Queueing Theory

Balakrishna Prabhu

Outline

4.1 Introduction

Introduction

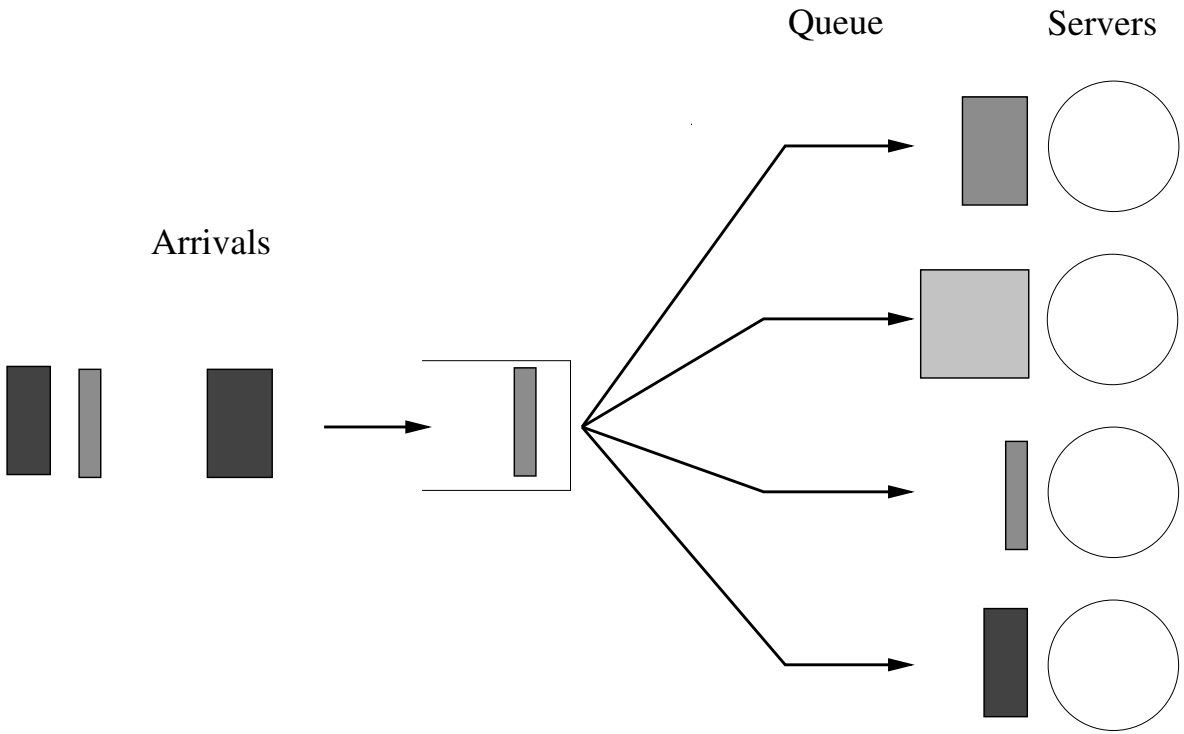


Figure: Servers with a common queue

Examples: airports, post-office,call centers. . .

Introduction

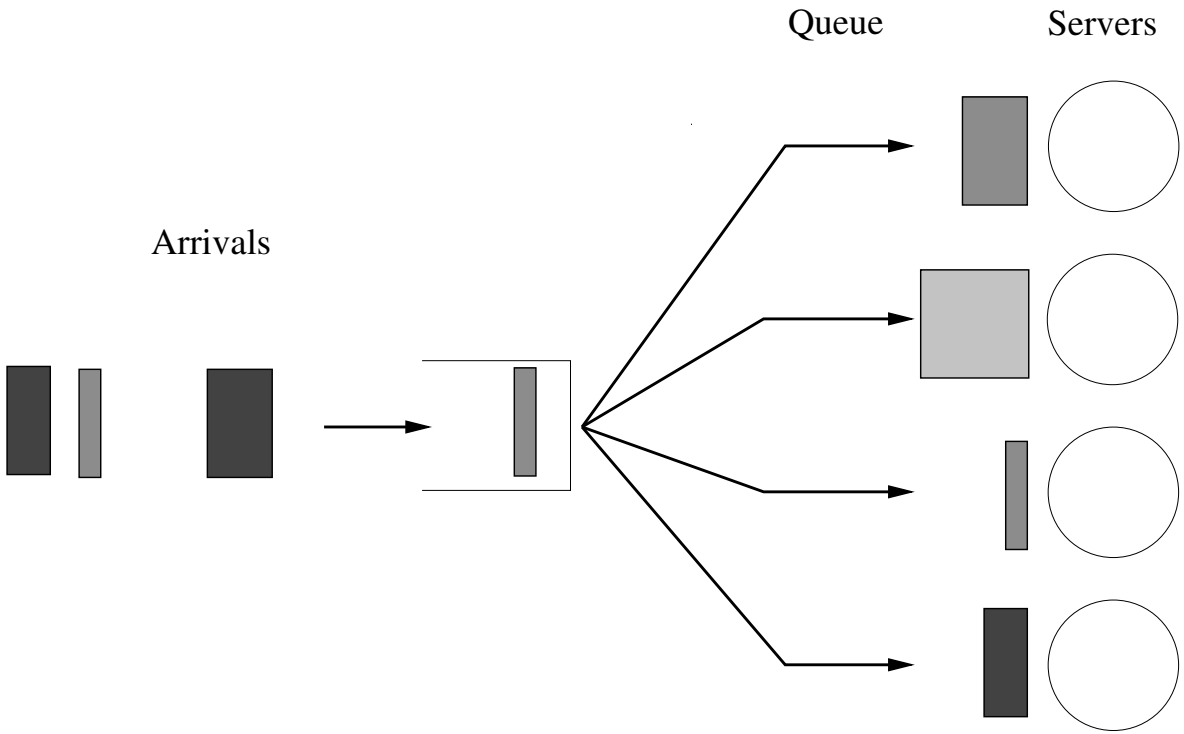


Figure: Servers with a common queue

Examples:supermarkets, data centers, . . .

Modelling a queue

Objective

Compute performance metrics: mean sojourn time, probability of waiting, etc.

Modelling a queue

Objective

Compute performance metrics: mean sojourn time, probability of waiting, etc.

Which parameters influence the most the performance of a queueing system?

Modelling a queue

Objective

Compute performance metrics: mean sojourn time, probability of waiting, etc.

Which parameters influence the most the performance of a queueing system?

- Arrival process (A)

Modelling a queue

Objective

Compute performance metrics: mean sojourn time, probability of waiting, etc.

Which parameters influence the most the performance of a queueing system?

- Arrival process (A)
- Service time distribution (S)

Modelling a queue

Objective

Compute performance metrics: mean sojourn time, probability of waiting, etc.

Which parameters influence the most the performance of a queueing system?

- Arrival process (A)
- Service time distribution (S)
- Number of servers (P)

Modelling a queue

Objective

Compute performance metrics: mean sojourn time, probability of waiting, etc.

Which parameters influence the most the performance of a queueing system?

- Arrival process (A)
- Service time distribution (S)
- Number of servers (P)
- System capacity (K)
- Service discipline (D)

Kendall Notation

Kendall Notation

A/S/P/K/D

Kendall Notation

Kendall Notation

A/S/P/K/D

Kendall Notation

Kendall Notation

A/S/P/K/D

- A can take values: M (Poisson process, Markovian), D (deterministic), G (general),...

Kendall Notation

Kendall Notation

A/S/P/K/D

- A can take values: M (Poisson process, Markovian), D (deterministic), G (general),...
- S can take values: M (exponential), D (deterministic), G (general),...

Kendall Notation

Kendall Notation

A/S/P/K/D

- A can take values: M (Poisson process, Markovian), D (deterministic), G (general),...
- S can take values: M (exponential), D (deterministic), G (general),...
- P is an integer ≥ 1

Kendall Notation

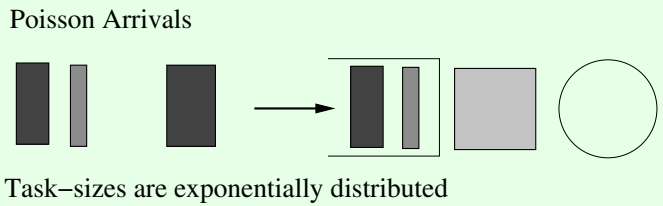
Kendall Notation

A/S/P/K/D

- A can take values: M (Poisson process, Markovian), D (deterministic), G (general),...
- S can take values: M (exponential), D (deterministic), G (general),...
- P is an integer ≥ 1
- K is an integer ≥ 1;
  - default value is ∞
- D can be: FIFO, LIFO, PS (Processor Sharing), Priority
  - default value is FIFO

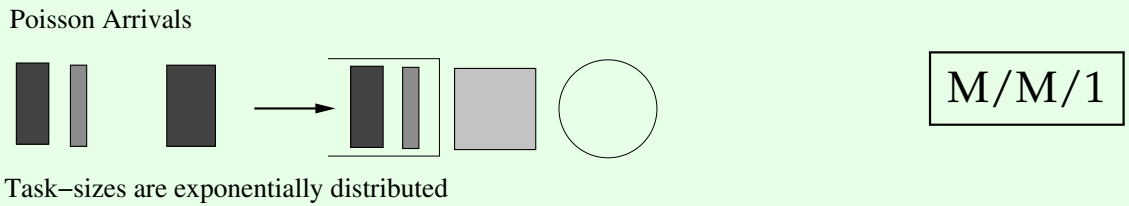
Kendall Notation

Example



Kendall Notation

Example



4.2 Two useful results

PASTA property

- Two different views of the performance measures

PASTA property

- Two different views of the performance measures
  - Time average vs Customer average

PASTA property

- Two different views of the performance measures
  - Time average vs Customer average
  - Not necessarily the same

PASTA property

- Two different views of the performance measures
  - Time average vs Customer average
  - Not necessarily the same

Example

D/D/1 queue: arrival every 2 seconds, service time is 1 second.

PASTA property

- Two different views of the performance measures
  - Time average vs Customer average
  - Not necessarily the same

Example

D/D/1 queue: arrival every 2 seconds, service time is 1 second.

Time average

Fraction of time the queue is empty

Customer average

Fraction of customers who see the queue empty

PASTA property

- Two different views of the performance measures
  - Time average vs Customer average
  - Not necessarily the same

Example

D/D/1 queue: arrival every 2 seconds, service time is 1 second.

Time average

Fraction of time the queue is empty = 0.5

Customer average

Fraction of customers who see the queue empty

PASTA property

- Two different views of the performance measures
  - Time average vs Customer average
  - Not necessarily the same

Example

D/D/1 queue: arrival every 2 seconds, service time is 1 second.

Time average

Fraction of time the queue is empty = 0.5

Customer average

Fraction of customers who see the queue empty= 1

In general, Time average $\neq$ Customer average

PASTA property

- Two different views of the performance measures
  - Time average vs Customer average
  - Not necessarily the same

Example

D/D/1 queue: arrival every 2 seconds, service time is 1 second.

Time average	Customer average
Fraction of time the queue is empty = 0.5	Fraction of customers who see the queue empty= 1

In general, Time average $\neq$ Customer average

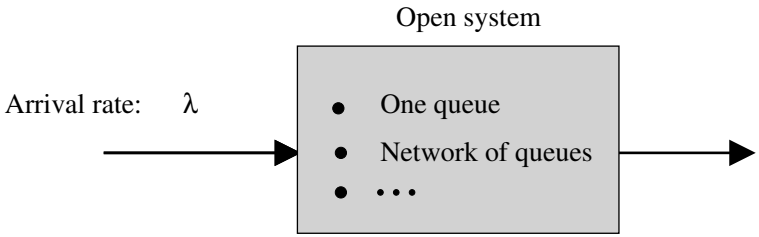
But...

PASTA (Poisson Arrivals See Time Averages) property

For Poisson arrivals, Time average = Customer average.

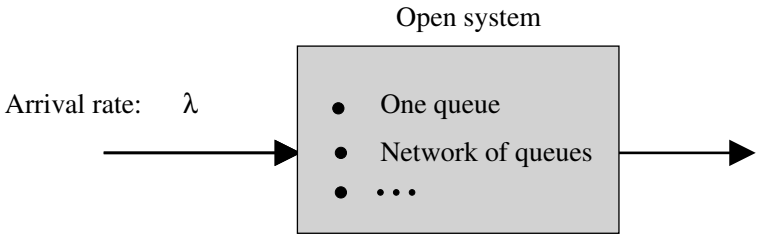
Little's law

- Sojourn times are real valued and cannot be modeled as Markov chains
- Little's law gives the relationship between the mean number in the system and the mean sojourn time



Little's law

- Sojourn times are real valued and cannot be modeled as Markov chains
- Little's law gives the relationship between the mean number in the system and the mean sojourn time



Little's law

$$\bar{T} = \frac{\bar{N}}{\lambda} \tag{1}$$

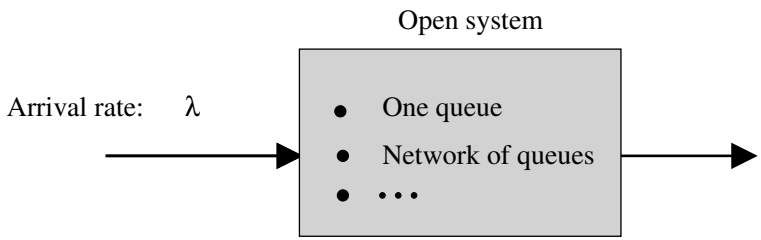
$\bar{T}$: mean sojourn time

$\bar{N}$: mean number in the system

λ : arrival rate

Little's law

- Sojourn times are real valued and cannot be modeled as Markov chains
- Little's law gives the relationship between the mean number in the system and the mean sojourn time



Little's law

$$\bar{T} = \frac{\bar{N}}{\lambda} \tag{1}$$

$\bar{T}$: mean sojourn time

$\bar{N}$: mean number in the system

λ : arrival rate

Observation

Valid under very weak assumptions on arrivals and services times. E.g., does not require Poisson arrivals or exponential service times

Recipe for analysing queues

1. Use Continuous Time Markov Chains (CTMCs) to model the number of customers/tasks/jobs in the system

Recipe for analysing queues

1. Use Continuous Time Markov Chains (CTMCs) to model the number of customers/tasks/jobs in the system

2. Determine the stationary distribution using Kolmogorov's theorem

Recipe for analysing queues

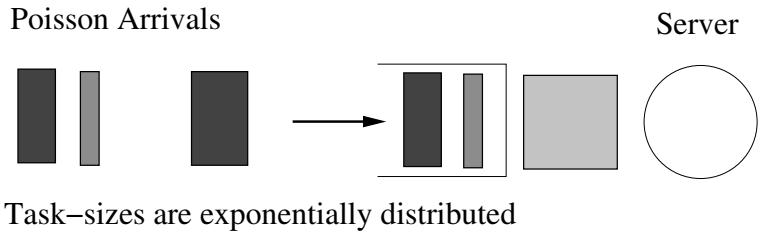
1. Use Continuous Time Markov Chains (CTMCs) to model the number of customers/tasks/jobs in the system

2. Determine the stationary distribution using Kolmogorov's theorem

3. Compute the performance measures
   - Deduce mean number in the system from the stationary distribution
   - Apply Little's law and/or PASTA as necessary to compute other performance metrics

4.3 Markovian queues

The M/M/1 queue



- Arrival process: Poisson of rate λ
- Job-size distribution: $\exp(\mu)$
- 1 server
- Infinite system capacity
- FIFO discipline

Difference between job-size and service time

- Job-size is the amount of work a customer brings (e.g., in a supermarket, job-size of a customer is he number of items in her caddie)
- Service time of a customer is the time the server spends to finish the work of this customer. For a given job-size, faster the server, lower is the service time.

Lemma (Service time distribution)

Let v be the server speed. If the job-size distribution is $\exp(\mu)$, then the service time distribution is $\exp(v\mu)$.

Assumption: Server speed is 1

The M/M/1 queue

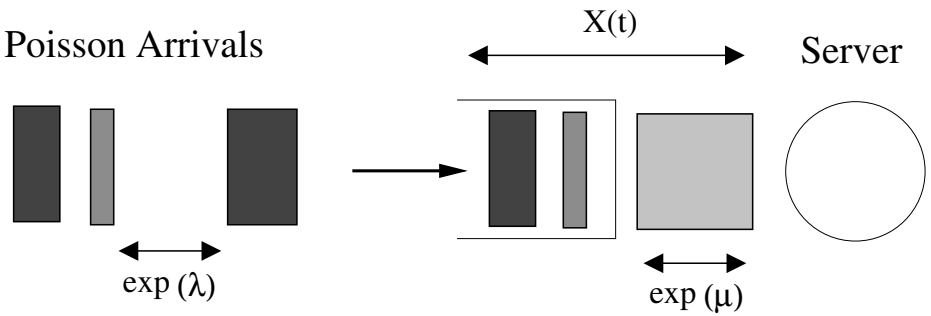
Performance measures

- Mean sojourn time
- Probability server is idle or is busy
- ...

Recipe for computing performance measures

1. Construct a Markov model
2. Determine stationary distribution
3. Compute performance measure

M/M/1 queue: Markov model



Recall: Poisson process

- Poisson process of rate λ means time between two successive arrivals is exponentially distributed with rate λ.
- Only one arrival at a time, i.e., no batch arrivals.

- Let $X(t)$ be the number of customers in the system at time t
  - $X(t)$ is a continuous-time stochastic process
  - $X(t)$ increases by 1 when a customer arrives
  - $X(t)$ decreases by 1 when a customer finishes service and leaves

Show $X(t)$ is a CTMC

When is a stochastic process a CTMC?

1. Countable state space
2. Time spent in each state is exponentially distributed
3. Probability of going to state j from state i does not depend on the history

M/M/1 queue: Markov model

Check $X(t)$ meets these conditions

1. $X(t)$ is a non-negative integer $\Rightarrow$ countable state space

M/M/1 queue: Markov model

Check $X(t)$ meets these conditions

1. $X(t)$ is a non-negative integer $\Rightarrow$ countable state space
2. Time spent in state i

State	Possible events	Distribution of time spent in this state
$i = 0$	a customer arrives	$\exp(\lambda)$
$i > 0$	a customer arrives and state goes to $i + 1$ a customer leaves and state goes to $i - 1$	$\exp(\lambda + \mu)$

M/M/1 queue: Markov model

Check $X(t)$ meets these conditions

1. $X(t)$ is a non-negative integer $\Rightarrow$ countable state space
2. Time spent in state i

State	Possible events	Distribution of time spent in this state
$i = 0$	a customer arrives	$\exp(\lambda)$
$i > 0$	a customer arrives and state goes to $i+1$ a customer leaves and state goes to $i-1$	$\exp(\lambda+\mu)$

3. Probability of going to state j from state i

Current state (i)	Next state (j)	Transition probability
$i = 0$	$j = 1$	1
$i > 0$	$j = i+1$	$\frac{\lambda}{\lambda+\mu}$
	$j = i-1$	$\frac{\mu}{\lambda+\mu}$

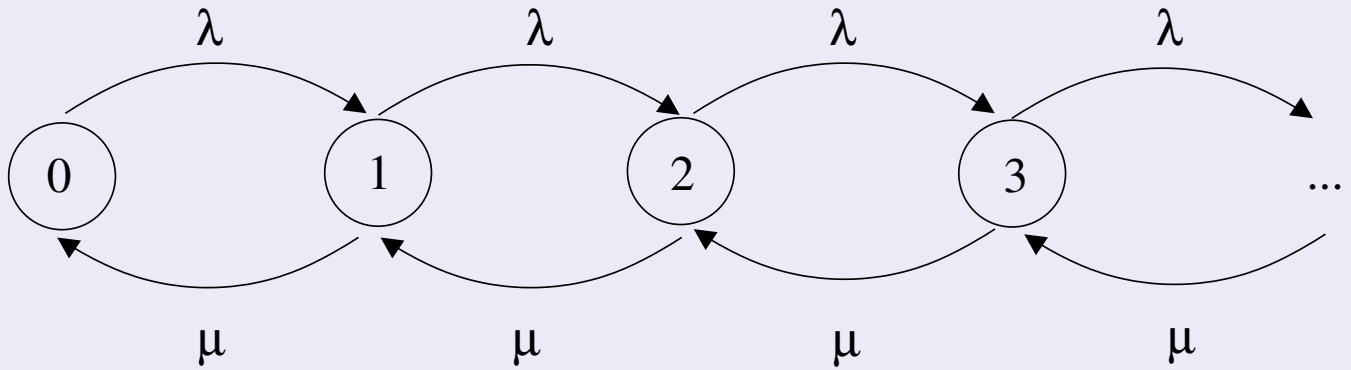
M/M/1 queue: Markov model

$X(t)$ is a CTMC

M/M/1 queue: Markov model

$X(t)$ is a CTMC

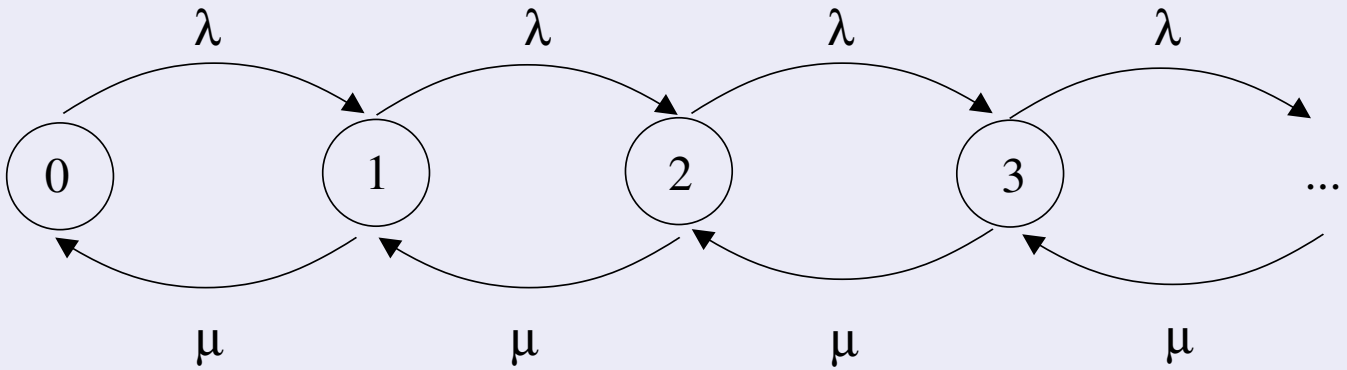
Transition diagram of $X(t)$



M/M/1 queue: Markov model

$X(t)$ is a CTMC

Transition diagram of $X(t)$



Transition matrix of $X(t)$

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & \cdots \end{array} \\ \begin{bmatrix} -\lambda & \lambda & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & \ddots \\ 0 & \mu & \ddots & \\ \vdots & \ddots & & \end{bmatrix} \end{array}$$

M/M/1 queue: Stationary distribution

$X(t)$ is a birth-death process with $\lambda_i = \lambda$ and $\mu_i = \mu$

Theorem (Stationary distribution of $X(t)$)

Let π_i be the stationary probability of finding i customers in the system. Then,

$$\pi_i = (1-\rho)\rho^i \tag{2}$$

where

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

is the load on the system.

Proof.

Apply the formula for birth-death processes with $\lambda_i = \lambda$ and $\mu_i = \mu$. □

Meaning of ρ

On average, λ customers arrive per unit time and each customer brings μ^{-1} amount of work on average. So, $\rho = \lambda\mu^{-1}$ is the average amount of works that arrives to the queue in a unit time.

M/M/1 queue: Performance measures

Stability

Stability means that the number in the queue does not grow to infinity. For an $M/M/1$ queue to be stable,

$$\boxed{\rho < 1} \tag{3}$$

Interpretation: ρ (which is the rate of work entering the queue) should be less than 1 (which is the rate at which server can work of). Recall, the speed of server is 1.

Probability server is busy

The server is busy when there is at least one customer in the queue. Therefore,

$$\boxed{P_{busy} = \sum_{i \geq 1} \pi_i = \rho.}$$

This is also the fraction of time the server is busy.

Probability a customer has to wait

A customer has to wait if the server is busy when she arrives. Using the PASTA property

$$\boxed{P_{wait} = P_{busy}}$$

M/M/1 queue: Performance measures

Mean number in the queue

Let $\bar{N}$ be the mean number in the queue. Then,

$$\bar{N} = \sum_{i \geq 0}^{\infty} i\pi_i$$

That is,

$$\boxed{\bar{N} = \frac{\rho}{1-\rho}} \tag{4}$$

Mean number waiting in the queue

This is the number waiting and excludes the customer in service. Let $\bar{N}_q$ be the mean number in the queue. Then,

$$\bar{N}_q = \sum_{i \geq 1} (i-1)\pi_i$$

That is,

$$\boxed{\bar{N}_q = \frac{\rho^2}{1-\rho}} \tag{5}$$

M/M/1 queue: Performance measures

Mean sojourn time

Let $\bar{T}$ be the mean number in the queue. From Little's law

$$\bar{T} = \frac{\bar{N}}{\lambda}$$

That is,

$$\boxed{\bar{T} = \frac{1}{\mu - \lambda}} \tag{6}$$

Mean waiting time

This is the mean time a customer has to wait before being taken into service Let $\bar{W}_q$ be the mean number in the queue.
From the Little's law applied to the waiting room

$$\bar{W} = \frac{\bar{N}_q}{\lambda}$$

That is,

$$\boxed{\bar{W} = \frac{\rho}{\mu - \lambda}} \tag{7}$$

M/M/1 queue: example

Example

- Given data

M/M/1 queue: example

Example

- Given data
  - Poisson process with 0.4 customers arriving on an average every minute.

M/M/1 queue: example

Example

- Given data
  - Poisson process with 0.4 customers arriving on an average every minute.

$$\Rightarrow \lambda = 0.4 \text{ per minute}$$

  - Exponential service time distribution with mean 2 minutes

M/M/1 queue: example

Example

- Given data
  - Poisson process with 0.4 customers arriving on an average every minute.

$$\Rightarrow \lambda = 0.4 \text{ per minute}$$

  - Exponential service time distribution with mean 2 minutes

$$\Rightarrow \frac{1}{\mu} = 2 \text{ minutes}$$

M/M/1 queue: example

Example

- Given data
  - Poisson process with 0.4 customers arriving on an average every minute.

$$\Rightarrow \lambda = 0.4 \text{ per minute}$$

  - Exponential service time distribution with mean 2 minutes

$$\Rightarrow \frac{1}{\mu} = 2 \text{ minutes}$$

M/M/1 queue: example

Example

- Given data
  - Poisson process with 0.4 customers arriving on an average every minute.

$$\Rightarrow \lambda = 0.4 \text{ per minute}$$

  - Exponential service time distribution with mean 2 minutes

$$\Rightarrow \frac{1}{\mu} = 2 \text{ minutes}$$

- Compute the load on the server

M/M/1 queue: example

Example

- Given data
  - Poisson process with 0.4 customers arriving on an average every minute.

$$\Rightarrow \lambda = 0.4 \text{ per minute}$$

  - Exponential service time distribution with mean 2 minutes

$$\Rightarrow \frac{1}{\mu} = 2 \text{ minutes}$$

- Compute the load on the server

$$\boxed{\rho = \frac{\lambda}{\mu} = 0.8}$$

$$\rho < 1 \Rightarrow \text{ queue is stable}$$

M/M/1 queue: example

Example (continued)

- Stationary distribution

$$\pi_i = (1-\rho)\rho^i = 0.2 \cdot 0.8^i$$

M/M/1 queue: example

Example (continued)

- Stationary distribution

$$\pi_i = (1-\rho)\rho^i = 0.2 \cdot 0.8^i$$

- Probability server is busy: $P_{busy} = \rho = 0.8$

M/M/1 queue: example

Example (continued)

- Stationary distribution
$$\pi_i = (1-\rho)\rho^i = 0.2 \cdot 0.8^i$$
- Probability server is busy: $P_{busy} = \rho = 0.8$
- Mean number in the queue
$$\bar{N} = \frac{\rho}{1-\rho} = 4$$

M/M/1 queue: example

Example (continued)

- Stationary distribution

$$\pi_i = (1-\rho)\rho^i = 0.2 \cdot 0.8^i$$

- Probability server is busy: $P_{busy} = \rho = 0.8$
- Mean number in the queue

$$\bar{N} = \frac{\rho}{1-\rho} = 4$$

- Mean number waiting in the queue

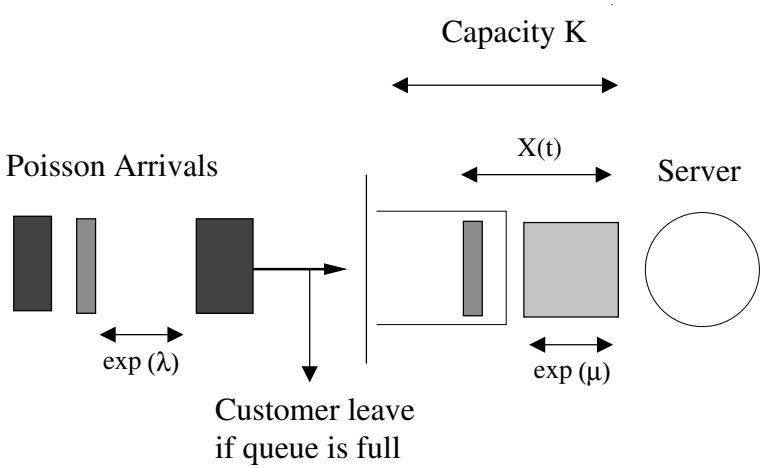
$$\bar{N}_q = \frac{\rho^2}{1-\rho} = 3.6$$

M/M/1 queue: example

Example (continued)

- Stationary distribution
$$\pi_i = (1-\rho)\rho^i = 0.2 \cdot 0.8^i$$
- Probability server is busy: $P_{busy} = \rho = 0.8$
- Mean number in the queue
$$\bar{N} = \frac{\rho}{1-\rho} = 4$$
- Mean number waiting in the queue
$$\bar{N}_q = \frac{\rho^2}{1-\rho} = 3.6$$
- Mean sojourn time
$$\bar{T} = \frac{1}{\mu - \lambda} = 10 \text{ minutes}$$
- Mean waiting time
$$\bar{W} = \frac{\rho}{\mu - \lambda} = 8 \text{ minutes}$$

The M/M/1/K queue



- Arrival process: Poisson of rate λ
- Job-size distribution: $\exp(\mu)$
- 1 server
- Capacity K
- FIFO discipline

Performance measures

- Probability a customer is rejected
- Mean sojourn time
- Probability server is idle or is busy
- ...

M/M/1/K queue: Markov model

- Let $X(t)$ be the number of customers in the system at time t
  - $X(t)$ is a continuous-time stochastic process
  - $X(t)$ increases by 1 when a customer arrives except when the queue is full
  - $X(t)$ decreases by 1 when a customer finishes service and leaves

Show $X(t)$ is a CTMC using the same argument that was used for the $M/M/1$ queue

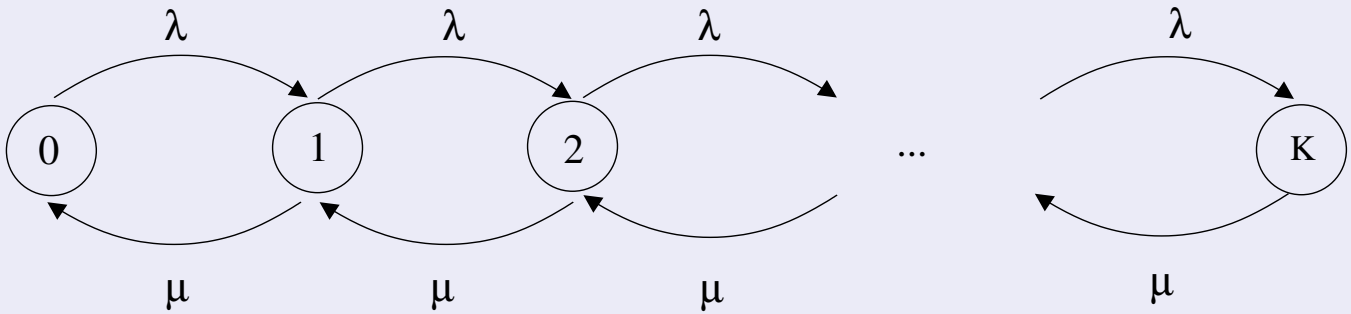
M/M/1/K queue: Markov model

$X(t)$ is a CTMC

M/M/1/K queue: Markov model

$X(t)$ is a CTMC

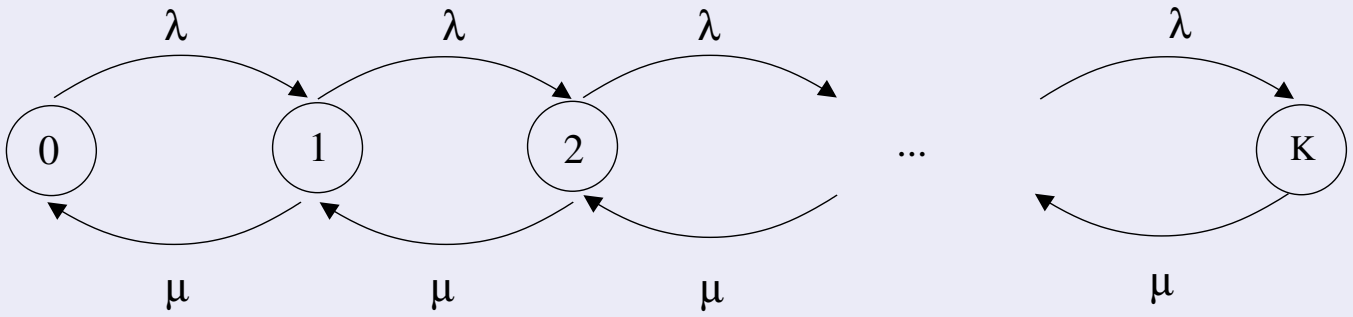
Transition diagram of $X(t)$



M/M/1/K queue: Markov model

$X(t)$ is a CTMC

Transition diagram of $X(t)$



Transition matrix of $X(t)$

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ K \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & K \end{array} \\ \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & \\ \mu & -(\lambda+\mu) & \lambda & \ddots & \\ 0 & \mu & \ddots & & \\ \vdots & \ddots & & & \\ \vdots & \ddots & & \mu & -\mu \end{bmatrix} \end{array}$$

M/M/1/K queue: Stationary distribution

$X(t)$ is a birth-death process with

$$\lambda_i = \lambda, \qquad i \leq K-1 \tag{8}$$

$$\mu_i = \mu, \qquad i \leq K \tag{9}$$

Theorem (Stationary distribution of $X(t)$)

Let π_i be the stationary probability of finding i customers in the system. Then,

$$\boxed{\pi_i = \frac{\rho^i}{\sum_{j=0}^{K} \rho^j}} \tag{10}$$

with $\rho = \lambda\mu^{-1}$.

M/M/1/K queue: Performance measures

Stability

This queue is always stable because the number of customer in the queue is always finite.

Probability queue is full

The queue is full when there are K customer in the queue.

$$\boxed{P_{full} = \pi_K}$$

This is also the fraction of time the queue is full.

Probability a customer is rejected

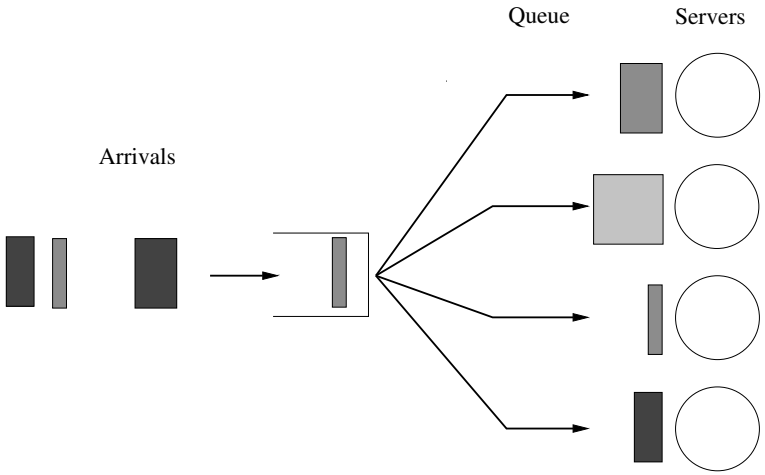
A customer is rejected when the queue is full. Using the PASTA property

$$\boxed{P_{reject} = P_{full} = \pi_K}$$

Other performance measures

Use the same method as for the $M/M/1$ queue.

The M/M/C queue



- Arrival process: Poisson of rate λ
- Job-size distribution: $\exp(\mu)$
- C servers
- Infinite system capacity
- FIFO discipline

Note: $C = 1$ gives the $M/M/1$ queue

Performance measures

- Probability a customer has to wait
- Mean sojourn time
- ...

M/M/C queue: Markov model

- Let $X(t)$ be the number of customers in the system at time t
  - $X(t)$ is a continuous-time stochastic process
  - $X(t)$ increases by 1 when a customer arrives
  - $X(t)$ decreases by 1 when a customer finishes service and leaves

Show $X(t)$ is a CTMC

M/M/C queue: Markov model

Check $X(t)$ meets the required conditions

1. $X(t)$ is a non-negative integer $\Rightarrow$ countable state space

M/M/C queue: Markov model

Check $X(t)$ meets the required conditions

1. $X(t)$ is a non-negative integer $\Rightarrow$ countable state space
2. Time spent in state i

State	Possible events i	Distribution of time spent in i
$i = 0$	a customer arrives	$\exp(\lambda)$
$0 < i \leq C$	a customer arrives and state goes to $i+1$ a customer leaves and state goes to $i-1$	$\exp(\lambda + i\mu)$
$i \geq C$	a customer arrives and state goes to $i+1$ a customer leaves and state goes to $i-1$	$\exp(\lambda + C\mu)$

M/M/C queue: Markov model

Check $X(t)$ meets the required conditions

1. $X(t)$ is a non-negative integer $\Rightarrow$ countable state space
2. Time spent in state i

State i	Possible events	Distribution of time spent in i
$i = 0$	a customer arrives	$\exp(\lambda)$
$0 < i \leq C$	a customer arrives and state goes to $i+1$ a customer leaves and state goes to $i-1$	$\exp(\lambda + i\mu)$
$i \geq C$	a customer arrives and state goes to $i+1$ a customer leaves and state goes to $i-1$	$\exp(\lambda + C\mu)$

3. Probability of going to state j from state i

Current state (i)	Next state (j)	Transition probability
$i = 0$	$j = 1$	1
$0 < i \leq C$	$j = i+1$	$\frac{\lambda}{\lambda + i\mu}$
	$j = i-1$	$\frac{i\mu}{\lambda + i\mu}$
$i \geq C$	$j = i+1$	$\frac{\lambda}{\lambda + C\mu}$
	$j = i-1$	$\frac{C\mu}{\lambda + C\mu}$

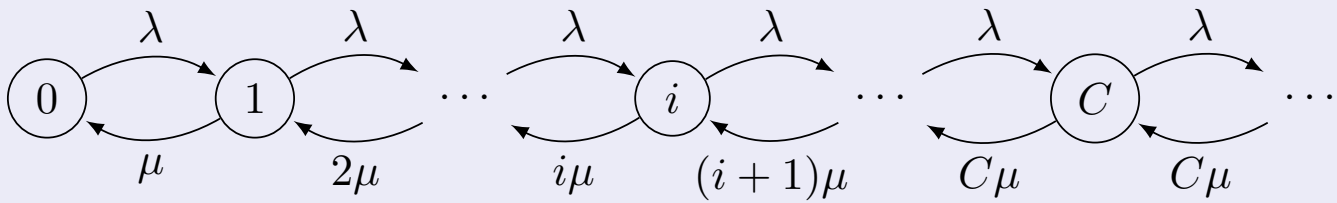
M/M/C queue: Markov model

$X(t)$ is a CTMC

M/M/C queue: Markov model

$X(t)$ is a CTMC

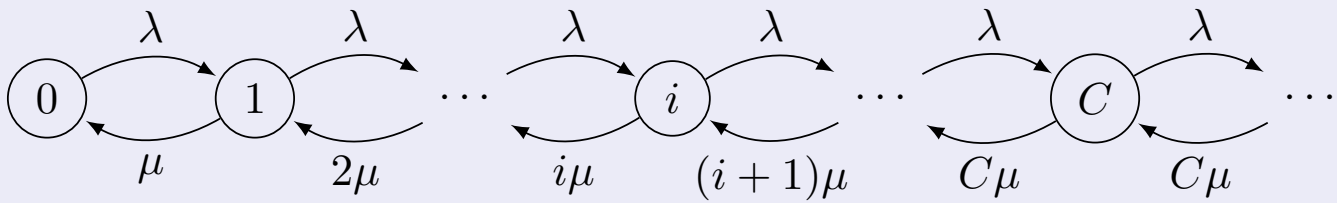
Transition diagram of $X(t)$



M/M/C queue: Markov model

$X(t)$ is a CTMC

Transition diagram of $X(t)$



Transition matrix of $X(t)$

Exercise!

M/M/C queue: Stationary distribution

$X(t)$ is a birth-death process with

$$\lambda_i = \lambda, \qquad i \geq 0 \tag{11}$$

$$\mu_i = \begin{cases} i\mu & 1 \leq i \leq C \\ C\mu & i > C \end{cases} \tag{12}$$

Theorem (Stationary distribution of $X(t)$)

Let $\rho = \lambda\mu^{-1}$. Then

$$\boxed{\pi_i = \begin{cases} \pi_0 \frac{\rho^i}{i!} & 0 \leq i \leq C \\ \pi_0 \frac{\rho^i}{C!C^{i-C}} & i > C \end{cases}} \tag{13}$$

with

$$\pi_0 = \left(\sum_{i=0}^{C} \frac{\rho^i}{i!} + \sum_{i>C} \frac{\rho^i}{C!C^{i-C}} \right)^{-1} \tag{14}$$

M/M/C queue: Performance measures

Stability

$$\boxed{\lambda < C\mu} \tag{15}$$

Probability a customer has to wait

A customer has to wait when all the servers are busy. Using the PASTA property,

$$P_{wait} = \sum_{i \geq C} \pi_i \tag{16}$$

That is,

$$\boxed{P_{wait} = \frac{\frac{\rho^C}{C!(1-\rho/C)}}{\sum_{i=0}^{C} \frac{\rho^i}{i!} + \sum_{i>C} \frac{\rho^i}{C!C^{i-C}}}} \tag{17}$$

$\boxed{P_{wait} \text{ is called the Erlang-C probability}}$

Other performance measures

Use the same method as for the $M/M/1$ queue.

M/M/C queue: Application to call-centers

Dimensioning a call-center (problem statement)

- **G**iven data
  - Poisson arrivals with rate 3 customers per minute
  - Call times are exponentially distributed with mean 5 minutes

M/M/C queue: Application to call-centers

Dimensioning a call-center (problem statement)

- **G**iven data
  - Poisson arrivals with rate 3 customers per minute
  - Call times are exponentially distributed with mean 5 minutes
- Objective

  Customer should not have to wait 99% of the time

M/M/C queue: Application to call-centers

Dimensioning a call-center (problem statement)

- **G**iven data
  - Poisson arrivals with rate 3 customers per minute
  - Call times are exponentially distributed with mean 5 minutes
- Objective

  Customer should not have to wait 99% of the time
- Question

  How many agents are required?

M/M/C queue: Application to call-centers

Dimensioning a call-center (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 3$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 5 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 15}$$

M/M/C queue: Application to call-centers

Dimensioning a call-center (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 3$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 5 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 15}$$
- **Objective**
  $$P_{wait} < 0.01$$

M/M/C queue: Application to call-centers

Dimensioning a call-center (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 3$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 5 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 15}$$
- **Objective**
  $$P_{wait} < 0.01$$
- Use Erlang-C forumla with $\rho = 15$

C	$\cdots$	23	24	25	26	$\cdots$
P_{wait}	$\cdots$	0.038	0.022	0.012	0.0068	$\cdots$

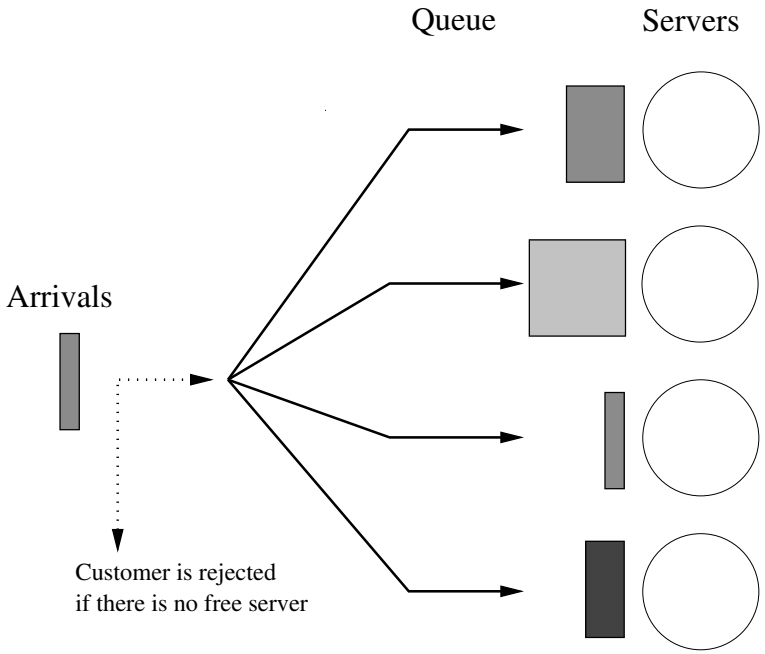
M/M/C queue: Application to call-centers

Dimensioning a call-center (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 3$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 5 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 15}$$
- **Objective**
  $$P_{wait} < 0.01$$
- Use Erlang-C forumla with $\rho = 15$

C	$\cdots$	23	24	25	26	$\cdots$
P_{wait}	$\cdots$	0.038	0.022	0.012	0.0068	$\cdots$

The M/M/C/C queue



- Arrival process: Poisson of rate λ
- Job-size distribution: $\exp(\mu)$
- C servers
- No waiting room
- FIFO discipline

Performance measures

- Probability a customer is rejected

M/M/C/C queue: Markov model

- Let $X(t)$ be the number of customers in the system at time t
  - $X(t)$ is a continuous-time stochastic process
  - $X(t)$ increases by 1 when a customer arrives
  - $X(t)$ decreases by 1 when a customer finishes service and leaves

Show $X(t)$ is a CTMC

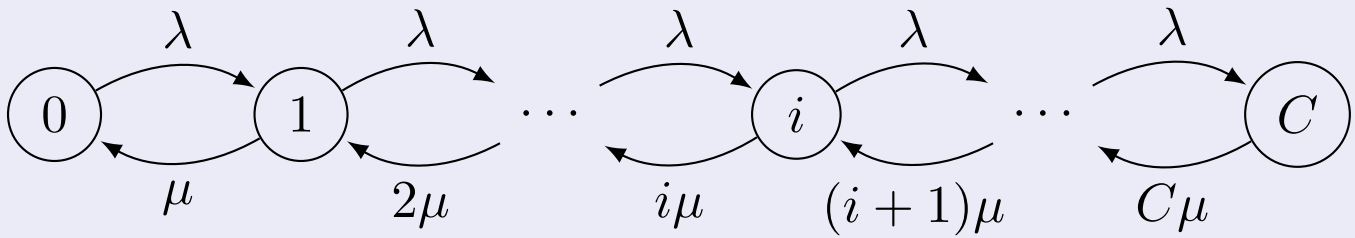
M/M/C/C queue: Markov model

$X(t)$ is a CTMC

M/M/C/C queue: Markov model

$X(t)$ is a CTMC

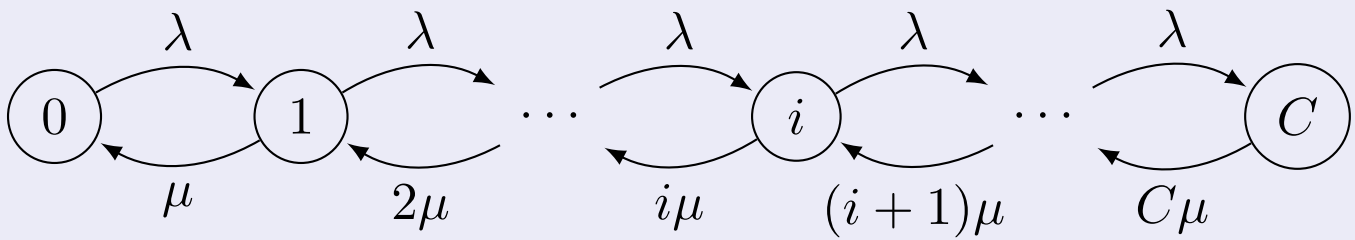
Transition diagram of $X(t)$



M/M/C/C queue: Markov model

$X(t)$ is a CTMC

Transition diagram of $X(t)$



Transition matrix of $X(t)$

Exercise

M/M/C/C queue: Stationary distribution

$X(t)$ is a birth-death process with

$$\lambda_i = \lambda, \qquad i \geq 0 \tag{18}$$

$$\mu_i = i\mu \qquad 1 \leq i \leq C. \tag{19}$$

Theorem (Stationary distribution of $X(t)$)

Let $\rho = \lambda\mu^{-1}$. Then

$$\boxed{\pi_i = \pi_0 \frac{\rho^i}{i!}, \quad 0 \leq i \leq C} \tag{20}$$

with

$$\pi_0 = \left(\sum_{i=0}^{C} \frac{\rho^i}{i!} \right)^{-1} \tag{21}$$

M/M/C/C queue: Performance measures

Stability

Always stable

Probability a customer is rejected

A customer is rejected when all the servers are busy. Using the PASTA property,

$$P_{reject} = \pi_C \tag{22}$$

That is,

$$\boxed{P_{reject} = \frac{\frac{\rho^C}{C!}}{\sum_{i=0}^{C} \frac{\rho^i}{i!}}} \tag{23}$$

$\boxed{P_{reject} \text{ is called the Erlang-B probability}}$

Insensitivity

Erlang-B formula is insensitive to the service-time distribution. That is, it is valid for any service-time distribution.

M/M/C/C queue: Application to telecommunication networks

Dimensioning a telecommunication network (problem statement)

- Given data
  - Poisson arrivals with rate 10 calls per minute
  - Call times are exponentially distributed with mean 5 minutes

M/M/C/C queue: Application to telecommunication networks

Dimensioning a telecommunication network (problem statement)

- **G**iven data
  - Poisson arrivals with rate 10 calls per minute
  - Call times are exponentially distributed with mean 5 minutes
- Objective

  Probability of rejecting a call should be less than 0.1%

M/M/C/C queue: Application to telecommunication networks

Dimensioning a telecommunication network (problem statement)

- **G**iven data
  - Poisson arrivals with rate 10 calls per minute
  - Call times are exponentially distributed with mean 5 minutes
- Objective

  Probability of rejecting a call should be less than 0.1%
- Question

  How many channels to buy? (Spectrum is costly...)

M/M/C/C queue: Application to telecommunication networks

Dimensioning a telecommunication network (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 10$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 4 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 40}$$

M/M/C/C queue: Application to telecommunication networks

Dimensioning a telecommunication network (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 10$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 4 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 40}$$
- **Objective**
  $$P_{reject} < 0.001$$

M/M/C/C queue: Application to telecommunication networks

Dimensioning a telecommunication network (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 10$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 4 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 40}$$
- **Objective**
  $$P_{reject} < 0.001$$
- Use Erlang-B forumla with $\rho = 50$

C	$\cdots$	57	58	59	60	$\cdots$
P_{wait}	$\cdots$	0.0022	0.0015	0.00102	0.00068	$\cdots$

M/M/C/C queue: Application to telecommunication networks

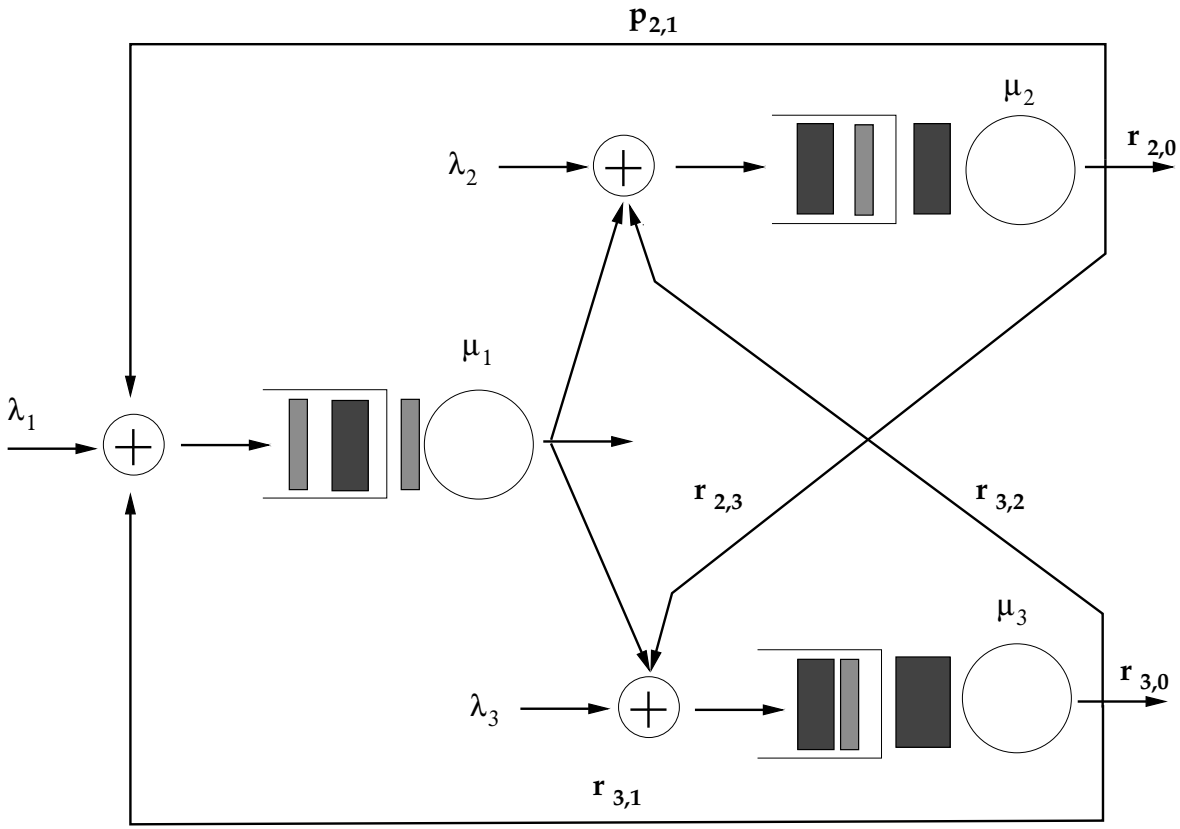
Dimensioning a telecommunication network (solution)

- **Given data**
  - Poisson arrivals with $\lambda = 10$ per minute
  - Exponential service times with
    $$\frac{1}{\mu} = 4 \text{ minutes}$$
  - Load
    $$\boxed{\rho = \frac{\lambda}{\mu} = 40}$$
- **Objective**
  $$P_{reject} < 0.001$$
- Use Erlang-B forumla with $\rho = 50$

C	$\cdots$	57	58	59	60	$\cdots$
P_{wait}	$\cdots$	0.0022	0.0015	0.00102	0.00068	$\cdots$

4.4 Networks of queues

Network of queues



Applications

Manufacturing facilities, shopping malls,. . .

Model description

M: number of queues in the network

At queue i:

- Arrival process: Poisson of rate λ_i
- Job-size distribution: $\exp(\mu_i)$
- C_i servers
- Infinite system capacity
- FIFO discipline
- $r_{i,j}$: probability of going to queue j after leaving queue i
- $r_{i,0}$; probability of leaving the network after finishing in queue i

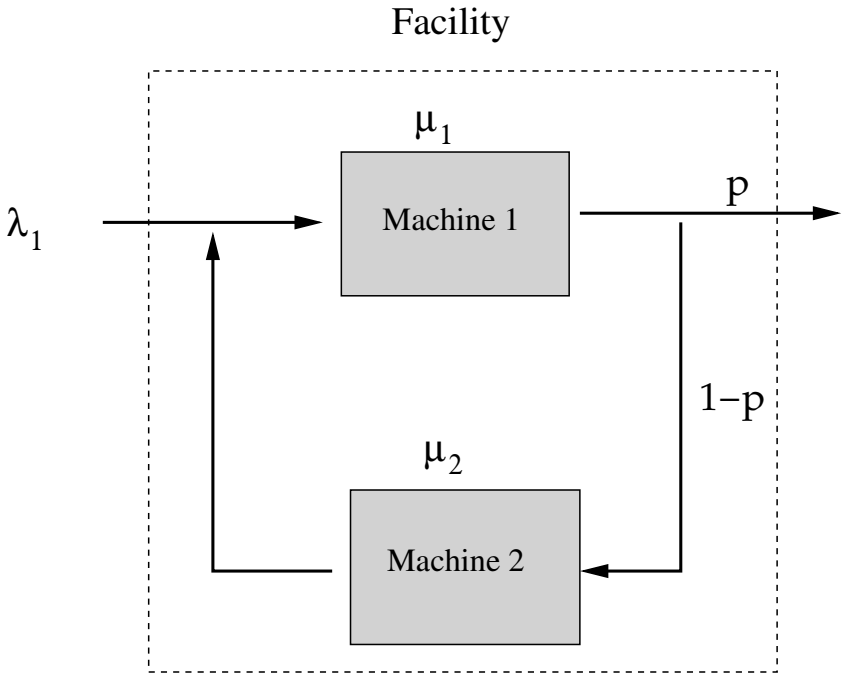
Remark

Networks can be defined for other disciplines and for finite capacity queue.

Model description

Example: a manufacturing facility

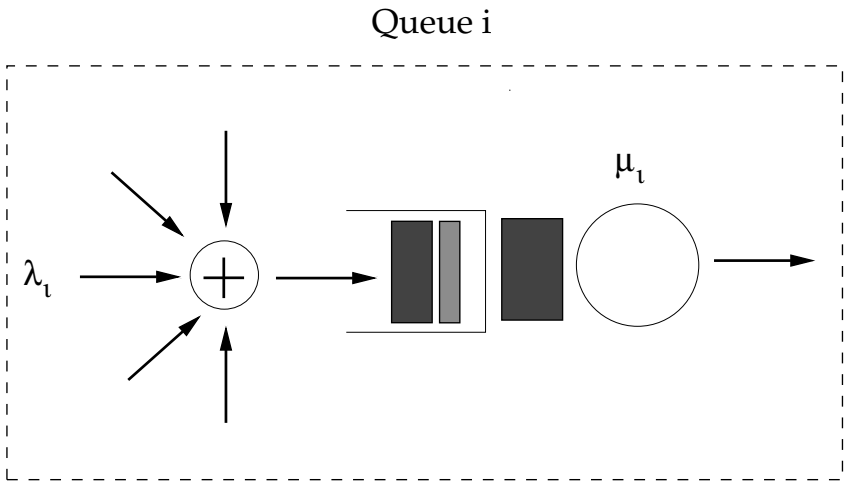
Consider a manufacturing facility with two machines. Orders arrive according to Poisson process of rate λ_1. It takes $\exp(\mu_1)$ distributed time to process an order in machine 1. When an order leaves machine 1 it is tested for quality control. If it passes the quality test, the order is delivered to the customer. Otherwise, it is sent to machine 2 where it is components are recovered (this takes time $\exp(\mu_2)$ after which it is sent back to machine 1 for processing.
Assume p is the probability of passing the quality test.



- $M = 2$
- $\lambda_1 = \lambda_1, \lambda_2 = 0$
- $C_1 = C_2 = 1$
- $r_{1,1} = 0, r_{1,2} = 1 - p, r_{2,1} = 1,$ $r_{2,2} = 0$
- $r_{1,0} = p, r_{2,0} = 0$

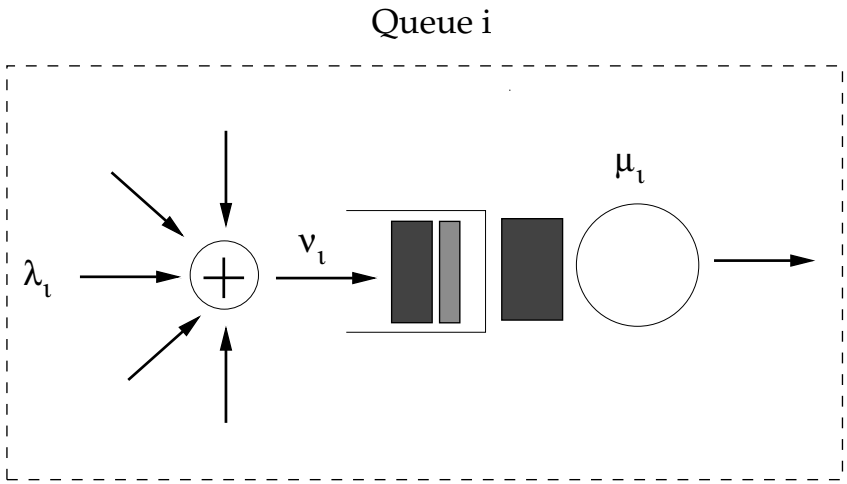
Effective arrival rate

- Customers arrive to a queue not only from outside the network but also from queue inside the network. The total arrival rate to queue i is this larger than λ_i.



Effective arrival rate

- Customers arrive to a queue not only from outside the network but also from queue inside the network. The total arrival rate to queue i is this larger than λ_i.



ν_i: effective arrival rate to queue i

Effective arrival rate

- Queue j is stable $\Rightarrow$

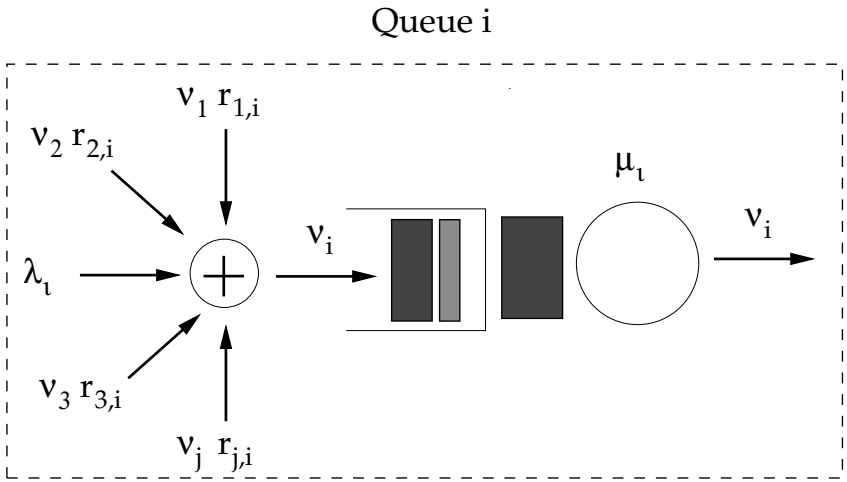
$$\begin{aligned}\text{rate of outflow from queue } j &= \text{ rate of inflow into queue } j \\ &= \nu_j\end{aligned}$$

- Rate of flow from queue j to queue i

$$\nu_j r_{j,i}$$

that is, what leaves queue j multiplied by the probability to going from j to i.

- Thus, effective arrival rate to queue i is



$$\nu_i = \lambda_i + \sum_{j=1}^{M} \nu_j r_{j,i}$$

Effective arrival rate

Some notation

- $\vec{\lambda}$: vector of external arrival rates

$$\vec{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_M]$$

- $\vec{\nu}$: Vector of external arrival rates
- R: Routing matrix

$$R = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & \cdots & M \end{array} \\ \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & \cdots & r_{1,M} \\ r_{2,1} & r_{2,2} & r_{2,3} & \ddots & r_{2,M} \\ r_{3,1} & r_{3,2} & \ddots & & \\ \vdots & \ddots & & & \\ r_{M,1} & \cdots & & \cdots & r_{M,M} \end{bmatrix} \end{array}$$

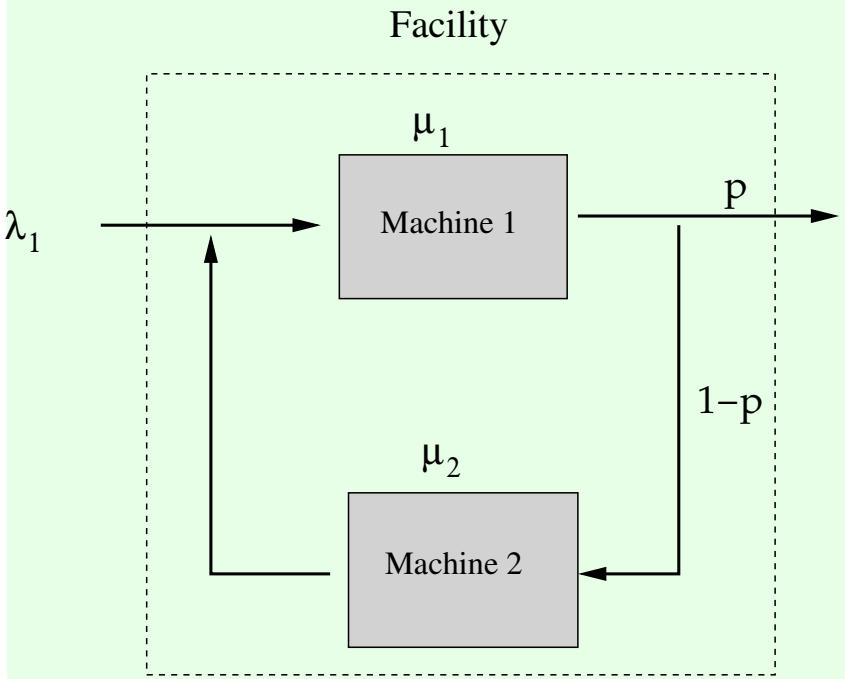
Theorem (Effective arrival rate)

The effective arrival rate is the solution of

$$\vec{\nu} = \vec{\lambda} + \vec{\nu}R \tag{24}$$

Effective arrival rate

Example: a manufacturing facility



- $\vec{\lambda} = [\lambda_1, 0]$
- Routing matrix

$$R = \begin{matrix} \\ 1 \\ 2 \end{matrix} \begin{matrix} \begin{matrix} 1 & 2 \end{matrix} \\ \begin{bmatrix} 0 & 1-p \\ 1 & 0 \end{bmatrix} \end{matrix}$$

- Effective arrival rate

$$[\nu_1, \nu_2] = \left[\frac{\lambda_1}{p}, \frac{\lambda_1(1-p)}{p}\right]$$

Performance analysis

- $\vec{X}(t) = [X_1(t), X_2(t), \ldots, X_M(t)]$, where X_i is number of customers in queue i

$$\boxed{\vec{X}(t) \text{ is a CTMC}}$$

- Markov-chain-based direct analysis can be done but is computationally expensive
- Jackson's theorem gives a simpler way

Performance analysis

- $\vec{X}(t) = [X_1(t), X_2(t), \ldots, X_M(t)]$, where X_i is number of customers in queue i

$$\boxed{\vec{X}(t) \text{ is a CTMC}}$$

- Markov-chain-based direct analysis can be done but is computationally expensive
- Jackson's theorem gives a simpler way

Theorem (Jackson)

Let $\pi^{(i)}(n_i)$ be the stationary probability of finding n_i customers in queue i assuming arrival rate ν_i and service rate μ_i when analysed independently of other queues. The stationary distribution of $\vec{X}(t)$ is given by

$$\pi(\vec{n}) = \prod_{i=1}^{M} \pi^{(i)}(n_i) \tag{25}$$

Performance analysis

- $\vec{X}(t) = [X_1(t), X_2(t), \ldots, X_M(t)]$, where X_i is number of customers in queue i

$$\boxed{\vec{X}(t) \text{ is a CTMC}}$$

- Markov-chain-based direct analysis can be done but is computationally expensive
- Jackson's theorem gives a simpler way

Theorem (Jackson)

Let $\pi^{(i)}(n_i)$ be the stationary probability of finding n_i customers in queue i assuming arrival rate ν_i and service rate μ_i when analysed independently of other queues. The stationary distribution of $\vec{X}(t)$ is given by

$$\pi(\vec{n}) = \prod_{i=1}^{M} \pi^{(i)}(n_i) \tag{25}$$

Interpretation of Jackson's theorem

- The network of M queues can be decomposed into M independent queues.
- Analyse each queue separately using the analysis for single queues taking the arrival rate at queue i to be ν_i and service times to be μ_i.

Recipe for analyzing a network

1. Determine the parameters: Routing matrix, arrival rate vector
2. Compute the effective arrival rates using

$$\vec{\nu} = \vec{\lambda} + \vec{\nu} R$$

3. Apply Jackson's theorem to obtain the stationary distribution
   - Calculate the stationary distribution of each queue independently of the others
4. Compute the performance measures

Performance analysis of the example

Recall

- $\nu_1 = \frac{\lambda_1}{p}, \nu_2 = \frac{\lambda_1(1-p)}{p}$
- From Jackson's theorem, machine i is an M/M/1 queue with arrival rate ν_i and service rate μ_i.

Performance analysis of the example

Recall

- $\nu_1 = \frac{\lambda_1}{p}, \nu_2 = \frac{\lambda_1(1-p)}{p}$
- From Jackson's theorem, machine i is an M/M/1 queue with arrival rate ν_i and service rate μ_i.

Performance measures at individual machines

Machine 1

- Load: $\rho_1 = \frac{\nu_1}{\mu_1}$
- Stationary distribution
$$\pi^{(1)}(n_1) = (1-\nu_1)\nu_1^{n_1}$$
- Mean number in the queue
$$\bar{N}_1 = \frac{\rho_1}{1-\rho_1}$$
- Mean sojourn time
$$\bar{T}_1 = \frac{1}{\nu_1 - \mu_1}$$

Machine 2

- Load: $\rho_2 = \frac{\nu_2}{\mu_2}$
- Stationary distribution
$$\pi^{(2)}(n_2) = (1-\nu_2)\nu_2^{n_2}$$
- Mean number in the queue
$$\bar{N}_2 = \frac{\rho_2}{1-\rho_2}$$
- Mean sojourn time
$$\bar{T}_2 = \frac{1}{\nu_2 - \mu_2}$$

Performance analysis of the example

Performance measures of the network

- Stationary distribution

$$\pi(n_1, n_2) = \pi^{(1)}(n_1)\pi^{(2)}(n_2)$$

- Probability there are no orders in the facility

$$\begin{aligned}\pi(0,0) &= \pi^{(1)}(0)\pi^{(2)}(0) \\ &= (1-\rho_1)\cdot(1-\rho_2)\end{aligned}$$

- Mean number in the facility

$$\bar{N} = \bar{N}_1 + \bar{N}_2$$

- Mean sojourn time of orders in the facility (apply Little's law to the facility)

$$\bar{T} = \frac{\bar{N}}{\lambda}$$

Important

Orders can go through the machines several times. Therefore,

$$\bar{T} \neq \bar{T}_1 + \bar{T}_2$$

Mean sojourn time in the facility is **not** the sum of the mean sojourn times in each of the machines.