi_1 &= \frac{\lambda}{\mu}\pi_0 \\
\pi_2 &= \frac{\lambda}{\mu}\pi_1 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0 \\
\pi_3 &= \frac{\lambda}{\mu}\pi_2 = \left(\frac{\lambda}{\mu}\right)^3 \pi_0 \\
&\vdots
\end{aligned}
$$

Recalling that $\lambda/\mu = \rho$, the traffic intensity, we can see that for an arbitrary state x_i:

$$\pi_i = \rho^i \pi_0$$

$M/M/1$: normalisation condition

By the normalisation condition we know that $\sum_{i=0}^{\infty} \pi_{\mathbf{i}} = 1$.

$M/M/1$: normalisation condition

By the normalisation condition we know that $\sum_{i=0}^{\infty} \pi_{\mathbf{i}} = 1$.

Substituting the expression for $\pi_{\mathbf{i}}$ we get:

$$\sum_{i=0}^{\infty} \pi_{\mathbf{i}} = \pi_{\mathbf{0}} \sum_{i=0}^{\infty} \rho^i = \pi_{\mathbf{0}} \frac{1}{1-\rho} \qquad \text{if } \rho < 1$$

$M/M/1$: steady state probabilities

From the normalisation condition we can deduce:

$$\boldsymbol{\pi_0} = 1 - \rho$$

and it follows that the probability of being in an arbitrary state i is

$$\boldsymbol{\pi_i} = (1 - \rho)\rho^i \qquad \text{for all } i > 0.$$

$M/M/1$: symbolic evaluation

This result means that we can deduce the steady state probability of being in an arbitrary state of a $M/M/1$ queue as soon as we know the arrival rate λ and the service rate μ.

$M/M/1$: symbolic evaluation

This result means that we can deduce the steady state probability of being in an arbitrary state of a $M/M/1$ queue as soon as we know the arrival rate λ and the service rate μ.

We do not need to carry out a numerical solution of the global balance equations.

$M/M/1$: symbolic evaluation

This result means that we can deduce the steady state probability of being in an arbitrary state of a $M/M/1$ queue as soon as we know the arrival rate λ and the service rate μ.

We do not need to carry out a numerical solution of the global balance equations.

Moreover, from this steady state distribution we can derive required performance measures directly in terms of ρ.

$M/M/1$: Utilisation, U

The queue is being utilised whenever it is non-empty; in other words the utilisation, U, is $1 - \pi_{\mathbf{0}}$.

Utlisation

$$U = \rho$$

$M/M/1$: Mean number of customers in the queue, N

This is the expectation of the number of customers in the service facility as a whole, i.e.

$$N = \sum_{n=1}^{\infty} n \times \boldsymbol{\pi_n} = \sum_{n=1}^{\infty} n \times (1-\rho)\rho^n = \frac{\rho}{(1-\rho)}$$

No. in queue

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

$M/M/1$: Mean number of customers waiting, N_b

This is the expectation of the number of customers in the buffer

$$N_b = \sum_{n=1}^{\infty}(n-1) \times \boldsymbol{\pi_n} = \sum_{n=1}^{\infty}(n-1) \times (1-\rho)\rho^n$$

$$= \rho \sum_{n=1}^{\infty} n \times (1-\rho)\rho^n = \frac{\rho^2}{(1-\rho)}$$

Number in buffer

$$N_b = \frac{\rho^2}{1-\rho}$$

$M/M/1$: Mean response time, R

Using Little's Law we can calculate the mean response time of the queue to be the mean number in the queue N, divided by the arrival rate λ,

$$R = N/\lambda = \frac{\rho}{1-\rho} \times \frac{1}{\lambda} = \frac{1/\mu}{1-\rho} = \frac{1}{\mu(1-\rho)}.$$

Response time

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

Other single queues

We can derive symbolic steady state distributions, and expressions for performance measures, for the other standard queues.

Other single queues

We can derive symbolic steady state distributions, and expressions for performance measures, for the other standard queues.

Thus, given almost any single queue model, in order to derive a performance measure it is only necessary to select the appropriate formula from a table and evaluate it using the parameters of your model.

Other single queues

We can derive symbolic steady state distributions, and expressions for performance measures, for the other standard queues.

Thus, given almost any single queue model, in order to derive a performance measure it is only necessary to select the appropriate formula from a table and evaluate it using the parameters of your model.

Some examples are in Lecture Note 6 and most textbooks on performance models will contain these formulae.

Example

- Consider again the wireless access gateway discussed previously.

Example

- Consider again the wireless access gateway discussed previously.
- Measurements have shown that packets arrive at a mean rate of 125 packets per second, and are buffered until they can be transmitted.

Example

- Consider again the wireless access gateway discussed previously.
- Measurements have shown that packets arrive at a mean rate of 125 packets per second, and are buffered until they can be transmitted.
- The gateway takes 2 milliseconds on average to transmit a packet.

Example

- Consider again the wireless access gateway discussed previously.
- Measurements have shown that packets arrive at a mean rate of 125 packets per second, and are buffered until they can be transmitted.
- The gateway takes 2 milliseconds on average to transmit a packet.
- The gateway currently has 13 places (including the packet being transmitted) and packets that arrive when the buffer is full are lost.

Example

- Consider again the wireless access gateway discussed previously.
- Measurements have shown that packets arrive at a mean rate of 125 packets per second, and are buffered until they can be transmitted.
- The gateway takes 2 milliseconds on average to transmit a packet.
- The gateway currently has 13 places (including the packet being transmitted) and packets that arrive when the buffer is full are lost.
- We wish to find out if the buffer capacity is sufficient to ensure that less than one packet per million gets lost.

Example

- We represent the gateway as a $M/M/1/13$ queue, with $\lambda = 125$ and $\mu = 1/0.002 = 500$.

Example

- We represent the gateway as a $M/M/1/13$ queue, with $\lambda = 125$ and $\mu = 1/0.002 = 500$.
- The utilisation of the gateway will be $\rho = \lambda/\mu = 0.25$.

Example

- We represent the gateway as a $M/M/1/13$ queue, with $\lambda = 125$ and $\mu = 1/0.002 = 500$.
- The utilisation of the gateway will be $\rho = \lambda/\mu = 0.25$.
- The loss rate is the arrival rate multiplied by the probability that the system is full, i.e. $\lambda \times \pi_{\mathbf{K}}$.

Example

- We represent the gateway as a $M/M/1/13$ queue, with $\lambda = 125$ and $\mu = 1/0.002 = 500$.
- The utilisation of the gateway will be $\rho = \lambda/\mu = 0.25$.
- The loss rate is the arrival rate multiplied by the probability that the system is full, i.e. $\lambda \times \pi_{\mathbf{K}}$.
- The proportion of lost packets is $\boldsymbol{\pi_K}$ (expected no. lost per time unit divided by the expected no. arriving per time unit)

$$\boldsymbol{\pi_K} = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} = (0.75 \times 0.25^{13})/(1 - 0.25^{14})$$
$$= (1.1176 \times 10^{-8})/0.99999999 = 1.12 \times 10^{-8}$$

Example

- We represent the gateway as a $M/M/1/13$ queue, with $\lambda = 125$ and $\mu = 1/0.002 = 500$.
- The utilisation of the gateway will be $\rho = \lambda/\mu = 0.25$.
- The loss rate is the arrival rate multiplied by the probability that the system is full, i.e. $\lambda \times \pi_K$.
- The proportion of lost packets is $\boldsymbol{\pi_K}$ (expected no. lost per time unit divided by the expected no. arriving per time unit)

$$\boldsymbol{\pi_K} = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} = (0.75 \times 0.25^{13})/(1-0.25^{14})$$
$$= (1.1176 \times 10^{-8})/0.99999999 = 1.12 \times 10^{-8}$$

- Thus the expected proportion of packets lost is 0.0112 every million packets, well within the requirement.

Networks of queues

If we consider a network of queues rather than a single queue the possible state space of the underlying Markov process become much more diverse.

Networks of queues

If we consider a network of queues rather than a single queue the possible state space of the underlying Markov process become much more diverse.

So we would not expect to derive symbolic steady state distributions of wide applicability in the same way as we have done for single queues.

Networks of queues

If we consider a network of queues rather than a single queue the possible state space of the underlying Markov process become much more diverse.

So we would not expect to derive symbolic steady state distributions of wide applicability in the same way as we have done for single queues.

However, for a large class of queueing networks a straightforward and efficient means of solving models has been found. These networks are known as product form networks.

Product form queueing networks

The term product form comes from the fact that the steady state distribution of these models can be derived as the product of the steady state distributions of the service centres in the network.

Product form queueing networks

The term product form comes from the fact that the steady state distribution of these models can be derived as the product of the steady state distributions of the service centres in the network.

In a queueing network the state of the system is characterised by the number of customers waiting at each of the service centres, usually represented as a tuple.

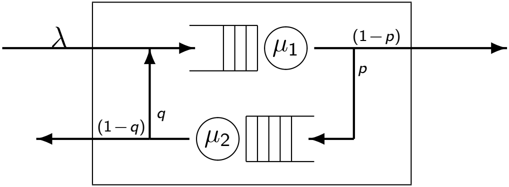
Product form queueing networks

The term product form comes from the fact that the steady state distribution of these models can be derived as the product of the steady state distributions of the service centres in the network.

In a queueing network the state of the system is characterised by the number of customers waiting at each of the service centres, usually represented as a tuple.

For example, in a simple queueing network with two service centres, the state (n_1, n_2) indicates that there are n_1 customers in service centre 1 (queueing or in service) and n_2 customers in service centre 2.

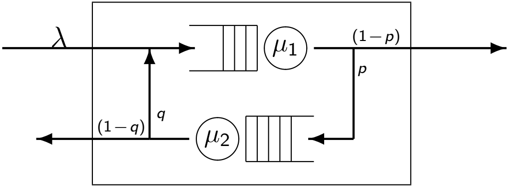
Product form distributions



For this model to have a product form steady state distribution means that the distribution can be expressed as a product of terms representing the steady states of each of the service centres considered in isolation, e.g.

$$\pi(n_1, n_2) = \boldsymbol{\pi_1}(n_1) \times \boldsymbol{\pi_2}(n_2)$$

Product form distributions



For this model to have a product form steady state distribution means that the distribution can be expressed as a product of terms representing the steady states of each of the service centres considered in isolation, e.g.

$$\pi(n_1, n_2) = \boldsymbol{\pi_1}(n_1) \times \boldsymbol{\pi_2}(n_2)$$

At steady state the service centres behave independently.

Traffic Equations

For a product form network the steady state distribution can be derived from the steady state distributions of the individual queues, which in turn can be derived from the traffic intensity.

Traffic Equations

For a product form network the steady state distribution can be derived from the steady state distributions of the individual queues, which in turn can be derived from the traffic intensity.

This means that we need to find the arrival rate λ_i at each queue i, in addition to the service rate μ_i which we are usually given.

Traffic Equations

For a product form network the steady state distribution can be derived from the steady state distributions of the individual queues, which in turn can be derived from the traffic intensity.

This means that we need to find the arrival rate λ_i at each queue i, in addition to the service rate μ_i which we are usually given.

The effective arrival rate at each queue must be derived taking into consideration any external arrivals and arrivals from other queues.

Traffic Equations

For a product form network the steady state distribution can be derived from the steady state distributions of the individual queues, which in turn can be derived from the traffic intensity.

This means that we need to find the arrival rate λ_i at each queue i, in addition to the service rate μ_i which we are usually given.

The effective arrival rate at each queue must be derived taking into consideration any external arrivals and arrivals from other queues.

A series of established theorems, the decomposition principle of Poisson/exponential distributions, and simple algebra, can help us to work out the arrival rate at each node in the network.

Burke's Theorem

Burke's Theorem

A Poisson arrival process at a service centre with exponential service rates generates a Poisson process of departures. Moreover the rate of departure is the same as the arrival rate.

Burke's Theorem

Burke's Theorem

A Poisson arrival process at a service centre with exponential service rates generates a Poisson process of departures. Moreover the rate of departure is the same as the arrival rate.

This result implies that each service centre in a chain of simple exponential service centres with Poisson arrivals can be analysed independently using results from simple queues.

Jackson's Theorem

Jackson generalised Burke's Theorem to an arbitrary network of exponential service centres each of which is driven by a Poisson arrival process.

Jackson's Theorem

Jackson generalised Burke's Theorem to an arbitrary network of exponential service centres each of which is driven by a Poisson arrival process.

Although the presence of feedback paths destroys the Poisson nature of the service centre arrivals, Jackson showed that they still behave as if they were Poisson driven.

Jackson's Theorem

Jackson generalised Burke's Theorem to an arbitrary network of exponential service centres each of which is driven by a Poisson arrival process.

Although the presence of feedback paths destroys the Poisson nature of the service centre arrivals, Jackson showed that they still behave as if they were Poisson driven.

Moreover the relationship between the arrival stream and the output stream is maintained. The networks considered by Jackson were open.

Jackson's Theorem

Jackson generalised Burke's Theorem to an arbitrary network of exponential service centres each of which is driven by a Poisson arrival process.

Although the presence of feedback paths destroys the Poisson nature of the service centre arrivals, Jackson showed that they still behave as if they were Poisson driven.

Moreover the relationship between the arrival stream and the output stream is maintained. The networks considered by Jackson were open.

Again, this implies that each service centre in the queueing network can be analysed independently.

Gordon and Newell's Theorem

Gordon and Newell considered a modification of Jackson's networks to networks which are closed rather than open.

Gordon and Newell's Theorem

Gordon and Newell considered a modification of Jackson's networks to networks which are closed rather than open.

Closed queueing networks are not driven by any external Poisson arrival process. Instead, the number of customers in the network is fixed, K.

Gordon and Newell's Theorem

Gordon and Newell considered a modification of Jackson's networks to networks which are closed rather than open.

Closed queueing networks are not driven by any external Poisson arrival process. Instead, the number of customers in the network is fixed, K.

Gordon and Newell's product form equation has the form:

$$\pi(n_1, n_2, \ldots, n_k) = \frac{1}{G(K)} \prod_{i=1}^{n} \pi_i(n_i)$$

where $G(K)$ is a normalisation constant chosen to ensure that the steady state probabilities sum to 1.

Traffic equations: implications of these results

If the arrival rate at queue i is λ_i, the departure rate will also be λ_i.

Traffic equations: implications of these results

If the arrival rate at queue i is λ_i, the departure rate will also be λ_i.

So, if all the departures from queue i go directly to queue $i+1$ the arrival rate at queue $i+1$ will also be λ_i, i.e. $\lambda_{i+1} = \lambda_i$.

Traffic equations: implications of these results

If the arrival rate at queue i is λ_i, the departure rate will also be λ_i.

So, if all the departures from queue i go directly to queue $i+1$ the arrival rate at queue $i+1$ will also be λ_i, i.e. $\lambda_{i+1} = \lambda_i$.

By the decomposition principle we know that if the departure stream is split and only goes to queue $i+1$ with probability p, then the arrival rate at queue $i+1$ will be $\lambda_{i+1} = p \times \lambda_i$

Traffic equations

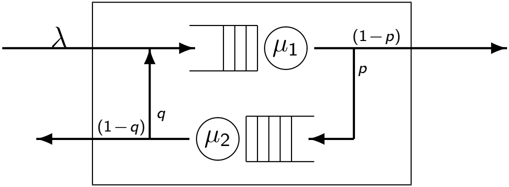
If we analyse all the service centres of a queueing network in this way, expressing its input stream as a sum of output streams from the environment or other service centres, we obtain what are known as the traffic equations.

Traffic equations

If we analyse all the service centres of a queueing network in this way, expressing its input stream as a sum of output streams from the environment or other service centres, we obtain what are known as the traffic equations.

If there are n service centres, we will have n equations in n unknown and solving the traffic equations we can find the arrival rate at each service centre.

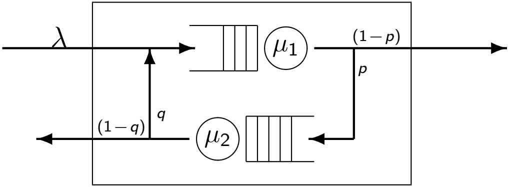
Traffic equations: example



For this simple network the traffic equations are:

$$\begin{aligned} \lambda_1 &= \lambda + q \times \lambda_2 \\ \lambda_2 &= p \times \lambda_1 \end{aligned}$$

Traffic equations: example

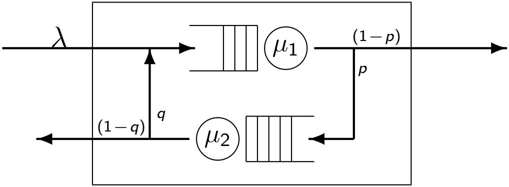


For this simple network the traffic equations are:

$$\lambda_1 = \lambda + q \times \lambda_2$$
$$\lambda_2 = p \times \lambda_1$$

Assume that $\lambda = 10$, $p = 0.5$ and $q = 0.4$.

Traffic equations: example

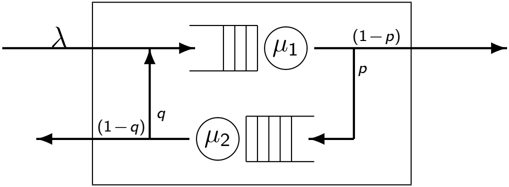


For this simple network the traffic equations are:

$$\begin{aligned}\lambda_1 &= \lambda + q \times \lambda_2 \\ \lambda_2 &= p \times \lambda_1\end{aligned}$$

Assume that $\lambda = 10$, $p = 0.5$ and $q = 0.4$.
Substituting $\lambda_2 = 0.5\lambda_1$ into the first equation: $\lambda_1 = 10 + 0.2\lambda_1$,

Traffic equations: example

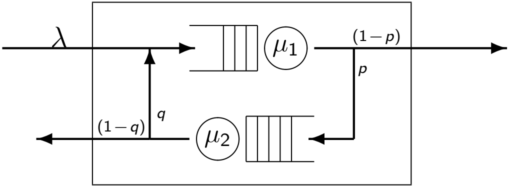


For this simple network the traffic equations are:

$$\lambda_1 = \lambda + q \times \lambda_2$$
$$\lambda_2 = p \times \lambda_1$$

Assume that $\lambda = 10$, $p = 0.5$ and $q = 0.4$.
Substituting $\lambda_2 = 0.5\lambda_1$ into the first equation: $\lambda_1 = 10 + 0.2\lambda_1$, i.e. $0.8\lambda_1 = 10$.

Traffic equations: example



For this simple network the traffic equations are:

$$\begin{aligned}\lambda_1 &= \lambda + q \times \lambda_2 \\ \lambda_2 &= p \times \lambda_1\end{aligned}$$

Assume that $\lambda = 10$, $p = 0.5$ and $q = 0.4$.
Substituting $\lambda_2 = 0.5\lambda_1$ into the first equation: $\lambda_1 = 10 + 0.2\lambda_1$, i.e. $0.8\lambda_1 = 10$.

Therefore $\lambda_1 = 12.5$ and $\lambda_2 = 6.25$.

Traffic equations vs. Global balance equations

The great advantage of solving traffic equations rather than the global balance equations is that the number of equations we need to solve grows linearly with the number of service centres, rather than exponentially, which is the case for global balance equations.

Assumptions

The assumptions are essentially those that we have seen previously with respect to Markov process, although they can be made more specific to the features of queueing networks.

Assumptions

The assumptions are essentially those that we have seen previously with respect to Markov process, although they can be made more specific to the features of queueing networks.

One exception is that we can now consider models with an infinite number of states since we do not need to numerically solve the global balance equations.

Assumptions

- Each service centre is flow balanced — the number of customers that arrive at each centre is equal to the number who depart.

Assumptions

- Each service centre is flow balanced — the number of customers that arrive at each centre is equal to the number who depart.
- The system exhibits one step behaviour — no two customers in the system "change state" at exactly the same time.

Assumptions

- Each service centre is flow balanced — the number of customers that arrive at each centre is equal to the number who depart.
- The system exhibits one step behaviour — no two customers in the system "change state" at exactly the same time.
- The system has routing homogeneity — the routing behaviour of customers is independent of the current queue lengths at both the source and the destination service centre.

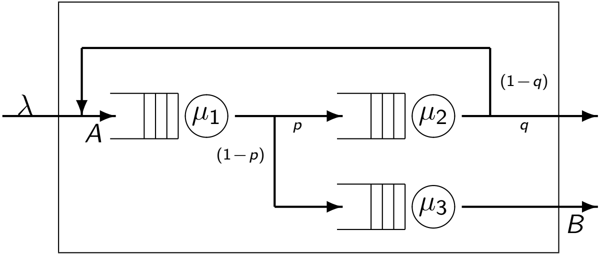
Assumptions

- Each service centre is flow balanced — the number of customers that arrive at each centre is equal to the number who depart.
- The system exhibits one step behaviour — no two customers in the system “change state” at exactly the same time.
- The system has routing homogeneity — the routing behaviour of customers is independent of the current queue lengths at both the source and the destination service centre.
- The system has device homogeneity — the service rate of customers at a service centre may depend on the number of jobs at that centre, but not more generally on the placement of customers within the network.

Assumptions

- Each service centre is flow balanced — the number of customers that arrive at each centre is equal to the number who depart.
- The system exhibits one step behaviour — no two customers in the system "change state" at exactly the same time.
- The system has routing homogeneity — the routing behaviour of customers is independent of the current queue lengths at both the source and the destination service centre.
- The system has device homogeneity — the service rate of customers at a service centre may depend on the number of jobs at that centre, but not more generally on the placement of customers within the network.
- The system experiences homogeneous external arrivals — the times at which arrivals from outside the network occur may not depend on the number or placement of customers within the network.

Exercise



1. Write down the traffic equations for the network.
2. If the value of λ is 9, $p = 0.2$ and $q = 0.5$, what is the effective arrival rate at point A?
3. Using the same values, what is the rate of external departures at point B in the network? Explain your reasoning.
4. If the service rate at service centre 1 is $\mu_1 = 20$, what is the probability that this queue is empty but the server is not idle? Explain your reasoning.