Network Performance Analysis

Guido Sanguinetti (based on notes by Neil Lawrence)

Autumn Term 2006-2007

Lectures will take place in the following locations:

Mondays 15:10 - 17:00 pm (St Georges LT8).

Tuesdays 16:10 pm - 17:00 pm (St Georges LT8).

Lab Classes In the Lewin Lab (Monday Weeks 3 and 9).

Week 1

Monday	Lectures 1 & 2	Networks and Probability Review
Tuesday	Lecture 3	Poisson Processes

Week 2

Monday	Lecture 4 & 5	Birth Death Processes, $M/M/1$ queues
Tuesday	Lecture 6	Little's Formula

Week 3

Monday	Lab Class 1	Discrete Simulation of the $M/M/1$ Queue
Tuesday	Tutorial 1	Probability and Poisson Processes

Week 4

Monday	No lecture	
Tuesday	No lecture	

Week 5

Monday	Lectures 7 & 8	Other Markov Queues and Erlang Delay
Tuesday	Tutorial 2	$M/M/1$ and Little's Formula

Week 6

Monday	Lectures 9 & 10	Erlang loss and Markov Queues review
Tuesday	Lecture 11	$M/G/1$

Week 7

Monday	Lectures 12 & 13	$M/G/1$ and $M/G/1$ with Vacations
Tuesday	Tutorial 3	Erlang Delay and Loss

Week 8

Monday	Lectures 14 & 15	Networks of Queues, Jackson's Theorem
Tuesday	Lecture 16	ALOHA

Week 9

Tuesday	Lab Class 2	Simulations of queues in NS2

Week 10

Exam Practice

Week 12

Monday	Past Paper Review	Review of the 2004-05 Exam Paper

- **Evaluation**: This course is 100% exam.

Reading List

The course is based primarily on [1]. The relevant parts of the book are as follows.

Chapter 1 For background reading and review.

Chapter 3 Section 3.1 – 3.5 For queueing models.

Chapter 4 In less detail: Section 4.1 – 4.4, Section 4.5.2, 4.5.3.

For additional background reading you may wish to consult some of the books listed below, [4] is the standard text for Introductory courses, but has little coverage on queueing theory. [2] is an alternative text for details on queues.

Your Notes

- Written by Neil Lawrence for this course in 2004.
- Designed to accompany [1], cover in extra detail areas of importance for the course and adding material where it is lacking (for example on simulation).
- First three chapters provide background material on networks and probabilities.
- Please inform me of any errors in these notes.

Introduction

- Queueing theory dates back to the beginning of the 20th Century and the pioneering work of Agner Krarup Erlang (1878-1929).
- Erlang was a Danish Telecoms Engineer who published the first paper on queueing theory [3].
- Initially, telephone call routing was achieved manually.
- The invention of the *automatic telephone exchange* (through the Strowger switch) in 1891 brought about the need for a theoretical study of the operation of telephone exchanges.
- Erlang's provided theoretical studies of these circuit switching networks.

Computer Networks

- Circuit switched networks were designed for telephone communication between two correspondents.
- By the early 1960s *packet switching networks* came into being.
- ARPANET was predecessor of the internet.

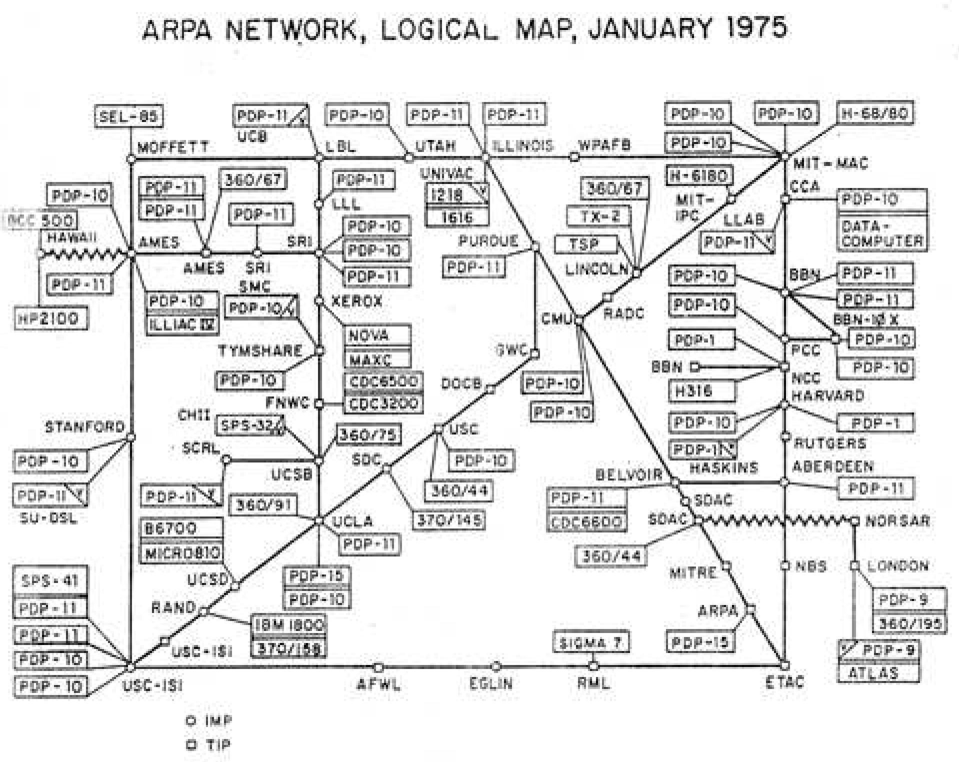


Figure 1: The ARPANET in 1975, image courtesy of the Computer History Museum, `http://www.computerhistory.org`.

Network Performance Analysis

- In this course we will introduce you to the tools for design and analysis of telecommunications networks.
- The core tools of this analysis are probability based: there is an assumption that the traffic in the system is stochastic.
- The techniques we will cover are applicable beyond the design and analysis of telecommunication networks. *e.g.* compute servers dealing with submitted jobs and databases handling queries, cashpoints queues, supermarket queues, telephone call centres *etc.*.

Practical Considerations

Broadly speaking we need to focus on two interrelated areas in our analysis:

1. **Design Issues**: when building or buying a router how much memory should we allow for? How much bandwidth do we need between routers? How much processing power should the router have so that it may exploit that bandwidth effectively?
2. **Quality Issues**: what will be the average delay for end users given our network design? What is the worst case delay? How likely is the system to drop packets?

These are important because resources are finite, there are limitations on the available bandwidth, processor speeds and memory.

Terminology

- A series of communications between two users/computers is known as a *session*.
- The interactions themselves consist of *messages* each of which may be split into *packets*.

Session Characteristics

- Characteristics of the session vary according to protocol:
  1. **Message arrival rate**. We can seek to model the arrivals as: *Poisson arrivals*, *deterministic arrivals*, *uniform arrivals* or as coming from a more general class of distributions.
  2. **Session holding time**: The length of a session.
  3. **Expected message length** and **length distribution**, message lengths can be modeled as *exponential distributions*, *uniform length*.
  4. **Allowable delay** - acceptable expected delays vary. The acceptable delay is low for VoIP and quite high for FTP.
  5. **Reliability** - FTP *must* be error free but VoIP can have some bit errors or packet loss without major consequences.
  6. **Message and packet ordering** - Some sessions will require that the packets arrive in the order they are sent: for example database requests.
- In this course we will be particularly interested in the first five of these characteristics.

Network Types

Circuit Switching: used for telephone routing. Guarantees an open ‘channel’ between two peers.

Packet Switching or Store and Forward Switching: used for data. May operate via

- **Dynamic Routing** Each packet of the message moves through the network independently.
- **Virtual Circuit Routing** A fixed path through the network is provided for the entire message.

The character of the two approaches are quite different.

Circuit Switching

- wasteful in terms of bandwidth. Nodes may not be using all the available bandwidth reserved for them.
- a guaranteed channel is provided with a known communication delay associated with it.
- the sum of rates for all sessions using a link cannot exceed the total capacity of the link. New sessions must be rejected when the link capacity is reached.
- used for telephone networks where the session rate, r_s, is the same for each session. Switching would be more complex for data networks and the nature of transmission would also mean that the links are inefficiently used.

Packet Switching

- Packet switching makes efficient use of bandwidth because, whilst communication between two nodes may be idle for a period of time, packets from another communication may be sent along the same line.
- Because it is unknown *a priori* how much bandwidth will be available it is difficult to guarantee a maximum delay.

Communication Media

- Circuit switching networks may share a communication through multiplexing (*e.g.* time division multiplexing or frequency division multiplexing.)
- Both are not 100% efficient (time multiplexing requires a small amount of time for switching to occur, and frequency division multiplexing requires a small 'buffer' band of frequencies between each session.
- Packet switching networks dedicate communication bandwidth to each channel as required. In a simple 'first in, first out'[1] (FIFO) system, packets arrive and, if the queue is empty, they are transmitted. Otherwise they are stored and await their turn for transmission.
- In packet switching there are two delays: a queueing delay and a transmission delay. The queueing delay is variable, transmission delay will be fixed .

[1] Also known as 'first come, first served' (FCFS).

The Road Ahead

Probability Review Review of probability theory.

Simple Queuing models Simple models of queues based on probability theory. Little's formula, $M/M/1$ queue, $M/M/\infty$, $M/M/m$, $M/M/m/m$ and $M/G/1$.

Simulation Simulation of queueing systems.

Examples

1. As a simple example let's consider a database (perhaps MySQL). We are constructing it from existing parts. We have two CPU/motherboard systems and two separate disk systems, their expected characteristics are shown in Table 2. Which should we match with which? Are we better off building one faster system and one slower system? Or should we build two systems that have a similar performance?

CPU	Service Time/ms	Disk	Service Time/ms
A	423	1	215
B	210	2	390

Table 2: Average times for a fictional database server for three different cases.

Probability Review

- Health warning: Mathematics and Probability are FUNDAMENTAL to this course.
- We are interested in trials which result in two random variables, X and Y, each of which has an 'outcome' denoted by x or y.
- We summarise the notation and terminology for these distributions in the following table.

Terminology	Notation	Description
Joint Probability	$P(X=x, Y=y)$	'The probability that $X=x$ and $Y=y$'
Marginal Probability	$P(X=x)$	'The probability that $X=x$ regardless of Y'
Conditional Probability	$P(X=x \mid Y=y)$	'The probability that $X=x$ given that $Y=y$'

Table 3: The different basic probability distributions.

A Pictorial Definition of Probability

- Here is a review of what probabilities are in a two dimensional graph.

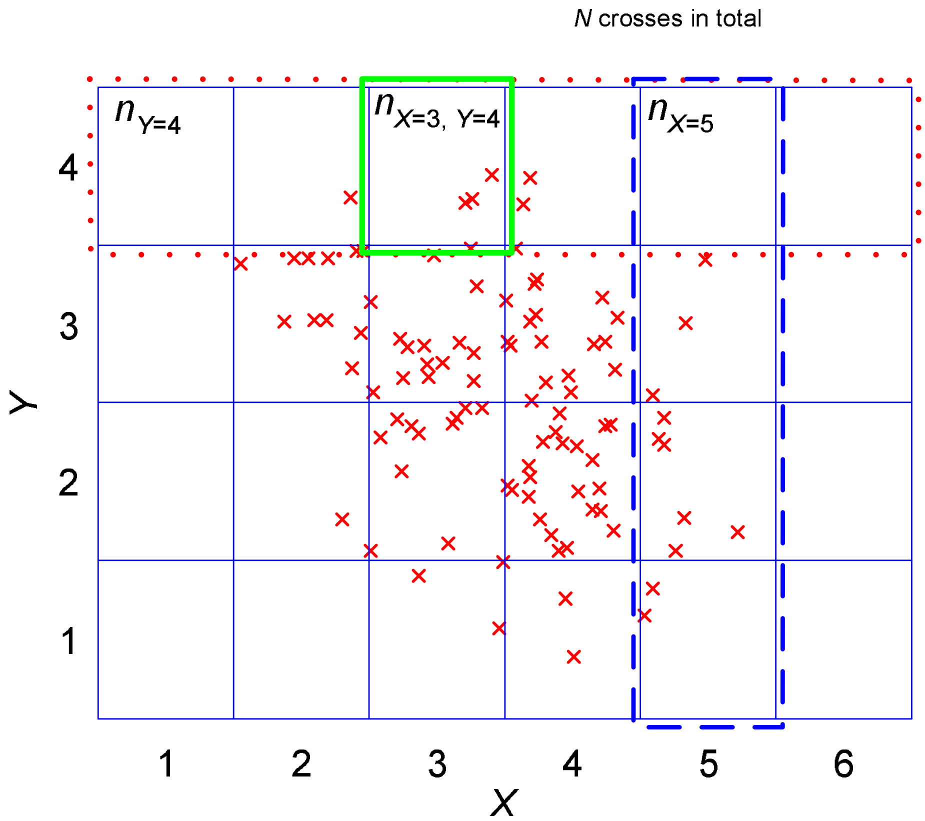


Figure 2: Representation of joint and conditional probabilities.

Different Distributions

Terminology	Definition
Joint Probability	$\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{N} = P\left(X=3, Y=4\right)$
Marginal Probability	$\lim_{N\to\infty} \frac{n_{X=5}}{N} = P\left(X=5\right)$
Conditional Probability	$\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{n_{Y=4}} = P\left(X=3 \vert Y=4\right)$

Table 4: Definition of probability distributions from Table 3 in terms of the system depicted in Figure 2.

Notational Details

- Typically we should write out $P(X = x, Y = y)$.
- In practice, we often use $P(x, y)$.
- This looks very much like we might write a multivariate function, *e.g.* $f(x, y) = \frac{x}{y}$.
  - For a multivariate function though, $f(x, y) \neq f(y, x)$.
  - However $P(x, y) = P(y, x)$ because $P(X = x, Y = y) = P(Y = y, X = x)$.
- We have denoted the probability distribution with a capital P.
- We will use capital Ps consistently to denote discrete distributions.
- For continuous probability density functions, we will turn to small letters.
- We now consider some 'rules of probability'.

Normalisation

All distributions are normalised. This is clear from the fact that $\sum_x n_{X=x} = N$, which gives

$$\sum_x P(x) = \lim_{N\to\infty} \frac{\sum_x n_{X=x}}{N} = \lim_{N\to\infty} \frac{N}{N} = 1.$$

A similar result can be derived for the marginal and conditional distributions.

The Sum Rule

- The marginal probability $P(y)$ is given by $\lim_{N\to\infty}\frac{n_{Y=y}}{N}$.
- The joint distribution $P(x,y)$ is given by $\lim_{N\to\infty}\frac{n_{X=x,Y=y}}{N}$.
- $n_{Y=y}=\sum_x n_{X=x,Y=y}$ so

$$\lim_{N\to\infty}\frac{n_{Y=y}}{N}=\lim_{N\to\infty}\sum_x\frac{n_{X=x,Y=y}}{N},$$

in other words

$$P(y)=\sum_x P(x,y).$$

This is known as the sum rule of probability.

The Product Rule

- $P\left(x|y\right)$ is given by $\lim_{N\rightarrow\infty}\frac{n_{X=x,Y=y}}{n_{Y=y}}$.
- $P\left(x,y\right)$ is given by

$$\lim_{N\rightarrow\infty}\frac{n_{X=x,Y=y}}{N}=\lim_{N\rightarrow\infty}\frac{n_{X=x,Y=y}}{n_{Y=y}}\frac{n_{Y=y}}{N}$$

or in other words

$$P\left(x,y\right)=P\left(x|y\right)P\left(y\right).$$

This is known as the product rule of probability.

Bayes' Rule

- From the product rule,

$$P(x, y) = P(y, x) = P(y|x)\, P(x)\,,$$

so

$$P(x|y)\, P(y) = P(y|x)\, P(x)$$

which leads to Bayes' Rule,

$$P(x|y) = \frac{P(y|x)\, P(x)}{P(y)}.$$

Independence

- The variable x is independent of y if the distribution over x is not affected by the value of y:

$$P(x) = P(x|y),$$

the conditional distribution of x given y is equal to the marginal distribution of x. This in turn implies that the joint distribution factorises,

$$P(x, y) = P(x|y)\,P(y) = P(x)\,P(y),$$

and furthermore that

$$P(y|x) = P(y).$$

The factorisation of the joint distribution when variables are independent turns out to be a very important property which makes a lot of analysis tractable.

Expectations

- A probabilistic model summarises our beliefs.
- Sometimes need to summarise the model by its expectations.
- Expected value of $f(x)$ under $P(x)$ is

$$\langle f(x) \rangle_{P(x)} = \sum_{x} P(x) f(x) .$$

Here we have introduced a particular notation for the expectation, you will also see expectations written in the form $E\{f(x)\}$.

Expectations under a Simple Distribution

- Define a distribution over x in Table 5.

x	0	1	2
$P(x)$	0.2	0.5	0.3
$f(x) = x^2$	0	1	4

Table 5: Distribution over x and associated values of $f(x)$.

- We can compute

$$\langle f(x) \rangle = \left\langle x^2 \right\rangle = 0 \times 0.2 + 1 \times 0.5 + 4 \times 0.3 = 1.7,$$

have dropped the subscript $P(x)$

- We can also see that the expectation of x is

$$\langle x \rangle = 0 \times 0.2 + 1 \times 0.5 + 2 \times 0.3 = 1.1.$$

- The expectation of x is known as the first moment of the distribution or its *mean*.
- The expectation of x^2 is known as the second moment — normally used to obtain the variance.
- Variance: given by the difference between the second moment and the square of the mean. It is often denoted by σ^2.

$$\sigma^2 = \left\langle x^2 \right\rangle - \langle x \rangle^2 .$$

For our example above the variance is given by

$$1.7 - (1.1)^2 = 0.49,$$

- Its square root is known as the standard deviation

$$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2},$$

$$\sqrt{1.7 - (1.1)^2} = 0.7.$$

Distributions as Functions

- Probability distribution with a simple table can be limiting.
- The Poisson Distribution — a distribution a a function
- First published by **Siméon Denis Poisson** (1781-1840) in 1837.
- Defined over the space of all non-negative integers.
- This set is countably infinite: impossible to summarise in a table!
- The Poisson distribution is therefore defined as

$$P\left(k|\mu\right) = \frac{\mu^k}{k!} \exp\left(-\mu\right). \tag{1}$$

where k is any integer from 0 to ∞, and μ is a parameter of the distribution.

A Poisson with $\mu = 2$

- To work out the probability of k in a Poisson $\mu = 2$ we can start filling a table.
- The values in a table are computed from (1)

k	0	1	2	$\ldots$
$P(x)$	0.135	0.271	0.271	$\ldots$

Table 6: Some values for the Poisson distribution with $\mu = 2$.

Mean and Variance

- Example 1, Chapter 3: Mean and Variance of Poisson Distribution

  As we mentioned in Section , the expectations under a distribution are important for summarising the distributions characteristics. Expectations of distributions that take the form of a function can also be computed. Typically we expect the resulting expectation to be a function of the distributions parameters.
- The variance of a Poisson is also μ — see Tutorial 1.

Continuous Variables

- So far discrete values of x or y.
- For continuous models we introduce the *probability density function* (PDF).
- Discrete case: defined probability distributions over a discrete number of states.
- How do we represent continuous as probability?
- Computer science student's height:
  - ➥ Develop a representation which could answer *any* question we chose to ask about a computer scientist's height.

Gaussian PDF

- One example is the Gaussian: we can query the distribution by integrating over regions of it.

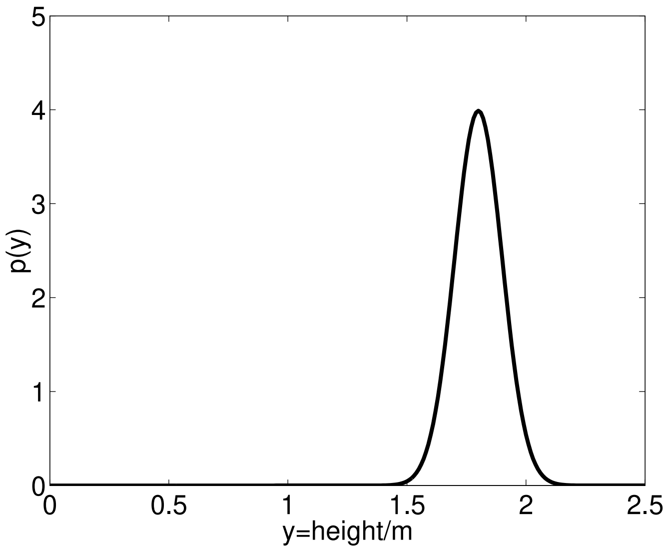


Figure 3: A Gaussian PDF representing heights of computer scientists.

- Represent a PDF with a small $p(\cdot)$ to differentiate it from a probability distribution: $P(\cdot)$.
- It is a positive function, integral over the region of interest is one[2].

[2] In what follows we shall use the word distribution to refer to both discrete probabilities and continuous probability density functions.

Manipulating PDFs

- Same rules for PDFs as distributions *e.g.*

$$p\left(y|x\right) = \frac{p\left(x|y\right)p\left(y\right)}{p\left(x\right)}$$

where $p\left(x,y\right) = p\left(x|y\right)p\left(y\right)$ and for continuous variables $p\left(x\right) = \int p\left(x,y\right)dy$.

- Expectations under a PDF

$$\left\langle f\left(x\right)\right\rangle_{p(x)} = \int f\left(x\right)p\left(x\right)dx$$

where the integral is over the region for which our PDF for x is defined.

Cumulative Distribution Functions

- PDF doesn't represent probabilities directly
- One very common question is: what is the probability that $x < y$?
- The cumulative distribution function (CDF) represents the answer for $-\infty < x < \infty$ the CDF is given by

$$P\left(x > y\right) = \int_{-\infty}^{y} p\left(x\right) dx,$$

for $0 \leq x < \infty$ then the CDF is given by

$$P\left(x > y\right) = \int_{0}^{y} p\left(x\right) dx.$$

Gaussian PDF and CDF

- In Figure 3 example of the CDF for the students' heights example.

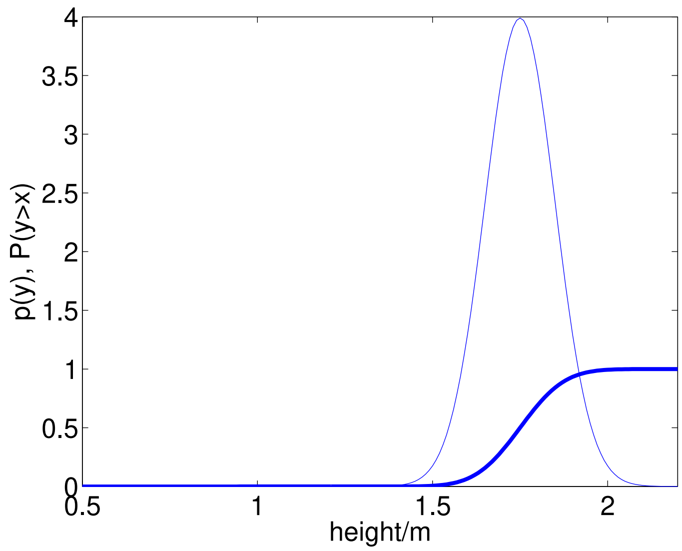


Figure 4: The cumulative distribution function (CDF) for the heights of computer science students. The thick curve gives the CDF and the thinner curve the associated PDF.

- The PDF can be recovered from the CDF through differentiation.

Sample Based Approximations

- Recall that expectation of any given function, $f(k)$, under a distribution, $P(k)$, can be written as

$$\langle f(k) \rangle_{P(k)} = \sum_k f(k) P(k)$$

Recall our definition of probability from Table 2.

$$P(k) = \lim_{S \to \infty} \frac{s_k}{S}$$

where we have now used S (instead of N)

- *Sample based approximations* hope that S is 'large enough' such that it to say

$$\langle f(k) \rangle_{P(k)} \approx \sum_k f(k) \frac{s_k}{S}.$$

Sample Based Approximations contd

- This formula may be rewritten as:

$$\sum_k f(k) \frac{s_k}{S} = \frac{1}{S} \sum_k f(k)\, s_k.$$

For $\mathbf{k}$ a vector of samples from $P(k)$with elements k_i and length S

$$\frac{1}{S} \sum_k f(k)\, s_k = \frac{1}{S} \sum_{i=1}^{S} f(k_i)$$

- This provides us with a sample based approximation

$$\langle f(k) \rangle_{P(k)} \approx \frac{1}{S} \sum_{i=1}^{S} f(k_i)\,.$$

- Special cases of this include the *sample mean,* often denoted by $\bar{k}$, and computed as

$$\bar{k} = \frac{1}{S} \sum_{i=1}^{S} k_i,$$

Sample Mean vs True Mean

- This is an approximation to the true distribution mean

$$\langle k \rangle \approx \bar{k}.$$

- The same approximations can used for continuous PDFs, so we have

$$\begin{aligned} \langle f(x) \rangle_{p(x)} &= \int f(x)\, p(x)\, dx \\ &\approx \frac{1}{S} \sum_{i=1}^{S} f(x_i), \end{aligned}$$

  where x_i are independently obtained samples from the distribution $p(x)$.

- Approximation gets better for increasing S and worse if the samples from $P(k)$ are *not* independent.

Note on Sample and Distribution Means

In most texts the word mean is used interchangeably to describe sample based means and the true mean of a distribution. However, you should always bear in mind the important difference between the two. Later in the course we will also come across time based averages and assumptions will be made that the time based average is equal to the average under a 'steady state distribution'.

Stochastic Processes

- Also known as probabilistic process, random process
- Can be viewed as a ‘random function’.
  - Normal function maps a set of inputs to a point value.
  - A probabilistic process maps a set of inputs to a probability distribution.
- *Wiener process* is the process behind Brownian motion.
- Stock market data is often modelled as a stochastic process.
- Definition: A random process

  May be defined as a *random function* which, for a particular value, is a random quantity.

Poisson Process

- A simple process which is related to the Poisson distribution: the Poisson process.
- Probabilities of a number of arrivals, k, as a function of time, t, given a 'rate', λ.
- $$P\left(k|t,\lambda\right) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}. \tag{2}$$
- The process is simply the distribution with μ substituted with λt.
- λ is now known as a rate, because product λt gives the expected number of arrivals at time t.
- The mean of the Poisson process (the first moment of k) is given by
$$\langle k \rangle = \lambda t.$$
- Poisson process is a form of counting process where the counts are expected to increase linearly with time.

Properties

- Superposition of Poissons
  If we combine N independent processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_N$ together the resulting process is also a Poisson process with a combined rate, λ, equal to the sum of each of the input rates: $\lambda = \sum_{n=1}^{N} \lambda_n$. This is shown pictorially in Figure 5.

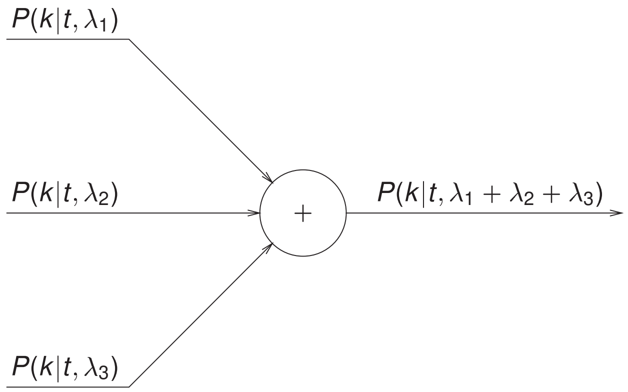


Figure 5: If independent Poisson processes form a single traffic flow overall process is a Poisson with a rate that s the sum of the rates of the separate processes.

Example of Poisson Superposition

- Example 1, Chapter 4

  There are three queues for registration of MSc students from different courses at the University of Sheffield. Students for the first course, Physics and Fine Art, arrive with a rate of 4 students per minute. Students for the second course, The Psychology of Sociology, arrive with a rate of 2 students per minute and students for the third course, Computer Science and Aromatherapy arrive with a rate of 5 students a minute.

  Assuming the student arrivals are independent and occur as a Poisson process what is the probability that after 15 minutes there have been **a)** 90 student arrivals? **b)** 180 arrivals?

Properties Contd

- Decomposition of Poisson Flow

  Assume we have one input channel where arrivals are occurring with a Poisson process with rate λ. This channel then splits into N separate output channels. Each input arrival is then independently assigned to output channel n with probability $P(n)$. In this case the output channels then behave like Poisson processes with rates given by $\lambda_n = P(n)\lambda$. This is shown in Figure 6.

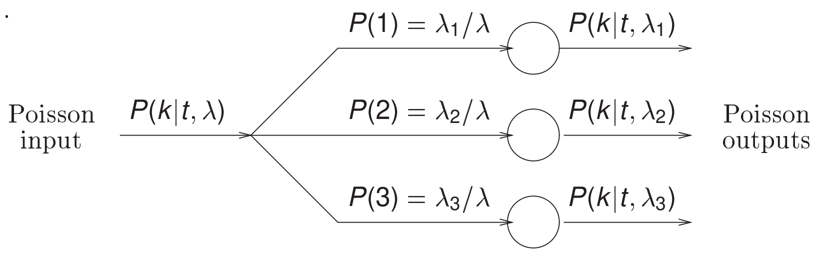


Figure 6: If a single Poisson process with rate λ is switched to different output channels, with each arrival being switched to channel n with probability $P(n) = \lambda_n/\lambda$, each output channel will be a Poisson process with rate λ_n.

Example of Poisson Decomposition

- Example 2, Chapter 4

  There is one registration for all MSc courses. Students arrive in the hall and are assigned to registration desks according to the first letter of their surname. Desk 1 ($d = 1$)takes students with names from A-G, desk 2 ($d = 2$) from H-M, desk 3 ($d = 3$) from N to S and desk 4 ($d = 4$) from R to Z.

  Surnames are distributed amoungst these four categories with the following probability distribution:

d	1	2	3	4
$P(d)$	0.20	0.27	0.25	0.28

  Assuming the students arrive independently of their surname as a Poisson process at a rate of 10 students per minute what is the probability that **a)** after 5 minutes 5 students have arrived at desk 1? **b)** after 10 minutes 20 students have arrived at desk 4?

Properties III

- Independence of Arrivals

  Consider a single Poisson process, if we look within this process at two *disjoint* time intervals (*i.e.* periods of time that don't overlap). Assume we have k_1 arrivals in the first interval and k_2 arrivals in the second interval. The distribution of arrivals in the two intervals are independent of each other, $P(k_1, k_2) = P(k_1)P(k_2)$.

Properties IV

- Similarity of Arrival Distributions for Identical Intervals

  Consider a single Poisson process with rate λ, let the number of arrivals between time s and $s+t$ be given by k_1. Now consider a second independent Poisson process with the same rate λ, let the number of arrivals between time r and $r+t$ be given by k_2. The distribution of k_1 and k_2 are then equal,

$$P\left(k_1\right)=P\left(k_2\right).$$

Interarrival Times and Poisson Processes

- Interarrival time: time between arrivals.
- Think of buses.
- X_i, are continuous values, what is their PDF?.

Interarrival Distribution

- Example 3, Chapter 4. Derivation of Interarrival Time Distribution

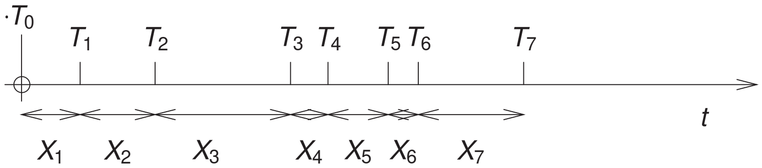


Figure 7: For a Poisson process, this figure shows the times of the arrivals, T_n, and the interarrival times, X_n.

In the figure above we show different arrival times $T_1, \dots, T_n$ for the process. You can think of these as the arrival times of the buses. We are interested in computing the distribution of the variables $X_i = T_i - T_{i-1}$: the times between bus arrivals. In this example you will derive this PDF for the interarrival times[3].

[3] The interarrival times, X_i, are continuous values, so they will be governed by a probability density function rather than a discrete probability distribution.

Properties of the Exponential Distribution

- The Memoryless property

$$P\left(X > s + t | X > t\right) = P\left(X > s\right)$$

- In English, the probability of the waiting time being greater than $s + t$ given that the waiting time is already greater than t is the same as the probability of the waiting time being greater than s.
- Mean and Variance

$$p(x) = \lambda e^{-\lambda x}$$

The mean of an exponential can be found to be,

$$\langle x \rangle_{p(x)} = \frac{1}{\lambda}$$

and the variance is

$$\left\langle x^2 \right\rangle_{p(x)} - \langle x \rangle^2_{p(x)} = \frac{1}{\lambda^2}.$$

These are easily confirmed — they are part of Tutorial 1.

Markov Chains and Markov Processes

Poisson process in a special case of the more general class of processes known as Markov processes. We will start by considering Markov processes as they operate in *discrete time*, otherwise known as *Markov chains*. Both Markov processes and Markov chains exhibit the *Markov property* which is an assumption about how future values drawn from the process depend on past values. We will first discuss the Markov property in the context of Markov chains, then we will show how a discrete system can be converted to a continuous system.

Markov Chains

- Consider a discrete representation of time $t = 0, 1, 2\ldots$. We associate a state variable, N, with this time such that we have values for N at $N(0)$, $N(1)$, $N(2)$, ...
- Turns in a board game.
- Markov Property
  - value at time t is dependent only on its value at time $t - 1$.
  - The future is conditional independent of the past (the conditioning being on the present).
  - State of the system in the future is *not* dependent on what happened in the past if we know the state at the present time.

Monopoly

- A real-life example of a Markov chain
- Forty squares on a Monopoly board.

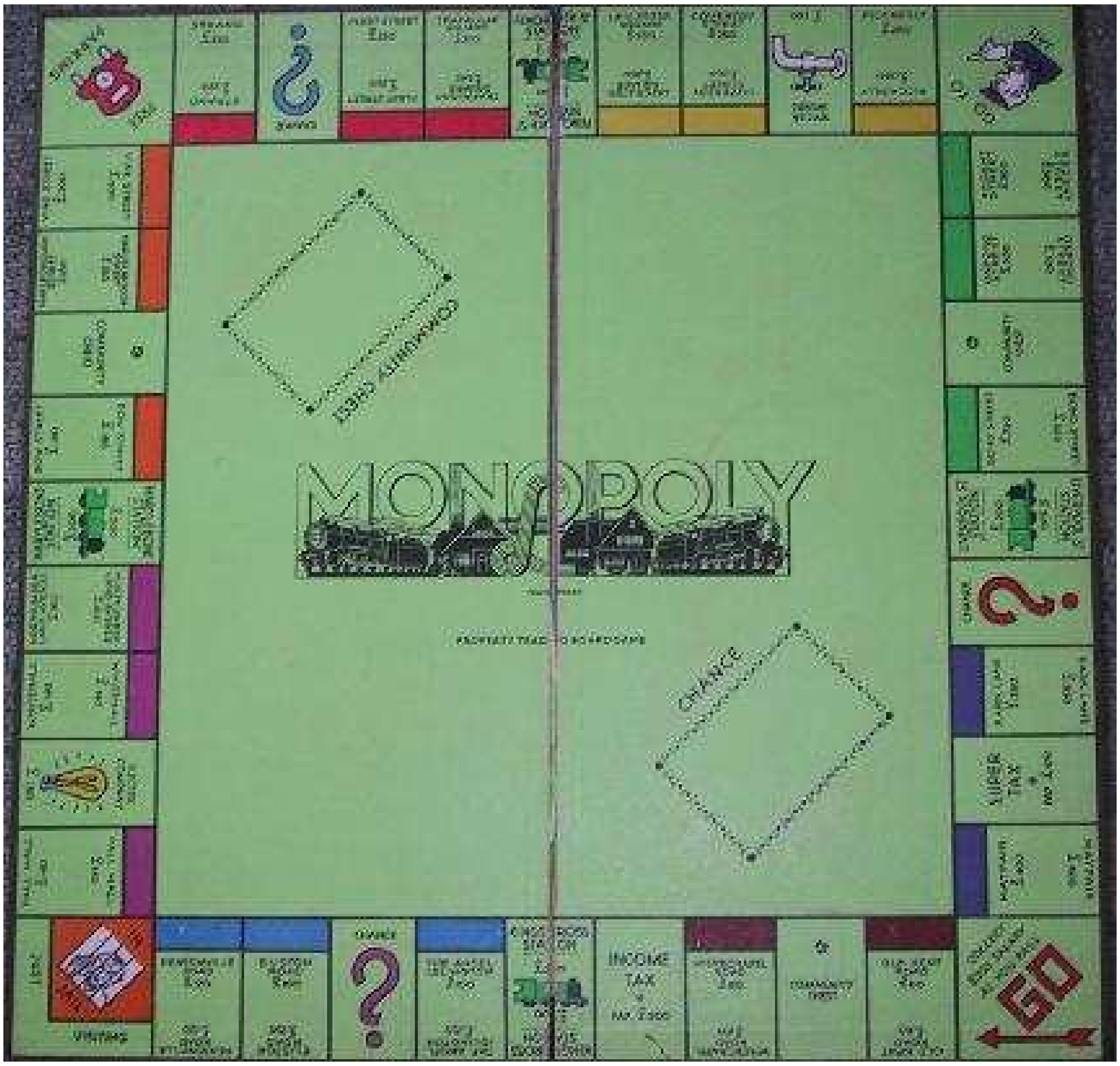


Figure 8: Image of a monopoly board.

Markov Chain Monopoly

- Player's state, $N(t)$, square number $+40\times$number of 'GO's

  Normally move through two six sided dice and adding them together. Denote the result of this die roll $d(t)$.
- State at time t is

$$N(t) = N(t-1) + d(t).$$

Monopoly exhibits the Markov property, future states are independent of your previous states given your current state.

State Diagrams

A Markov chain can be represented by a state diagram just as finite state machines are often represented in this manner. One important difference is that Markov chains are not necessarily finite, they can have infinity states! A further difference is that the transitions for Markov chains are associated with probabilities (or in the case of Markov processes rates).

Poisson Process as a Markov Process

We've introduced the game of Monopoly as a Markov chain. In this next session we shall relate Markov chains, where time is treated as a discrete variable, to Markov processes, where time is a continuous variable. We will show how the Poisson process is a special case of a Markov process where a constrained set of transitions are allowed. To do this we will first introduce a new version of Monopoly that we call 'Coin Monopoly'.

Coin Monopoly

- Player tosses the coin when it is their turn.
- We take $d(t) = 0$ if a coin is tails and $d(t) = 1$ if it is heads. This is depicted as a state diagram in Figure 9.

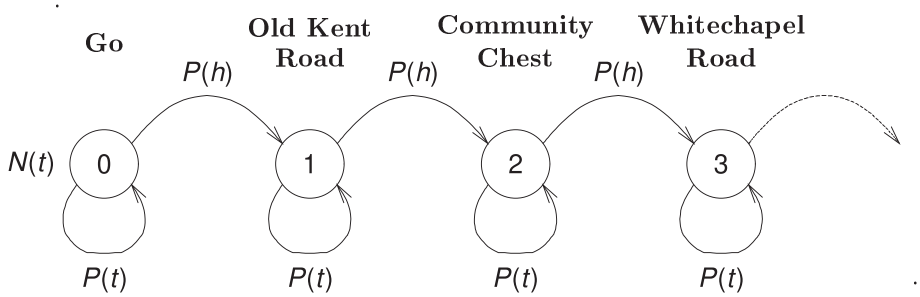


Figure 9: State diagram for coin monopoly. Each node (circle) represents a different possible state value. The state changes (or doesn't change) with each time step according the probabilities associated with the arrows.

- Takes a long time to play ... let's speed it up.

Fast Coin Monopoly

- Time, t, originally changes by $\Delta t = 1$.
- Special coin with $P(h) = \lambda \Delta t$ (which must be less than 1), giving $P(t) = 1 - \lambda \Delta t$.
- Use the *binomial distribution* to compute how many heads.

Binomial Distribution for n Trials

- Example 1, Chapter 5. It is possible to compute a probability distribution for a number of successes given a number of trials. This distribution is known as the binomial. For n trials (equivalent to n coin tosses or n turns for Coin Monopoly) the probability of k successes (equivalent to k moves forward) is given by

$$P(k) = \left(\begin{array}{c} n \\ k \end{array} \right) P(h)^k (1 - P(h))^{n-k} \qquad (3)$$

where $\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$ is known as the binomial coefficient.

By considering the limiting case as the probability of a success (or heads) goes to zero but the number of trials goes to ∞ show that a Poisson process can be recovered.

Markov Process State Diagrams

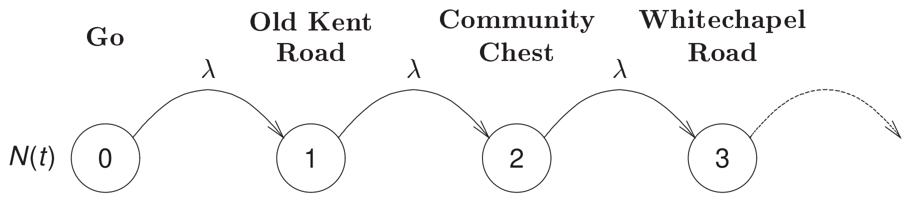


Figure 10: State diagram for the Markov process given by Coin Monopoly in the limit as $\Delta t \to 0$. This is the state diagram for a Poisson process. Notice in comparison to Figure 9, the probabilities have been replaced by rates and there are no self-transition arrows.

Summary

You've now seen that a Poisson process can be derived starting from a simple Markov chain. This gives some deeper insight to a Poisson process. In the coin monopoly model there is a fixed (constant) probability of moving forward in every turn and whether a move forward occurs in each turn is completely independent of what happened in previous turns. The Poisson process is the continuous analogue of this stem. There is a constant probability of an arrival occurring at any time which is independent of what has gone before. This gives you an intuitive explanation of the two properties of the Poisson process — similarity and independence.

Steady State

- Steady state behaviour vs the transient dynamics.
- What do we mean by steady state?
  1. A stable condition that does not change over time or in which change in one direction is continually balanced by change in another.[4]
  2. A steady state is a dynamic equilibrium.
- A Markov process *can* reach a steady state distribtuion.

[4] http://www.dictionary.com.

Steady State contd

- A probabilistic process $p(k|t)$ is a 'noisy function' of time.
  - If process converges to a distribution as time $\longrightarrow \infty$ then it has a steady state.

$$\lim_{t\to\infty} p(k|t) = p(k)$$

    the average value of k at steady state is then given by

$$\langle k \rangle_{p(k)} = \sum_{k=0}^{\infty} p(k)\, k.$$

- A Poisson process doesn't have a steady state distribution.

Coin Monopoly Revisited

- 'Coin Monopoly' Markov chain (as we described it) has no *steady state* distribution
  - ➥ Over time the value of k will always increase.
- In this lecture introduce a (slightly) more complex Markov process: the *birth-death processes*.
- A birth-death process can have a steady state distribution.
- We use detailed balance to discover this distribution.

Birth Death Process

- A model for births and deaths in a population.
- Recall the Markov chain underlying 'Coin Monopoly'.
  - Modify this diagram: allow backwards moves.
  - Prevent state from going negative.
  - This system is known as a birth-death process.

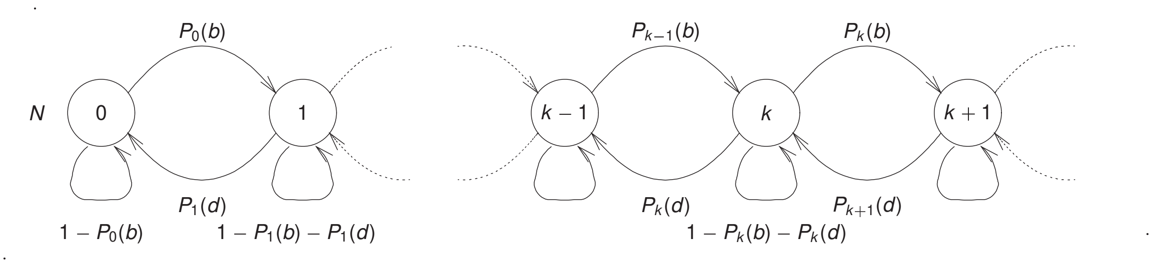


Figure 11: State diagram for a birth-death chain.

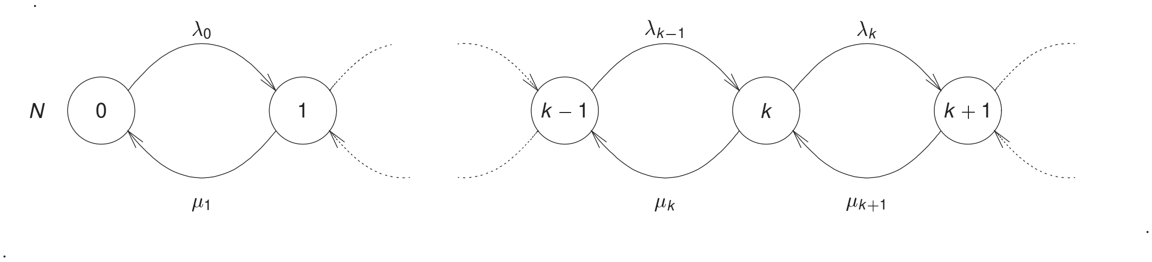


Figure 12: State diagram for a birth-death process.

Kendall's Notation

- Simple way of summarising the characteristics of a queue.

Arrival characteristics	/	Departure characteristics	/	number of servers	/	max number of customers

The arrival and departure characteristics are summarised by

- M — Markov — Poisson distributed arrivals, exponentially distributed service times.
- D — Deterministically distributed times.
- G — General distribution for inter-arrival/service times.

Global Balance

- Seek a stationary distribution through global balance equations.
- At equilibrium the frequency of transitions out of state k equals the frequency of transitions into k [1, pg 260-262].
- Markov chain can be defined by a transition probability matrix.
  - Gives the probability of transition from any state to any other.

$$\mathbf{P} = \begin{bmatrix} P_{11} & \cdots & P_{i1} & \cdots & P_{N1} \\ \vdots & \ddots & & & \vdots \\ P_{1j} & & P_{ij} & & P_{Nj} \\ \vdots & & & \ddots & \vdots \\ P_{1N} & \cdots & P_{iN} & \cdots & P_{NN} \end{bmatrix}.$$

  where probability of transition from state i to state j is given by P_{ij}.

Global Balance II

- If $P(j)$ is steady state probability for state j, global balance says

$$P(j)\sum_{i=0}^{\infty} P_{ji} = \sum_{i=0}^{\infty} P(i)\, P_{ij}$$

"the probability of leaving state j should be equal to the probability of arriving in state j."

- Markov Process: rates of transfer.
- Rates of transfer between states are given by a matrix with elements q_{ij} for transfer between i and j. Global balance equations give

$$P(j)\sum_{i=0}^{\infty} q_{ji} = \sum_{i=0}^{\infty} P(i)\, q_{ij}.$$

Detailed Balance

- For the birth death processes transitions only between state k and $k+1$ or $k-1$.
  - ➥ $P_{ij} = 0$ if $i > j+1$ and $i < j-1$.
- Special case of the global balance: detailed balance equations. Looks at the flow between two states.
- For the discrete case

$$P(k-1)\, P_{k-1,k} = P(k)\, P_{k,k-1}$$

- For the continuous case

$$P(k-1)\, q_{k-1,k} = P(k)\, q_{k,k-1}.$$

- With our notation rate from $q_{k-1,k} = \lambda_{k-1}$ and $q_{k,k-1} = \mu_k$ giving

$$P(k-1)\, \lambda_{k-1} = P(k)\, \mu_k.$$

Boundary Flow

- Consider the flow rates at a *boundary* between two states, $k - 1$ and k.

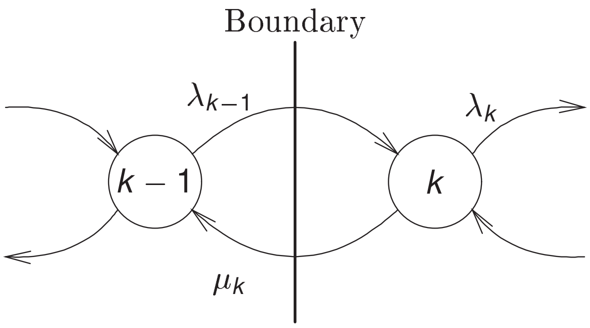


Figure 13: Flow rates at a Boundary for a birth-death process.

- At steady state
  - flow from $k - 1$ to k is $\lambda_{k-1} P(k-1)$,
  - flow from k to $k - 1$ is $\mu_k P(k)$.
- For flow to be balanced we have

$$\lambda_{k-1} P(k-1) = \mu_k P(k) .$$

- For different systems birth rate and death rate will differ. First we consider $M/M/1$ queue.

Kendall's Notation — $M/M/1$

Arrival characteristics	/	Departure characteristics	/	number of servers	/	max number of customers

$$M/M/1$$

- M — Markov — Poisson distributed arrivals, exponentially distributed service times.
- D — Deterministically distributed times.
- G — General distribution for inter-arrival/service times.

$M/M/1$ Queue

- $M/M/1$ arrivals occur as a Poisson process, service times are exponentially distributed.
- A birth-death process
  - constant birth rate $\lambda_k = \lambda$ (from the Poisson process)
  - constant death rate (from the exponentially distributed service times) $\mu_k = \mu$.
- State diagram is

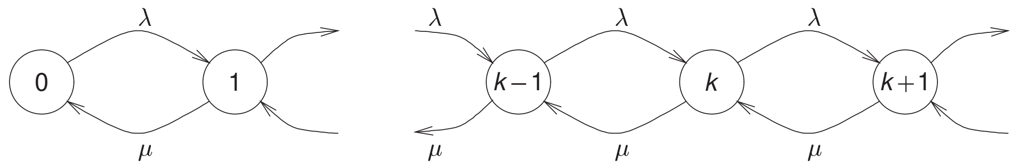


Figure 14: State diagram for $M/M/1$ queue.

- Derive stationary distribution is derived using detailed balance.

$M/M/1$ Stationary Distribution

- The state diagram highlights several points.
  1. There are no transitions to a state below 0.
  2. There is no maximum value for the state shown, this implies that k is unbounded.
  3. The birth rates and death rates for each state are indicated.
  4. Transitions are only possible between neighbouring states.
- Flow at boundary: analyse between k and $k-1$.

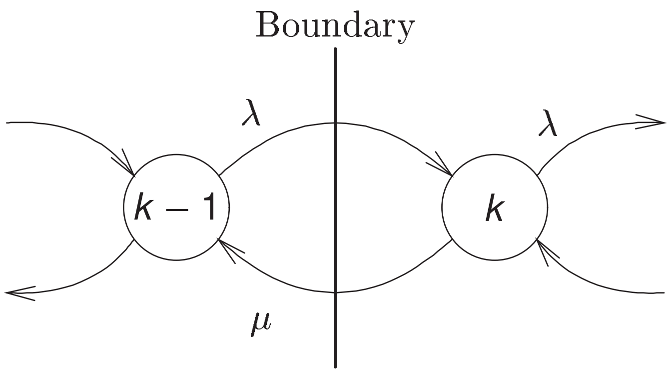


Figure 15: Flow rates at a boundary.

- At steady state,

$$\lambda P\left(k-1\right) = \mu P\left(k\right).$$

$M/M/1$ — Stationary Distribution II

- At steady state,

$$\lambda P(k-1) = \mu P(k).$$

- Leading to

$$P(k) = \frac{\lambda}{\mu} P(k-1).$$

- Thus

$$\begin{aligned} P(k) &= \left(\frac{\lambda}{\mu}\right)^k P(0) \\ &= \rho^k P(0), \end{aligned}$$

where we have defined $\rho = \frac{\lambda}{\mu}$, which, for the $M/M/1$ queue, is called the *utilisation factor*.

Normalisation

- To find $P(0)$ we use the normalisation condition.

$$\sum_{k=0}^{\infty} P(k) = 1$$

and because $P(k) = \rho^k P(0)$,

$$1 = P(0) \sum_{k=0}^{\infty} \rho^k$$

giving

$$P(0) = \frac{1}{\sum_{k=0}^{\infty} \rho^k}. \tag{4}$$

- The infinite series in the denominator converges if $\rho < 1$ (when $\lambda < \mu$),

$$\sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho} \quad \text{for} \quad \rho < 1,$$

so $P(0) = (1-\rho)$ and the full distribution for the *steady state* probabilities is

$$P(k) = \rho^k (1-\rho).$$

Faster Arrivals than Service

Notice that if the arrival rate (birth rate) is greater than the service rate (death rate) then $\rho > 1$ and the series in (4) doesn't converge. This is indicative of the fact that if the birth rate is greater than the death rate there is no steady state distribution. The queue continues to grow in size.

Average Number in System

- Network design: interested in the average queue length.
- We use the steady state probabilities to compute this quantity.
- The average number of customers in the system *is not* the same as the average queue length
  - ➥ a customer being served (*e.g.* a packet under transmission) is not in the queue.
- Average number of customers in the system is the expected state of the system.

$$\langle k \rangle_{P(k)} = \sum_{k=0}^{\infty} kP(k)$$

Average Number in System II

Example 1, Chapter 6. The average number in the system for the $M/M/1$ queue is given by considering the steady state distribution. Compute this value.

The variance is computed as an excercise in the Tutorial Sheet.

$$\mathrm{var}\left[k\right] = \left\langle k^2 \right\rangle - \left\langle k \right\rangle^2 = \frac{\rho}{(1-\rho)^2}.$$

Discrete Simulation of $M/M/1$

- Next week: first lab class.
- Discrete simulation of an $M/M/1$ queue: easy to run because it takes the form of a Markov chain.
- State diagram for this system.

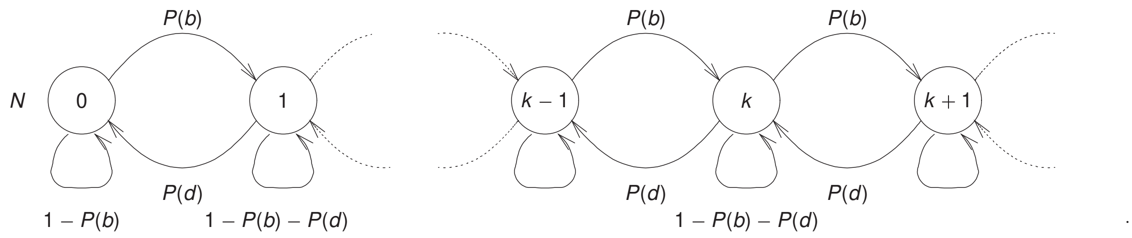


Figure 16: State diagram for the discrete version of the $M/M/1$ queue.

Hints for the Simulation

- Use a loop over discrete time steps.
- At each time step there are three possible events:
  - the state of the system increases by one;
  - the state of the system decreases by one and
  - the state of the system stays the same.
- Modelling the queue is therefore a simple matter of testing which of these outcomes is true.

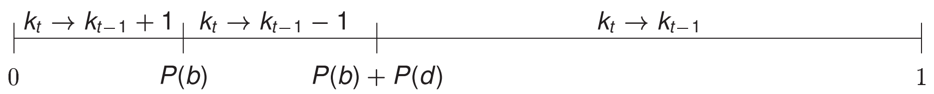


Figure 17: Testing for one of three outcomes. Generate a random number and see where it lands.

Approximating a Continuous $M/M/1$ Queue

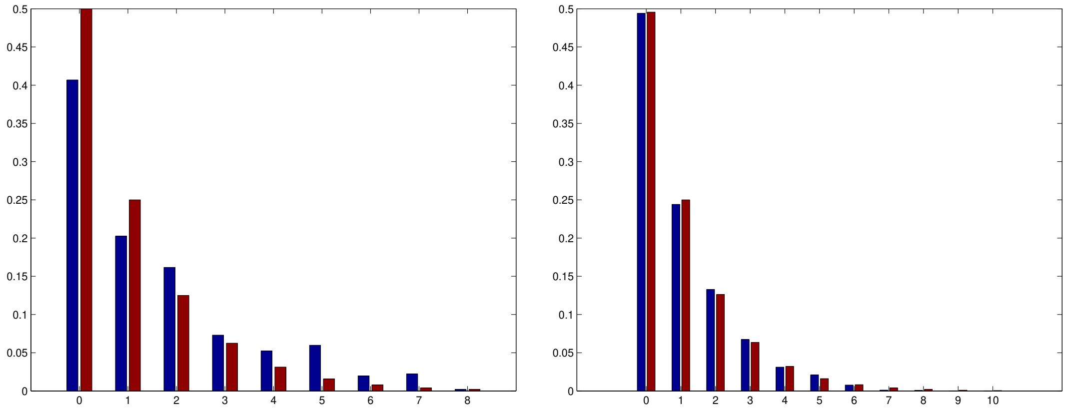
- We know from last week that a Poisson process can be derived by considering the limit of a simple Markov chain.
- A continuous $M/M/1$ queue is the limit of the discrete model we are using for simulation
- In your lab class you can make ths approximation by taking Δt to be much smaller than λ or μ.
- Then you may set $P\,(b) = \lambda\Delta t$ and $P\,(d) = \mu\Delta t$.

Demonstration

- Lab class: run a simulation of the discrete queue.
- Answer questions on the analysis of that queue.
- Now we demonstrate a simulation in the lecture.
- We also give you histograms that are similar to those you should be producing in the lab class.

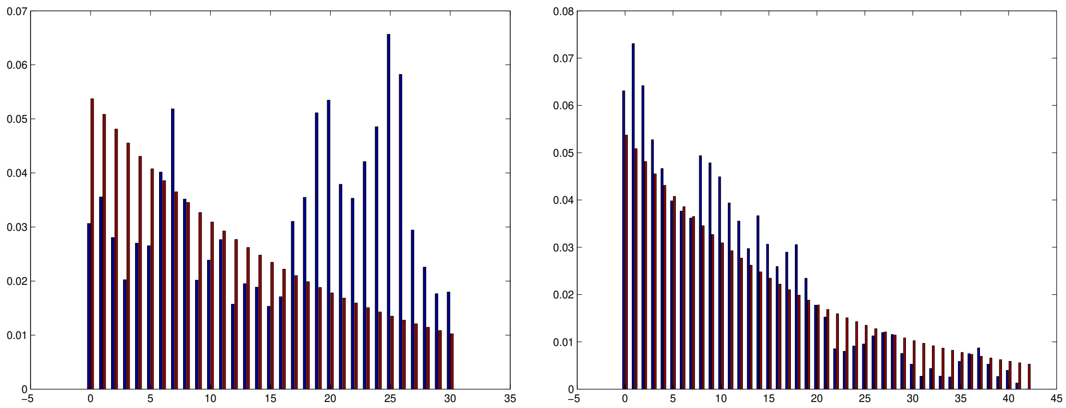
Histograms

[th]



Left figure: steady state probabilities after the first 100 seconds for $\rho = 0.5$. Right figure: steady state probabilities after the first 1000 seconds for $\rho = 0.5$.

[th]



Left figure: steady state probabilities after the first 100 seconds for $\rho = 0.95$. Right figure: steady state probabilities after the first 1000 seconds for $\rho = 0.95$.

Little's Formula

- Important quality of service issue is the average delay.
- We would also like to be able to compute
  - the average number in the queue and
  - the average time spent queueing.
- We introduce Little's formula, a very general formula.
- Can be applied to $G/G/1$ queues.
- Gives the relationship between
  - Steady-state average size of a queue, $\bar{N}$.
  - The steady-state arrival rate, λ.
  - The steady-state customer delay, $\bar{T}$.

Time and Steady State Averages

- We have already introduced the distribution means and sample based means.
- Time based averages are a further type of average.
- The time average is the average value of a particular function over time.

$$\bar{f}_t = \frac{1}{t}\int_0^t f(\tau)\, d\tau.$$

- The derivation of Little's formula relies on an assumption.
  - Assume that averages under the steady state distribution are equivalent to the time average in the limit as $t \longrightarrow \infty$,
  - Average number in the system as a function of time is given by

$$\bar{N}_t = \frac{1}{t}\int_0^t N(\tau)\, d\tau$$

    and the time average after in the limit as $t \longrightarrow \infty$ is given by

$$\bar{N} = \lim_{t\longrightarrow\infty} \bar{N}(t).$$

    which we assume is equivalent to $\langle k \rangle_{p(k)} \equiv \bar{N}$. See pg 156-157 in [1].

The Formula

- Applies under steady state conditions.
- Under the steady state assumptions it states,

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

Example: Nightclub queue.

You are driving past a nightclub, wondering how long you will have to wait to go in. You notice on Thursday nights that the queue typically contains 20 people and that people arrive at a rate of 40 per hour.

According to Little's formula, how long would you expect to have to wait to get in?

Answer

The waiting time is given by Little's formula:

$\bar{N} = 20$, $\lambda = 40$ and $\bar{T} = \frac{\bar{N}}{\lambda} = \frac{20}{40} = \frac{1}{2}\text{hour}.$

Intuition

If we see there are twenty people in a queue and they enter a building at a rate of 40 per hour, we might expect to have to wait half-an-hour in the queue. Formally though, we will build upon the intuition with a simple graphical 'proof' of the formula based upon that given in [1].

Graphical Proof I

Example 1, Chapter 7. Graphical Proff of Little's Formula.

Little's Formula

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

- Applies regardless of the arrival and service time distributions.
- Relies on the system being in *steady state*.
- Can be used at the level of the system.
- Or the level of the queue.
- Or the level of service.

Average Delay in $M/M/1$

- The mean of the steady state distribution for an $M/M/1$ queue is

$$\bar{N} = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}.$$

- Can use Little's formula to tell us the average delay.
- From $\bar{N} = \lambda \bar{T}$ we rearrange to recover

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

- This is not the time spent in the queue!
- This is the time spent in the system: it includes *service time.*
  - arises from *propagation delay.*
  - length of phone call *etc.*

Little's Formula for Queueing Time and Transmission Time

- The total delay can be divided into
  - queueing time, W,
  - and service time X.

$$T = X + W.$$

- Similarly the average delay may be decomposed as

$$\bar{T} = \bar{X} + \bar{W}.$$

- The queue and transmission may the be analysed independently using Little's formula.

The Transmission

- Assume that customers/packets do not leave the queue.
- Rate of arrivals at the queue must equal the rate of arrivals at transmission.
- We define the the average number under transmission as ρ
$$\rho = \lambda \bar{X},$$
where $\bar{X}$ is the average time spend under transmission for each packet.
- For the $M/M/1$ queue the average transmission time is given by $\bar{X} = \frac{1}{\mu}$,
- So for the $M/M/1$ queue we have
$$\rho = \frac{\lambda}{\mu},$$
a formula which we have already used in the last chapter.

Utilisation Factor

- ρ is the utilisation factor.
- Represents the fraction of time that the transmission line (or server) is in use.
- It is a measure of efficiency.
- It is also the steady state probability that the transmission line is in use.
- For a lossless system this probability is given by $1 - P(0)$.

Utilisation Example

Example 2, Chapter 7. Simple example of Little's formula. You are at a queue in a bar and notice that the bar tender spends 10% of her time cleaning between customers. If the typical number of people waiting to be served is 3 what is the ratio of time spent serving to time being served?

The Queue

- For arrival rate λ, and $\bar{N}_Q$, the average number in queue, and $\bar{W}$, the average queueing (waiting) time.

$$\bar{N}_Q = \lambda \bar{W}$$

- For the $M/M/1$ queue

$$\bar{T} = \bar{X} + \bar{W} = \frac{1}{\mu} + \bar{W}$$

so the average time spent waiting in the queue is given by

$$\bar{W} = \bar{T} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}.$$

- Average number in the queue is then

$$\bar{N}_Q = \frac{\rho^2}{1-\rho}.$$

Other Markov Queues

- The $M/M/1$ queue is perhaps the simplest Markov queue.
- We will introduce two related Markov queues and compute their stationary distributions.
- A simple modification to the $M/M/1$ queue: a *finite length* modification $M/M/1/m$
- The *infinite server* queue, $M/M/\infty$.

Finite Length $M/M/1/m$

- We have assumed that $M/M/1$ queue can have an infinite length.
- In practice a restricted buffer size — perhaps queue length is also restricted.
- We assume the system can contain at maximum m customers (*i.e.* the queue can contain $m-1$).

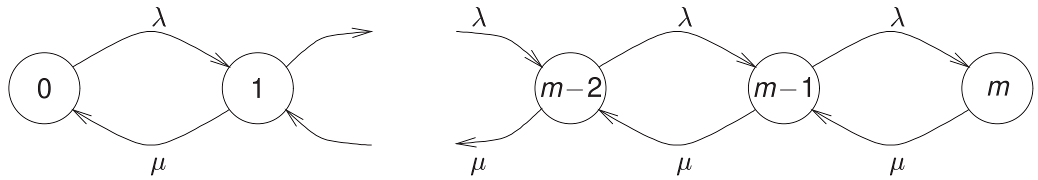


Figure 18: State diagram for the $M/M/1/m$ queue. Note that there are no transitions to states above the mth state.

Examples

- Example 1: Derive the stationary distribution for this queue.
- Example 2: Blocking probabilities

Infinite Serving Capacity — $M/M/\infty$

- The $M/M/1/m$ queue covers the case where the queue length is bounded.
- We can also derive when the number of servers is *unbounded*.
- In the $M/M/1$ queue we assume that there is only one server.
- It can be useful to consider an unlimited number of servers:
  - ➥ One server for every customer in the system.
- The state diagram for this system can be drawn.

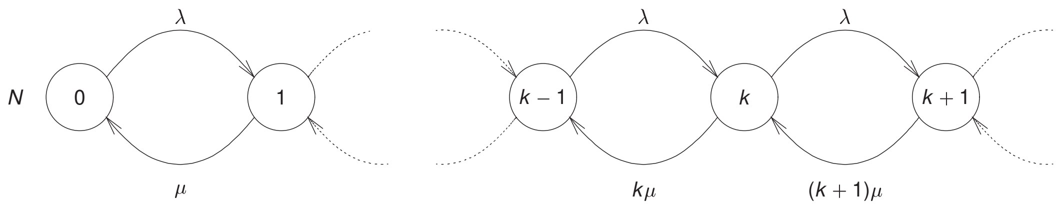


Figure 19: State diagram for the $M/M/\infty$ queue.

Detailed Balance

- Make use of the standard detailed balance equations for this queue.

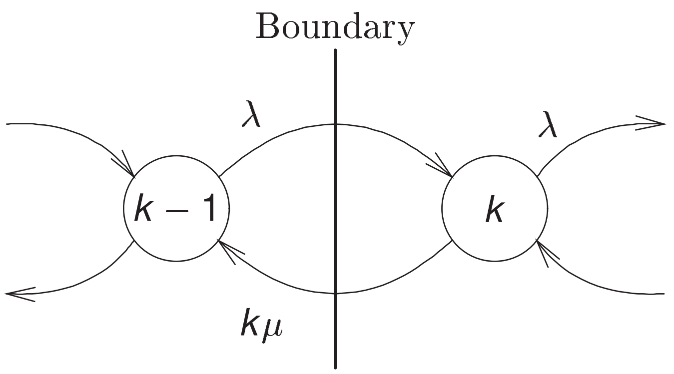


Figure 20: Flow rates at a boundary for $M/M/\infty$ queue.

- From the flow rates at the boundary we have in general
- $$\lambda_{k-1} P\left(k-1\right) = \mu_k P\left(k\right)$$
  - There is infinite service capacity in the queue the service rate at state k is simply $\mu_k = k\mu$.
  - The rate of arrivals stays constant, $\lambda_k = \lambda$, so for this queue detailed balance gives.

$$\lambda P\left(k-1\right) = k\mu P\left(k\right)$$

- Reorganising in terms of $P(k)$ then gives

$$P(k) = \frac{\lambda}{k\mu} P(k-1)$$

and therefore

$$P(k) = P(0) \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}.$$

Normalisation Condition

- Using the normalisation condition for this queue,

$$\sum_{k=0}^{\infty} P(k) = 1,$$

we obtain $P(0)$ as

$$P(0) = \left[\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^k \frac{1}{k!} \right]^{-1}. \qquad (5)$$

The Offered Load

- It is convenient to define the offered load,

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

  as arrival rate divided by service rate.
- This is a dimensionless constant, but has international standard units of Erlangs (after A. K. Erlang).
- Substituting this definition back into (5) we have,

$$P(0) = \left[\sum_{k=0}^{\infty} \frac{a^k}{k!} \right]^{-1} .$$

- Recalling that

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = \exp(x)$$

  we have

$$P(0) = \exp(-a)$$

  and the steady state distribution is then

$$P(k) = \frac{a^k}{k!} \exp(-a)$$

- Which is recognised as a Poisson distribution with parameter a.

Utilization Factor

- Notice that the utilization factor for this system, ρ, doesn't really exist.
- For the $M/M/\infty$ there is infinite capacity so we can't ask what percentage is in use!
- The proporation of time that some part of the system is in use is

$$1 - P(0) = 1 - \exp(-a).$$

Erlang Delay $M/M/m$

- We have considered the situation were there are as many servers as there are customers.
- More realistic real world situations involve limited numbers of servers.
- For example, call centres.
- The system is known as the *Erlang delay* system. In Kendall's notation it is written $M/M/m$.
- The 'average holding time', $\tau = \frac{1}{\mu} \equiv \bar{X}$, which, for a network involving telephones is the time a call is open.

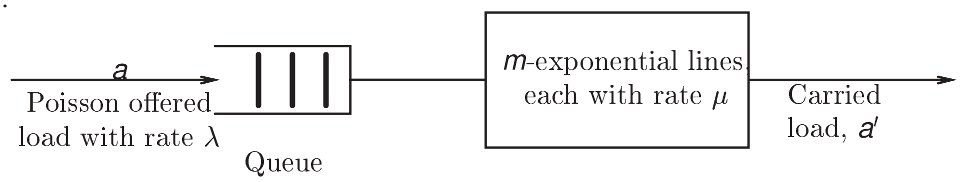


Figure 21: Queueing model of the Erlang delay system.

State Diagram

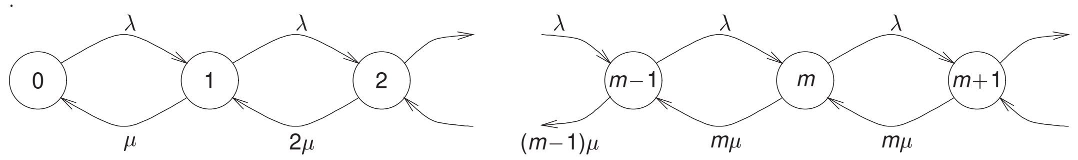


Figure 22: State diagram of an Erlang delay system.

- The ‘death rate’ μ_k has different definitions in different parts of the state space.
- We need to consider detailed balance for each part in turn.

Detailed Balance

- The Erlang delay system may be analysed as a birth-death process with birth rate

$$\lambda_k = \lambda$$

and death rate

$$\mu_k = \begin{cases} k\mu & k = 0, 1, \ldots, m-1 \\ m\mu & k > m \end{cases}$$

The Two Situations

- This means we have to consider two situations in detailed balance
  1. When the system's state, k, is less than or equal to the number of servers.
  2. When the system's state, k, is greater than the number of servers.

  We will cover this as an example.

Stationary Distribution

- The overall stationary distribution is then

$$P(k) = \begin{cases} P(0)\frac{(m\rho)^k}{k!}, & k \le m \\ P(0)\frac{m^m \rho^k}{m!} & k > m \end{cases}$$

$$P(0) = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!\,(1-\rho)}\right]^{-1}.$$

Note that this relies on $\rho < 1$, if $\lambda \ge m\mu$ the system has no stationary distribution.

Probability of Waiting

- For the $M/M/m$ queue a key quantity of interest is the probability of of a customer queueing, P_Q.
- What is the probability, when the system is at steady state, that a customer will arrive and find all the m servers busy?

$$P(k \geq m) = \sum_{k=m}^{\infty} P(k)$$

$$P(k \geq m) = P(0) \sum_{k=m}^{\infty} \frac{m^m \rho^k}{m!}$$

$$P(k \geq m) = P(0) \frac{(\rho m)^m}{m!} \sum_{k=m}^{\infty} \rho^{k-m},$$

Probability of Waiting contd

- Substituting $\dot{k} = k - m$ in the sum gives

$$P\left(k \geq m\right) = P\left(0\right) \frac{\left(\rho m\right)^m}{m!} \sum_{\dot{k}=1}^{\infty} \rho^{\dot{k}},$$

- Making the usual substitution this can be re-written

$$C\left(m, a\right) = P\left(k \geq m\right) = P\left(0\right) \frac{\left(\rho m\right)^m}{m!\left(1-\rho\right)}. \quad (6)$$

Erlang C Formula

- This is known as the *Erlang C formula*, or Erlang's second formula after A.K. Erlang. By making use of our definition of a we can rewrite the offered load as

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

and substituting into (6) we have

$$C(m,a) = \frac{P(0)\,a^m}{m!\left(1 - \frac{a}{m}\right)}.$$

Computing the Erlang C Formula

- Can be computed on a computer: but
  - some care must be taken when computing it.
  - when m is large overflow problems will occur for a^m and $m!$.
- Compute the formula in log space,

$$\ln C\left(m, a\right) = \ln P\left(0\right) + m \ln a - \ln m! - \ln\left(1 - \frac{a}{m}\right).$$

  Still problems with $\ln P\left(0\right)$ term, but each of the other terms can be computed with appropriate functions in MATLAB.
- When we need to compute the Erlang C curve, for the purposes of exams and exercises, we will use the old fashioned method of curves.

Erlang C Curves

- We have provided you with two sets of curves in your notes.
- In an exam, if they are need curves will be provided with the exam paper.
- Make sure you mark the curves and hand them in with your paper.

Examples

- Examples 2 and 3 at the end of the chapter cover the Erlang C curves.

Average number in System.

- We can also compute the average number of calls in the Erlang Delay system,

$$\bar{N} = \frac{\rho}{1-\rho} C(m,a) + m\rho$$

- These two terms represent the average queue length and the average number of busy lines.
- The average queue length is given by

$$\bar{N}_Q = \frac{\rho}{1-\rho} C(m,a)$$

Average Queueing Time

- Using Little's formula the average waiting time is simply

$$\bar{W} = \frac{\bar{N}_Q}{\lambda} = \frac{C(m,a)}{m\mu(1-\rho)}.$$

- Note that this is the average queueing time for all customers (whether they queued or not).

Queueing Time for those who Queue

- Use the fact that the average queueing time for those who don't queue is zero.
- The average queueing time can be decomposed as

$$\bar{W} = C(m,a)\bar{W}_Q + (1 - C(m,a))\,\bar{W}_{\sim Q}$$

  - $\bar{W}_Q$ is the average queueing time for those who do queue
  - $\bar{W}_{\sim Q}$ is the average queueing time for those who don't queue.
- Since $\bar{W}_{\sim Q} = 0$ the second term in this sum is zero we can rewrite in terms of $\bar{W}_Q$ as

$$\bar{W}_Q = \frac{\bar{W}}{C(m,a)} = \frac{1}{m\mu(1-\rho)} = \frac{1}{\mu(m-a)}$$

  so the average queueing time for those who do queue is independent of the probability of queueing.

Example

- Example 4 at the end of the chapter

So what is the average waiting time for the telephone service centre example? What is the queueing time for those who have to queue?

Erlang Loss $M/M/m/m$

- The Erlang delay system contained m servers.
- It was assumed that, if all servers were occupied, the customer would be placed on hold until a server became available.
- In the $M/M/m/m$ system there is a maximum number of customers.

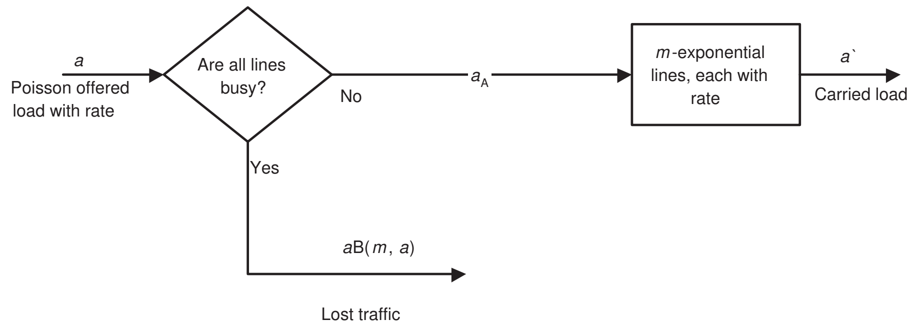


Figure 23: Flow chart of the Erlang Loss System.

State Diagram

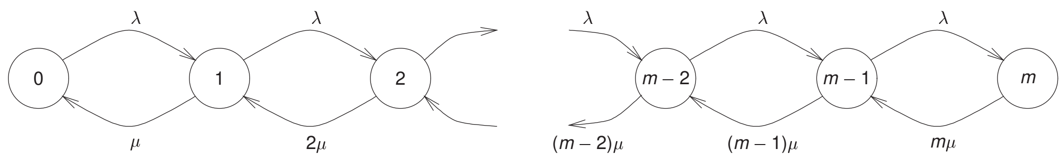


Figure 24: State Diagram for the Erlang loss system, $M/M/m/m$

- The offered load, $a = \frac{\lambda}{\mu}$, is again dimensionless with units of Erlangs.
- The Erlang loss system may be analysed as a birth-death process with birth rate

$$\lambda_k = \begin{cases} \lambda & k = 0, 1, \ldots, m-1 \\ 0 & k \geq m \end{cases}$$

- And death rate

$$\mu_k = \begin{cases} k\mu & k = 0, 1, \ldots, m \\ 0 & k > m \end{cases}$$

Detailed Balance Equations

- When the system's state, $k \leq m$, we have boundary flow as below

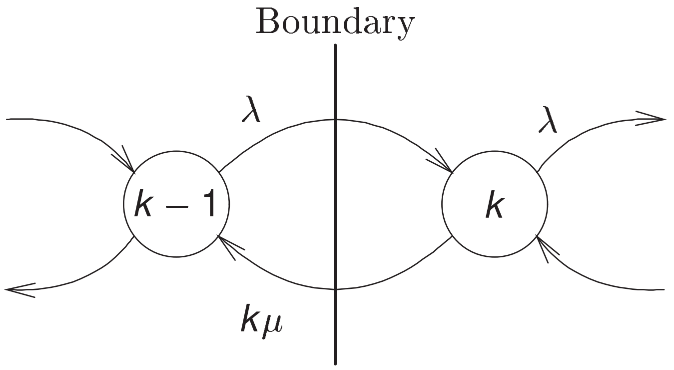


Figure 25: Flow rates at a boundary.

- This leads to the following balance equation

$$\lambda P(k-1) = k\mu P(k).$$

- So for $k = 0 \ldots m$ we may write

$$P(k) = \frac{\lambda}{k\mu} P(k-1). \tag{7}$$

- The system state cannot go above m so we have no need to consider the region where $k > m$. Substituting in the offered load we have,

$$P(k) = \frac{a}{k} P(k-1),$$

which means that

$$\begin{aligned} P(k) &= \prod_{n=1}^{k} \frac{a}{n} P(0) \\ &= \frac{a^k}{k!} P(0). \end{aligned}$$

Normalisation Condition

- The normalisation condition gives us

$$P(0)\sum_{k=0}^{m}\frac{a^k}{k!}=1$$

$$P(0)=\frac{1}{\sum_{k=0}^{m}\frac{a^k}{k!}},$$

- Which leads to the truncated Poisson distribution

$$P(k)=\frac{\frac{a^k}{k!}}{\sum_{n=0}^{m}\frac{a^n}{n!}}$$

- *Note if $m\rightarrow\infty$ we recover the $M/M/\infty$ queue as we would expect.*

Probability of Blocking

- For the $M/M/m$ queue we were interested in the probability that a customer would have to wait in the queue for service.
- In the $M/M/m/m$ system there is no queue because the maximum state of the system is the same as the number of servers.
- The probability of waiting is therefore zero.
- The probability of blocking is of interest: the probability that when we arrive the system is full and we are denied service.
- As we discussed for the $M/M/1/m$ queue the blocking probability is given by $P(m)$

$$B(m,a) = P(m) = \frac{\frac{a^m}{m!}}{\sum_{n=0}^{m} \frac{a^n}{n!}}.$$

- Use Erlang B curves to compute.

Carried Load and Utilization Factors

- Since traffic is lost in $M/M/m/m$ the carried load, a', is smaller than the offered load.
- The carried load is given by the offered load multiplied by the probability that the packet is not blocked.

$$a' = a(1 - P(m)) = a(1 - B(m, a))$$

- The carried load per line is the trunk occupancy (the utilization facctor),

$$\rho = \frac{a'}{m}.$$

- Note the contrast with the lossless systems where

$$\rho = \frac{a}{m}$$

  because the carried load is equal to the offered load $a' = a$.
- This also applies for the other system with loss we have seen, the $M/M/1/m$ queue.
- The rate of service in the system is no longer given by the arrival rate, λ, we need to consider those packets

that are lost.

$$\dot{\lambda} = \lambda(1 - P(m))$$

- This has effects when studying the queue using Little's formula.

Load Tripling

- **Question**

1. For a blocking probability of 0.01, what is the number of trunks needed for 6 Erlangs of offered load?
2. For a blocking probability of 0.01, what is the number of trunks needed for 18 Erlangs?
3. Comment on the increase in load compared to the increase in the number of trunks.

Answer

1. 13
2. 28
3. A tripling in load only leads to a (just over doubling) of the resources required. To understand this further we look at the utilization factors. For the first system the utilisation factor is

$$\rho_1 = \frac{6}{13}(1 - 0.005) = 0.459$$

whereas for the second system the utilisation factor is

$$\rho_2 = \frac{18}{28}(1 - 0.007) = 0.6384$$

The 30 trunk system is thus more efficient. This explains why it is cheaper to call the U.S.A. than Italy.

Example $M/M/m/n$

This example is from the Exam paper in 2002-3.

- Question

1. A call centre employs thirty operators selling tickets for a major sporting event. The call centre has sixty lines in total. When all operators are active new calls are placed in a queue for the first available operator. The call centre is to be modelled as an $M/M/m/n$ queue, *i.e.* a queue in which there are m servers and a maximum state of n.

   (a) Draw a state diagram representing the queueing system.

   (b) Write down the detailed balance equations for the two cases where the queue state $k < m$ and when $m \leq k < n$.

   (c) For each of the two cases outlined in the previous part write down the steady state distribution as a function of $P(0)$, the probability of the system being empty.

   (d) Use the fact that the steady state distribution must be normalised to show that

$$P(0) = \left[\sum_{k=0}^{m} \frac{a^k}{k!} + \sum_{k=m+1}^{n} \frac{a^k}{m! m^{k-m}} \right]^{-1}$$

where $a = \frac{\lambda}{\mu}$, λ is the arrival rate and μ is the rate of the service time distribution.

$M/G/1$ Queue

- Service times are *generally* distributed: cannot use the birth death process to analyse the system.
- We introduce a different type of analysis based on the mean residual service time.
- We will derive the Pollaczek-Khinchin formula.
- This gives the expected queueing time for any distribution of service times with *finite variance*.

Pollaczek-Khinchin formula

The Pollaczek-Khinchin formula is given as

$$\bar{W} = \frac{\lambda \left\langle X^2 \right\rangle}{2(1-\rho)}$$

where

- $\bar{W}$ is expected waiting time,
- $\left\langle X^2 \right\rangle = \left\langle X^2 \right\rangle_{p(X)}$ is the second moment under the service time distribution,
- utilisation factor is given as $\rho = \lambda \bar{X}$ and
- $\bar{X} = \langle X \rangle_{p(X)}$ is the mean under the service time distribution.

Proof of Pollaczek-Khinchin Formula

- Proof of the Pollaczek-Khinchin formula.
- In our proof we will make three assumptions
  1. The queue is FIFO queue, *i.e.* customers are served in the order they arrive. This assumption may be relaxed but it will simplify our proof.
  2. The arrival times are statistically independent of the service times.
  3. The service time distribution has a variance which is finite.
- To formulate the proof we introduce a concept known as the mean residual service time.

The Residual Service Time

- The residual service time is

  > ... the amount of service time that remains for the customer under service from the perspective of a customer who has just arrived in the queue.

- The first part of the proof occurs from the perspective of the arriving customer.
- The arriving customer is considered to be the ith customer.

Simple Definitions and Schematic

- Example 1 in the Notes
- We introduce some simple definitions of terms.

$$
\begin{aligned}
W_i &= \text{Waiting time in queue of } i\text{th customer.}\\
R_i &= \text{`Residual service time' seen by the } i\text{th customer.}\\
X_i &= \text{Service time of the } i\text{th customer.}\\
k_i &= \text{Number of customers in the system when the } i\text{th custome}
\end{aligned}
$$

This enables us to draw a simple schematic of the system as it is perceived by the ith arrival. This is shown in Figure 27.

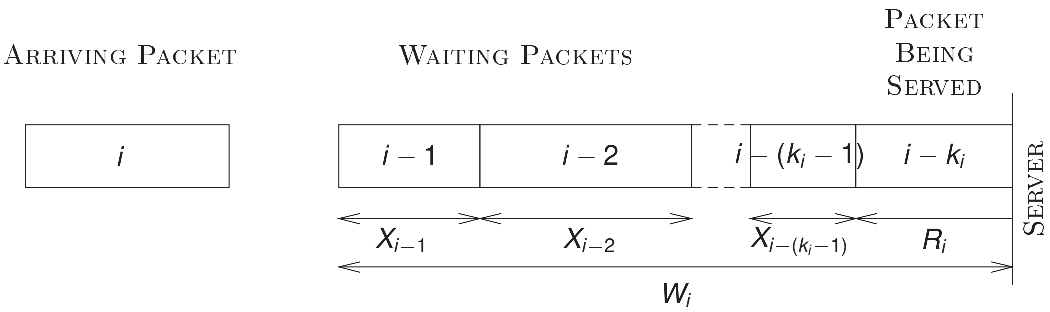


Figure 26: Schematic representation of residual service time.

Wait for ith Customer

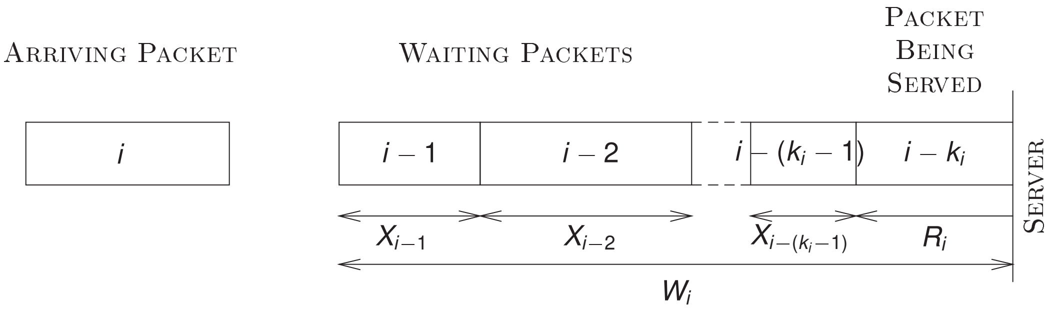


Figure 27: Schematic representation of residual service time.

- From the figure the wait for the ith customer will be given by

$$W_i = R_i + \sum_{j=i-(k_i-1)}^{i-1} X_j$$

- Summing over both sides (*i.e.* for all the arriving customers up to the Ith)

$$\sum_{i=1}^{I} W_i = \sum_{i=1}^{I} R_i + \sum_{i=1}^{I} \sum_{j=i-(k_i-1)}^{i-1} X_j$$

- Dividing through by I gives us

$$\sum_{i=1}^{I} \frac{W_i}{I} = \sum_{i=1}^{I} \frac{R_i}{I} + \sum_{i=1}^{I} \sum_{j=i-(k_i-1)}^{i-1} \frac{X_j}{I}$$

- In the limit as $I \longrightarrow \infty$ this becomes

$$\lim_{I\to\infty}\sum_{i=1}^{I}\frac{W_i}{I} = \lim_{I\to\infty}\sum_{i=1}^{I}\frac{R_i}{I} + \lim_{I\to\infty}\sum_{i=1}^{I}\frac{(k_i-1)X_i}{I}$$

- The terms of which are recognised as *time averages*.

Time Average Equivalence

- If we assume *time averages* are equivalent to *steady state averages* this gives

$$\langle W \rangle = \langle R \rangle + \langle (k-1)\, X \rangle \, .$$

- The length of the queue as seen by the ith arrival is independent of the ith arrival's length (our second assumption)
  - ➥ this means $\langle (k-1)X \rangle = \langle k-1 \rangle \langle X \rangle$ and since $(k-1)$ is the length of the queue we may write

$$\bar{W} = \bar{R} + \bar{N}_Q \bar{X}$$

- Using Little's theorem for the queue we have

$$\bar{N}_Q = \lambda \bar{W}$$

- Substituting back in we have

$$\bar{W} = \bar{R} + \lambda \bar{X} \bar{W} \, .$$

Wait in Terms of Residual Service Time

- From the previous slide:

$$\bar{W} = \bar{R} + \lambda \bar{X} \bar{W}.$$

- And the fact that $\rho = \lambda \bar{X}$ (which stems from Little's formula for the service) gives us

$$\bar{W} = \bar{R} + \rho \bar{W}$$

- Leading to

$$\bar{W} = \frac{\bar{R}}{1 - \rho} \tag{8}$$

- We now have the average queueing time in terms of the mean residual service time. To complete the Pollaczek-Khinchin formula we simply need to calculate $\bar{R}$.

Mean Residual Service Time from Departures Perspective

- We again turn to a graphical proof to obtain the value of the mean residual service time.

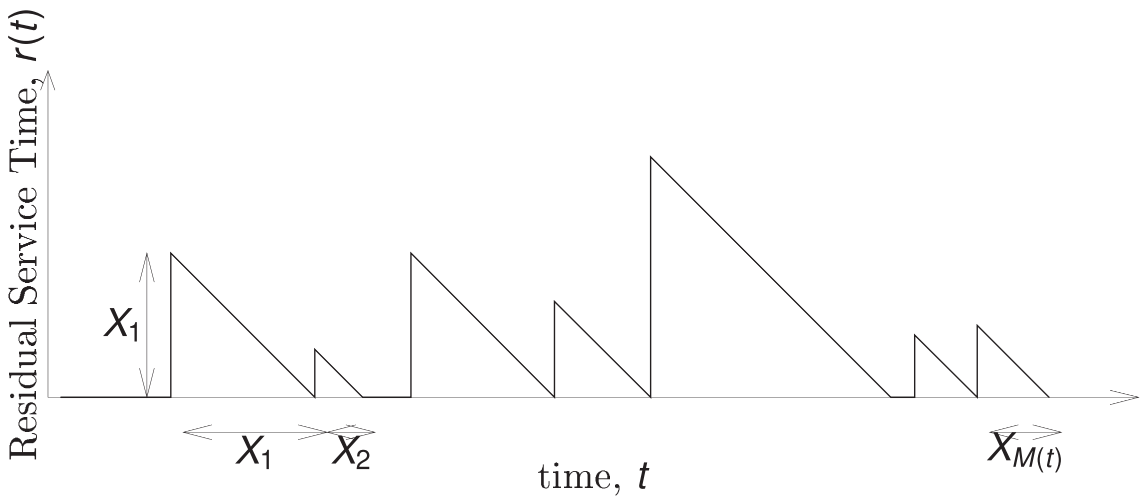
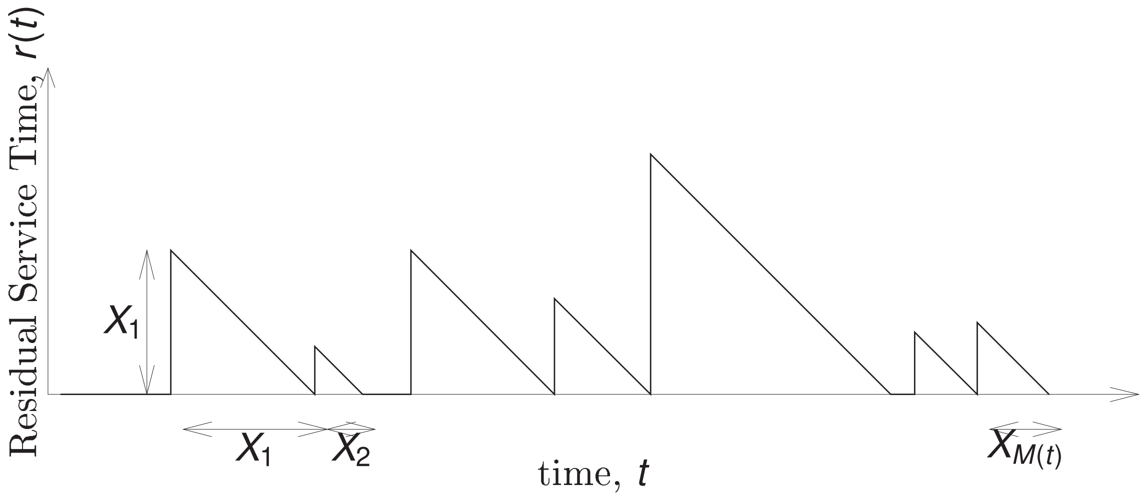


Figure 28: Graphical computation of mean residual time. $M\ (t)$ is the number of service completions in the interval $[0,\ t]$.

Residual Service Time



- The time average of the mean residual service time can then be calculated in terms of the area under the curve, $A(t)$.
- This is given by the area under each of the triangles,

$$A(t) = \frac{1}{2}\sum_{i=1}^{M(t)} X_i^2.$$

Residual Service Time

- The average service time is the area divided by time, t,

$$A(t) = \frac{1}{2}\sum_{i=1}^{M(t)} X_i^2.$$

$$\begin{aligned}\frac{A(t)}{t} &= \frac{M(t)}{t}\frac{A(t)}{M(t)} \\ &= \frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)}.\end{aligned}$$

↬ we have introduced the number of arrivals after time t, as $M\,(t)$.

↬ As $t \to \infty$, *if the variance is finite*, we find that $M(t) \to \infty$ and we obtain

$$\bar{R} = \frac{1}{2}\lambda \left\langle X^2 \right\rangle.$$

Pollaczek-Khinchin Formula

- Again assuming that time averages and ensemble averages are equal we substitute back into (8) to obtain the Pollaczek-Khinchin formula

$$\bar{W} = \frac{\lambda \left\langle X^2 \right\rangle}{2(1-\rho)}.$$

- This implies that the total waiting time in the system is

$$\bar{T} = \bar{X} + \frac{\lambda \left\langle X^2 \right\rangle}{2(1-\rho)}$$

and through Little's formula we can obtain the expected queue length, $\bar{N}_Q$, and the expected number in the system, $\bar{N}$.

Sanity Check — Markov Service times

- The exponentially distributed service times,

$$P(X) = \mu \exp(-\mu X),$$

have a mean of

$$\bar{X} = \frac{1}{\mu}$$

and a variance of $\frac{1}{\mu^2}$ leading to a second moment

$$\left\langle X^2 \right\rangle = \frac{2}{\mu^2}$$

- Substituting into the Pollaczek-Khinchin formula we obtain

$$\begin{aligned}\bar{W} &= \frac{\rho}{\mu(1-\rho)} \\ &= \frac{\rho}{\mu - \lambda}\end{aligned}$$

which is identical to those we derived for the $M/M/1$ queue.

Deterministic Service — $M/D/1$

- Some systems have deterministic service times. For example, transmission of fixed length packets over a channel with a fixed bandwidth.
- We represent a 'deterministic distribution' as

$$P(X) = \delta\left(X - \frac{1}{\mu}\right)$$

  where $\delta(\cdot)$is the Dirac delta function.
- For this system

$$\langle X \rangle = \frac{1}{\mu}$$

  and

$$\left\langle X^2 \right\rangle = \frac{1}{\mu^2}$$

  i.e. the variance is zero.

P-K Formula for Deterministic Systems

- For this system
$$\langle X \rangle = \frac{1}{\mu}$$
and
$$\left\langle X^2 \right\rangle = \frac{1}{\mu^2}$$
i.e. the variance is zero.
- Substituting into the Pollaczek-Khinchin formula we obtain
$$\bar{W} = \frac{\rho}{2\mu\left(1-\rho\right)}.$$

Table of Delays and Average Numbers in System

- In fact we may derive the full range of results for these distributions.

Expected value	General	Exponential	Deterministic
$\bar{W}$	$\frac{\lambda}{2(1-\rho)}\left\langle X^2\right\rangle$	$\frac{\rho}{\mu(1-\rho)}$	$\frac{\rho}{2\mu(1-\rho)}$
$\bar{T}=\bar{W}+\frac{1}{\mu}$	$\frac{\lambda}{2(1-\rho)}\left\langle X^2\right\rangle+\frac{1}{\mu}$	$\frac{1}{\mu(1-\rho)}$	$\frac{2-\rho}{2\mu(1-\rho)}$
$\bar{N_Q}=\lambda\bar{W}$	$\frac{\lambda^2}{2(1-\rho)}\left\langle X^2\right\rangle$	$\frac{\rho^2}{(1-\rho)}$	$\frac{\rho^2}{2(1-\rho)}$
$\bar{N}=\lambda\bar{W}+\rho=$	$\frac{\lambda^2}{2(1-\rho)}\left\langle X^2\right\rangle+\rho$	$\frac{\rho}{1-\rho}$	$\frac{\rho(2-\rho)}{2(1-\rho)}$

Example — Uniform Service Times

- Uniform Service Times — $M/U/1$.
- Consider a uniform probability distribution is defined between 0 and b for $X > 0$ as

$$U\left(X|b\right) = \begin{cases} \frac{1}{b} & 0 < X \leq b \\ 0 & X > b \end{cases}$$

show that the man of this uniform distribution is given by $\frac{b}{2}$ and its variance by $\frac{b^2}{12}$ and find the expected waiting time for an $M/U/1$ queue?

Answer

- For the mean we have

$$\langle X \rangle = \int_0^\infty X p(X)\, dX$$

Answer

- For the mean we have

$$\begin{aligned}\langle X \rangle &= \int_0^\infty X p(X)\, dX \\ &= \int_0^\infty X U(X|b)\, dX\end{aligned}$$

Answer

- For the mean we have

$$\begin{aligned}\langle X\rangle &= \int_0^\infty Xp\left(X\right)dX \\ &= \int_0^\infty XU\left(X|b\right)dX \\ &= \int_0^b \frac{X}{b}dX\end{aligned}$$

Answer

- For the mean we have

$$\begin{aligned}
\langle X \rangle &= \int_0^\infty X p(X)\, dX \\
&= \int_0^\infty X U(X|b)\, dX \\
&= \int_0^b \frac{X}{b} dX \\
&= \left[\frac{X^2}{2b}\right]_0^b
\end{aligned}$$

Answer

- For the mean we have

$$
\begin{aligned}
\langle X \rangle &= \int_0^\infty X p\left(X\right) dX \\
&= \int_0^\infty X U\left(X|b\right) dX \\
&= \int_0^b \frac{X}{b} dX \\
&= \left[\frac{X^2}{2b}\right]_0^b \\
&= \frac{b^2}{2b} - 0
\end{aligned}
$$

Answer

- For the mean we have

$$
\begin{aligned}
\langle X \rangle &= \int_0^{\infty} X p\left(X\right) dX \\
&= \int_0^{\infty} X U\left(X | b\right) dX \\
&= \int_0^{b} \frac{X}{b} dX \\
&= \left[\frac{X^2}{2b}\right]_0^b \\
&= \frac{b^2}{2b} - 0 \\
&= \frac{b}{2}
\end{aligned}
$$

Answer

- For the mean we have

$$
\begin{aligned}
\langle X \rangle &= \int_0^{\infty} X p\,(X)\, dX \\
&= \int_0^{\infty} X U\,(X|b)\, dX \\
&= \int_0^{b} \frac{X}{b} dX \\
&= \left[\frac{X^2}{2b}\right]_0^b \\
&= \frac{b^2}{2b} - 0 \\
&= \frac{b}{2}
\end{aligned}
$$

Variance

- For the variance we also require

$$\left\langle X^2 \right\rangle = \int_0^\infty X^2 p(X)\, dX$$

Variance

- For the variance we also require

$$\begin{aligned}\left\langle X^2 \right\rangle &= \int_0^\infty X^2 p\left(X\right) dX \\ &= \int_0^b \frac{X^2}{b} dX\end{aligned}$$

Variance

- For the variance we also require

$$
\begin{aligned}
\left\langle X^2 \right\rangle &= \int_0^{\infty} X^2 p\left(X\right) dX \\
&= \int_0^{b} \frac{X^2}{b} dX \\
&= \left[\frac{X^3}{3b}\right]_0^b
\end{aligned}
$$

Variance

- For the variance we also require

$$
\begin{aligned}
\left\langle X^2 \right\rangle &= \int_0^\infty X^2 p\left(X\right) dX \\
&= \int_0^b \frac{X^2}{b} dX \\
&= \left[\frac{X^3}{3b}\right]_0^b \\
&= \frac{b^3}{3b} - 0
\end{aligned}
$$

Variance

- For the variance we also require

$$\begin{aligned}\left\langle X^2\right\rangle &= \int_0^\infty X^2 p\left(X\right) dX \\ &= \int_0^b \frac{X^2}{b} dX \\ &= \left[\frac{X^3}{3b}\right]_0^b \\ &= \frac{b^3}{3b} - 0 \\ &= \frac{b^2}{3}\end{aligned}$$

Variance

For the variance we also require

$$
\begin{aligned}
\left\langle X^2 \right\rangle &= \int_0^\infty X^2 p\left(X\right) dX \\
&= \int_0^b \frac{X^2}{b} dX \\
&= \left[\frac{X^3}{3b}\right]_0^b \\
&= \frac{b^3}{3b} - 0 \\
&= \frac{b^2}{3}
\end{aligned}
$$

The variance is $\left\langle X^2 \right\rangle - \left\langle X \right\rangle^2 = \frac{b^2}{3} - \left(\frac{b}{2}\right) = \frac{b^2}{12}$.

Putting it Together

- For simplicity we consider $\mu = \frac{1}{\langle X \rangle}$. So $\mu = \frac{2}{b}$ and thus $b = \frac{2}{\mu}$.
- So we can write the second moment in terms of μ. $\langle X^2 \rangle = \frac{b^2}{3}$.
- Substituting in $b = \frac{2}{\mu}$ we have $\langle X^2 \rangle = \frac{4}{3\mu^2}$.
- Substituting in to the Pollaczek-Khinchin formula we recover

$$\bar{W} = \frac{2\rho}{3\mu\,(1-\rho)}.$$

Automatic Repeat Request (ARQ) System.

- The $M/G/1$ queue system can be used for modelling various systems of interest.
- In an ARQ system packets are transmitted and the system waits for an acknowledgement.
- If no acknowledgement is received the system retransmits.
- In this section we review Example 3.15 from Bertsekas and Gallager.
- A schematic of the system is given in Figure 29.

ARQ Schematic

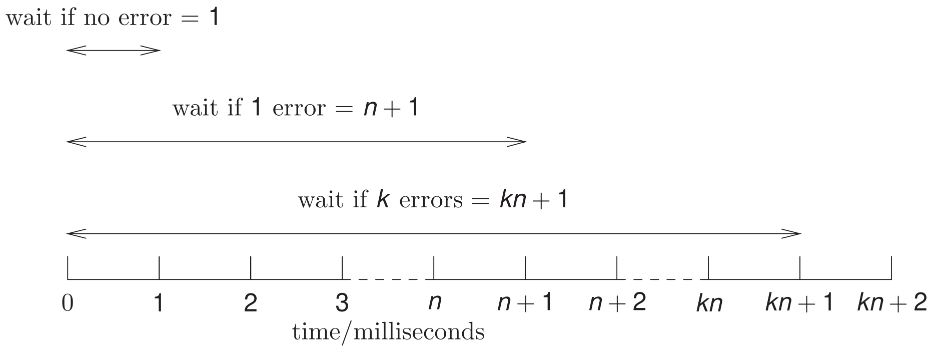


Figure 29: The ARQ System.

- For the purposes of this analysis, we will assume
  1. That each packet takes 1 millisecond to be transmitted.
  2. The system waits a maximum of $n-1$ milliseconds for an acknowledgement. If there is no acknowledgement the system retransmits.
  3. The acknowledgement occurs if the packet is received without errors. Otherwise acknowledgement doesn't occur.
  4. The probability of an error in a packet is taken to be p and we take errors to occur independently.
  5. Assume that packets arrive as a Poisson process with rate λ.

ARQ Contd

- In practice, the fourth assumption is probably unrealistic (independence of errors).
- If an error occurs due to problems in the transmission channel, it is likely that several packets in a row will not be received, so the errors do not occur independently.
- The time interval between transmission and re-transmission will be $1 + kn$ milliseconds with probability $(1 - p)\, p^k$.
- This time interval is associated with k retransmissions and one successful transmission.
- The probability of this being the service time is given by

$$P\,(X = 1 + kn) = (1 - p)\, p^k$$

which is recognised as a geometric distribution.

Average Service Time

- Using this distribution we can compute the expected service time which is given by

$$
\begin{aligned}
\langle X \rangle &= \sum_{k=0}^{\infty} (1+kn)(1-p)\,p^k \\
&= (1-p)\left(\sum_{k=0}^{\infty} p^k + n\sum_{k=0}^{\infty} kp^k\right) \\
&= 1 + n\frac{p}{1-p}
\end{aligned}
$$

since $\sum_{k=0}^{\infty} p^k = \frac{1}{1-p}$ and $\sum_{k=0}^{\infty} kp^k = \frac{p}{(1-p)^2}$.

ARQ Utilization Factor

- We can therefore write the utilization factor, $\rho = \lambda \langle X \rangle$, as
$$\rho = \lambda + \lambda n \frac{p}{1-p}.$$
- This has to be less than 1.
- Recall that the utilisation factor is
$$\rho = 1 - P(0).$$
- This implies that
$$P(0) = 1 - \lambda - \lambda n \frac{p}{1-p}$$

ARQ Second Moment

- The second moment of the service time is

$$\begin{aligned}
\left\langle X^2 \right\rangle &= \sum_{k=0}^{\infty} (1+kn)^2 (1-p)\, p^k \\
&= (1-p) \left(\sum_{k=0}^{\infty} p^k + 2n \sum_{k=0}^{\infty} kp^k + n^2 \sum_{k=0}^{\infty} k^2 p^k \right) \\
&= 1 + \frac{2np}{1-p} + \frac{n^2 \left(p+p^2\right)}{(1-p)^2}
\end{aligned}$$

since

$$\sum_{k=0}^{\infty} k^2 p^k = \frac{p+p^2}{(1-p)^3}$$

- These moments can be used with the Pollaczek-Khinchin formula to obtain statistics for the queue.

$M/G/1$ Queues with Vacations

- Consider a server which, when it is free, attends to some other business. For example it could transmit control and record keeping traffic. These periods are sometimes known as vacations (which is an American word that means holidays). The vacation periods are characterised by
  1. If a new arrival arrives during a vacation it must wait until the vacation is over before being dealt with.
  2. If a vacation ends and the server is idle it immediately goes on a new vacation.

$M/G/1$ with Vacations

- Example 5 In the Notes

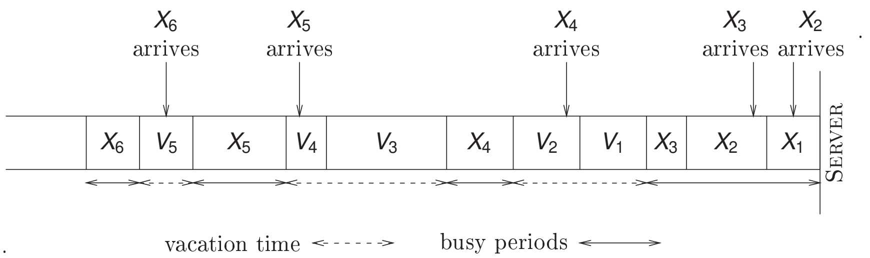


Figure 30: Schematic of $M/G/1$ queue

- We denote the length of these successive vacations to be V_1, V_2, ... We assume that V_1,V_2, . . . are i.i.d. random variables which are also independent of the customer times and the service times. We can then attempt to compute the mean residual service time for this queue for a *Poisson (Markov) arrival* system with *general service times* and *generally distributed vacation times*.

Residual Service Time

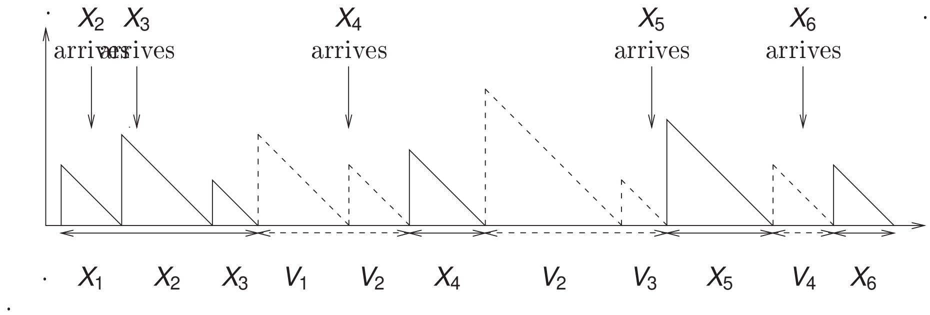


Figure 31: Computation of residual service time for system with vacations.

- For this system

$$\begin{aligned}\frac{1}{t}\int_0^t r(\tau)\,d\tau &= \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 + \frac{1}{t}\sum_{i=1}^{L(t)}\frac{1}{2}V_i^2 \\ &= \frac{M(t)}{t}\frac{1}{2}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} + \frac{L(t)}{t}\frac{1}{2}\frac{\sum_{i=1}^{L(t)} V_i^2}{L(t)}\end{aligned}$$

arrival process, $M(t)$, vacation process, $L(t)$.

Average of Vacation Distribution

- This means we have to deal with the average number of vacations. On average the proportion of time available for vacations is $(1-\rho)$, *i.e.* that time where the transmission line is not in use.
- If the average length of a vacation at time t is $\bar{V}_t$ then the number of vacations that will have arrived at that time is
$$L(t) = \frac{t(1-\rho)}{\bar{V}_t}.$$
Recalling
$$\frac{1}{t}\int_0^t r(\tau)\, d\tau = \frac{M(t)}{t}\frac{1}{2}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} + \frac{L(t)}{t}\frac{1}{2}\frac{\sum_{i=1}^{L(t)} V_i^2}{L(t)}$$
- Taking the limit of as $t \to \infty$ we can write
$$\bar{R} = \frac{1}{2}\lambda\left\langle X^2\right\rangle + \frac{(1-\rho)\left\langle V^2\right\rangle}{2\bar{V}}.$$

P-K Formula

- So given that the assumptions that we used to develop the formula
$$\bar{W} = \frac{\bar{R}}{1-\rho}$$
are still valid, we may write
$$\bar{W} = \frac{\lambda \left\langle X^2 \right\rangle}{2\left(1-\rho\right)} + \frac{\left\langle V^2 \right\rangle}{2\bar{V}}.$$

$M/G/1$ with Vacations — Simple example

A bar has Poisson arrivals and one member of staff serving. When there are no customers at the bar, the staff member leaves the bar to check on the kitchen. Assuming these checks are exponentially distributed in length, with mean $\bar{V}$ and that customer service times are exponentially distributed with mean $\bar{X}$ what is the average waiting time in the queue?

Exponential standard deviation $= \frac{1}{\mu}$, mean$= \frac{1}{\mu}$. So if mean is $\bar{V}$, standard deviation is $\bar{V}$ and variance is $\bar{V}^2$. Thus second moment is $\langle V^2 \rangle = 2\bar{V}^2$. Similarly for second moment of service time $\langle X^2 \rangle = 2\bar{X}^2$. Defining $\rho = \lambda \bar{X}$ we have

$$\bar{W} = \frac{\lambda \bar{X}}{(1-\rho)} + \bar{V} = \frac{\rho}{(1-\rho)} + \bar{V}.$$

Deterministic vacations would have lead to

$$\bar{W} = \frac{\lambda \bar{X}}{(1-\rho)} + \frac{\bar{V}}{2} = \frac{\rho}{(1-\rho)} + \frac{\bar{V}}{2}.$$

Networks of Queues and Simulation

- So far in the course we have looked at queueing models in isolation.
- We will now turn briefly to some of the issues that arise when they are combined.
- We will start by considering a simple combination of two $M/M/1$ queues.

'$M/M/1$' Queues in Tandem

- Consider a simple system involving two transmission lines in tandem.

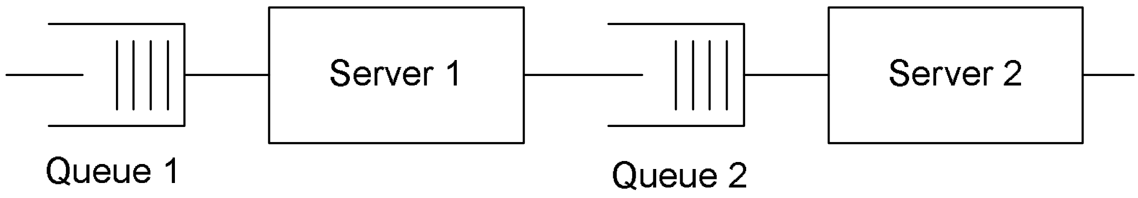


Figure 32: Two queues in tandem

- Packets have a fixed length and their transmission times are proportional to their length.
- Packet lengths are exponentially distributed and arrivals occur at the first queue as a Poisson process.

Implications

- The first queue may be modelled as $M/M/1$, but the second queue cannot be.
  - A packet which is large will take a long time in the first queue. During this time the second queue will have a chance to clear.
  - The result will be that for the second queue larger packets are likely to arrive when the queue is smaller.
  - This means that a large packet will typically experience less delay at the second queue than a small one.

Analogy

- One analogy for this situation is a slow moving vehicle on a road.
- From this vehicle's driver's perception there is very little traffic on the road because the slow moving vehicle never catches up with any.
- A similar thing occurs with $M/M/1$ queues in tandem.
- A large packet arrives at the second queue to find it has less packets because the smaller packets have had time to be processed.

Lack of Independence

- The first queue destroys the independence of the packet length and the state of the second queue.
- This problem is difficult to deal with analytically.
  - ➥ To proceed with our analysis we will need to break this dependency.
  - ➥ This can be achieved by considering service times for each packet to be statistically independent at each queue.

Time Reversibility — Burke's Theorem

- In the analysis of the Markov chain systems we have seen so far we relied on detailed balance.
- Detailed balance equations are a special case of global balance equations which arise when transitions only occur between neighbouring states in the Markov chain.
- Detailed balance can be applied to find the stationary distribution in any birth-death processes.
- Burke's theorem leads to an additional special property of birth-death processes known as *time reversibility*.

Time Reversibility

- The principle with time reversibility is as follows:
  - for a birth-death process at steady state the arrival process is Poisson with rate λ. Time reversibility states that the *departure process* will also be Poisson with rate λ.
- A proof is given in Section 3.7 of Bertsekas and Gallager.

Time Revesibility Intuition

- Consider the following 'intuition' for the $M/M/1$ queue.
- The service rate for the system is given by μ.
- The service rate is *not* the departure rate because a departure can only occur when the system has a state $k > 0$.
- This condition is satisfied for a fraction of the time which is given by the utilization factor ρ.
- The departure rate is therefore given by

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

.

Burke's Theorem

- Burke's theorem follows on from time reversibility. It states that
  - if an $M/M/1$, an $M/M/m$ or an $M/M/\infty$ queue at its stationary distribution has a Poisson arrival process with a rate λ
  - then the departure process is also Poisson with rate λ.
  - Furthermore the number of the customers in the system at time t is *independent* of the sequence of departure times prior to t.
- Queues with loss are slightly more complex because the arrival process does not have rate λ due to blocked packets.

Burke's Theorem

- The second property is worthy of comment as it seems counterintuitive.
- You might expect that if there were a rapid series of departures from the queue then the queue would be shorter than normal,
- Burke's theorem tells us that this is not the case.
- From this we can infer that the state of a queue *before* a rapid series of departures is about to occur must be abnormally high.
- Otherwise it *would* be abnormally low after the departures which Burke's theorem tells us is not the case.

$M/M/1$ Queues in Tandem Revisited

- Lets return to the $M/M/1$ queues in tandem. — Example 2 in notes.
- Burke's theorem tells us that the departure process from the first queue will be a Poisson with rate λ.
- We allow service times at the second queue to be independent of service times at the first queue.
  - This breaks the dependence of the queue length at the second queue on the service time at the second queue.
  - In this case the system can be modelled as two $M/M/1$ queues.

M/M/1 in Tandem

- If the service time distributions at the two queues are governed by exponential distributions with rate parameters μ_1 and μ_2.
  - ➥ This leads to utilization factors for each queue as $\rho_1 = \frac{\lambda}{\mu_1}$ and $\rho_2 = \frac{\lambda}{\mu_2}$.
  - ➥ The distributions for the states of the systems are given by
$$p(k_1) = \rho_1^{k_1}(1-\rho_1)$$
$$p(k_2) = \rho_2^{k_2}(1-\rho_2)\,.$$
  - ➥ The joint distribution for both systems is then
$$\begin{aligned} p(k_1, k_2) &= p(k_1)\,p(k_2) \\ &= \rho_1^{k_1}(1-\rho_1)\,\rho_2^{k_2}(1-\rho_2) \end{aligned}$$
    because the two queues can be viewed independently.
  - ➥ As a result the average delay in the system can be computed as the sum of the average delays in each queue
$$\bar{T} = \bar{T}_1 + \bar{T}_2 = \frac{1}{\mu_1 - \lambda} + \frac{1}{\mu_2 - \lambda}.$$

Relation to Reality

- This will typically be larger than the actual delay because
  - ↬ when the queue length is correlated with packet size
    - ∗ a large packet will find the second queue to be smaller than normal
    - ∗ and would therefore spend less time than might be expected in the system.

Kleinrock's Independence Approximation

- Kleinrock's independence approximation suggests assuming independence for every node in the network.
- The approximation is good if if large numbers of packets from outside the network are introduced at every queue.
- Consider a general network with labelled nodes and links between these nodes.
- There are several packet streams in the network, each stream has a rate of x_s.
- We define $f_{ij}(s)$ as the fraction of packets from stream s that cross the link between i and j.
- The rate associated with a link between node i and j is then

$$\lambda_{ij} = \sum_s f_{ij}(s)\, x_s.$$

Kleinrock's Independence Approximation

- Kleinrock suggested that merging traffic streams on a transmission line has the effect of restoring independence of arrival and service times.
- The approach is to assume that each link can be treated as an $M/M/1$ queue. This is a good approximation for
  1. arrivals which occur as a Poisson process
  2. exponentially distributed packets
  3. densely connected networks
  4. moderate-heavy traffic.
- Given this assumption, from the $M/M/1$ queue, we know that the average number at each link is

$$\bar{N}_{ij} = \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}}$$

The Average in the System

- Thus the average number in the entire system is

$$\begin{aligned}\bar{N} &= \sum_{ij} \bar{N}_{ij} \\ &= \sum_{ij} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}}\end{aligned}$$

- Average delay at each link is given by Little's theorem

$$\bar{T}_{ij} = \frac{1}{\mu_{ij} - \lambda_{ij}}.$$

Example — A Simple Network

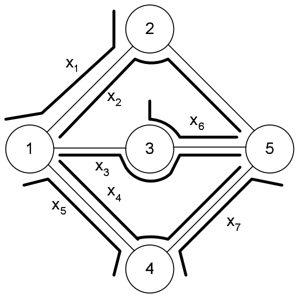


Figure 33: A small network with five nodes.

Question

Figure 33 shows a simple five node network. In the network flow only occurs from lower numbered nodes to higher numbered nodes. The connections in the network have average service times in milliseconds as follows:

$\bar{X}$	1	2	3	4	5
1	-	-	-	-	-
2	0.05	-	-	-	-
3	0.05	-	-	-	-
4	0.0333	-	-	-	-
5	-	0.05	0.05	0.033	-

Different routes through the network are indicated with thick lines. Each route has a rate associated with it as follows

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
λ	5	10	7	12	4	8	4

According to the Kleinrock independence assumption, what is the most congested link? What is the quickest route between node 1 & 5?

Answer

From the diagram we can see which links traffic flows down. Let us take each link in turn. Using the superposition of Poisson flows we have

λ	1	2	3	4	5
1	-	-	-	-	-
2	$x_1 + x_2 = 15$	-	-	-	-
3	$x_3 = 7$	-	-	-	-
4	$x_4 + x_5 = 16$	-	-	-	-
5	-	$x_2 = 10$	$x_3 + x_6 = 15$	$x_4 + x_7 = 16$	-

$\rho = \lambda \bar{X}$	1	2	3	4
1	-	-	-	-
2	$15 \times 0.05 = \frac{3}{4}$	-	-	-
3	$7 \times 0.05 = \frac{7}{20}$	-	-	-
4	$16 \times 0.0333 = \frac{8}{15}$	-	-	-
5	-	$10 \times 0.05 = \frac{1}{2}$	$15 \times 0.05 = \frac{3}{4}$	$16 \times 0.033 = \frac{8}{15}$

Answer Contd

$1-\rho$	1	2	3	4	5
1	-	-	-	-	-
2	$\frac{1}{4}$	-	-	-	-
3	$\frac{13}{20}$	-	-	-	-
4	$\frac{7}{15}$	-	-	-	-
5	-	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{7}{15}$	-

$N=\frac{\rho}{1-\rho}$	1	2	3	4	5
1	-	-	-	-	-
2	3	-	-	-	-
3	$\frac{7}{13}$	-	-	-	-
4	$\frac{8}{7}$	-	-	-	-
5	-	1	3	$\frac{8}{7}$	-

The most congested links are from 1 to 2 and 3 to 5.

Answer Contd

To compute the answer to the second part we need the average delay at each link

$\bar{T} = \frac{\bar{X}}{1-\rho}$	1	2	3	4
1	-	-	-	-
2	$4 \times 0.05 = \frac{1}{5}$	-	-	-
3	$\frac{20}{13} \times 0.05 = \frac{1}{13}$	-	-	-
4	$\frac{15}{7} \times 0.033 = \frac{1}{14}$	-	-	-
5	-	$2 \times 0.05 = \frac{1}{10}$	$4 \times 0.05 = \frac{1}{5}$	$\frac{15}{7} \times 0.033 = \frac{1}{14}$

This means that to go via node 2 takes $\frac{1}{5} + \frac{1}{10} = 0.3$ via node 3 takes $\frac{1}{13} + \frac{1}{5} = 0.277$ via node 4 it takes $\frac{1}{14} + \frac{1}{14} = \frac{1}{7} = 0.1429$ according to Kleinrock's independence assumption.

Here we have assumed that there is no further delay associated with each link (propagation, processing). In practice such delays may arise however we shall not consider them. See Bertsekas & Gallager Section 3.6.1 for more details.

Jackson's Theorem

- While Kleinrock's independence assumption approximates the activities at each queue as $M/M/1$ queues, Jackson's theorem provides a result for the number in the queueing system as a whole.
- Some analysis of the system is given in Section 3.8 of Bertsekas and Gallager.
- For our purposes it is sufficient to give the assumptions and the result.
  - there is no correlation between service times and queue lengths.
  - randomisation is used to divide the traffic along different routes through the network.
- Given these assumptions the average number of packets in the network is given by

$$N = \sum_{i,j} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}}.$$

Network Simulation

- The approximations we have outlined so far may be useful for 'back of the envelope' style calculations.
- In practice though it is highly likely that you will wish to simulate the system.
- There are a range of simulation packages available.
- We will introduce a simple one based on the Tcl scripting language called 'The Network Simulator' or ns-2 (`http://www.isi.edu/nsnam/ns/`.
- It is a discrete event simulator aimed at networking research.
- It supports a range of protocols including TCP and multicast protocols and it is straightforward to extend using C++.
- It is a powerful research tool which has the added advantage of being freely available.
- It is installed in the Lewin Laboratory on the Linux systems.
- If you require a personal installation, it can be installed through the Cygwin (`http://www.cygwin.com`) API.

Network Simulation with NS2

- As a lab class in week 9 you will simulate some simple queues in NS2 and compare the simulations with some results obtained using Kleinrock's independence assumption.
- The lab class will be completed in Linux so please make sure you know your Unix password for the DCS system before the lab start.
- The lab sheet has some notes on simple commands you will need.
- If you are unfamiliar with Linux it is suggested that you try them out before starting the lab.

Multiaccess Systems

- Finally we briefly review multiaccess systems.
- These are systems where there is one communication channel and multiple stations (or nodes) making use of it.
- Examples of multiaccess systems include: Satellite systems; packet radio networks; a multitapped bus.

LANs

- A very common example of multi-access channels is the local area network.
  - In a local area network transmission rates are around 10Mbps to 1Gbps.
  - Transmission distances range from a few meters to several kilometres.
  - The required error rates are low: typically from 10^{-8} to 10^{-11}, that's about 1 error every Gigabyte.
- Typically we are interested in the throughput of the network — the amount of data or messages passing through the network each second.
- If the number of overhead bits (associated with network administration *etc.*) is low then this is approximately equal to the utilization factor ρ.

Channel Multiplexing

- Until the introduction of the ALOHA network the standard approach to splitting a channel was through multiplexing.
- We will now briefly review this approach to multiplexing using Example 3.9 from Bertsekas and Gallagher.
- Assume we have m statistically independent Poisson streams.
- Each has arrival rate $\frac{\lambda}{m}$.
- The packets to be transmitted are assumed to have an exponentially distributed length.
- For a given bandwidth this leads to a service time distribution for the whole transmission line which is an exponential with a rate parameter μ.

Statistical Multiplexing

- System 1: merge the Poisson streams together, through statistical multiplexing.
- The superposition property of the Poisson distribution tells us that they become a Poisson stream with rate λ.
- If the transmission times are exponentially distributed the average delay, from the $M/M/1$ queue, will be

$$\bar{T} = \frac{1}{\mu - \lambda}.$$

Time or Frequency Division Multiplexing

- System 2: provide each communication channel with a separate 'circuit'.
- The transmission line needs to be split into m separate channels.
- Each of these separate channels will now have service time distributions with a reduced rated $\frac{\mu}{m}$.
- Delay in this system will now be

$$\bar{T} = \frac{1}{\frac{\mu}{m} - \frac{\lambda}{m}} = \frac{m}{\mu - \lambda}$$

which is m times larger than that for statistical multiplexing.

Why does this happen?

The reduction in delay arises from the fact that when only one of the m channels is transmitting with statistical multiplexing it can use the entire bandwidth of the channel but with time or frequency division multiplexing $\frac{m-1}{m}$ of the bandwidth is wasted.

Multiaccess Systems

- In multiaccess systems splitting the channel using time division multiplexing is known as time division multiple access (TDMA).
- Similarly splitting the channel using frequency division multiplexing is known as frequency division multiple access (FDMA).
- We have already shown that statistical multiplexing can make more efficient use of bandwidth than time or frequency based multiplexing.
- However in a multiaccess system it can be difficult to implement in the manner discussed above.

Multiplexing for Multiaccess

- We assumed that the all the transmissions could be queued in a single $M/M/1$ queue.
- In a local area network the transmitting stations (or nodes) are distributed across different portions of the network as are the receiving nodes.
- Forming a single queue is not possible without some form of centralised control.
- TDMA and FDMA can be implemented by by dedicating one channel for each station to transmit on, but allowing each station to receive on any channel.
- This allows for an analysis of the type we described above. However, it is not clear how to perform statistical multiplexing in these circumstances.

Random Access

- When we have stations distributed across a single communication channel would be to have a centralised computer attempting to control which computer is able to access our transmission lines at what times.
- However, we may wish communication to occur without centralised control. In this case we may choose a *random access* solution.
- We will consider random access networks where there is one communication channel on which all the nodes communicate[5].
- Each station sends a packet immediately whenever it has a packet to send. This means that packets can collide and may have to be resent.

[5] It could be the case that only a subset of the nodes have access to the channel.

Pure ALOHA Networks

- ALOHA network was developed at the University of Hawaii.
- Packet radio network which communicated between different stations on the Hawaiian islands.
- For the ALOHA network
  - all stations attempt to communicate on the same channel.
  - When a station attempts to transmit, it detects if its transmission has interfered with an existing transmission: this is known as a *collision*.
  - If a collision occurs the station becomes *backlogged* and attempts to transmit at a later time.
  - If both stations attempted to retransmit after the same period of time they would simply collide again.
  - To avoid this problem the stations normally wait for a random period of time before retransmitting.

ALOHA

- For this network we assume that each station can tell whether the transmitted packets were correctly received.
- Furthermore, we assume that each packet requires 1 'time unit' for transmission. This means that if a packet of interest is sent at time t, then any other packet sent between time $t-1$ and time $t+1$ will collide with our packet.

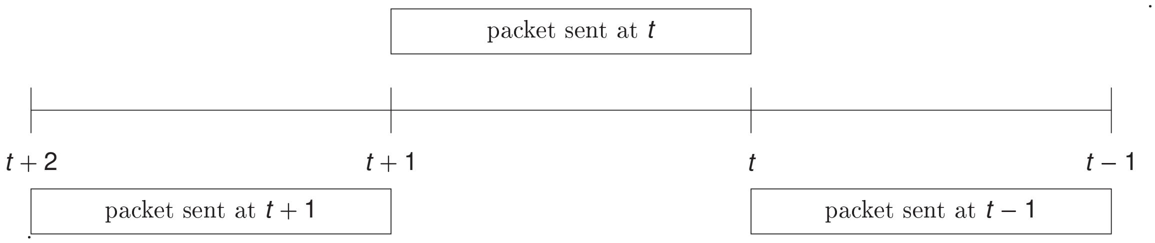


Figure 34: Pure ALOHA, packets collide with any other packet transmitted within one time unit of their transmission. The diagram shows a packet sent at t and packets sent at $t+1$ and $t-1$. Any packet sent after $t-1$ and before $t+1$ will collide with the packet sent at t.

Simple Analysis

- The arrival rate for packets at all stations is a Poisson process with rate λ.
- We assume that the number of stations is very large (infinite) so that an individual station's packet rate is an insignificant proportion of λ, *i.e.*

$$\frac{\lambda_n}{\lambda} = \lim_{\delta \to 0} \delta = 0.$$

- We denote the number of backlogged stations at any particular time as n.

Backlogged Stations

- A station is considered backlogged between the time that it became aware that its previous transmission failed and the time of a successful transmission.
- Once a station is aware that a transmission was unsuccessful it waits for T_i time units to retransmit where the distribution over these waits is given by

$$p\left(T_i\right) = \mu \exp\left(-\mu T_i\right).$$

- If n nodes are backlogged then the arrivals from backlogged nodes is a Poisson process with rate $n\mu$.
- The overall 'arrival process' for the system is Poisson with a rate $G\left(n\right) = \lambda + n\mu$.

Pure ALOHA

- The probability that there are no arrivals in an interval of 2 time units for a Poisson process with rate $G(n)$ is given by

$$p(k = 0|G(n), 2) = \exp(-2G(n)).$$

- Therefore this is the probability that there is no collision between the packet of interest and any other packets in the network.
- It is therefore the probability of a successful transmission.

Throughput

- The rate of attempted transmissions is $G(n)$ so the rate of successful transmissions is given by

$$G'(n) = G(n)\exp(-2G(n)),$$

- This is known as the throughput.
- It is the number of packets that successfully reach their destination. The throughput is plotted as a function of G .

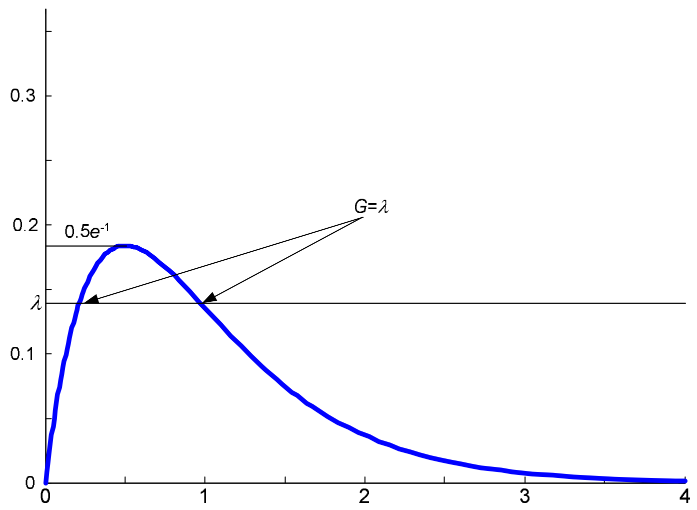


Figure 35: Throughput for pure ALOHA as a function of G.

Maximum Throughput

- The curve goes up to a maximum point and drops down thereafter.
- We can find the maximum of the throughput by looking for points where the gradient of the throughput is zero.
- Taking the gradient of G' with respect to G gives

$$\frac{dG'}{dG} = \exp(-2G) - 2G\exp(-2G).$$

Stationary Points

- To find the stationary points we set this equation to zero and solve for G. The equation becomes

$$\exp(-2G) - 2G\exp(-2G) = 0$$

$$(1 - 2G)\exp(-2G) = 0$$

  - So either $\exp(-2G) = 0$, which in turn implies that $G \longrightarrow \infty$ or
  - $1 - 2G = 0$ which implies $G = \frac{1}{2}$.
  - The first solution is a minimum (and the throughput is zero) and the second solution is a maximum.

- The maximum is the maximum possible throughput for the system and is given when $G = \frac{1}{2}$.

Maximum Throughput

- Substituting into the equation for the throughput we find that the maximum is $\frac{1}{2}\exp(-1) = \frac{1}{2e} \approx 0.184$. This maximum throughput shows that at best the channel is in use only about 18% of the time[6].
- The system will be in equilibrium if the throughput equals λ.

[6] Note that since the service time is fixed at $X_n = 1$ for all n then the utilisation for this system is given by $\rho = \lambda \bar{X} = \lambda$.

Analysis of Backlog Delay

- Recall our assumption that a backlogged node waits for a time, T_i, before attempting to retransmit,

$$p\left(T_i\right) = \mu \exp\left(-\mu T_i\right).$$

- Given this assumption, can we compute the delay associated with this backlogged node?
- Since the node may experience several collisions, the wait associated with the backlog will involve a sum over the random variable T_i. It turns out to be an *Erlang distribution*.

Sums of Exponential Variables and the Erlang Distribution

- If a random variable, x, is exponentially distributed with rate parameter b. The sum of a random variables, each of which is exponentially distributed with rate parameter b is distributed as an *Erlang distribution.*
- Consider variables a set of a variables$(x_1 \ldots x_a)$,

$$p\left(x_i|b\right) = b \exp\left(-bx_i\right).$$

- We define the sum of the values to be, $z = \sum_{i=1}^{a} x_i$.
- The distribution governing z is then an *Erlang distribution*, given by

$$p\left(z|a,b\right) = \frac{b^a}{(a-1)!} z^{a-1} \exp\left(-bz\right).$$

Notes on Erlang

- Note that this distribution is a generalisation of the exponential distribution because the exponential is the special case where $a = 1$.
- This clearly should be the case because if $a = 1$ then the 'sum' of that variable will simply be the variable itself.

Erlang Distribution Mean and Variance

The mean of the Erlang distribution is given by

$$\langle z \rangle_{p(z|a,b)} = \frac{a}{b}$$

and the variance is

$$\operatorname{var}(z) = \frac{a}{b^2}.$$

Delay Associated with Backlogs

- If the node retransmits k times then the distribution over the total time required, $W = \sum_{i=1}^{k} T_i$, comes from the sum of k independent exponentially distributed variables.
- We know that the resulting variable X is Erlang distributed with parameters μ and k,

$$p\left(W|k,\mu\right) = \frac{\mu^k}{(a-1)!} W^{k-1} \exp\left(-\mu W\right).$$

- The probability of transmitting after k attempts is given by

$$p\left(k\right) = q\left(1-q\right)^{k-1}.$$

  where q is the probability of a successful transmission[7]. For pure ALOHA $q = \exp\left(-2G\right)$.

[7] Compare this with the ARQ system we analysed in the chapter on $M/G/1$ queues. There we were considering errors (unsuccessful transmissions) so $p = (1-q)$.

Distribution for W

- From the product rule and the sum rule we know that the probability distribution for W is given by

$$p(W|\mu) = \sum_{k=1}^{\infty} p(W|k,\mu)\, p(k),$$

$$p(W|\mu) = \sum_{k=1}^{\infty} q(1-q)^{k-1} \frac{\mu^k}{(k-1)!} W^{k-1} \exp(-\mu W),$$

substituting $\dot{k} = k - 1$

$$p(W|\mu) = \sum_{\dot{k}=0}^{\infty} q(1-q)^{\dot{k}} \frac{\mu^{\dot{k}+1}}{\dot{k}!} X^{\dot{k}} \exp(-\mu W),$$

extracting the q, the exponential term and one of the μs from the product we rewrite as,

$$p(X|\mu) = q\mu \exp(-\mu W) \sum_{\dot{k}=0}^{\infty} \frac{(\mu(1-q)W)^{\dot{k}}}{\dot{k}!}, \tag{9}$$

The term inside the sum is then recognised as being of the form

$$\exp(x) = \sum_{\dot{k}=0}^{\infty} \frac{x^{\dot{k}}}{\dot{k}!}$$

$$\sum_{k=0}^{\infty} \frac{(\mu(1-q)W)^{k}}{k!} = \exp(\mu(1-q)W),$$

which can be substituted back into (9) to obtain

$$\begin{aligned} p(W|\mu) &= q\mu \exp(-\mu W)\exp(\mu(1-q)W), \\ &= q\mu \exp(-q\mu W) \end{aligned}$$

which is recognised as an exponential distribution with rate parameter $\mu' = q\mu = \mu \exp(-2G)$.

- Using the fact that G is a function of μ we can write

$$\mu' = \mu \exp(-2(\lambda + n\mu)). \tag{10}$$

- So the service rate is a function of λ, n and μ.

Optimum value of μ

- For a given value of λ and n there is a optimal value for μ.
- This value of μ can be found by seeking the maximum rate, this is done by differentiating with respect to μ

$$\frac{d\mu'}{d\mu} = \exp\left(-2\left(\lambda + n\mu\right)\right) - 2n\mu \exp\left(-2\left(\lambda + n\mu\right)\right)$$

and setting the result to zero to find the stationary points,

$$\exp\left(-2\left(\lambda + n\mu\right)\right) - 2n\mu \exp\left(-2\left(\lambda + n\mu\right)\right) = 0,$$

$$\left(1 - 2n\mu\right) \exp\left(-2\left(\lambda + n\mu\right)\right) = 0.$$

Solutions for μ

- There are two solutions to this equation. The first is $\exp\left(-2\left(\lambda+n\mu\right)\right)=0$, meaning that $\lambda+n\mu\rightarrow\infty$. This solution is a minimiser . The other solution is given when
$$0=1-2n\mu$$
or
$$\mu=\frac{1}{2n},$$
this turns out to be a maximum of (10).

Notes on Optimum

- By maximising the 'rate' parameter μ' we minimise the average delay associated with the backlog, which is given by $\bar{W} = \frac{1}{\mu'}$.
- By setting $\mu = \frac{1}{2n}$ we find a minimum for the average delay:

$$\bar{W}_{\min} = 2n \exp\left(2\lambda + 1\right)$$

.

- In practice it may be difficult to set the rate μ to something that is dependent on the number of backlogged nodes.
- This is because the number of backlogged nodes can only be known by nodes communicating their status over the network, we have not allowed for such a mechanism in our analysis.
- Note that the minimum average delay goes up linearly with the number of backlogged nodes and *exponentially* with the rate of arrivals for the system, λ.

Slotted ALOHA

- A simple improvement to the ALOHA system comes from constraining when packets can be sent.
- Instead of allowing stations to transmit immediately they receive a packet each station waits for the beginning of a ‘slot’ to start transmitting.
- These slots are synchronised across the entire network.
- In this system a collision occurs if another station attempts to transmit in the same slot.
- This happens if another station received a packet for transmission in the preceding slot.
- If each slot is one time unit in length the probability of this event is given by

$$p\left(k=0|G\left(n\right),2\right)=\exp\left(-G\left(n\right)\right).$$

Slotted ALOHA

- The probability is smaller than for pure ALOHA because the ‘vulnerable period’ for a collision is smaller (half the length of before).
- A collision now only occurs if another station attempts to transmit in the same slot. As a result the rate of successful transmissions is given by

$$G'(n) = G(n)\exp(-G(n)),$$

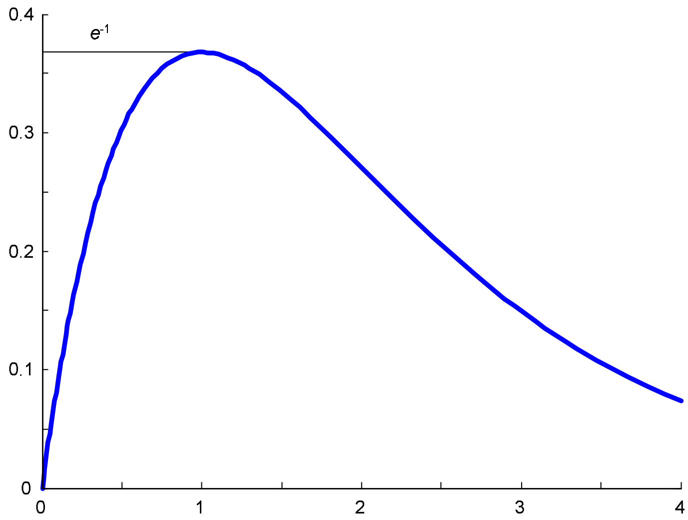


Figure 36: Throughput for slotted ALOHA as a function of G.

Smaller Vulnerable period

The smaller vulnerable period leads to a greater maximum throughput which is now found at $G = 1$. The maximum throughput is $\frac{1}{e}$. Slotted ALOHA therefore can be twice as efficient as pure ALOHA.

References

References

[1] D. P. Bertsekas and R. G. Gallager. *Data Networks*. Prentice-Hall, Upper Saddle River, NJ, second edition, 1992.

[2] W.-C. Chan. *Performance Analysis of Telecommunications and Local Area Networks*. Kluwer, Dordrecht, The Netherlands, 2000.

[3] A. K. Erlang. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, 20, 1909.

[4] A. S. Tanenbaum. *Computer Networks*. Prentice-Hall, Upper Saddle River, NJ, fourth edition, 2003.