Imperial College
London

Traffic Theory & Queueing Systems

Javier A. Barria
j.barria@imperial.ac.uk
1009a DEEE

Back to Curriculum

Course Module

PG Streams
MSc A&DIC
MSc Control
MSc C&SP

EE9SO7 Traffic Theory & Queueing Systems

Lecturer(s): Dr Javier Barria

Aims:
The aim of this course is to provide students with the opportunity to develop a conceptual framework for modelling and analysing different communication networks (e.g. circuit-switched and packet-switched networks). The course will show, firstly, how to set up such models and, secondly, how to use them in the performance (e.g. QoS) analysis of communication systems.

Learning Outcomes:
-Describe and discuss the validity of different performance modelling/assessment techniques.
-Identify different communication system QoS related problems, and the appropriate solution techniques for these problems.
-Describe and discuss the underlying assumptions of the studied modelling/assessment techniques.
-Determine the conditions in wich a communication system is operating.
-State and solve communication system performance related problems.
-Derivate and evaluate related performance analytical expressions.

Syllabus:
Introduction to teletraffic analysis. Mathematical basis of traffic theory: Markov processes. Loss-system analysis: route congestion in circuit-switched systems; models for overflow traffic; restricted availability; congestion in circuit switches. Delay-system analysis: introduction to queueing theory;
congestion in message-switched systems and packet-switched systems; queueing network models. Analysis of random-access protocols; Traffic characterisation of Broadband Services; Admission and Access control in Broadband networks; Routing in ATM networks. Performance/Reliability (Performability) models.

Assessment:

Coursework contribution: 0%

Term: Spring **Slot:** 4C

Closed or Open Book (end of year exam): Closed

Coursework Requirement
nil

Oral Exam Required (as final assessment): no

Prerequisite: None required

Course Homepage: unavailable

Lecturer Management

Traffic Theory & Queueing Systems

E4.05 - S07: Part 1

Javier A. Barria
j.barria@imperial.ac.uk
1009a DEEE

LECTURE ON
Distribution functions, stochastic processes and Markov chains.

Basic definitions

The theory of probability begins with the idea of a random experiment and its sample space, which is the set of possible outcomes.

A random variable is a function mapping each element of a sample space to a real number. The reason for having random variables is that numbers are easier to work with than elements like, for example, heads and tails.

Discrete random variables

Definition: Repeated independent trials are called Bernoulli trials, if there are only two possible outcomes: Success - p and Failure - q.

- **Geometric random variable**

  Is the number of Bernoulli trials up to and including the first success:

$$P(X = x) = q^{x-1}p \tag{1}$$

- **Binomial random variable**

  A binomial random variable, is the number of success in a series of n Bernoulli trials. For x success and $n - x$ failures:

  – The probability of any outcome is $p^x q^{n-x}$

  – The number of ways of distributing x success in n trials is: $\binom{n}{x}$

  Thus the probability that an event occurs with x success and $n - x$ failure is:

$$P(X = x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \tag{2}$$

Distribution functions

The cumulative distribution function (CDF), of the random variable X, sometimes called just the distribution function, is defined by:

$$F_X(x) = Pr[X \leq x] \tag{3}$$

where Pr is read the "probability that". A CDF is discrete, continuous or mixed according to the type of its random variable. When there is no ambiguity, the subscript X of the CDF is generally not used.

When working with random variables, we are often interested in the "average" outcome of the underlying random experiment. The formalism for the average is the expected value or mean of a random variable, defined by:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{4}$$

- **Poisson distribution**

This distribution can be seen as the limiting case of the binomial distribution.

$X =$ number of success in a large number of trials, where the probability of success is very small, that is, $n \to \infty$, $p \to 0$ and using the notation $\lambda = np$.

$$\begin{aligned} P(X=i) &= \frac{n(n-1)\dots(n-i+1)}{i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} \\ &= \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-i+1}{n}\right)\frac{\lambda^i}{i!}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^i} \\ &= e^{-\lambda}\frac{\lambda^i}{i!} \end{aligned} \quad (5)$$

since as $n \to \infty$ for finite i, $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$.

- **Poisson distribution (alternative)**

Alternative, lets assume λ to be the mean arrival rate of process, i.e., $\frac{numberofevents}{unitinterval}$.

Moreover, lets also consider that,

- Probability of one event in $(t, t+h)$ is λh.
- Probability of more that one event is very small.
- Probability of event occurring in two non-overlapping intervals are independent.

Then,

- Probability that i customers arrive in m subintervals is a binomial distribution.

$$p(X=i) = \binom{m}{i}[\lambda h + 0(h)]^i[1-\lambda h+0(h)]^{m-i} \quad (6)$$

- Taking the limit as $h \to 0$ and as $m \to \infty$ obtains the Poisson distribution.

$$p(X=i) = \frac{(\lambda h)^i}{i!}e^{-\lambda h} \quad (7)$$

- **The exponential distribution**

The most commonly used distribution function in reliability and performance modelling is the exponential distribution. Its CDF and Probability Density Function (pdf), respectively, are

$$F(t) = 1 - e^{-\lambda t} \quad (8)$$

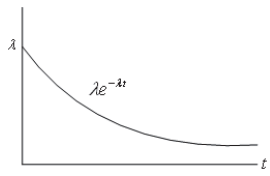
$$f(t) = \lambda e^{-\lambda t} \quad (9)$$

The exponential distribution has what is called the memoryless property. Informally, this means that if you are waiting for something to happen, the remaining time you have to wait does not depend on how long you have already been waiting. In a reliability setting, it means that the distribution of the remaining life of a component does not depend on how long it has been working.

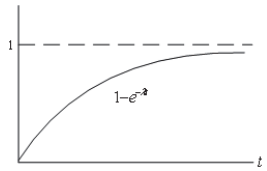
Consider an exponentially distributed random variable X and a time t. The random variable Y, defined by $Y = X - t$, is the remaining life of X at time t. Let $G_t(y)$ be the conditional probability that $Y \le y$ given $X > t$. We have:

$$\begin{aligned} G_t(y) &= P(Y \le y | X > t) \\ &= P(X - t \le y | X > t) \\ &= P(X \le y + t | X > t) \\ &= \frac{P(t < X \le y+t)}{P(X>t)} = \frac{\int_t^{y+t} f(t)dt}{\int_t^{\infty} f(t)dt} \\ &= \frac{\int_t^{y+t}\lambda e^{-\lambda t}dt}{\int_t^{\infty}\lambda e^{-\lambda t}dt} \\ &= \frac{e^{\lambda t}(1-e^{-\lambda y})}{e^{-\lambda t}} \\ &= 1 - e^{-\lambda y} \end{aligned} \quad (10)$$

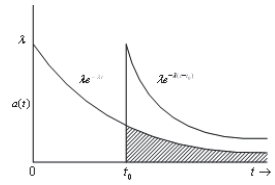
The distribution function for the remaining life of X is the same as the original distribution function for X. This proves that the remaining life of X does not depend on the time that has passed so far.



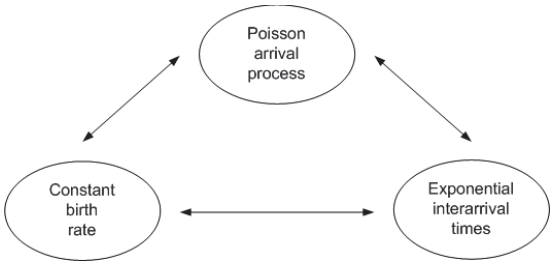
Probability Density Function (pdf) Exponential distribution



Probability Distribution Function (PDF) Exponential distribution



The memoryless property of the exponential distribution



The memoryless triangle

- **Relationship between Poisson process and Exponential distribution**

  Let τ to be the time between adjacent arrival $PDF = A(t)$ and $pdf = a(t)$. Then, $A(t)$ the probability that the time between arrivals is $\leq t$, will be given by $A(t) = 1 - P[\tau > t]$.
  But $P[\tau > t]$ is just the probability that no arrivals occurs in $(0, t)$, that is, $P(0)$. Therefore we have

  $$A(t) = 1 - P(0) = 1 - e^{-\lambda t} \qquad (11)$$

  and

  $$a(t) = \lambda e^{-\lambda t}, t \geq 0 \qquad (12)$$

  This means that

  - For a Poisson arrival process, the time between arrival is exponentially distributed.
  - Poisson arrival process has exponential inter arrival times.

Little's Theorem: $N = \lambda T$

$N(t)$ = number of customers in the system at time t.
$\alpha(t)$ = number of customer who arrived in $[0, t]$.
T_i = time spend in the system by the $i-$th arriving customer.

1. Time average of $N(\tau)$ up to time t:

   $$N_t = \frac{1}{t}\int_0^t N(\tau)d\tau, \qquad \lim_{t\to\infty} N_t = N$$

2. Time average arrival rate over $[0, t]$:

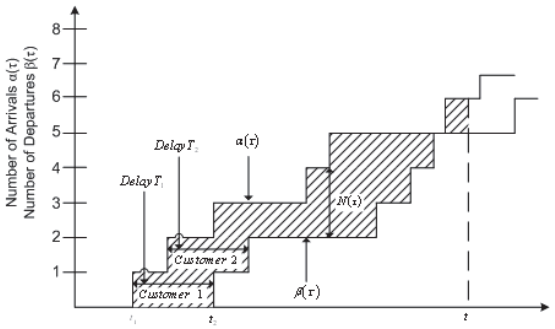
   $$\lambda_t = \frac{\alpha(t)}{t}, \qquad \lim_{t\to\infty} \lambda_t = \lambda$$

3. Time average of customer delay up to time t:

   $$T_t = \frac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)}, \qquad \lim_{t\to\infty} T_t = T$$

4. N, λ, T can be related by a simple formula:

   $$N = \lambda T$$



Littles Theorem

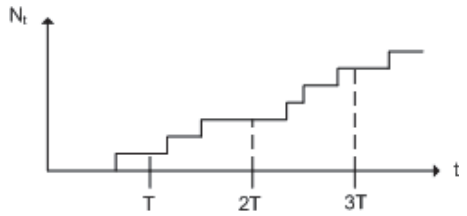
Proof of the Little's Theorem. If the system is empty at time t $[N(t) = 0]$, the shaded area can be expressed both as $\int_0^t N(\tau)d\tau$ and as $\sum_{i=1}^{\alpha(t)} T_i$. Dividing both expressions by t, equating them, and taking the limit as $t \to \infty$ gives Little's Theorem. If $N(t) > 0$, we have

$$\sum_{i=1}^{\beta(t)} T_i \leq \int_0^t N(\tau)d\tau \leq \sum_{i=1}^{\alpha(t)} T_i$$

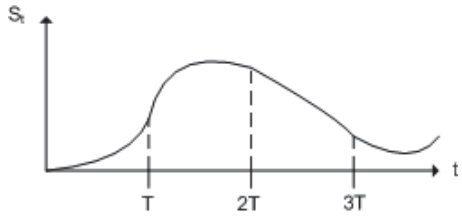
and assuming that the departure rate $\frac{\beta(t)}{t}$ up to time t tends to the steady-state arrival rate λ, the same argument applies.

LECTURE ON
Stochastic (Random) Process

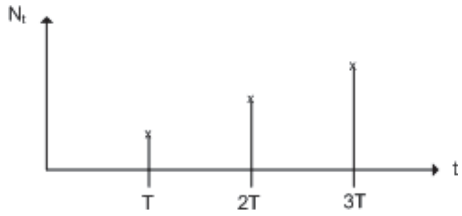
A stochastic process is a mathematical model of a randomly-varying function of time (or some other independent variable).
Examples:
Traffic census: Suppose we measure (over a specific period) at a given point:
N_t = the total vehicle count in the interval $[0, t]$
S_t = the sound intensity at time t



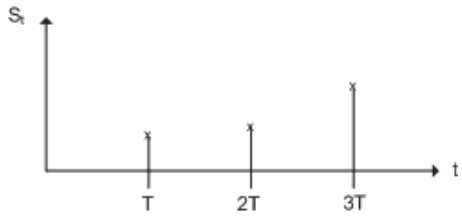
Continuous-time. Discrete-state.



Continuous-time. Continuous-state.



Discrete-time. Discrete-state.



Discrete-time. Continuous-state.

Basic Approach

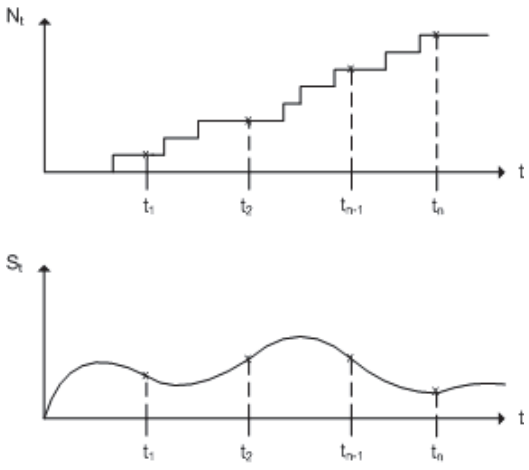
Think of the time histories of N_t and S_t as being generated by a random experiment, Φ. Then,

1. Each trial (i.e. repetition of the experiment) will yield a pair of records $(\{N_t\}, \{S_t\})$: the pair will be different for each trial.

2. Since Φ is a random experiment we expect *statistical regularity*: that is, at any chosen time t_1, the observed values of N_t and S_t should show statistical regularity.

3. Therefore we can view N_t and S_t as *random variables*.

4. More generally, for any set of times $(t_1, t_2, \ldots, t_n)$ we have a set of random variables $(N_{t_1}, \ldots, N_{t_n})$ jointly distributed on the probability space $\{\Omega, F, P\}$ representing Φ.



Definition
Suppose we are given:

1. A random experiment Φ represented by a *probability model* $\{\Omega, F, P\}$

2. A *set of times*, T

3. For each time $t \in T$ a RV on $\{\Omega, F, P\}$, $X_t : \Omega \to \mathrm{R}$

Then we call the indexed collection of RV $\{X_t : t \in T$ a *(scalar) stochastic process* on $\{\Omega, F, P\}$.

Comments:

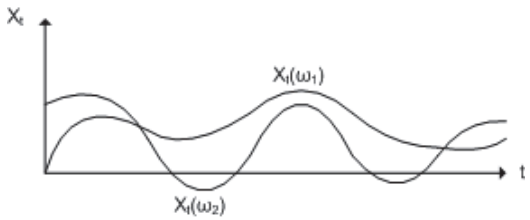
1. The set T is the *index set* of the process $\{X_t\}$
   Usually $T = [0, \infty)$ or $(-\infty, -\infty)$ (Continuous-index process)
   or $T = \{0, 1, 2, \dots\}$ or $\{\dots, -2, -1, 0, 1, 2, \dots\}$ (Discrete-index process)

2. We can think of the outcome, ω, as the record of the behaviour of all the variables of interest in Φ. Thus, for example, if the only variable of interest are X_t, Y_t, Z_t, we could take ω as the time history of the vector (X_t, Y_t, Z_t). Then, as required for any RV:
   Complete knowledge of ω in a trial implies knowledge of the values of $X_t(\omega), Y_t(\omega), Z_t(\omega)$, for each time instant t.

3. For a particular outcome, ω, the value of the process $\{X_t : t \in T\}$ at time t is a *real number*, $X_t(\omega)$, which is a function of *two* variables: ω, t.

Interpretation

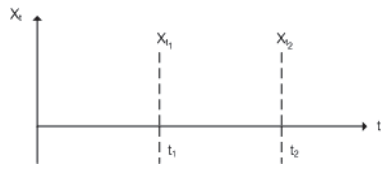
1. Suppose we fix ω (i.e. the outcome ω is known): then $X_t(\omega)$ is a *known function* of t.



One function for each ω

The functions $X_t(\omega_i)$ are called **Sample Paths** (or *records* or *realisations* or *time histories*) of the process $\{X_t\}$ and a collection of them is called an **ensemble**.

2. Suppose we fix t: then, by definition X_t is a RV on $\{\Omega, F, P\}$

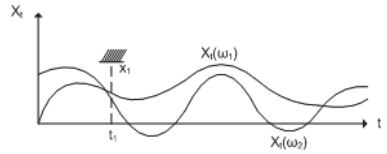


One RV for each t

Probability Specification of $\{X_t\}$

In order to derive a probabilistic specification of a general stochastic process $\{X_t\}$, we can proceed as follows:

1. Consider a fixed time t_1
Then X_t is a RV: denote its distribution by F_1
Then $F_1(x_1; t_1) \triangleq P[X_{t_1} \leq x_1]$, all $x_1 \in$ R
F_1 is called the first-order distribution of $\{X_t\}$ at time t_1

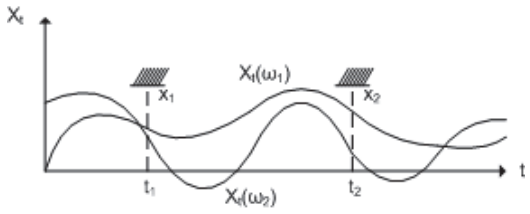


Note: F_1 gives no information about the dynamics of $\{X_t\}$

2. Now consider two fixed time t_1, t_2
Then (X_{t_1}, X_{t_2}) is a 2-dimensional random vector on $\{\Omega, F, P\}$ with joint-distribution F_2 where:

$$F_2(x_1, x_2; t_1, t_2) \triangleq P[X_{t_1}(\omega) \leq x_1, X_{t_2}(\omega) \leq x_2], \quad \text{all } (x_1, x_2) \in \mathrm{R}^2 \quad (13)$$

F_2 is called the second-order distribution at (t_1, t_2)



Note: F_2 gives *some* information about the dynamics of $\{X - t\}$

Definition
The *nth-order distribution* of $\{X_t\}$ at $(t_1, \dots, t_n)$ is the joint distribution, F_n, given by:

$$F_n(x_1, \dots, x_n; t_1, \dots, t_n) = P[X_{t_1} \leq x_1, \dots, X_{t_n} \leq x_n], \quad \text{for all } (x_1, \dots, x_n) \in \mathrm{R}^n \quad (14)$$

The collection of *all* such F_n for the process $\{X_t\}$ is called the **family of finite-order distribution (FFD)** for the stochastic process $\{X_t\}$.

Comments:

1. Almost all the probabilistic properties of $\{X_t\}$ are determined by the FFD.

2. The FFD clearly satisfy the *consistency relations*:

   (a) $F_n(x_1, \ldots, x_{n-1}, \infty) = F_{n-1}(x_1, \ldots, x_{n-1})$, etc.

   (b) $F_n(y_1, \ldots, y_n; u_1, \ldots, u_n) = F_n(x_1, \ldots, x_n; t_1, \ldots, t_n)$

   where $(y_1, \ldots, y_n)$, $(u_1, \ldots, u_n)$ are identical permutations of $(x_1, \ldots, x_n)$, $(t_1, \ldots, t_n)$ respectively.

3. The second-order family $\{F_2\}$ *completely characterises* two important classes of stochastic processes: these are

   (a) *Gaussian Processes*

   (b) *Markov Processes*

4. *Spectral Analysis* of $\{X_t\}$ depends only on F_2.

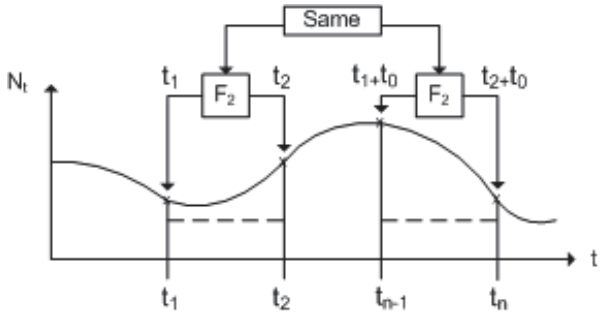
LECTURE ON
Stationary Process

These are stochastic processes which represent systems or processes in *statistical equilibrium*: the probability distributions do not change with time.
Example: the thermal noise voltage in a resistor at constant temperature.

Definition

Let $\{X_t : t \in T\}$ be a process with a *linear index* set T.
Means that if $t_1, t_2 \in T$ then $(t_1 + t_2) \in T$. The process $\{X_t\}$ is *stationary of order* k iff:

$$F_k(x_1, \ldots, x_k; t_1, \ldots, t_k) = F_k(x_1, \ldots, x_k; t_1 + t_0, \ldots, t_k + t_0)$$
for all $(x_1, \ldots, x_k) \in \mathrm{R}^k$
and all $t_1, \ldots, t_k, t_0 \in T$ (15)



2-stationary Process

Comments:

1. From the consistency relations, if $\{X_t\}$ is k-stationary then $\{X_t\}$ is j-stationary, for all $j < k$.

2. If $\{X_t\}$ is k-stationary for *all* k, we say it is **strictly stationary**.

3. If $\{X_t\}$ is stationary, applying a time shift of $t_0 = -t_1$ gives the following relations:

$$\begin{cases} F_1(x_1; t_1) = F_1(x_1; 0) \\ \text{(i.e. a constant amplitude distribution)} \\ F_2(x_1, x_2; t_1, t_2) = F_1(x_1, x_2; 0, t_2 - t_1) \\ \text{(depends only on time interval } \tau = t_2 - t_1) \end{cases}$$

So we can use the notation

$$\begin{cases} F_1(x_1) \\ F_2(x_1, x_2; \tau) \\ etc.etc. \end{cases}$$

if the process $\{X_t\}$ is stationary.

4. For Gaussian process and Markov process:
2-stationarity implies *strictly stationarity*.

Examples:

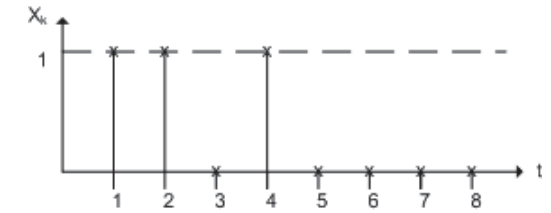
1. Bernoulli Process

This is simply an infinite sequence of Bernoulli trials, with

$X_k =$ number of successes at kth trial

and $P(X_k = 1) = p$

$P(X_k = 0) = q = 1 - p$



Bernoulli Process. An example of a pure noise process

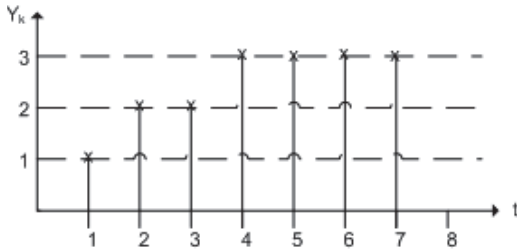
Note that $\{X_k\}$ is simply a sequence of iid RV and so the FFD are completely specified by the (time-invariant) 1st-order distribution F_1, defined here by $(p_0 = p; p_1 = q)$. It follows that $\{X_k\}$ is *strictly stationary*.

2. Binomial Process

Given a Bernoulli process $\{X_k\}$ we can generate a *Binomial Process*, $\{Y_k\}$, by the relation

$$Y_k = \sum_{i=1}^{k} X_i \qquad \text{for } k = 1, 2, \ldots \qquad (16)$$

In words, $\begin{cases} Y_k = & \text{number of sucesses in first } k \text{ trials} \\ = & \text{binomial RV (parameters } k, p) \end{cases}$



Binomial Process.

NB:

(1) Since Y_k is binomial (k, p), the FFD all depend on k; hence $\{Y_k\}$ is *non-stationary*.

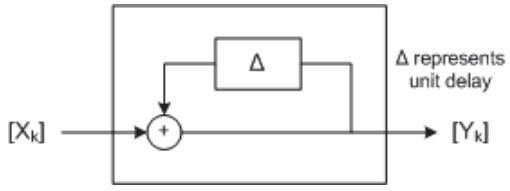
(2) The *increments* $\triangle Y_r = (Y_{k+r} - Y_k)$ on non-overlapping intervals are *independent* RV.

It follows that the FFD are completely specified by the 2nd-order distributions F_2. In fact $\{Y_k\}$ is example of a special-type of **Markov process** called an *independent-increment process*.

(3) Alternative definition of $\{Y_k\}$ is by means of the difference equations

$$Y_k = Y_{k-1} + X_k \qquad (17)$$

where $\{X_k\}$ is Bernoulli process and $Y_0 = 0$.
A system representation of (17) is given by the diagram.



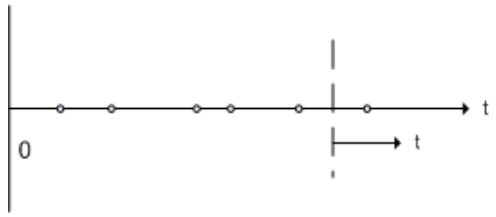
NB:

(i) Equation (17) is a special case of a *first-order difference equation* driven by an iid input sequence: the output of any such system is a **discrete-time Markov process**.

(ii) Equation (17) is a *linear equation* with iid input: the output of such a system is called a **linear process**.

3. Poisson Process

Consider a random stream of events:



Let N_t = number of events in $(0, t]$, for all $t \geq 0$
Then $\{N_t : t \geq 0\}$ is a **Poisson process** with *intensity* μ, iff:

(i) $N_0 = 0$

(ii) For any finite set of times $t_1, t_2, \ldots, t_n$ such that $0 < t_1 < t_2 \cdots < t_n$, the random variables $N_{t_1}, (N_{t_2} - N_{t_1}), \ldots$ are **independent Poisson RV**, with parameters $\mu t_1, \mu(t_2 - t_1), \ldots$ respectively.

Comments:

1. Putting $t_1 = t; t_2 = t + h$ in condition (ii) of a Poisson process (see above) gives:

$$P[(N_{t+h} - N_t) = k] = \frac{(\mu h)^k}{k!} e^{-\mu h}, \; k = 0, 1, \ldots$$
$$\text{when } h \to 0 = \mu h + o(h), \; k = 1$$
$$= (1 - \mu h) + o(h), k = 0$$

2. $\{N_t\}$ is a continuous-time example of,
(a) an *independent-increment process*
(b) a *Markov process*.

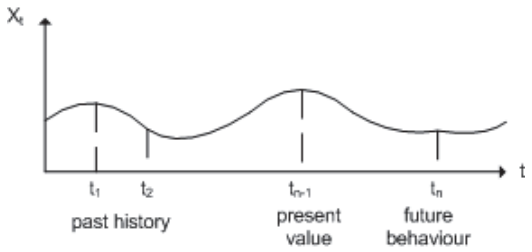
LECTURE ON
Markov Processes.

Definition:

A stochastic process $\{X_t : t \in T\}$ is a *Markov process* iff for every $n > 1$, for every set of times $(t_1, t_2, \ldots, t_n) \in T^n$, and for every $(x_1, \ldots, x_n) \in \mathsf{R}^n$.

$$P[X_{t_n} \leq x_n | X_{t_{n-1}} = x_{n-1}, \ldots, X_{t_1} = x_1] =$$
$$P[X_{t_n} \leq x_n | X_{t_{n-1}} = x_{n-1}] \quad (18)$$

In words, $\{X_t\}$ is a Markov process if and only if the conditional distribution of X_t given the past history $(X_{t_1}, \ldots, X_{t_{n-1}})$ is the same as the conditional distribution of X_{t_n} given the most recent value $X_{t_{n-1}}$.



NB: Since the future behaviour $\{X_t : t > t_{n-1}\}$ is completely determined by the present value X_t the latter is called the state of $\{X_t\}$ at time t_{n-1}.

Comments:

1. The basic property specified in the definition is called the *Markov property*: it is usually written in the form

$$P[X_{t_n}|X_{t_{n-1}},\dots,X_{t_1}] = P[X_{t_n}|X_{t_{n-1}}] \qquad (19)$$

2. Consider the three values X_s, X_t, X_u, where t is the present time and $s < t < u$, and suppose that the value of X_t is known. Then:

$$\begin{aligned} P(X_s, X_u|X_t) &= P(X_s|X_t)P(X_u|X_s, X_t) \\ &= P(X_s|X_t)P(X_u|X_t) \qquad (20) \end{aligned}$$

The last two expressions are equal by the Markov property.
i.e. Given the present state X_t, the past and future behaviour of the process are statistically independent (This is an alternative definition of a Markov process).

3. *Discrete-state* Markov processes are usually called **Markov chains**. Examples are the Binomial processes and the Poisson processes.

4. Probability specification of a Markov process.
Consider a Markov chain $\{X_t : t \in T\}$, with state set E= $\{1, 2, \dots\}$ and denote the probability

$$P[X_t = i_n|X_t = i_{n-1},\dots,X_t = i_1] \qquad (21)$$

by

$$p[i_n|i_{n-1},\dots,i_1; t_1,\dots,t_n] \qquad (22)$$

Then the nth-order distribution F_n is determined by the joint probability mass function:

$$\begin{aligned} &P[X_t = t_1,\dots,X_t = i_n] \\ &\quad = p(i_1)p(i_2|i_1)p(i_3|i_2,i_1) - - - -p(i_n|i_{n-1},\dots,i_1) \\ &\quad = p(i_1)p(i_2|i_1)p(i_3|i_2) - - - -p(i_n|i_{n-1}) \end{aligned}$$

by the Markov property

So the nth-order distribution $F_n(t_1,\dots,t_n)$ is completely specified by:

(i) the *initial distribution*, $p(i_1; t_1)$

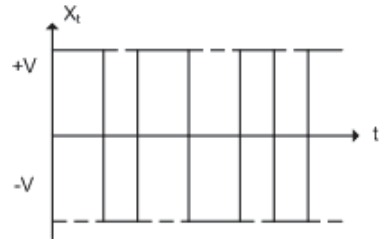
(ii) the set of *conditional distributions*,

$$p(i_k|i_{k-1}; t_{k-1}, t_k) \qquad \text{with } k = 2,\dots,n \qquad (23)$$

The latter are called the *transition distributions* of $\{X_t\}$

NB: Both (i) and (ii) are obtainable from the family F_2.

Example: Random Telegraph Process
This is a binary (2-level) process which switches between the two levels at the event times of a Poisson process.



We can define the random telegraph process $\{X_t\}$ as follows:

(i) X_0 is a binary RV with probability distribution

$$P[X_0 = +V] = q \qquad (24)$$
$$P[X_0 = -V] = q = 1 - p \qquad (25)$$

(ii) For $t > 0$

$$X_t = X_0(-1)^{N_t} \qquad (26)$$

where $\{N_t\}$ is a Poisson process with rate μ

Now since $\{X_t\}$ is a transformed Poisson process it follows that $\{X_t\}$ processes the Markov property and hence must be completely characterised by the distribution F_1 and F_2.

Probability specification of the RT process

(i) *1st-order distribution* $F_1(t)$

Use the relation

$$P(X_t) = \sum_{X_0} P(X_0, X_t) = \sum_{X_0} P(X_0)P(X_t|X_0) \tag{27}$$

Write

$p_+(t) = P[X_t = +V]$

$p_{+|+}(t) = P[X_t = +V|X_0 = +V]$ etc, etc.

Then,

$$\begin{aligned} p_{+|+}(t) &= P[\text{Even no. of jumps in } (0,t)] \\ &= P[N_t \text{is even}] = e^{-\mu t}\cosh \mu t \end{aligned}$$

Similarly,

$$p_{-|+}(t) = e^{-\mu t}\sinh \mu t \tag{28}$$

and it is easy to see that

$$\begin{aligned} p_{+|-}(t) &= p_{-|+}(t) \\ p_{-|-}(t) &= p_{+|+}(t) \end{aligned}$$

Also, we are given that $p_+(0) = p; p_-(0) = q$

So, finally, the first-order distribution is given by (using Equation (27)):

$$\begin{cases} p_+(t) = \frac{p}{2}(1 + e^{-2\mu t}) + \frac{q}{2}(1 - e^{-2\mu t}) \\ p_-(t) = \frac{p}{2}(1 - e^{-2\mu t}) + \frac{q}{2}(1 + e^{-2\mu t}) \end{cases}$$

NB:

1. Note that F_1 is time-dependent. However it can be seen that
$$\lim_{t\to\infty} p_+(t) = \lim_{t\to\infty} p_-(t) = \frac{1}{2}$$
So $\{X_t\}$ becomes 1-stationary for large t

2. In the special case when $p = q = \frac{1}{2}$, we get
$p_+(t) = p_-(t) = \frac{1}{2}$, for all t, so that $\{X_t\}$ is 1-stationary in this case.

(ii) 2nd-order distribution $F_2(t_1, t_2)$

In this case we can use $P(X_t, X_t) = P(X_t)P(X_t|X_t)$ where $P(X_t)$ is given by the first-order distribution $F_1(t_1)$

Then, as before,

$$p_{+|+}(t_1, t_2) = p_{-|-}(t_1, t_2) = \frac{1}{2}(1 + e^{-2\mu(t_2 - t_1)}) \tag{29}$$

$$p_{-|+}(t_1, t_2) = p_{+|-}(t_1, t_2) = \frac{1}{2}(1 - e^{-2\mu(t_2 - t_1)}) \tag{30}$$

so that, putting $\tau = t_2 - t_1$,

$$p(+V, +V; t_1, t_1 + \tau) = p_+(t_1)[\frac{1}{2}(1 + e^{-2\mu\tau})] \tag{31}$$

together with similar relations for $p(-V, +V)$; $p(+V, -V)$; $p(-V, -V)$

NB: When $p = q = \frac{1}{2}$ we get (see above), $p_+(t_1) = p_-(t_1) = \frac{1}{2}$ and then F_2 depends only on τ, i.e. $\{X_t\}$ is 2-stationary.

LECTURE ON
Finite-state Markov chains

These form an important class of stochastic process, widely used because they are easy to analyse. The state set E of a finite Markov chain is *finite*, and we can always take

$$E = \{1, 2, \ldots, N\} \tag{32}$$

1. Discrete-time Case:

Now the index-set T is *discrete* and we can assume

$$T = \{0, 1, 2, \ldots\} \tag{33}$$

Let $\{X_t : t \in T\}$ be a discrete-time, finite-state Markov chain. Then we know that $\{X_t\}$ is characterised by (i) an initial state distribution and (ii) a set of transition distributions.

Define:

$$p_{ij}^n(t) = P[X_{t+n} = j|X_t = i] \tag{34}$$

the n-step transition probabilities at time t.

The most important of these are the one-step transition probabilities:

$$p_{ij}(t) = p_{ij}^1(t) \tag{35}$$

The $p_{ij}(t)$ contain all the information we need about the dynamics of the Markov chain $\{X_t\}$.

NB:

1. For fixed i, $p_{ij}^n(t)$ defines a *conditional probability distribution*, so

$$\sum_j p_{ij}^n = 1 \tag{36}$$

2. The set of transition probabilities can be regarded as the elements of an NxN matrix:

$$\mathbf{P}_t^n = [p_{ij}^n(t)]_{NxN} \tag{37}$$

$\mathbf{P}_t^n$ is the n-step transition matrix for $\{X_t\}$

Notice that $P_t^{(n)}$ is a matrix with (a)non-negative elements, (b) unit row sums.
Such a matrix is called **a stochastic matrix**.

Now define the **state distribution** at time t by:

$$p_i(t) = P[X_t = i] \tag{38}$$

and also write

$$\mathbf{p}(t) = \begin{bmatrix} p_1(t) \\ . \\ . \\ . \\ . \\ p_N(t) \end{bmatrix} \tag{39}$$

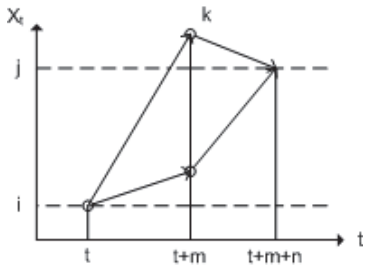
Then, by the basic Markov property,

$$p_j(t+n) = \sum_i p_i(t) p_{ij}^n(t) \tag{40}$$

or, in matrix form,

$$\mathbf{p}^T(t+n) = \mathbf{p}^T(t)\mathbf{P}_t^n \tag{41}$$

Calculation of $\mathbf{P}_t^{(n)}$



Now, since all paths from i to j must pass through some state k at time $t+m$:

$$\begin{aligned} p_{ij}^{m+n}(t) &= \sum_k P[X_{t+m} = k, X_{t+m+n} = j | X_i = i] \\ &= \sum_k p_{ik}^m(t) p_{kj}^n(t+m), \text{for each } i, j \end{aligned}$$

or, in matrix form,

$\mathbf{P}_t^{m+n} = \mathbf{P}_t^m \mathbf{P}_{t+m}^n$ Chapman-Kolmogorov equation

NB:

1. Transition matrices must always satisfy this condition

2. The $C-K$ equation enables everything to be computed from the 1-step transition matrices.

LECTURE ON
Homogeneous Markov Chains

A Markov chain $\{X_t\}$ is said to be homogeneous (in t) when its transition matrices are time-invariant, i.e. when

$$\mathbf{P}_{t+s}^{(n)} = \mathbf{P}_t^{(n)} \qquad \text{for all } n,\ t,\ s$$

We can then write

$$\mathbf{P}_t^{(n)} = \mathbf{P}^{(n)} \tag{42}$$

and the *Chapman-Kolmogorov* equation becomes

$$\mathbf{P}^{(m+n)} = \mathbf{P}^{(m)}\mathbf{P}^{(n)} \tag{43}$$

In particular,

$$\mathbf{P}^{(m+1)} = \mathbf{P}^{(m)}\mathbf{P} \qquad \text{where } \mathbf{P} = \mathbf{P}^{(1)}$$

so that, by recursion,

$$\mathbf{P}^{(m)} = \mathbf{P}^m \tag{44}$$

Now apply this result to Equation (27) . We get:

$$\mathbf{p}^T(t+n) = \mathbf{p}^T(t)\mathbf{P}^n \tag{45}$$

and, putting $t = 0$,

$$\mathbf{p}^T(n) = \mathbf{p}^T(0)\mathbf{P}^n \tag{46}$$

Equations (44) and (46) show that the transition distributions $\mathbf{P}^{(m)}$ and all the state distribution $\mathbf{p}^T(n)$ are determined by the 1-step matrix **P**.

Equation (46) **is the basic relation for homogeneous Markov chains.**

LECTURE ON
Stationary Distributions (for finite-state/homogeneous/discrete-time Markov chains)

The Markov chain $\{X_t\}$ will be stationary iff its state distribution $\mathbf{p}^T(n)$ is independent of n: any such state distribution π^T must satisfy equation (46) above, so that:

$$\pi^T = \pi^T\mathbf{P} \tag{47}$$

Such a π^T is a **stationary distribution** for $\{X_t\}$: choosing the initial state distribution, $\mathbf{p}^T(0) = \pi^T$ makes $\{X_t\}$ a stationary process.

NB: In many cases (depending on the transition structure of $\{X_t\}$ it is found that the n-step transition matrix $\mathbf{P}^n$ converges (as $n \to \infty$) to a matrix with identical rows.

It can be shown that, if this is so, the rows of $\mathbf{P}^n$ all tend to π^T.

$$\lim_{n\to\infty} \mathbf{P}n = [\mathbf{e}\pi^T], \qquad \text{where } \mathbf{e} = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \\ 1 \end{bmatrix}$$

In such a case, for sufficiently large n we may write

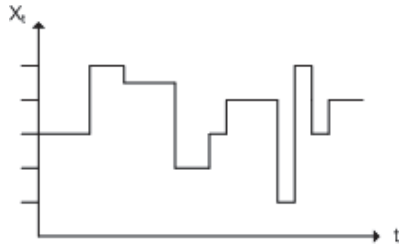
$$\mathbf{p}^T(n) = \mathbf{p}^T(0)\mathbf{P}^n = \mathbf{p}^T(0)[\mathbf{e}\pi^T] = \pi^T \tag{48}$$

i.e. the state distribution tends to the stationary distribution π for all initial state distributions $\mathbf{p}^T(0)$.

LECTURE ON
Continuous-Time Markow Chains

Many of the random processes occurring in communications systems can be represented by continuous-time Markov chains - that is, Markov processes with a discrete state space and continuous time set. In these notes the general properties of continuous-time Markov chains are reviewed and the relations between some of these properties are indicated.

Consider a general continuous-time Markov chain $\{X_t\}$ with state space $E = \{1, 2, \ldots, N, \ldots\}$ and with index set (time set) $T = [0, \infty)$. The sample paths (time histories) of $\{X_t\}$ will have the following general form:



Examples of such processes are: Poisson processes, birth-death processes, M/M/K queueing processes, etc. By definition, $\{X_t\}$ exhibits the basic **Markov property**:

$$P(X_k|X_{k-1}) = P(X_k|X_{k-1}, \ldots, X_1) \qquad (49)$$

for any choice of times $t_1, t_2, \ldots, t_k$ in T. Here, X_k denotes X_t at time t_k.

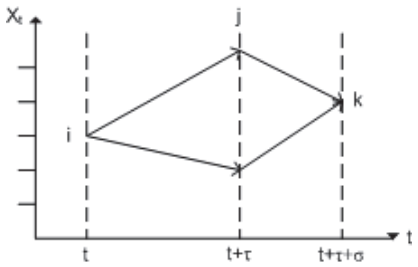
Hence we can analyse the dynamic behaviour of $\{X_t\}$ in terms of transition probabilities, p_{ij}, defined by

$$p_{ij}(\tau; t) = P(X_{t+\tau} = j|X_t = i) \qquad (50)$$

In almost all cases of interest we can take $\{X_t\}$ to be a homogeneous chain, in which case the transition probabilities are independent of t, so that we can write

$$p_{ij}(\tau; t) = p_{ij}(\tau) \qquad (51)$$

As in the case of discrete-time Markov chains, the transition probabilities p_{ij} satisfy the Chapman-Kolmogorov (C-K) equations.



Summing over all paths from i to k:

$$p_{ik}(\tau + \sigma) = \sum_j p_{ij}(\tau) p_{jk}(\sigma) \qquad (52)$$

for each pair of states (i, k).

As in the discrete-time case, it is convenient to introduce vector/matrix notation and to define the following terms:

$$p_i(t) = P[X_t = i] \qquad (53)$$

$$\mathbf{p}_t = \begin{bmatrix} p_1(t) \\ p_2(t) \\ \cdot \\ \cdot \end{bmatrix}$$

(State probability vector)

$$\mathbf{P}(\tau) = [p_{ij}(\tau)]$$

(Transition probability matrix for interval τ)

Using the above definitions, the C-K equations can be written:

$$\mathbf{P}(\tau + \sigma) = \mathbf{P}(\tau)\mathbf{P}(\sigma) \qquad (54)$$

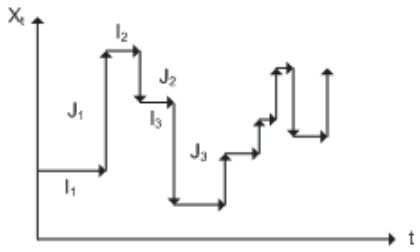
The transition probability matrix $\mathbf{P}(\tau)$ is a matrix function of the time interval τ: each of its elements, $p_{ij}(\tau)$, tells us how the probability of transition $i \to j$ varies with the length of the transition interval τ.

Now in the discrete-time case we know that the k-step transition matrix $\mathbf{P}^{(k)}$ is easily computed from the 1-step matrix $\mathbf{P}$: in fact $\mathbf{P}^{(k)} = \mathbf{P}^k$. In a similar way, for a continuous-time chain we compute $\mathbf{P}(\tau)$ from a knowledge of the local properties of $\{X_t\}$.

i.e. from the transition properties of $\{X_t\}$ on an infinitesimal interval. To do so it is convenient first to introduce an alternative representation of a general continuous-time chain.

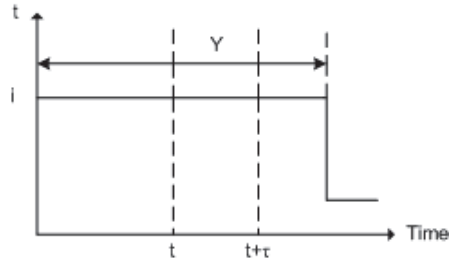
Alternative Representation of $\{X_t\}$

Fact: Any continuous-time Markov chain has sample paths which consist of a sequence of intervals $(I_1, I_2, \dots)$ interleaved with a sequence of jumps $(J_1, J_2, \dots$ i.e. each sample path can be expressed as an alternating sequence $(I_1, J_1, I_2, J_2, \dots)$.



By the basic Markov property, the intervals $I_1, I_2, \dots$ must be *independent* random variables. But if $\{X_t\}$ is to be a Markov process the Markov property is required to hold at every time t, and for this to be so *every interval must be an exponential random variable*.

Proof: For a given state i, let Y denote the sojourn time in i (time between entry into and exit from i) and let t and $t+\tau$ be fixed times, measured from the time of entry into i:



By the Markov property

$$P(Y > t+\tau | Y > t) = P(Y > \tau) \qquad (55)$$

since if $Y > t$ we can re-set the clock to 0 at time t. [The fact that the state i was occupied from 0 to t is irrelevant if we know that $X_t = i$]. Rewriting (55),

$$\frac{P(Y > t+\tau, Y > t)}{P(Y > t)} = P(Y > \tau) \qquad (56)$$

$$\text{ie. } P(Y > t+\tau) = P(Y > t)P(Y > \tau) \qquad (57)$$

So, writing $F_Y^c(t) = P(Y > t)$, the function F_Y^c must satisfy

$$F_Y^c(t+\tau) = F_Y^c(t)F_Y^c(\tau) \qquad (58)$$

with boundary conditions

$$F_Y^c(0) = 1 \text{ and } F_Y^c(\infty) = 0 \qquad (59)$$

It is easily verified that the exponential function

$$F_Y^c(t) = e^{-\mu t}, \qquad t \geq 0 \qquad (60)$$

satisfies (59). Furthermore it may be shown that this is the only non-negative, monotonic solution of (59). Thus the probability distribution function for Y must have the form:

$$F_Y(t) = 1 - F_Y^c(t) = 1 - e^{-\mu t}, \qquad t \geq 0 \quad (61)$$

and so Y is an **exponential RV**.

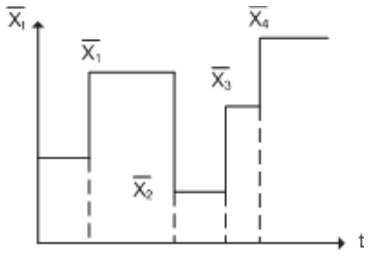
Note that the parameter μ need not have the same value in very state: in general μ will be state-dependent and so we denote its value in state i by μ_i. Thus we can say that, for any state i, the sojourn time Y is an exponential RV with parameter μ_i. It follows that, at any time t,

$$P(X_{t+\triangle t} \neq i | X_t = i) = \mu_i \triangle t + o(\triangle t) \qquad (62)$$

i.e. that the probability of a jump out of state i in any short interval $(t, t+\triangle t)$ is $\mu_i \triangle t$. The parameter μ_i is the *jump rate* out of state i.

Jump Probabilities

Associated with any continuous-time Markov chain $\{X_t\}$ is a discrete-time chain $\{\overline{X_k} : k = 0, 1, 2, \dots\}$ which specifies the state immediately after a jump, i.e. $\overline{X_k}$ specifies the *destination* of the kth jump.



The discrete-time chain $\{\overline{X_k}\}$ has various names: it is known as the *jump chain* or the *next-state chain*, of the continuous-time chain $\{X_t\}$. [It is an example of an *embedded Markov chain*: this is the discrete-time chain obtained from a given continuous-time stochastic process by restricting the time-set to a sequence of special times $(t_0, t_1, t_2, \dots)$]. Now consider the 1-step transition probabilities for $\{\overline{X_k}\}$, defined by:

$$P_{ij} = P(\overline{X}_{k+1} = j | \overline{X}_k = i) \tag{63}$$

Note that the transition probabilities P_{ij} are the *next-state probabilities* for the original continuous-time chain $\{X_t\}$; in words:

P_{ij} =[probability, given that the *present state* is i, that the *next state* will be j]

Clearly the dynamics of the jump chain $\{\overline{X_k}\}$ are completely determined by the P_{ij}. Note also that $\{\overline{X_k}\}$ tell us about the behaviour of the *sequence of states* in the original continuous-time chain $\{X_t\}$, but nothing about the time spent in each state.

Transition Rates

In order to get a complete description of the behaviour of the continuous-time chain $\{X_t\}$, we must put together the next-state probabilities P_{ij} (which determine the state sequence) and the jump rates μ_i (which determine the times spent in each state). To do this, note that, when the present state of X_t is i, the probability that a jump will occur in the next short interval $\triangle t$ and that it will be to state j, is given by

$$\begin{aligned} P(X_{t+\triangle t} = j | X_t = i) &= P(X_{t+\triangle t} \neq i | X_t = i) \\ .P(X_{t+\triangle t} = j | X_t = i, X_{t+\triangle t} \neq i) & \\ &= (\mu_i \triangle t).P_{ij} + o(\triangle t) \\ &= (\mu_i P_{ij})\triangle t + o(\triangle t) \end{aligned} \tag{64}$$

For any two states i, j we now define the *transition rate* (or *transition intensity*) from $i \to j$ as the coefficient q_{ij} given by

$$q_{ij} = \mu_i P_{ij} \tag{65}$$

so that, for any state $j \neq i$, equation (64) can be written as:

$$p_{ij}(\triangle t) = q_{ij}\triangle t + o(\triangle t) \tag{66}$$

The set of transition rates q_{ij} gives a *complete description* of the stochastic behaviour of the continous-time Markov chain $\{X_t\}$. For example we can easily determine the jump rates μ_i and next-state probabilities P_{ij} from the transition rates, using the relations:

$$\mu_i = \sum_{j \neq i} q_{ij} \tag{67}$$

$$P_{ij} = \left[\frac{q_{ij}}{\mu_i}\right] = \left[\frac{q_{ij}}{\sum_{j \neq i} q_{ij}}\right] \tag{68}$$

Note that the q_{ij} are defined only for changes of state, i.e. for $j \neq i$. However, for no change we have:

$$\begin{aligned} P(X_{t+\triangle t} = i | X_t = i) &= 1 - P(X_{t+\triangle t} \neq i | X_t = i) \\ &= 1 - \mu_i \triangle t + o(\triangle t) \end{aligned} \tag{69}$$

For reasons which will emerge later it is convenient to define the *self-transition* rates q_{ii} by

$$q_{ii} = -\mu_i \tag{70}$$

so that, using (67)

$$q_{ii} = -\sum_{j \neq i} q_{ij} \tag{71}$$

In terms of the self-transition rates we can rewrite (69) in the form:

$$p_{ii}(\triangle t) = 1 + q_{ii}\triangle t + o(\triangle t) \tag{72}$$

Transition Rate Matrix

It is sometimes convenient (particularly when the state space E is *finite*) to collect all the transition rates together as the elements of a matrix **Q**:

$$\mathbf{Q} = [q_{ij}] \tag{73}$$

Q is called the *transition rate matrix* for $\{X_t\}$ (or the *infinitesimal generator* of $\{X_t\}$). It has several obvious properties:

1. all off-diagonal elements of **Q** are non-negative
2. the diagonal elements of **Q** are all negative
3. the row sums of **Q** are all zero. i.e.

$$\sum_j q_{ij} = 0, \qquad \text{for each } i \in E$$

Kolmogorov Differential Equations

If the transition rates q_{ij} are all known, equations (66) and (72) show how to evaluate the transition probabilities $p_{ij}(\tau)$ when τ is very small. A method for evaluating the $p_{ij}(\tau)$ for any interval τ can be derived by using the Chapman-Kolmogorov relations.

In equation (52) take $\sigma = \triangle\tau$ (i.e. a small increment in τ): then

$$p_{ik}(\tau + \triangle\tau) = \sum_j p_{ij}(\tau)p_{jk}(\triangle\tau) \tag{74}$$

$$= \sum_{j\neq k} p_{ij}(\tau).q_{jk}\triangle\tau + p_{ik}(\tau)[1 + q_{kk}\triangle\tau] \tag{75}$$

on using equations (66) and (72). Thus, for any pair of states (i, k),

$$p_{ik}(\tau + \triangle\tau) = p_{ik}(\tau) + \sum_j p_{ij}(\tau)q_{jk}\triangle\tau \tag{76}$$

Now subtract $p_{ik}(\tau)$ from both sides, divide by $\triangle\tau$, and allow $\triangle\tau$ to shrink to zero; the result is a set of first-order differential equations for the $p_{ij}(\tau)$:

$$\frac{dp_{ik}(\tau)}{d\tau} = \sum_j p_{ij}(\tau)q_{jk} \tag{77}$$

for every pair of states (i, k).

This set of coupled differential equations (called *Kolmogorov's forward equation*), together with the initial conditions

$$p_{ik}(0) = 0, k \neq i \tag{78}$$

$$= 1, k = i \tag{79}$$

is sufficient to define the behaviour of the $p_{ij}(\tau)$. On replacing τ by $\triangle\tau$ and σ by τ in equation (52), and following the same procedure as above, we obtain an alternative set of differential equations for the $p_{ij}(\tau)$:

$$\frac{dp_{ik}(\tau)}{d\tau} = \sum_j q_{ij}p_{jk}(\tau) \tag{80}$$

for every pair of states (i, k).

This set (*Kolmogorov's backward equations*) is again sufficient to determine the $p_{ij}(\tau)$. The most convenient way of expressing equations (77) and (80) is in matrix form:

$$\frac{d\mathbf{P}(\tau)}{d\tau} = \mathbf{P}(\tau).\mathbf{Q} \tag{81}$$

$$\frac{d\mathbf{P}(\tau)}{d\tau} = \mathbf{Q}.\mathbf{P}(\tau) \tag{82}$$

with, in both cases, *initial condition* $\mathbf{P}(0) = \mathbf{I}$. It is easily verified that a formal solution to equation (81) [or to Equation (82)] with $\mathbf{P}(0) = \mathbf{I}$ is:

$$\mathbf{P}(\tau) = exp^{\mathbf{Q}\tau} \tag{83}$$

where the matrix exponential $exp^{\mathbf{M}}$ is defined for any square matrix **M** by the usual power series:

$$exp^{\mathbf{M}} = \mathbf{I} + \mathbf{M} + \frac{1}{2}!\mathbf{M}^2 + \ldots \tag{84}$$

Provided that $\{X_t\}$ has no instantaneous states i.e. provided that the jump-rates μ_i are all finite. It may be shown that the exponential solution in equation (83) is indeed the required solution to equation (81) [or (82)].

LECTURE ON
Continuous-time Markov chains: Equilibrium Behaviour

Consider a continuous-time Markov chain $\{X_t\}$ with *space state* $E = \{1, 2, 3, \ldots, N\}$, *transition rates* q_{ij}, and transition probabilities $p_{ij}(\tau)$.
Given the *state probability distribution* $(p_i(t) : i = 1, 2, 3, \ldots)$ at any time t, we can compute the distribution at any later time $(t + \tau)$ by the usual Markov equations:

$$p_j(t+\tau) = \sum_{i \in E} p_i(t) p_{ij}(\tau), \qquad j = 1, 2, \ldots \tag{85}$$

where the transition probabilities $p_{ij}(\tau)$ can, in principle, be computed using the Kolmogorov differential equations. In fact, in matrix form we have:

$$\begin{aligned} \mathbf{P}(\tau) &= exp^{\mathbf{Q}\tau} \\ &= \mathbf{I} + \tau\mathbf{Q} + \frac{\tau^2}{2!}\mathbf{Q}^2 + \ldots \end{aligned} \tag{86}$$

The Markov chain $\{X_t\}$ is said to be *irreducible* if every pair of states (i, j) are *intercommunicating*, that is, if, for each pair (i, j), we have:

$$p_{ij}(\tau) > 0 \qquad \text{for some} \qquad \tau > 0$$

Thus an irreducible Markov chain is simply one in which it is possible to get from any state i to any other state j. It may be shown that if every state $i \in E$ is *stable* (this means that the q_{ii} is *finite* for each i) then

$$\mathbf{P}(\tau) \to I \qquad \text{as} \qquad \tau \to \infty$$

Such a transition matrix is said to be *standard*.

Stationary Distributions

In most cases of practical interest $\{X_t\}$ settles down to statistical equilibrium as $t \to \infty$, i.e. for large t, $\{X_t\}$ is a *stationary process* with a time-invariant state probability distribution (which may or may not depend on the initial state X_0). If $\{X_t\}$ is an *irreducible* chain this equilibrium distribution will be unique.

Suppose that $(\pi_i : i = 1, 2, \ldots)$ is the equilibrium distribution of $\{X_t\}$. Then, since (π_i) is time-invariant, the equilibrium version of equation (85) is:

$$\pi_j = \sum_i \pi_i p_{ij}(\tau), \text{ for each } i \in E \text{ and every } \tau > 0$$

or, in matrix form,

$$\begin{aligned} \underline{\pi}^T &= \underline{\pi}^T P(\tau) \\ &= \underline{\pi}^T e^{Q\tau} \\ &= \underline{\pi}^T \sum_{k=0}^{\infty} \frac{(Q\tau)^k}{k!}, \text{ on using (86)} \end{aligned} \tag{87}$$

Clearly a distribution $\underline{\pi}$ will satisfy (87) if and only if it satisfies the equation:

$$\underline{\pi}^T \mathbf{Q} = \underline{0}^T \tag{88}$$

or, in scalar form:

$$\sum_{i \in E} \pi_i q_{ij} = 0, \text{ each } i \in E \tag{89}$$

These are the so-called *equilibrium equations* for the continuous-time Markoc chain $\{X_t\}$ and any distribution satisfying (88) is called a *stationary* (or invariant) *distribution* for $\{X_t\}$.

Note that we always need the normalisation condition.

$$\sum_i \pi_i = 1 \tag{90}$$

to turn a solution of (88)into a proper probability distribution.

Fact: If $\{X_t\}$ is *irreducible*, with a *standard* transition matrix $P(\tau)$, then, for every initial state i,

$$p_{ij}(\tau) \to \pi_j, \text{ as } \tau \to \infty \text{ for each } j \in E$$

i.e. the equilibrium distribution (π_j) is independent of the initial state i

Balance Equations

Equations (89) show that, when the Markov chain $\{X_t\}$ is in statistical equilibrium, each state $j \in E$ is characterised by the condition:

$$\sum_{i \in E} \pi_i q_{ij} = 0 \tag{91}$$

or, if we separate out the the *self-transition* term q_{jj}:

$$\sum_{i \neq j} \pi_i q_{ij} = -\pi_j q_{jj}, \text{ for each state } j \tag{92}$$

But, by definition, $-q_{jj} = +\sum_{i \neq j} q_{ji}$ and so equation (92) can be written:

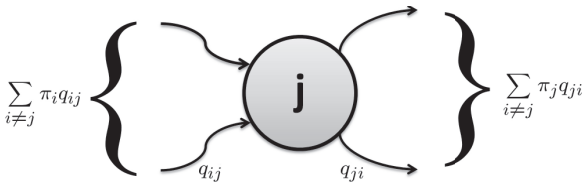
$$\sum_{i \neq j} \pi_i q_{ij} = \sum_{i \neq j} \pi_j q_{ji}, \text{ for each state } j \qquad (93)$$

These equations (one for each state $j \in E$) are called the **global balance equations** for $\{X_t\}$. They are simply the equilibrium equations (89) re-written with the self-transition terms q_{jj} eliminated. The equilibrium state distribution $(\pi_i : i = 1, 2, \dots)$ must satisfy equation (93).

Interpretation of (93): If we inspect the chain $\{X_t\}$ in equilibrium, the probability that we shall find $\{X_t\}$ in state i at time t and making a transition to state j in $(t, t + \triangle t)$ is $(\pi_i q_{ij})\triangle t$. i.e. is proportional to $\triangle t$. The coefficient $(\pi_i q_{ij})$ is sometimes called the *probability flux* from state i to state j: it is the *rate* at which $i \to j$ transitions occur at time t. Using this concept, equation (93) can be expressed in words as:

At equilibrium:

Total probability flux into state j =
Total probability flux out of state j



Although the global balance equations are simply a set of linear algebraic equations, they are usually not easy to solve, especially when the state space is large. Fortunately, in some cases it is possible to use a set of simpler equations.

Local Balance

The global local balance equation for a state j expresses the fact that, in equilibrium, there is a balance between the *total* flux into j and the *total* flux out of j. In certain types of continuous-time Markov chain a stronger set of conditions apply: there must be *flux balance between each pair of states*. In such case we must have:

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for each pair } (i, j \neq i) \qquad (94)$$

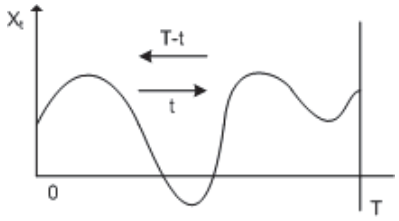
These equations are called the **local balance equations** for $\{X_t\}$. They are clearly much easier to solve than the corresponding global balance equations.

Comments:

(i) It is clear that local balance is a stronger requirement than global balance. For any state j, summing equations (94) over all $i \neq j$ yields equation (93). It follows than any solution to the local balance equations is also a solution to the global balance equations, i.e. is a stationary distribution of $\{X_t\}$.
(ii) As already noted, a Markov chain may not satisfy local balance. It is not easy to identify the class of chains for which local balance holds; one characteristic of such chains is that they are *reversible*.

Reversible Processes

If $\{X_t\}$ is a stochastic process (not necessarily Markov) on the time interval $[0, T]$, then the process $\{X_{T-t}\}$ is a time-reversed process on $[T, 0]$, i.e. the time index runs backwards from time T to time 0.



Question: How are the properties of $\{X_{T-t}\}$ related to those of $\{X_t\}$?

The main results are as follows:

Property A: If $\{X_t\}$ is a Markov process then so is $\{X_{T-t}\}$.

Proof: For any three times $s < t < u$ in $[0, T]$, we have

$$\begin{aligned} P[X_s | X_t, X_u] &= \frac{P[X_s, X_t, X_u]}{P[X_t, X_u]} \\ &= \frac{P[X_s]P[X_t, X_u | X_s]}{P[X_t, X_u]} \\ &= \frac{P[X_s]P[X_t | X_s]P[X_u | X_t]}{P[X_t]P[X_u | X_t]} \end{aligned}$$

by the Markov property

Then, cancelling $P[X_u|X_t]$ and using Bayes' Rule:

$$P[X_s|X_t, X_u] = P[X_s|X_t] \quad (95)$$

which is the Markov property for the reversed process $\{X_{T-t}\}$.

Property B: If $\{X_t\}$ is a *stationary* Markov chain, with equilibrium distribution (π_i), then $\{X_{T-t}\}$ is also stationary with equilibrium distribution (π_i). This follows immediately from the fact that if $\{X_t\}$ is stationary then, (by definition of stationarity)

$P[X_t = i] = P[X_{T-t} = i]$, for each state i.

Now denote the *transition probabilities* of the reversed chain by $p'_{ij}(\tau)$ and the *transition rates* by q'_{ij}. Then, if $\{X_t\}$, and hence $\{X_{T-t}\}$, are both *stationary*, we have, by definition, for any $\tau > 0$:

$$\begin{aligned} p'_{ij}(\tau) &= P[X_t = j|X_{t+\tau} = i] \\ &= \frac{P[X_t = j]P[X_{t+\tau} = i|X_t = j]}{P[X_{t+\tau} = i]} \\ &\text{, by Bayes Rule} \\ &= \frac{\pi_j p_{ji}(\tau)}{\pi_i} \end{aligned}$$

Then, letting $\tau \to 0$, we get

$$q'_{ij} = \frac{\pi_j}{\pi_i} q_{ji} \quad (96)$$

i.e. the transition rate from $i \to j$ for the reversed chain is $\frac{\pi_j}{\pi_i}$ times the transition rate from $j \to i$ for the forward chain.

Definition: A stationary Markov chain $\{X_t\}$ is **reversible** if the corresponding time-reversed chain $\{X_{T-t}\}$ is statistically identical to $\{X_t\}$.
For such a chain we must clearly have

$$q'_{ij} = q_{ij}, \text{ all } (i, j) \quad (97)$$

and hence, using (96), all the q_{ij} must satisfy

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ all } (i, j) \quad (98)$$

But these are just the local balance equations, (94). Thus we conclude that:

$\{X_t\}$ is reversible $\Leftrightarrow$ Local balance holds for $\{X_t\}$

Comments:
(i) It is easy to show that any *birth/death process* satisfies the conditions for local balance and hence is reversible.
(ii) Clearly a *necessary* condition for reversibility is that if a transition $i \to j$ is feasible then the reverse transition $j \to i$ must also be feasible. This condition is not of course *sufficient* to guarantee reversibility.

Partial Balance
Suppose that for each state $j \in E$ we can partition the remaining set of states $\{i : i \neq j\}$ into a collection of disjoint subsets $E_j^1, E_j^2, \ldots, E_j^m$ such that the equilibrium distribution (π_i) satisfies the following balance equations for the state j:

$$\sum_{i \in E_j^k} \pi_i q_{ij} = \sum_{i \in E_j^k} \pi_j q_{ji}, \text{ for each subset } E_j^k \quad (99)$$

with a similar set of equations for every state j. Then the chain $\{X_t\}$ is said to satisfy partial balance and equations (99) are **partial balance equations** for $\{X_t\}$.

Comments:
(i) Partial balance equations such as (99) are intermediate between local balance equations, (94), and global balance equations (93). In fact it is easy to see that

(π_i) satisfies Equation (94) $\Rightarrow$ (π_i) satisfies Equation (99) $\Rightarrow$ (π_i) satisfies Equation (93)

(ii) The advantage of using partial balance, where it is applicable, is that equations (99) will be easier to solve that equations (93). Unlike local balance, it is not possible to characterise the existence of partial balance for a chain by an an equivalent property such as reversibility; however the communication structure of the chain sometimes suggests that partial balance may apply.

Footnote:

Evolution of the State Probability Distribution

The notion of *probability flux* is useful when $\{X_t\}$ is not in equilibrium. Then, the net probability flux into (our out of) a state gives the rate of change of the state probability for that state. Too see this, note that, by the usual arguments:

$$p_j(t+\triangle t) = \sum_{i \neq j} p_i(t)(q_{ij}\triangle t) + p_j(t)\left[1 - \sum_{i \neq j}(q_{ij}\triangle t)\right] \quad (100)$$

so that, subtracting $p_j(t)$ from each side, dividing by $\triangle t$, and letting $\triangle t \to 0$, we get:

$$\frac{dp_j(t)}{dt} = \sum_{i \neq j} p_i(t)q_{ij} - \sum_{i \neq j} p_j(t)q_{ji}, \text{ for each } j \in E \quad (101)$$

and the right-hand side of (101) is just the difference between the probability fluxes into and out of state j.

Note:

(i) The global balance equations for the equilibrium follow immediately on setting $\frac{dp_j}{dt} = 0$

(ii) Equations (101) are essentially the *Kolmogorov forward differential equations* in slightly disguised form.

References:

- Kleinrock, L. 1975. Queuing Systems, Vol.1, New York: Willey.
- Cooper, R. B. 1990. Introduction to queueing theory, CEE Press.
- Girard, A. 1990. Routing in dimensioning in circuit-switched networks, Addision-Wesley.
- Harrison P. G. and Patel N. M. 1993, Performance modelling of communication networks and coputer architectures, Addision-Wesley.
- Schwartz, M. 1988. Telecommunication networks, Addision-Wesley.
- Bertsekas, D. P. and Gallager R. 1992. Data Networks, Englewood Cliffs, NJ: Prentice Hall.
- Ross, K. W. 1995. Multiple loss model for broadband telecommunication networks, Springer-Verlag.
- Schwartz, M. 1996. Broadband integrated networks. Prentice Hall. - Sahner R. A., Trivedi, K. S. and Puliafito A. 1995. Performance and reliability analysis of computer systems, Kluwer Academic Publ.
- Kershenbaum A. 1994. Telecommunication network design algorithms, McGraw-Hill.

Further reading:

- Hui, J. Y. 1990. Switching and traffic theory for integrated broadband networks, Kluger Press.
- King, P. J. B. 1990. Computer and communication systems performance modelling, Prentice-Hall.
- Acampora, A. S. 1994. An introduction to broadband networks. New York: Plenum Press.
- Stallings, W. 1992. ISDN and Broadband ISDN, New York: Macmillan.
- Tanenbaum, A. S. 1989. Computer Networks, Englewood Cliffs, NJ: Prentice-Hall.
- Stallings, W. 1985. Data Computer Communications, New York: Macmillan.
- Shcwartz, M. 1987. Telecommunication Network Protocols, Modelling, and Analysis, Reading: Addison Wesley.
- Kleinrock, L. 1976. Queueing Systems, Vol.2, New York: Willey.

Traffic Theory & Queueing Systems

E4.05 - S07: Part 2

Javier A. Barria
j.barria@imperial.ac.uk
1009a DEEE

LECTURE ON
Birth/death Process

Consider a discrete-state, continuous-time Markov chain $\{N_t\}$ with *state space*

$$E = \{0, 1, 2, \ldots, N\}$$

Note: we can have $N = \infty$.

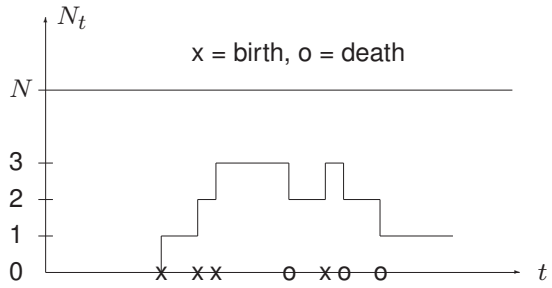
Then $\{N_t\}$ is a *birth/death process* iff

$$P\left(N_{t+\Delta t} = j \middle| N_t = i\right) = \begin{cases} \lambda_i \Delta t, & j = i+1 \\ \mu_i \Delta t, & j = i-1 \\ 0, & |j-i| > 1 \end{cases}$$

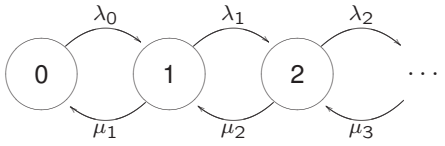
- Future of $\{N_t\}$ is regardless of its past history.
- The only possible transitions from any state i are $\boxed{i \to i+1}$ and $\boxed{i \to i-1}$.
- λ_i = birth coefficient (rate) in state i
  μ_i = death coefficient (rate) in state i

Of course, at the state space boundaries, we must have

$$\lambda_N = 0 \quad \text{since } N_t \le N$$
$$\mu_0 = 0 \quad \text{since } N_t \ge 0$$



The dynamics of $\{N_t\}$ can be represented by a *state transition diagram*, which shows the possible transitions between states:



Birth/death Equations

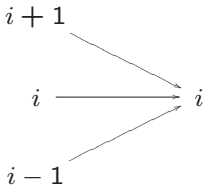
Denote the *state probabilities* at time t by

$$p_i(t) = P(N_t = i), \quad \text{for } i = 0, 1, \ldots, N$$

Since $\{N_t\}$ is Markov, we can compute the state probabilities $p_i(t)$ from knowledge of the initial distribution $\{p_i(0)\}$ plus the transition rates λ_i, μ_i.

Consider the event $\left\{N_{t+\Delta t} = i\right\}$.

On the short interval $[t, t+\Delta t]$, there are only 3 possible transitions into $\left\{N_{t+\Delta t} = i\right\}$ — as shown in the figure (assuming that i is not a boundary state).



time t time $t + \Delta t$

So

$$\begin{aligned} & P\left(N_{t+\Delta t}=i\right) \\ &= P\left(N_t=i, N_{t+\Delta t}=i\right) \\ &\quad +P\left(N_t=i-1, N_{t+\Delta t}=i\right) \\ &\quad +P\left(N_t=i+1, N_{t+\Delta t}=i\right) \end{aligned}$$

That is,

$$\begin{aligned} & p_i(t+\Delta t) \\ &= p_{i-1}(t)(\lambda_{i-1}\Delta t)+p_{i+1}(t)(\mu_{i+1}\Delta t) \\ &\quad +p_i(t)(1-\lambda_i\Delta t-\mu_i\Delta t) \end{aligned}$$

Now substract $p_i(t)$ from each side, divide throughout by Δt, and let $\Delta t \to 0$. The result is

$$\begin{aligned} \dot{p}_i(t) &= \lambda_{i-1}p_{i-1}(t) \\ &\quad -(\lambda_i+\mu_i)p_i(t) \\ &\quad +\mu_{i+1}p_{i+1}(t) \end{aligned} \qquad (1)$$

for all non-boundary states, i.e., $0 < i < N$.

Apply similar argument at the boundary states, 0 and N.

At $i = 0$:

$$\dot{p}_0(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t) \qquad (2)$$

At $i = N$ when $N < \infty$:

$$\dot{p}_N(t) = \lambda_{N-1} p_{N-1}(t) - \mu_N p_N(t) \qquad (3)$$

- (1)–(3) are called *birth/death equations*
- Solution to this set of simultaneous linear differential equations gives the state probability distribution at time t
- *Analytical solution* is only possible when the coefficients λ_i and μ_i have a simple form but *numerical solution* is always possible for $N < \infty$

Equilibrium Behaviour of $\{N_t\}$

Since the transition rate $\{\lambda_i, \mu_i\}$ are time-invariant, we might expect $\{N_t\}$ to settle down to *statistical equilibrium* as t increases.

For this to happen $\{N_t\}$ must not grow without bound as $t \to \infty$, and to guarantee this we must have (roughly speaking):

$$\mu_i > \lambda_i, \quad \text{for all large } i$$

More precisely, define the *stability parameter*:

$$\boxed{S = 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2} + \cdots}$$

Then $\{N_t\}$ is a *stable process* (i.e. $\to$ equilibrium) iff

$$\boxed{S < \infty} \qquad (4)$$

Note: this condition is automatically satisfied if $N < \infty$ (and the rates λ_i, μ_i are all finite).

Now suppose that (4) holds. Then, *for each state* i,

$$p_i(t) \xrightarrow{t} \text{constant probability, } \pi_i$$

That is,

$$\dot{p}_i(t) \xrightarrow{t} 0$$

The birth/death equations then become *algebraic equations*:

$$\begin{aligned} -\lambda_0\pi_0 + \mu_1\pi_1 &= 0 \\ \lambda_0\pi_0 - (\lambda_1+\mu_1)\pi_1 + \mu_2\pi_2 &= 0 \\ \lambda_1\pi_1 - (\lambda_2+\mu_2)\pi_2 + \mu_3\pi_3 &= 0 \\ &\vdots \end{aligned}$$

These equations are called the *equilibrium equations*.

The equilibrium equations can be simplified by adding the first i equations together.

This gives

$$\boxed{\mu_i \pi_i = \lambda_{i-1} \pi_{i-1} \quad \text{for } i = 1, 2, \ldots}$$

Recursive solution, from $i = 0$, gives

$$\boxed{\pi_i = \frac{\lambda_{i-1} \lambda_{i-2} \cdots \lambda_0}{\mu_i \mu_{i-1} \cdots \mu_1} \pi_0}$$

Note: π_0 is the probability that the system is *empty*.

π_0 is given by the *normalising condition*:

$$\boxed{\sum_i \pi_i = 1}$$

That is,

$$\begin{aligned} \pi_0 &= \frac{1}{1 + \sum_{i \geq 1} \frac{\lambda_{i-1} \cdots \lambda_0}{\mu_i \cdots \mu_1}} \\ &= \boxed{1/S} \end{aligned}$$

Note: from the stability condition, we must have $\pi_0 > 0$ (and hence $\pi_i > 0$ for all i) in any stable birth/death process.

LECTURE ON
Erlang model

The most basic model for an *unbuffered* link.

Assumptions

- The total arrival stream (i.e., stream of arriving demands) is a *Poisson process* with rate λ.
- The channel holding times are *independent and exponential* random variable with mean holding time of $1/\mu$.
- The access switch gives *full availability*, i.e., each traffic source has access to *all* the channels on the link.

Now let N_t = number of busy channels on the link at time t.

Then $\{N_t\}$ is a *birth/death process*.

Birth coefficients:

Since arrival process is Poisson at rate λ,

$$P\left(N_{t+\Delta t} = i + 1 \middle| N_t = i\right) = \lambda \Delta t, \quad \text{for } i < N$$

That is,

$$\boxed{\lambda_i = \lambda, \quad \text{for } i < N}$$

Death coefficients:

Suppose i channels are busy at time t.

Each of these channels can be viewed as a Bernoulli trial with success probability:

$$P\,(\text{busy} \rightarrow \text{free}) = \mu \Delta t$$

And since the channels are acting independently, the probability that exactly k channels will become idle in $(t, t + \Delta t)$ is *binomial*:

$$\begin{aligned} &P(k \text{ channels} \rightarrow \text{idle}) \\ &= \binom{i}{k} (\mu \Delta t)^k (1 - \mu \Delta t)^{i-k} \end{aligned}$$

So

$$\begin{aligned} &P(1 \text{ channel} \rightarrow \text{idle}) \\ &= \binom{i}{1} (\mu \Delta t)(1 - \mu \Delta t)^{i-1} \\ &= i\mu \Delta t + o(\Delta t) \end{aligned}$$

where $o(\Delta t)$ = terms of order $(\Delta t)^2$ etc.

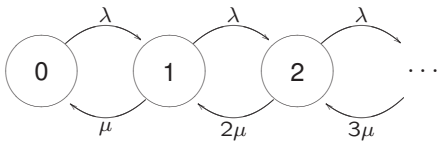
That is,

$$\boxed{\mu_i = i\mu, \quad \text{for } i > 0}$$

Probability Distribution for N_t
Case: Infinite number of channels $(N = \infty)$

In this case, all arriving traffic is carried.

State transition diagram is:



Stability parameter is now

$$\begin{aligned} S &= 1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{2\mu}\right) + \cdots \\ &= 1 + \rho + \frac{1}{2!}\rho^2 + \cdots \text{ where } \boxed{\rho = \frac{\lambda}{\mu}} \\ &= \boxed{e^{\rho}} < \infty \end{aligned}$$

So process $\{N_t\}$ is stable for all values of λ and μ.

The equilibrium state probability distribution is now given by the equilibrium equations

$$\pi_i = \left(\frac{\lambda_{i-1}}{\mu_i}\right)\pi_{i-1} = \left(\frac{\lambda}{i\mu}\right)\pi_{i-1} = \left(\frac{\rho}{i}\right)\pi_{i-1}$$

Recursive solution gives

$$\pi_i = \left(\frac{\rho^i}{i!}\right)\pi_0,\ i = 1, 2, 3, \ldots$$

where

$$\pi_0 = \left(\frac{1}{S}\right) = e^{-\rho}$$

So

$$\boxed{\pi_i = \frac{\rho^i}{i!}e^{-\rho},\ i = 0, 1, 2, \ldots}$$

The number of busy channels (π_i) has Poisson distribution.

- It follows immediately that

$$\begin{aligned} \text{mean carried traffic, } E[N_t] &= \rho \\ \text{variance of carried traffic, } Var(N_t) &= \rho \end{aligned}$$

  Traffic for which variance = mean is sometimes called *pure chance trafic*.

- We could of course have computed $E[N_t]$ from the *traffic equation*:

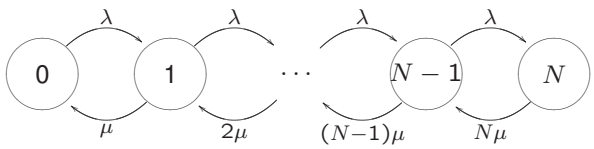
$$\begin{aligned} \text{mean carried traffic} &= \text{(mean arrival rate)} \\ &\quad \cdot\text{(mean holding time)} \\ E[N_t] &= \frac{(\lambda)}{(\mu)} \\ &= \rho \quad \text{[erlang]} \end{aligned}$$

- In practice, the link size N is *finite*: we can then interpret $\{N_t\}$ as *the traffic that would be carried if the link size N was infinite*. This amount of traffic is called the offered traffic.

Probability Distribution for N_t
Case: Finite number of channels $(N < \infty)$

In this case, the link can become *saturated* at $N_t = N$.

So the transition diagram must be truncated:



$\{N_t\}$ is guaranteed to be *stable* since N is finite.

In fact,

$$S = 1 + \rho + \frac{\rho^2}{2!} + \cdots + \frac{\rho^N}{N!}$$

As before, the solution of the equilibrium equations is

$$\begin{aligned} \pi_i &= \left(\frac{\rho^i}{i!}\right)\pi_0,\ i = 0, 1, \ldots, N \\ \pi_0 &= \frac{1}{S} \end{aligned}$$

So

$$\boxed{\pi_i = \frac{\rho^i/i!}{\sum_{j=0}^{N} \rho^j/j!}, \quad i = 0, 1, \ldots, N}$$

That is, π_i has *truncated* Poisson distribution.

Note: This is one of the standard models for traffic carried by an *unbuffered* communication link.

Erlang Loss Formula

When $N_t = N$ (i.e., all channels are busy) the link is *saturated* and no further traffic can be accepted until N_t falls below N.

$$\begin{aligned} P(\text{link saturation}) &= P(N_t = N) \\ &= \pi_N \\ &= \frac{\rho^N/N!}{S(N,\rho)} \end{aligned} \qquad (5)$$

where

$$S(N,\rho) = \sum_{j=0}^{N} \frac{\rho^j}{j!}$$

- Right hand side of (5) is called the Erlang loss formula and denoted by $E_N(\rho)$.
- Note that $E_N(\rho)$ is a function of (i) link size N and (ii) mean offered traffic ρ. Its value is very well tabulated and graphed.

Recursive Evaluation of Erlang Loss Formula

Recursive evaluation is possible by using:

$$\boxed{E_N(\rho) = \frac{\rho E_{N-1}(\rho)}{N + \rho E_{N-1}(\rho)}}$$

with

$$\boxed{E_0(\rho) = 1}$$

Time vs *Call* Congestion

- π_N is the probability that, if we inspect the link at an arbitrary time, we will find it saturated.

  By the ergodic property, this is also *the proportion of time for which the link is saturated*.

  So, π_N is called the time congestion of the link, denoted in general by B_T.

- In practice, we normally want to know *the fraction of arriving demands which will meet saturation (and hence be blocked)*

  This is called the call congestion, denoted by B_C.

Now when the arrival stream is a *Poisson* stream, the arrival rate (= birth rate, λ_i) is independent of the link state i.

It follows that the arrivals see an *unbiassed sample* of the link state distribution.

In particular,

$$P(\text{call meets saturation}) = P(\text{system is saturated})$$

That is,

$$\boxed{B_C = \pi_N = B_T}$$

for a Poisson arrival stream.

Thus, for the Erlang model,

$$\text{call congestion, } B_C = E_N(\rho)$$

This is the probability that a call is *blocked*.

Offered, Carried and Lost Traffics

From the traffic equation,

$$\text{mean offered traffic, } \rho_o = \lambda\left(\frac{1}{\mu}\right) = \boxed{\rho}$$

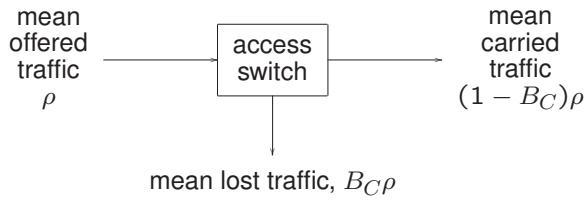
Also,

$$\text{arrival rate for } \textit{accepted} \text{ demands} = (1 - B_C)\lambda$$

So, again from the traffic equation,

$$\begin{aligned}\text{mean carried traffic, } \rho_c &= (1 - B_C)\lambda\left(\frac{1}{\mu}\right) \\ &= \boxed{(1 - B_C)\rho}\end{aligned}$$

Finally, mean lost traffic, $\rho_l = \boxed{B_C\rho}$



$E[N_t]$ and $Var(N_t)$

We can of course compute the mean and variance of the carried traffic as $E[N_t]$ and $Var(N_t)$.

In this case, we get

$$\begin{aligned}E[N_t] &= (1 - B_C)\rho \\ Var(N_t) &= E[N_t] - \rho B_C \underbrace{(N - E[N_t])}_{\substack{\text{mean no. of} \\ \textit{idle} \text{ channels}}}\end{aligned}$$

Note that

$$\boxed{Var(N_t) < E[N_t]}$$

That is, the carried traffic is *smoother* than the pure chance traffic.

Mean Channel Occupancy

The traffic $\{N_t\}$ is carried on N channels.

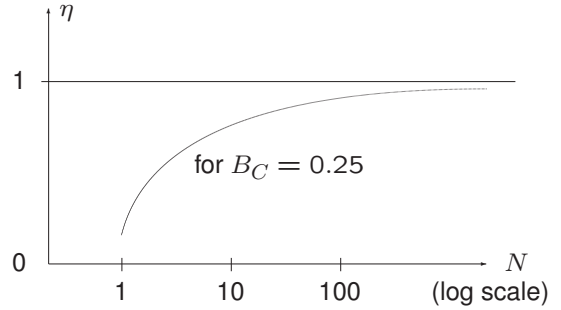
Thus, the *mean channel occupancy* (η) is obtainable from

$$\begin{aligned}\eta &= \text{mean carried traffic / channel} \\ &= \boxed{\frac{(1 - B_C)\rho}{N}}\end{aligned}$$

Note that, necessarily,

$$\boxed{0 < \eta < 1}$$

For fixed B_C, we would expect that $\eta \to 1$ as $N \to \infty$



Hunting Procedure

The traffic distribution π_i for the Erlang model does *not* depend on the *channel selection procedure* (also called the *hunting procedure*)—the method of searching for an idle channel for a new call arrival.

There are two main methods:

Sequential hunting:

Channels are tested in numerical order $(1, 2, \ldots)$ until an idle channel is found. This method gives *non-uniform* channel loading—the lowest numbered channels carry the most traffic.

Random hunting:

Channels are tested in random order. This method gives *uniform* channel loading—each channel carries the same average traffic.

If

$$\eta_i = \text{mean occupancy of } i^{\text{th}} \text{ channel}$$

then for sequential hunting

$$\eta_1 > \eta_2 > \cdots > \eta_N$$

and for random hunting

$$\eta_1 = \eta_2 = \cdots = \eta_N$$

We can calculate the *overall mean channel occupancy* (η) from

$$\eta = \frac{1}{N} \sum_i \eta_i$$

Important fact: π_i remains a truncated Poisson distribution even when the holding times are *not* exponential. This is called the *insensitivity property*, which makes the Erlang model very useful in practice.

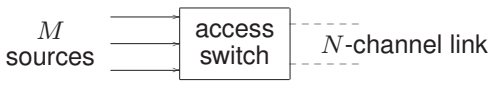
LECTURE ON
Engset Model

The Erlang model assumes that the total input stream is Poisson. This is only valid when the arrival rate is independent of the the link state, N_t.

In practice, we can usually assume this when the number of sources is much greater than the number of channels.

In this lecture, we want to consider *limited-source* traffic: called the *Engset* model

Suppose that we have M sources (acting independently) and a link with N channels.



First note that, since there is no buffering,

$$\boxed{j \textit{ channels} \text{ busy} \leftrightarrow j \textit{ sources} \text{ busy}}$$

So the total arrival rate to the link will *fall* as N_t *increases*—and the total offered traffic will be less than that predicted by the Erlang model.

Engset Model's Assumptions

- Each *idle* source is a *Poisson source*:

  This means that

  $$P(\text{source generates new demand in } (t, t+\Delta t) | \text{source is idle}) = \lambda \Delta t$$

- Channel holding times are *exponential* with mean $(1/\mu)$.

- Full-availability access (complete sharing).

Derivations

Let N_t = number of busy channels on the link at time t.

Then $\{N_t\}$ is a *birth/death process* with upper boundary at

$$i_{\max} = \begin{cases} M, & \text{when } N \geq M \\ N, & \text{when } M \geq N \end{cases}$$

Birth coefficients:

Since the number of *idle* sources in state i is $(M-i)$, we have

$$\begin{aligned} \lambda_i &= (M-i)\lambda, \quad i < i_{\max} \\ &= (1 - i/M)M\lambda \\ &= (1 - i/M)\lambda_o \end{aligned}$$

where we denote the *calling rate when all sources are idle* as

$$\boxed{\lambda_o = M\lambda}$$

Death coefficients:

As in Erlang model,

$$\mu_i = i\mu, \quad \text{for } i > 0$$

Thus λ_i *decreases* linearly as i increases and μ_i *increases* linearly as i increases.

In the following, we consider two cases:

- $N \geq M$: no congestion—all offered traffic is carried.

- $M > N$: congestion will occur.

Probability Distribution for N_t
Case: $(N \geq M)$

In this case, $i_{\max} = M$.

System is *finite*, hence equilibrium is guaranteed.

As usual,

$$\begin{aligned} \pi_i &= \left(\frac{\lambda_{i-1}}{\mu_i}\right)\pi_{i-1} \\ &= \left(\frac{(M-(i-1))\lambda}{i\mu}\right)\pi_{i-1} \end{aligned}$$

Recursive solution gives

$$\pi_i = \binom{M}{i}\alpha^i \pi_0, \; i = 1, 2, \ldots, M$$

where

$$\boxed{\alpha = (\lambda/\mu)}$$

is the *offered traffic / idle source*.

Normalisation gives

$$\pi_0 = \left[\sum_{j=0}^{M}\binom{M}{j}\alpha^j\right]^{-1}$$

Now introduce the parameter

$$\boxed{p = \frac{\alpha}{1+\alpha}}$$

from which

$$\alpha = \frac{p}{1-p}$$

Then, multiplying numerator and denominator of π_i by the factor $(1-p)^M$, we get

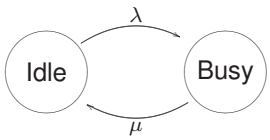
$$\pi_i = \frac{\binom{M}{i} p^i (1-p)^{M-i}}{\sum_{j=0}^{M} \binom{M}{j} p^j (1-p)^{M-j}}, i = 0, 1, \ldots, M$$

From binomial distribution, the denominator $= 1$. So

$$\boxed{(\pi_i) \text{ is binomial } (M, p)}$$

This is to be expected, since each source acts independently (no interaction due to system congestion).

Source Model



In equilibrium,

$$\begin{aligned} P(\text{source is busy}) &= \left(\frac{\lambda}{\lambda+\mu}\right) \\ &= \left(\frac{\alpha}{1+\alpha}\right) = p \end{aligned}$$

Thus, we have M non-interacting sources, each with busy probability $= p$.

$\Rightarrow$ Number of busy *sources* is binomial(M, p). $\Rightarrow$ Number of busy *channels* is binomial(M, p).

- No congestion:

  offered traffic = carried traffic

  That is,

  $$\rho_o = \rho_c = E[N_t] = \boxed{Mp} = \boxed{\frac{M\alpha}{1+\alpha}}$$

  And $E[N_t]$ = mean number of busy *sources*

- From binomial distribution,

  $$\begin{aligned} E[N_t] &= Mp \\ Var(N_t) &= Mp(1-p) \end{aligned}$$

  Thus, we have

  $$\boxed{\text{variance} < \text{mean}}$$

  Traffic with this property is called the *smooth traffic*.

Probability Distribution for N_t
Case: $(M > N)$

In this case, $i_{\max} = N$.

Same analysis as before gives

$$\pi_i = \frac{\binom{M}{i} p^i (1-p)^{M-i}}{\sum_{j=0}^{N} \binom{M}{j} p^j (1-p)^{M-j}}, i = 0, 1, \ldots, N$$

Note that the denominator now < 1 since $N < M$.

State distribution π_i is now a *truncated binomial distribution*, also called the *Engset distribution*.

Engset Loss Formula

Since $N < M$, congestion will now occur from time to time.

Time congestion is now given by

$$B_T = P(N_t = N) = \pi_N(M, p) = e_N$$

This is called the Engset loss formula.

For fixed (M, p), recursive evaluation is possible by using:

$$\boxed{e_N = \frac{(M - N + 1)\alpha e_{N-1}}{N + (M - N + 1)\alpha e_{N-1}}}$$

with

$$\boxed{e_0 = 1}$$

Call congestion is now *less* than B_T because the arrival rate falls as the link gets busier.

Analysis

By symmetry, all sources see a statistically identical arrival pattern. Any given *idle* source (say source number j) will see the link occupancy pattern *generated by the remaining* $(M - 1)$ *sources*.

Thus, an idle source can be regarded as an external observer of a system driven by $(M - 1)$ sources, and the state distribution seen by the idle source will be $(\widehat{\pi}_i)$, where

$$\boxed{\widehat{\pi}_i(M, p) = \pi_i(M - 1, p), \quad i = 0, 1, \ldots, N}$$

So an arriving demand will find the link in a state i with probability $\widehat{\pi}_i(M, p) = \pi_i(M - 1, p)$. Then, call congestion is given by

$$\boxed{B_C = \pi_N(M - 1, p)}$$

and hence $B_C < B_T$, as expected in this case.

Mean Carried Traffic

For $N \geq M$, we know that

$$\rho_c = E[N_t] = \boxed{Mp}$$

This is the traffic that would be carried if no congestion—called the *intended offered traffic*.

But when there is congestion (i.e. $N < M$), some of this offered load is not carried. In fact, for a fraction π_N of the total time, the link is *saturated*, and so the offered load (which, during saturation, is $(M - N)p$) is rejected:

$$E[N_t] = Mp - \pi_N(M, p)(M - N)p$$

That is,

$$\boxed{\rho_c = E[N_t] = Mp\left[1 - \left(1 - \frac{N}{M}\right)\pi_N\right]}$$

An alternative expression for ρ_c is [see Problem Sheet 1]:

$$\rho_c = \left[\frac{(1 - B_C)\alpha M}{1 + (1 - B_C)\alpha}\right] = Mp\left[\frac{1 - B_C}{1 - B_C p}\right]$$

And since

$$\rho_c = (1 - B_C)\rho_o$$

$$\rho_o = \left[\frac{\alpha M}{1 + (1 - B_C)\alpha}\right] = \left[\frac{Mp}{1 - B_C\rho}\right]$$

- The actual offered traffic is *greater* than the intended offered traffic.
- The actual offered traffic depends on B_C.
- Why?

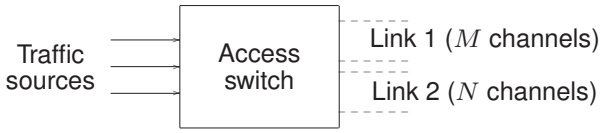
- The variance $\boxed{Var(N_t)}$ is awkward to evaluate.

- *Important*: When $M \gg N$, the arrival rate in state i is $\lambda_i = (1 - i/M)\lambda_o \simeq \lambda_o$ for all i and so we are back to the *Erlang model*.

Important fact: The Engset distribution still applies even when (a) the holding times are non-exponential and/or (b) the source idle times are non-exponential.

LECTURE ON
Overflow Traffic

Consider the following arrangement:



Link 1 is the *first-choice* link. Link 2 is the *second-choice* or *overflow* link.

Traffic is offered to Link 2 only when Link 1 is saturated.

Example: automatic alternative routing

The traffic offered to Link 2 (called *overflow traffic*) comes *in bursts*—no overflow traffic arrives when Link 1 is not saturated.

This sort of traffic has a high variance because of its bursty characteristic.

Birth/death Overflow Traffic Model

A simple birth/death model for such traffic is achived by making the arrival rate to the overflow link *increases* as the overflow link gets busier.

Assumptions

- Overflow arrival stream is a variable-rate Poisson stream in which the arrival rate when i overflow channels are busy is $(K + i)\lambda$, where K, λ are parameters.

- Channel holding times are exponential with mean $(1/\mu)$.

- Full-availability access to overflow link.

Define

$N_t =$ number of busy overflow channels at time t

$\{N_t\}$ is a birth/death process, with

Birth coefficients:

$$\lambda_i = (K + i)\lambda, \quad i \geq 0$$

Death coefficients:

$$\mu_i = i\mu, \quad i > 0$$

We consider two cases:

- $N = \infty$

- $N =$ finite

Case: Infinite overflow link ($N = \infty$)

The traffic will settle to statistical equilibrium if the stability parameter S is finite.

Now

$$S = 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2} + \cdots$$

That is,

$$\begin{aligned} S &= 1 + K\left(\frac{\lambda}{\mu}\right) + \frac{K(K+1)}{2!}\left(\frac{\lambda}{\mu}\right)^2 + \cdots \\ &= 1 + K\alpha + \frac{K(K+1)}{2!}\alpha^2 + \cdots \quad \text{with } \alpha = \frac{\lambda}{\mu} \\ &< \infty \quad \text{iff } \boxed{\alpha < 1} \end{aligned}$$

in which case

$$\boxed{S = (1-\alpha)^{-K}}$$

With $\alpha < 1$, the equilibrium distribution of $\{N_t\}$ is solution of

$$\pi_i = \left(\frac{\lambda_{i-1}}{\mu_i}\right)\pi_{i-1} = \left(\frac{K+i-1}{i}\right)\alpha\pi_{i-1}$$

from which, by recursion,

$$\pi_i = \binom{K+i-1}{i}\alpha^i\pi_0$$

and

$$\pi_0 = S^{-1} = (1-\alpha)^K$$

so that

$$\boxed{\pi_i = \binom{K+i-1}{i}\alpha^i(1-\alpha)^K, \quad i = 0, 1, \ldots}$$

This is *negative binomial (or Pascal) distribution*.

- From the distribution π_i, we can show that

$$\begin{aligned} E[N_t] &= \boxed{\frac{K\alpha}{1-\alpha}} = \rho_c = \rho_o \\ Var(N_t) &= \boxed{\frac{K\alpha}{(1-\alpha)^2}} \end{aligned}$$

  Hence (since $\alpha < 1$)

  $$\boxed{\text{variance} > \text{mean}}$$

  and the traffic is called *rough traffic*.

- The parameters (K, α) have no obvious system interpretation.

  In practice, they are chosen to match the values of $E[N_t]$ and $Var(N_t)$ to the required values.

Case: Finite overflow link (N finite)

The same analysis as before shows that the state distribution (π_i) is now a *truncated Pascal distribution*.

Congestion is now possible on the overflow link and it can be shown that

$$\boxed{B_T = \pi_N(K, \alpha)}$$

and

$$\boxed{B_C = \pi_N(K+1, \alpha)}$$

So now

$$\boxed{B_C > B_T}$$

This is as expected since the arrival rate *increases* with i.

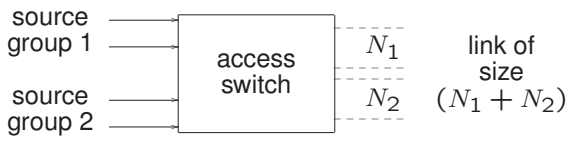
For this model, the mean carried traffic is

$$\rho_c = \frac{(1-B_C)\alpha K}{1-(1-B_C)\alpha}$$

LECTURE ON
Restricted Availability

In practice, the access switch may not give full-availability access to the link—then an arriving demand may be blocked even when the link is not saturated.

Example: grouped sources



In *birth/death models* for traffic on a link, the effect of restricted availability can be modelled by means of *loss factors* , β_i, defined as follows:

$$\beta_i = P(\text{arrival is blocked} | N_t = i), \quad i = 0, 1, \ldots$$

By definition, we will always have

$$\beta_N = 1, \quad \text{(all channels are busy)}$$

For full-availability access

$$\beta_i = 0, \quad i < N$$

For restricted-availability access

$$0 < \beta_i < 1, \quad \text{for at least one } i < N$$

Now consider a birth/death model with full-availability access, and with birth coefficients λ_i. The corresponding restricted-availability version of the system is a birth/death model with *new* birth coefficients:

$$\lambda_i' = (1 - \beta_i)\lambda_i$$

since

$$P(\text{ arrival in } (t, t + \Delta t) \text{ and not blocked } | N_t = i) = (\lambda_i \Delta t)(1 - \beta_i)$$

Using the new birth coefficients in the equilibrium equations gives a *modified equilibrium distribution* (π_i').

The overall blocking probability B_C is then computed from the call congestion:

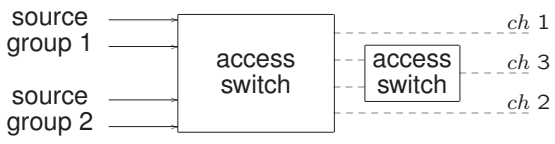
$$B_C = \sum_{i=0}^{N} \pi_i' \left(\frac{\lambda_i}{\overline{\lambda}} \right) \beta_i$$

where $\overline{\lambda}$ is the overall mean arrival rate.

Determination of the β_i

In practice, they can be estimated by experiments for a given access configuration. For certain access schemes, the values of β_i can be computed theoretically.

Example:



By inspection, assuming *balanced traffic*,

$$\begin{aligned} \beta_3 &= 1 \\ \beta_2 &= 1/3 \\ \beta_1 &= 0 \end{aligned}$$

But, in general, very complex combinatorial calculations are needed. So the following *approximation* is often used:

- Compute $\boxed{\beta_{N-1}}$ (usually the easiest).
- Then set

$$\boxed{\tilde{\beta}_i = (\beta_{N-1})^{N-i}, \quad i = 1, 2, \ldots, N-2}$$

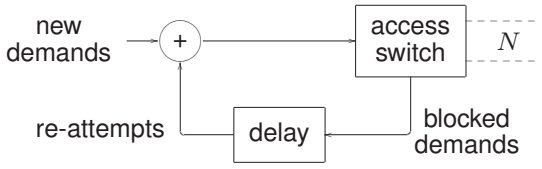
The resulting model is called a *geometric group* model for the given access scheme—gives reasonable approximation to (π_i') in most cases.

LECTURE ON
Repeated Attempts

The birth/death models considered so far ignore the effect of *re-attempts*, i.e., the resubmission of blocked arrivals at a later time.

This effect is important under heavy-traffic conditions, but is difficult to analyse, since re-attempts normally occur soon after the initial blocking (e.g., in automatic redialling systems in telephony).

This results in a *non-Poisson* arrival stream.



Simple Re-attempt Model

Assumptions

- A blocked demand is re-submitted with probability p.
- Re-submission occurs after a long interval.

With these assumptions, the re-attemps can be treated as *additional new demands* to the system. That is, the feedback effect simply increases the offered load.

If $B_C =$ call congestion (with re-attempts) then

$$P(\text{demand is blocked } \textit{and} \text{ re-submited}) = B_C p$$

So

$$P(\text{demand is submited } j \text{ times in total}) = (B_C p)^{j-1}(1 - B_C p)$$

This is *geometric distribution*.

From geometric distribution,

$$\text{mean number of attempts / demand} = \frac{1}{1 - B_C p} = \bar{N}$$

Thus, if

$$\text{offered traffic } \textit{without} \text{ re-attempts} = \rho_o$$

then

$$\text{total offered traffic } \textit{including} \text{ re-attempts} = \bar{N}\rho_o = \frac{\rho_o}{1 - B_C p}$$

This depends on B_C, which in turn depends on the total offered traffic.

- The call congestion B_C computed from this model is the probability that an arrival (whether a new demand or a re-attempt) wil be blocked.
- The probability that a new demand will *eventually* be accepted is *not* $(1 - B_C)$ but

$$(1 - B_C)/(1 - B_C p)$$

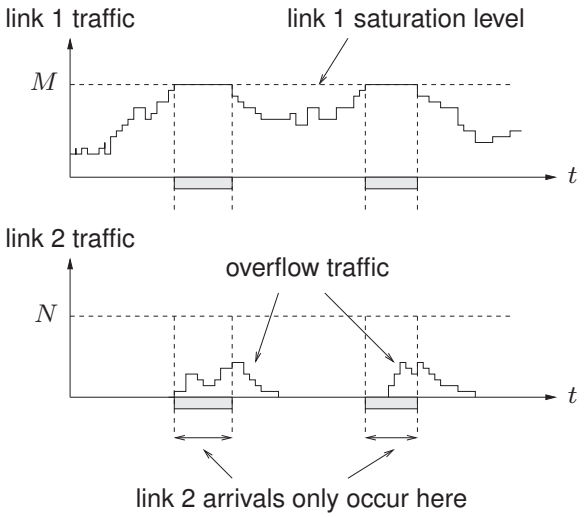
This is sum of acceptance probabilities for 1^{st}, 2^{nd}, ... attempts.

LECTURE ON
Models for Overflow Traffic

The basic arrangement in which overflow traffic occurs is the following:

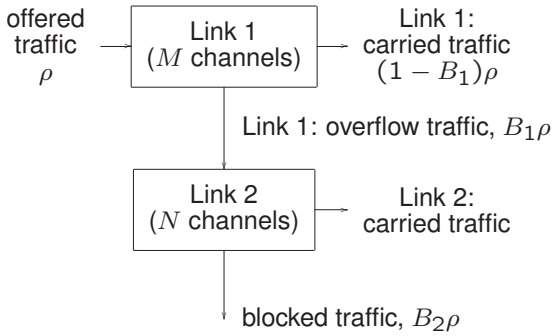
Link 1 is the *first-choice* link. Link 2 is the *second-choice* or *overflow* link. Traffic is offered to Link 2 only when Link 1 is saturated. That is, the *blocked traffic* from Link 1 is the *offered traffic* to Link 2. This traffic is called *overflow traffic* —it comes in bursts and hence has a large variance.

Example of Time History



The statistical characteristics of the overflow traffic depend partly on the nature of the traffic offered to Link 1. However, there are certain basic results which hold for all forms of offered traffic.

An alternative picture of the basic set-up is as follows:



where

$$\begin{aligned} B_1 &= \text{Link 1 blocking probability} \\ &= P(\text{arrival meets saturation on Link 1}) \\ B_2 &= \text{Link 2 blocking probability} \\ &= P(\text{arrival meets saturation on both Links 1 and 2}) \end{aligned}$$

Note that B_2 is not the probability that Link 2 is saturated—nor is it the call congestion for Link 2.

If λ is the mean arrival rate to Link 1 and ρ is the offered traffic level, then we have:

- Link 1 call congestion $= \boxed{B_1}$
- Link 2 call congestion

$$\begin{aligned} &= \frac{\text{arrival rate to Link 2 for blocked arrivals}}{\text{total arrival rate to Link 2}} \\ &= \frac{\lambda B_2}{\lambda B_1} = \boxed{\frac{B_2}{B_1}} \end{aligned}$$

- Link 1 mean carried traffic $= \boxed{(1-B_1)\rho}$
- Link 2 mean carried traffic

$$\begin{aligned} &= (B_1\rho) - (B_2\rho) \\ &= \boxed{(B_1 - B_2)\rho} \end{aligned}$$

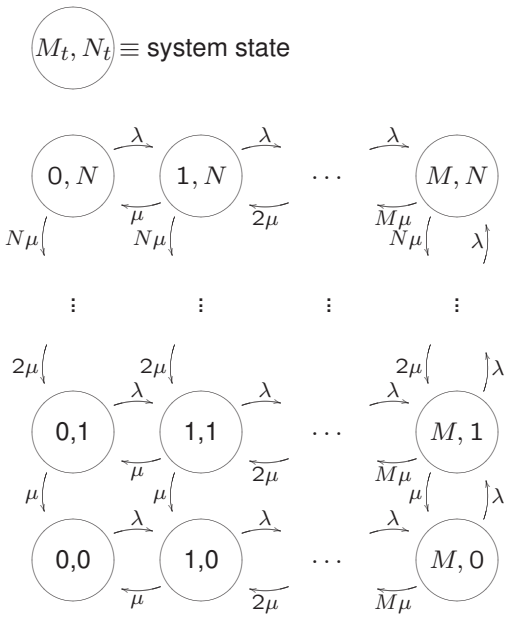
In order to compute the blocking probabilities, B_1 and B_2, the characteristics of the original offered traffic must be specified. The simplest situation is when the offered traffic is *pure chance traffic* (i.e. Poisson arrivals and exponential holding times), in which case the given 2-link set-up can be described by a *2-dimensional birth/death process*.

2-D Birth/death Model for Overflow Traffic

Assume that the offered traffic is pure chance traffic with parameters (λ, μ). Let $\{(M_t, N_t)\}$ is a 2-D birth/death process with state space

$$\boxed{E = \{(i,j) | 0 \leq i \leq M, 0 \leq j \leq N\}}$$

State transition diagram



Note: Clearly *local balance* does not hold. As usual, there are two cases to consider (i.e. infinite/finite overflow link).

Case: Infinite overflow link $(N = \infty)$ "Kosten model"

By using probability generating functions, it is possible to obtain expressions for the equilibrium state probability distribution (π_{ij}) and for the various moments of the distribution. The key results are that the *mean* and *variance* of the overflow traffic are given by:

$$m = E[N_t] = \rho B_1 = \rho E_M(\rho)$$

$$v = Var(N_t) = m\left(1 - m + \frac{\rho}{M + m + 1 - \rho}\right)$$

It is easily verified that $v > m$, i.e. as expected, the overflow traffic is *rough* traffic. In practice, the overflow link is finite (N finite), in which case we can regard the above as a model for *offered* overflow traffic. It is sometimes called the *Kosten model* for such traffic.

Case: Finite overflow link ($N < \infty$) "Brockmeyer model"

The key results in this case are:

$$\begin{aligned} m &= E[N_t] = \rho(B_1 - B_2) \\ &= \rho[E_M(\rho) - E_{M+N}(\rho)] \\ v &= Var(N_t) \\ &= m(1-m) \\ &\quad +\rho\left(\frac{\rho(B_1 - B_2)}{M+1-\rho(1-B_1)} - NB_2\right) \end{aligned}$$

The above formulae reduce to the Kosten expressions when $B_2 = 0$. Also, we already know that

$$\text{Link 2 call congestion} = \frac{B_2}{B_1} = \boxed{\frac{E_{M+N}(\rho)}{E_M(\rho)}}$$

This model, which gives us the mean and variance of the carried traffic on a finite overflow link, is sometimes called the *Brockmeyer model*.

Note

- The blocked traffic in this system has a mean value of

  $\rho B_2 = \rho E_{M+N}(\rho)$.

- In some networks, this blocked overflow traffic is allowed to overflow on to a *second overflow link*. We can use a Kosten model with a Link 1 size of $(M + N)$ to find the variance of this second overflow.

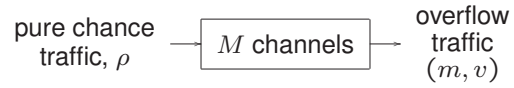
Moment-matching Models

The Kosten/Brockmeyer traffic models are unnecessarily detailed for many applications. Frequently, all we require is a traffic model which generates traffic with a given mean m and given variance $v > m$. There are 3 widely-used *approximate models* for offered overflow traffic, all based on the idea of matching moments:

- Birth/death process with linearly-increasing birth rate (which has been seen earlier)
- Equivalent random traffic model
- Interrupted Poisson process model

Equivalent Random Traffic (ERT) Model

The Kosten model gives formulae for the mean m and variance v of the overflow traffic produced when ρ erlangs of *pure chance* traffic is offered to an M-channel link, i.e. an Erlang model (M, ρ). Diagrammatically:



In fact, *for any rough traffic process*, with given (m, v) and $v > m$, there is a unique pair of Erlang-model parameters $(\widehat{M}, \widehat{\rho})$ such that the following *fictitious model* gives the required values (m, v):

$< \cdots$ fictitious Erlang model $\cdots >$

This is the so-called *equivalent random traffic model* for the given rough traffic—and $\widehat{\rho}$ is the "equivalent random traffic".

- The ERT model is a "moment-matching" model. The parameters $(\widehat{M}, \widehat{\rho})$ are chosen to give the required mean m and variance v. The match will not normally be exact (since $\widehat{M}$ must be integer).

- If the given rough traffic is offered to an N-channel link, we can now easily compute the *call congestion* as:

$$\boxed{B_C = \frac{E_{\widehat{M}+N}(\widehat{\rho})}{E_{\widehat{M}}(\widehat{\rho})}}$$

  Think: WHY?

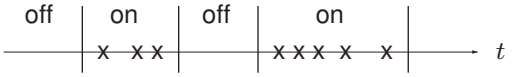
Interrupted Poisson Process (IPP) Model

In the Kosten model, the arrival process of the overflow traffic consists of a *Poisson arrival stream* which is

- ON: when the first-choice link is saturated

- OFF: when the first-choice link is not saturated

That is, the Poisson stream is switched ON and OFF by the state of the first-choice link:



- The ON periods are just the periods when $M_t = M$ on the first-choice link; we know that these periods are *exponential* random variable with mean $1/(M\mu)$.

- The OFF periods are, however, *not* exponential random variable. ← Why?

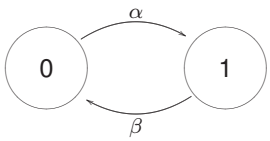
Suppose we now pretend that the OFF periods are exponential. This will give a good approximation to the true overflow arrival process. With this approximation, the ON/OFF switching process can be represented by a *2-state Markov process* $\{Y_t\}$: Let

$$Y_t = \begin{cases} 0, & \text{arrival stream is OFF} \\ 1, & \text{arrival stream is ON} \end{cases}$$

and suppose that

$$\begin{aligned} \text{mean OFF period} &= 1/\alpha \\ \text{mean ON period} &= 1/\beta \end{aligned}$$

Then $\{Y_t\}$ has state transition diagram:



So, at equilibrium,

$$\begin{aligned} \pi_{\text{off}} = \pi_0 &= \frac{\beta}{\alpha+\beta} \\ \pi_{\text{on}} = \pi_1 &= \frac{\alpha}{\alpha+\beta} \end{aligned}$$

Hence, mean overflow arrival rate is

$$\boxed{\bar{\lambda} = \left(\frac{\alpha}{\alpha+\beta}\right)\lambda}$$

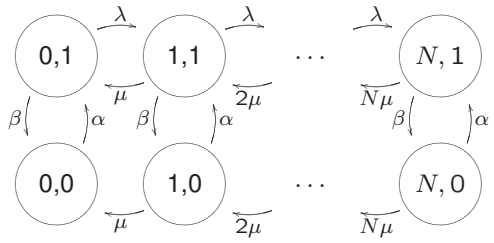
and mean offered overflow traffic is

$$\boxed{\bar{\rho} = \left(\frac{\alpha}{\alpha+\beta}\right)\rho}$$

where $\rho = \lambda/\mu$.

Now suppose that this traffic is offered to an overflow link of size N, and let N_t be the number of busy channels on this overflow link. Then the *joint process* $\{N_t, Y_t\}$ is a 2-dimensional birth/death process called an *interrupted Poisson process*.

State transition diagram



Clearly, *local balance* does not hold—and the equilibrium does not have a product form. As usual, there are 2 cases (infinite/finite overflow link).

Case: Infinite overflow link $(N = \infty)$
"IPP model"

Solution by probability generating functions leads to:

$$\mu_{(k)} = \left[\frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+k-1)}\right]\rho^k$$

for $k = 1, 2, \ldots$,

where

$$\mu_{(k)} = E[N_t(N_t-1)\cdots(N_t-k+1)]$$

is the k^{th} factorial moment. In particular, we find that

$$m = E[N_t] = \left(\frac{\alpha}{\alpha+\beta}\right)\rho$$
$$v = Var(N_t) = m\left[1 + \frac{(\beta/\alpha)m}{\alpha+\beta+1}\right]$$

Note that $v > m$, as required for overflow traffic.

- The IPP model is a "moment-matching" approximation to the proper (Kosten) model for offered overflow traffic. There are *3 free parameters* (λ, α, β), since we do not have to make λ equal to the "true" arrival rate in the Kosten model.

- In theory, we can choose (λ, α, β) so as to match *the first 3 moments* of any given rough traffic. But in practice, we may only know (m, v) and so there is a spare degree of freedom.

Case: Finite overflow link $(N < \infty)$
"IPP model"

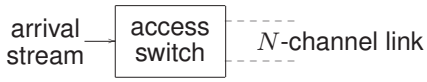
In practice, N is finite and there is a possibility of blocking on the overflow link. The most direct way of finding this blocking probability is to *solve numerically the global balance equations* of the IPP state transition diagram for the equilibrium distribution (π_{ij}) and then:

call congestion, B_C
$= P(N_t = N|\text{arrival stream is ON})$
$= P(N_t = N|Y_t = 1)$

LECTURE ON
Buffered Systems

When a stream of demands is given unbuffered access to a communication link, blocked demands (i.e. demands arriving to find the link saturated) must be rejected by the system.

Such system is called *loss-type* system.



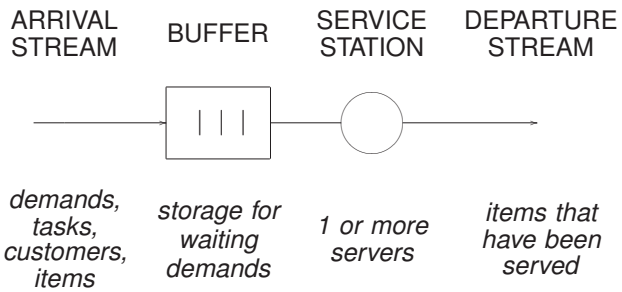
If a *buffer* is introduced between the arriving stream of demands and the link, blocked demands need no longer be rejected (unless the buffer is full) but instead can be *held* (i.e. stored) until they can be transmitted over the link. Such blocked demands are therefore delayed by the amount of time for which they must wait in the buffer.

Such system is called *delay-type* system.

The analysis of the behaviour of systems like this is the main objective of *queuing theory*, which is basically the study of *buffered-access service systems, i.e., queuing sytems*.

General Queuing System

Representation



Specification

- Arrival pattern
  - Interarrival-time distribution
  - Batch-size distribution

- Service pattern
  - Number of servers
  - Service-time distbution
  - Batch-size distribution
- Buffer arrangements
  - Number of buffers and how they are connected
  - Buffer sizes
- Queue discipline
  - Rules for selecting next item for service (including pre-arrigned priority classes)
  - Rules for premature exit (time-outs, etc.)

Kendall's Notation

The following shorthand notation is almost universal:

$$\boxed{A/S/K/N/QD}$$

where

A	$=$	Type of arrival-time distribution
S	$=$	Type of service-time distribution
K	$=$	Number of servers
N	$=$	System capacity (number of items in system when saturated)
QD	$=$	Queue discipline

Examples

- K-server queue with Poisson arrivals, exponential service times, infinite buffer and service in arrival order:

$$\boxed{M/M/K/\infty/\text{FIFO}}$$

- Erlang loss system:

$$\boxed{M/M/K/K}$$

Performance Analysis

As with unbuffered systems, we are concerned with two questions:

Stability

Under what conditions does the system settle down to statistical equilibrium?

1. Finite-buffer systems are always stable—we get statistical equilibrium with

   $$\boxed{\text{offered traffic = carried traffic + lost traffic}}$$

   where the lost traffic is due to buffer saturation.

2. Infinite-buffer systems have no lost traffic—*all* of offered traffic must be either carried or stored. So we get stability if and only if

   $$\boxed{\text{total arrival rate} < \text{max. available service rate}}$$

Equilibrium Performance

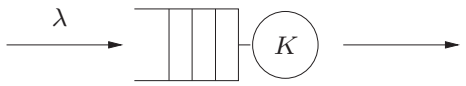
In a stable queuing system, the main performance measures of interest are (ideally) the distributions or (at least) the means and variances of:

1. Queue length
2. Waiting time

LECTURE ON $M/M/K$ Queuing System

Defined by:

- Poisson arrival stream (assume mean rate $= \lambda$)
- Exponential service times (assume mean service time $= 1/\mu$)
- K servers
- Infinite buffer
- FIFO queue discipline
- Full availability access (this is implied by convention, except explicitly stated otherwise)



Let

Q_t = number of items in buffer at time t
N_t = number of items in system at time t

Then

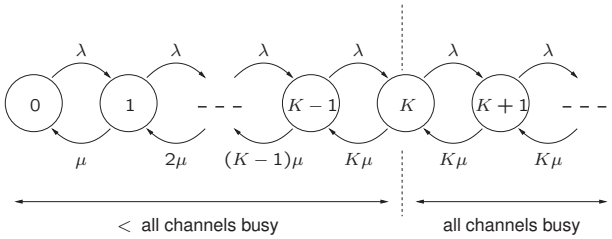
$$Q_t = \begin{cases} N_t - K, & N_t \geq K \\ 0, & N_t \leq K \end{cases}$$

and so the stochastic process $\{Q_t\}$ and $\{N_t\}$ are related.

However,

$\{Q_t\}$ is *not* a Markov process
$\{N_t\}$ is a Markov process

$\{N_t\}$ is a birth/death process with state diagram:



Stability

For this process, the stability parameter is

$$S = 1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{2\mu}\right) + \cdots + \frac{1}{K!}\left(\frac{\lambda}{\mu}\right)^K \left[1 + \left(\frac{\lambda}{K\mu}\right) + \left(\frac{\lambda}{K\mu}\right)^2 + \cdots\right]$$

Now set

$$A = \left(\frac{\lambda}{\mu}\right) \Leftarrow \boxed{\text{offered traffic}}$$
$$\rho = \left(\frac{A}{K}\right) \Leftarrow \boxed{\text{offered traffic / server}}$$

Then

$$S = 1 + A + \frac{A^2}{2!} + \cdots + \frac{A^K}{K!}\underbrace{\left(1 + \rho + \rho^2 + \cdots\right)}_{\rightarrow \left(\frac{1}{1-\rho}\right) \text{ iff } \rho < 1}$$

And so

$$\boxed{S < \infty \text{ iff } \rho < 1}$$

That is, the system settles down to equilibrium if and only if the offered traffic / server is less than 1 erlang.

Equilibrium Distribution

Local balance holds and the balance equations are

$$\pi_i = \left(\frac{\lambda}{i\mu}\right)\pi_{i-1} = \left(\frac{A}{i}\right)\pi_{i-1}, \quad \text{for } i \leq K$$

$$\pi_i = \left(\frac{\lambda}{K\mu}\right)\pi_{i-1} = \rho\pi_{i-1}, \quad \text{for } i \geq K$$

Recursive solution then gives

$$\pi_i = \begin{cases} \left(\frac{A^i}{i!}\right)\pi_0, & \text{if } i \leq K \\ \left(\frac{A^K}{K!}\right)\rho^{i-K}\pi_0, & \text{if } i \geq K \end{cases} \quad (1)$$

Note that

$$\pi_K = \left(\frac{A^K}{K!}\right)\pi_0$$

as in unbuffered Erlang model.

Also, for $i \geq K$, we can write $i = K + j$ (with $j \geq 0$) and then

$$\pi_{K+j} = \left(\frac{A^K}{K!}\right)\rho^j\pi_0 = \rho^j\pi_K \qquad (2)$$

Normalisation gives, as usual, $\pi_0 = \frac{1}{S}$ where

$$S = \sum_{i=0}^{K}\left(\frac{A^i}{i!}\right) + \left(\frac{A^K}{K!}\right)\left(\frac{\rho}{1-\rho}\right)$$

and so

$$\boxed{\pi_0 = \frac{1}{\left(A^K/\,K!\right)}\left[\frac{(1-\rho)E_K(A)}{(1-\rho)+\rho E_K(A)}\right]}$$

where $E_K(A)$ is the *Erlang loss* for A erlangs.

Note The equilibrium state distribution (π_i) is *independent of the queue discipline in operation*.

Erlang Delay Formula

Let W = waiting time of a typical arrival (= time between entry into and exit from the buffer).

Then

$$P(\text{delay}) = P(W > 0) = P(N_t \geq K)$$

which is the probability that all K servers are busy.

Furthermore,

$$\begin{aligned} P(N_t \geq K) &= \sum_{i=K}^{\infty}\pi_i \\ &= \sum_{j=0}^{\infty}\pi_{K+j} \\ &= \sum_{j=0}^{\infty}\pi_K\rho^j \quad \text{from (2)} \end{aligned}$$

So

$$P(\text{delay}) = \frac{\pi_K}{1-\rho} = \frac{\left(A^K/\,K!\right)\pi_0}{1-\rho}$$

That is, using (2),

$$\boxed{P(\text{delay}) = \frac{E_K(A)}{(1-\rho)+\rho E_K(A)}}$$

This is called *Erlang delay formula*, $D_K(A)$.

Note

- $P(\text{delay})$ is *independent of queue discipline*.
- Note that for a given offered traffic level A,

  $$D_K = \frac{E_K}{(1-\rho)+\rho E_K} = \frac{E_K}{1-\rho(1-E_K)}$$

  so

  $$\boxed{D_K > E_K}$$

  as expected (Why?).

Queue-length Distribution

The best approach to find the equilibrium distribution of Q_t is to look first at the queue-length distribution *for delayed arrivals*, i.e., to evaluate $P(Q_t = i|\text{all servers busy})$.

We have

$$\begin{aligned} P(Q_t = i|N_t \geq K) &= \frac{P(N_t = K+i)}{\sum_{j=0}^{\infty}P(N_t = K+j)} \\ &= \frac{\pi_K\rho^i}{\sum_{j=0}^{\infty}\pi_K\rho^j} \end{aligned}$$

where the second equality follows since

$$\pi_{K+i} = \pi_K\rho^i$$

So

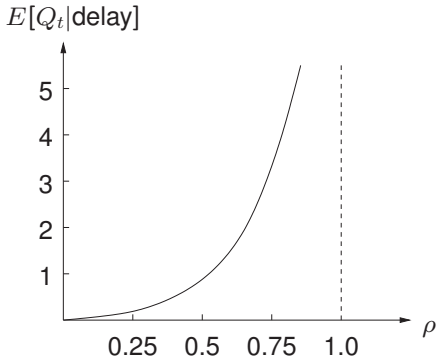
$$\boxed{P(Q_t = i|\text{delay}) = (1-\rho)\rho^i, \quad i = 0, 1, 2, \ldots}$$

which is geometric distribution.

We deduce immediately that

$$\begin{aligned} E[Q_t|\text{delay}] &= \frac{\rho}{1-\rho} \\ Var(Q_t|\text{delay}) &= \frac{\rho}{(1-\rho)^2} \end{aligned}$$

Notice the sensitivity to ρ as $\rho \to 1$.



Unconditional Queue-length Distribution

We know that:

- *Delayed arrivals* see a *geometric* queue length.
- *Non-delayed arrivals* see *zero* queue length.

So, for *all* arrivals, the queue-length distribution is

$$\begin{aligned} P(Q_t = i) &= P(\text{delay})P(Q_t = i|\text{delay}) \\ &\quad + P(\text{no delay})\underbrace{P(Q_t = i|\text{no delay})}_{=\begin{cases} 1, & \text{if } i = 0 \\ 0, & \text{if } i > 0 \end{cases}} \end{aligned}$$

But $P(\text{delay}) = D_K(A)$, and so

$$P(Q_t = i) = \begin{cases} D_K(A)(1-\rho)\rho^i, & \text{if } i > 0 \\ 1 - \rho D_K(A), & \text{if } i = 0 \end{cases} \quad (3)$$

From (3), it is left as an excercise to derive the *mean* and *variance* of Q_t.

$$\begin{aligned} E[Q_t] &= D_K(A)\left(\frac{\rho}{1-\rho}\right) \\ Var(Q_t) &= E[Q_t]\left[\left(\frac{1}{1-\rho}\right) + \right. \\ &\qquad \left. [1 - D_K(A)]\left(\frac{\rho}{1-\rho}\right)\right] \end{aligned}$$

- Remember that all of these results are *equilibrium results* which are only valid when the offered traffic / channel, $\rho < 1$, i.e. when the total offered traffic, $A < K$. In this case, *all* the offered traffic is carried and so A is also the *mean carried load*.
- Recall that the Erlang loss model has a state distribution (truncated Poisson) which remains valid for non-exponential service times. However, for the Erlang delay model ($M/M/K$ system) this is *not* the case—the geometric form of the queue-length distribution holds only for *exponential* service times.
- Single-server queue ($K = 1$)

  It is easily verified in this case that

  $$D_K(A) = D_1(A) = \rho$$

  and all the above results can be simplified.

Waiting-time Distribution

The waiting time W of an arbitrary arrival depends on:

- the number of items already waiting in the buffer
- the queue discipline

waiting time of FIFO queue

Consider *delayed* arrivals. For such arrivals, $N_t \geq K$ and so all K servers are *continuously busy* until the end of the waiting period.

- Departure rate from the system = $\boxed{K\mu}$ throughout the waiting period.
- Interdeparture intervals (until the test arrival enters service) are *iid exponential random variable* with common mean $(K\mu)^{-1}$.
- Departure stream is a *Poisson stream* with rate $K\mu$.

Now an arrival joining a queue of length i must wait for $(i+1)$ departures from the system before entering service.

For *delayed* arrivals, we therefore have

$$\begin{aligned} &P(W > \tau | Q_t = i) \\ &\quad = P(< (i+1) \text{ departures in } (0,\tau)) \\ &\quad = \sum_{j=0}^{i} \frac{(K\mu\tau)^j}{j!} e^{-K\mu\tau} \end{aligned}$$

Also, in equilibrium, we know that for *delayed* arrivals:

$$P(Q_t = i) = (1-\rho)\rho^i$$

Therefore, for delayed arrivals,

$$\begin{aligned} &P(W > \tau) \\ &\quad = \sum_{i=0}^{\infty} P(Q_t = i) P(W > \tau | Q_t = i) \\ &\quad = \sum_{i=0}^{\infty} (1-\rho)\rho^i \sum_{j=0}^{i} \frac{(K\mu\tau)^j}{j!} e^{-K\mu\tau} \\ &\quad = e^{-K\mu(1-\rho)\tau} \end{aligned}$$

where the last equality follows from interchanging summation order and using $\lambda = K\mu\rho$.

That is,

$$\boxed{P(W \leq \tau | W > 0) = 1 - e^{-K\mu(1-\rho)\tau}, \quad \tau \geq 0}$$

That is, for *delayed* arrivals, the waiting time W is exponentially-distributed with mean $[K\mu(1-\rho)]^{-1}$.

Of course, for *non-delayed* arrivals, $W = 0$, and so the waiting-time distribution for *all* arrivals is

$$\begin{aligned} &P(W \leq \tau) \\ &= \underbrace{[1 - D_K(A)]}_{\text{Prob. of no delay}} \\ &\quad + \underbrace{[D_K(A)]}_{\text{Prob. of delay}} \underbrace{[1 - e^{-K\mu(1-\rho)\tau}]}_{\text{Exponential wait}} \end{aligned}$$

- The *geometric* distribution of Q_t and the *exponential* distribution of W are closely related. [The sum of a geometric number of iid exponential random variables is itself exponential.]
- Remember that W depends on the queue discipline; the above results are valid for the FIFO case only.

Waiting Time of other Queue Disciplines

A queue disciplines said to be *non-biassed* if the order in which items are selected for service does not depend on the items' service times. Examples of non-biassed queue disciplines are:

- First-in / first-out (FIFO)
- Last-in / first-out (LIFO)
- Service in random order (SIRO)

It may be shown (e.g. R. B. Cooper's book) that

$$E[W_{\text{FIFO}}] = E[W_{\text{SIRO}}] = E[W_{\text{LIFO}}] \qquad (4)$$

and

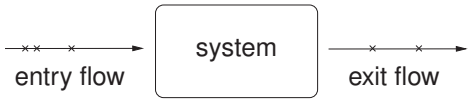
$$Var(W_{\text{FIFO}}) < Var(W_{\text{SIRO}}) < Var(W_{\text{LIFO}}) \qquad (5)$$

Result (4) suggests that the *mean waiting time* $E[W]$ *is independent of queue discipline*. This is in fact true for all non-biassed queue disciplines.

Little's Theorem

This very general mean-value result states that for a wide range of systems, we have, at equilibrium:

mean number of items in system
= mean entry rate into system
× mean sojoirn time in system



Let

A_t = number of items entering system in $[0, t]$
N_t = number of items in system at tine t
T_i = sojourn time (time between entry and exit) of i^{th} item to enter system

and

$$\bar{\lambda} = \lim_{t\to\infty}\left(\frac{A_t}{t}\right) \Leftarrow \text{mean entry rate}$$

$$\bar{N} = \lim_{t\to\infty}\left[\frac{1}{t}\int_0^t N_t\,dt\right] \Leftarrow \text{mean number in system}$$

$$\bar{T} = \lim_{t\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n} T_i\right] \Leftarrow \text{mean sojourn time}$$

Then, Little's theorem states that, if $\bar{\lambda}$ and $\bar{T}$ exist, then so does $\bar{N}$ and

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

- In most applications, the system is an ergodic stochastic system and equation (1) can be written

  $$\boxed{E[N_t] = \lambda E[T]}$$

  where λ is the mean arrival rate; $E[T]$ is the equilibrium mean sojourn time; and $E[N_t]$ is the equilibrium mean number of items in the system.

- Little's theorem is a generalisation of the *basic traffic equation*. In applications, "system" can be interpreted as, for example:
  - a buffer
  - a complete queuing system
  - a complete network
- Remember: Little's theorem applies to *mean values only!*

Heuristic Proof of Little's Theorem (P. Burke at Bell Labs)

For each item i, think of the sojoirn time T_i as a quantity of work, brought into the system by item i and used up at unit rate while item i is in the system.

Then

- Mean rate at which *new work* is brought into the system $= \lambda E[T]$.
- Mean rate at which *total workload* is being reduced
  $=$ (mean rate / item)
  $\times$(mean number of items in system)
  $= (1) \times E[N_t]$
  and at equilibrium, these two rates must be equal.

LECTURE ON $M/M/K/N$ Queuing System

This lecture focuses on studying the effect of a finite buffer.

Now suppose that

$$\boxed{\text{buffer capacity} = B, \quad B < \infty}$$

so that

$$\boxed{\text{system size} = N = K + B}$$

It is now possible for an arrival to find the *buffer full*; such an arrival must be rejected by the system.

Thus the system is a *mixed (loss/delay) system*.

Congestion Analysis

As before, $\{N_t\}$ is a birth/death process, but now the state space is truncated at $N_t = N = K + B$.

Thus the stability is guaranteed for all values of ρ, including $\rho > 1$.

As before, the *local balance equations* give:

$$\begin{aligned} \pi_i &= \left(\frac{A^i}{i!}\right)\pi_0, \quad 0 \le i \le K \\ &= \left(\frac{A^K}{K!}\right)\rho^{i-K}\pi_0, \quad K \le i \le K+B \end{aligned}$$

However, the *normalisation* is now different. We find that

$$\pi_0 = S^{-1}$$

where now

$$S = \frac{A^K}{K!}\left[E_K^{-1}(A) + \frac{\rho(1-\rho^B)}{1-\rho}\right], \quad \rho \neq 1$$

so that

$$\boxed{\pi_0 = \frac{1}{\left(A^K/K!\right)}\left[\frac{(1-\rho)E_K(A)}{(1-\rho)+\rho(1-\rho^B)E_K(A)}\right]}$$

for $\rho \neq 1$.

From the equilibrium distribution (π_i), it is easy to compute the two main congestion probabilities.

These are:

$$\begin{aligned} P(\text{delay}) &= P(\text{all } K \text{ servers busy}|\text{buffer not full}) \\ &= P(K \le N_t < K+B) \\ &= \boxed{\pi_K \left(\frac{1-\rho^B}{1-\rho}\right)} \\ P(\text{loss}) &= P(\text{buffer full}) \\ &= P(N_t = K+B) \\ &= \boxed{\pi_K \rho^B} \end{aligned}$$

where, in each case,

$$\pi_K = \left(\frac{A^K}{K!}\right)\pi_0$$

Note All the results quoted above are valid only when $\rho \neq 1$. When $\rho = 1$ exactly, the results must be modified accordingly.

Queue-length Distribution

The queue length seen by *rejected* arrivals is of course $Q_t = B$ (i.e. buffer full). The queue length seen by arrivals which are *accepted but delayed* has a *truncated geometric distribution*, since

$$\begin{aligned} &P(Q_t = i|\text{delay}) \\ &= P(Q_t = i | K \le N_t < K+B) \\ &= \frac{P(N_t = K+i)}{P(\text{delay})} \\ &= \frac{\pi_K \rho^i}{\pi_K \left(\frac{1-\rho^B}{1-\rho}\right)}, \quad i = 0,1,\ldots,B-1 \\ &= \rho^i \left(\frac{1-\rho}{1-\rho^B}\right), \quad i = 0,1,\ldots,B-1 \end{aligned}$$

Mean Queue Length

From the queue length distribution, it is straightforward to show that

$$E[Q_t|\text{delay}] = \left(\frac{\rho}{1-\rho}\right) - \left(\frac{B\rho^B}{1-\rho^B}\right)$$

This is mean queue length seen by *delayed* arrivals.

The *unconditional* mean queue length is then given by

$$\begin{aligned} E[Q_t] &= P(\text{delay})E[Q_t|\text{delay}] \\ &\quad + P(\text{loss}) \underbrace{E[Q_t|\text{loss}]}_{=B} \end{aligned}$$

Mean Waiting Time

Use Little's theorem—but we must be careful to use the correct entry rate.

For items *accepted* into the buffer (i.e. not rejected) the *entry rate* is

$$\boxed{\lambda_A = \lambda[1 - P(\text{loss})]}$$

Then applying Little's theorem to the buffer,

$$\boxed{E[W] = \left(\frac{1}{\lambda_A}\right) E[Q_t]}$$

Note For *delayed* arrivals, we shall have

$$E[W|\text{delay}] = \left(\frac{1}{\lambda_A}\right) E[Q_t|\text{delay}]$$

since

$$\begin{aligned} E[W] &= P(\text{delay})E[W|\text{delay}] \\ E[Q_t] &= P(\text{delay})E[Q_t|\text{delay}] \end{aligned}$$

LECTURE ON
Non-Markov Queuing System

The $M/M/K$ system and its variants are the only queuing systems in which the number of items in system at time t, N_t, is a continuous-time Markov process.

However, there are many queuing systems for which the $M/M/K$ system is not a suitable model. In particular, it is often unreasonable to assume exponential service times.

It is therefore important to consider systems of the $M/G/K$ type, where G denotes that the service times are iid random variables with a *general* (i.e. unspecified) distribution.

Unfortunately, $M/G/K$ systems are very difficult to analyse in the multichannel case ($K > 1$). And most of the available results are for the single-channel (or single server) case ($K = 1$).

$M/G/1$ Queuing System

This is a single-server queue with

- a *Poisson* arrival stream
- a *general* service-time distribution
- an *infinite-capacity* buffer

Because of the (in general) non-exponential service time, the probability of a service-completion in any short interval $(t, t + \Delta t)$ is *not* independent of the past history of the system. In fact, it depends on the *elapsed service time* (time since the current service started), ΔS_t.

The *residual service time* (time until the *end* of the current service) thus depends not only on N_t but also on the elapsed service time.

There are several ways of dealing with this complication:

- Define a 2-component stochastic process

  $$\{(N_t, \Delta S_t), t \leq 0\}$$

  in which the *state* $(N_t, \Delta S_t)$ contains the relevant bit of past history.

  Such a process is, once again, a Markov process (of a type called a *piecewise-deterministic process (PDP)*).

- Concentrate on the values of $\{N_t\}$ at the instants immediately after a departure occurs—the so-called *departure instants*. The sequence of values

  $$\{N_i : i = 1, 2, \ldots\}$$

  where N_i is the number of items left behind at the i^{th} departure instant, forms a *discrete-time Markov chain*, $\{N_i\}$.

  This is because the only events that can occur between 2 successive departures are arrivals. And these arrivals are Poisson and hence memoryless.

  The discrete-time chain $\{N_i\}$ is called an *embedded Markov chain* for the given $M/G/1$ system.

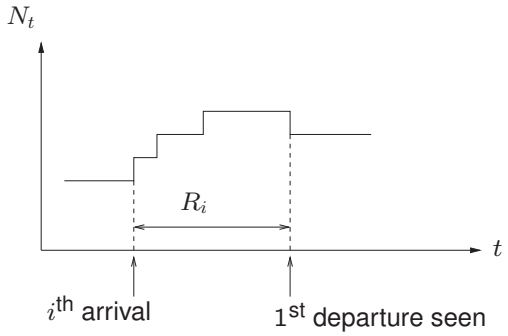
  It may be shown that the equilibrium state distribution of the continuous-time process $\{N_t\}$ is the same as that of the embedded chain $\{N_i\}$. This is why analysis of the latter gives us the information needed.

- The third approach is to confine attention to *mean value analysis*. This is the easiest method and it is also applicable to *priority queuing systems* (to be seen later).

Mean-value Analysis of $M/G/1$ system

Suppose the system is in equilibrium and, for the i^{th} arrival to the system, let

$$\begin{aligned} R_i &= \text{residual service time} \\ &= \text{time until first departure} \\ &\quad \text{seen by } i^{\text{th}} \text{ arrival} \end{aligned}$$



Now assume FIFO queue discipline.

Then, for the i^{th} arrival, if

$$\begin{aligned} S_i &= \text{service time} \\ W_i &= \text{waiting time} \\ Q_i &= \text{queue length found on arrival} \end{aligned}$$

we have

$$W_i = R_i + \sum_{j=1}^{Q_i} S_{i-j}$$

which is the total time to serve all items ahead of the i^{th} arrival.

Taking expectations throughout:

$$\begin{aligned} E[W_i] &= E[R_i] + E\left[\sum_{j=1}^{Q_i} S_{i-j}\right] \\ &= E[R_i] + E\left[E\left[\sum_{j=1}^{Q_i} S_{i-j}\middle| Q_i\right]\right] \end{aligned}$$

where the second term above is conditioning on Q_i and $\sum_{j=1}^{Q_i} S_{i-j}$ is sum of Q_i iid RVs.

It follows that

$$\begin{aligned} E[W_i] &= E[R_i] + E\left[Q_i E[S]\right] \\ &= E[R_i] + E\left[Q_i\right] E[S] \end{aligned}$$

But since Poisson arrivals see an unbiassed sample of queue behaviour,

$$E[Q_i] = E[Q_t]$$

and so

$$E[W] = E[R] + E[Q]E[S] \qquad (6)$$

But, by Little's formula,

$$E[Q] = \lambda E[W]$$

and so (6) becomes

$$E[W] = E[R] + \rho E[W]$$

since $\rho = \lambda E[S]$.

That is, for an $M/G/1$/FIFO system:

$$\boxed{E[W] = \frac{E[R]}{1-\rho}} \qquad (7)$$

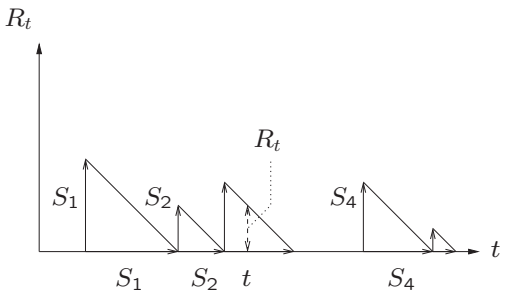
Note Of course, we must have $\rho < 1$ for stability.

Computation of $E[R]$

Now let

R_t = residual service time seen by a *virtual arrival* at time t

Then, at equilibrium, $\{R_t\}$ is a continuous-time stochastic process which looks like:



Now, assuming that $\{R_t\}$ is *ergodic* (in mean):

$$\begin{aligned} E[R_t] &= \lim_{T\to\infty} \frac{1}{T}\int_0^T R_t\,dt \\ &= \lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{M_T}\left(\frac{1}{2}S_i^2\right) \end{aligned}$$

where

$\frac{1}{2}S_i^2$ = area of the i^{th} triangle

M_t = number of completed service in $[0, T]$

Rewriting this equation gives

$$E[R_t] = \lim_{T\to\infty} \frac{1}{2}\left(\frac{M_T}{T}\right)\left[\frac{1}{M_T}\sum_{i=1}^{M_T} S_i^2\right]$$

where we note that

$$\begin{aligned} \frac{M_T}{T} &= \text{service completion rate} \\ &= \text{mean arrival rate, } \lambda \\ \frac{1}{M_T}\sum_{i=1}^{M_T} S_i^2 &= \text{mean square service time} \\ &= E[S^2] \end{aligned}$$

It then follows that, for the $M/G/1$ system,

$$\boxed{E[R_t] = \frac{1}{2}\lambda E[S^2]} \qquad (8)$$

- This result is *independent of the queue discipline*.
- $E[R_t]$ depends on the *second moment* of S.

Finally, putting (7) and (8) together, we get

$$\boxed{E[W] = \frac{\lambda E[S^2]}{2(1-\rho)}} \qquad (9)$$

This is the well-known *Pollacek-Khinchin formula* for the mean waiting time of a $M/G/1$ queuing system.

Note We know that $E[W]$ is the same for all unbiassed queue disciplines. Hence, (9) is valid for all such queue discipline. Furthermore, (8) is independent of queue discipline. It then follows that (7) is *also* independent of queue discipline, i.e., (7)–(9) *all hold for any unbiassed queue discipline*.

Example: $M/M/1$ System

$$E[S^2] = Var(S) + (E[S])^2 = \frac{2}{\mu^2}$$

$$\Rightarrow E[W] = \left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{\mu}\right) = \left(\frac{\rho}{1-\rho}\right)E[S]$$

Example: $M/D/1$ System

Here, D means "deterministic": it represents the case when

$$S = \text{constant} = h \quad \text{(say)}$$

$$E[S^2] = Var(S) + (E[S])^2 = h^2$$

$$\Rightarrow E[W] = \left[\frac{\rho}{2(1-\rho)}\right]h = \left[\frac{\rho}{2(1-\rho)}\right]E[S]$$

Note that this case (for which $Var(S) = 0$) provides a *lower bound* for $E[W]$ for all $M/G/1$ systems. It also shows that assuming exponential S can lead to a reduction in $E[W]$ of up to 100%.

Queue Length in $M/G/1$ System

Applying Little's theorem to the Pollacek-Khinchin formula (9) gives

$$E[Q_t] = \frac{\lambda^2 E[S^2]}{2(1-\rho)} \qquad (10)$$

This is the mean queue length seen by *all* arrivals (including those which are not delayed).

The queue-length *distribution* is much harder to determine. Its form will depend on the service-time distribution of the system being studied.

LECTURE ON
Priority Queuing Systems

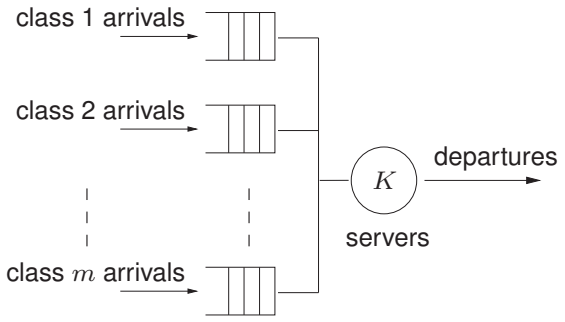
These are queuing systems in which the order of entry into service is determined by *pre-assigned priority classes*.

Suppose there are m such priority classes in use, numbered $1, 2, \ldots, m$, with the convention that

$$\text{Class } 1 \succ \text{Class } 2 \succ \cdots \succ \text{Class } m$$

where $\succ$ means has priority over.

Then we can visualise such a priority queuing system as:



and items in a given buffer have priority over items in all lower buffers.

There are 2 main types of priority in common use:

1. Non-preemptive priority
   Service of an item may *not* be interrupted by a higher-priority arrival.

2. Preemptive priority
   Service is interrupted by any higher-priority arrival. In this case, there are 2 possibilities.

   (a) Preemptive resume
   After the period of interruption, the interrupted service is resumed at the point where it was interrupted.

   (b) Preemptive restart
   After the period of interruption, the interrupted service starts again from the beginning.

Multiserver systems with priority queue discriplines are very difficult to analyse. The main results available are mean-value results for $M/G/1$ systems.

$M/G/1$ Queuing System with Non-preemptive Priority

Suppose there are m priority classes and let

$$\begin{aligned} \lambda_k &= \text{arrival rate} \\ S_k &= \text{service time} \\ W_k &= \text{waiting time} \\ Q_k &= \text{queue-length seen on arrival} \end{aligned}$$

for a typical arrival in class k.

Then, the *offered traffic* in class k is

$$\boxed{\rho_k = \lambda_k E[S_k]} \qquad (11)$$

and the total offered traffic is

$$\rho = \sum_{k=1}^{m} \rho_k$$

For stability, we require $\rho < 1$, i.e.

$$\boxed{\sum_{k=1}^{m} \rho_k < 1}$$

If this condition does not hold, then one or more of the *lower* priority classes will not be adequately served.

To derive for the *mean waiting time*, we can use the same sort of mean-value analysis as before.

Let

$$R = \text{\textit{residual service time} seen by an arbitrary arrival}$$

THEN

For class 1 arrivals

$$\begin{aligned} E[W_1] &= E[R] + E[Q_1]E[S_1] \\ &= E[R] + \lambda_1 E[W_1]E[S_1] \\ &\quad \text{by Little's theorem} \end{aligned}$$

$$\Rightarrow \boxed{E[W_1] = \frac{E[R]}{1-\rho_1}} \qquad (12)$$

For class 2 arrivals

Now the mean wait is made up of 3 components:

$$E[W_2] = E[R] + \underbrace{(E[Q_1]E[S_1] + E[Q_2]E[S_2])}_{\textit{backlog} \text{ to be cleared}} + \underbrace{(\lambda_1 E[W_2])\, E[S_1]}_{\textit{new work} \text{ due to class 1 arrivals during } W_2}$$

Then, using Little's theorem on each side:

$$E[W_2] = E[R] + (\rho_1 E[W_1] + \rho_2 E[W_2]) + \rho_1 E[W_2]$$

$$\Rightarrow E[W_2] = \frac{E[R] + \rho_1 E[W_1]}{1 - \rho_1 - \rho_2} = \frac{E[W_1]}{1 - \rho_1 - \rho_2} \text{ using (12)}$$

$$\Rightarrow \boxed{E[W_2] = \frac{E[R]}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}} \qquad (13)$$

For class k arrivals

We can now *lump together* all higher-priority classes into a single class with offered traffic level:

$$\boxed{\sigma_{k-1} = \sum_{i=1}^{k-1} \rho_i}$$

Then, using the same argument as for class 2, we get:

$$\boxed{E[W_k] = \frac{E[R]}{(1 - \sigma_{k-1})(1 - \sigma_k)}} \qquad (14)$$

This is the basic result for non-preemptive priority systems of $M/G/1$ type.

NOTES

(i) To evaluate $E[R]$, note that

$$E[R] = \frac{1}{2}\lambda E[S^2] = \frac{1}{2}\lambda \left[\left(\frac{\lambda_1}{\lambda}\right) E[S_1^2] + \cdots + \left(\frac{\lambda_m}{\lambda}\right) E[S_m^2]\right] = \frac{1}{2}\sum_{k=1}^{m} \lambda_k E[S_k^2]$$

(ii) The *overall mean waiting time* is clearly:

$$\boxed{E[W] = \sum_{k=1}^{m} \left(\frac{\lambda_k}{\lambda}\right) E[W_k]} \qquad (15)$$

(iii) One way of reducing $E[W]$ is to give priority to items with shorter expected service times.

The extreme version of this idea is the *shortest job first (SJF)* queue discipline, in which the next item chosen for service is the one with the shortest service time. Clearly, SJF is only feasible if each service time is known in advance.

It can be shown (quite easily) that for SJF:

$$E[W_T] = \frac{E[R]}{(1 - \sigma_T)^2} \qquad (16)$$

with

$$\sigma_T = \lambda_T E[S_T]$$

where

λ_T = arrival rate

$E[S_T]$ = mean service time

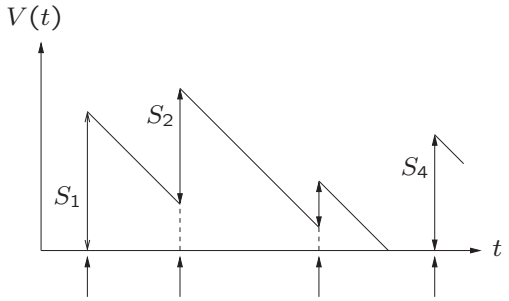
for items with $S \leq T$.

Note that $E[W_T]$ increases from $E[R]$, for very short jobs, to $E[R] \big/ (1 - \rho)^2$ for very long jobs.

(iv) Kleinrock's conservation law

Although we can reduce the *overall* mean wait by suitable scheduling, we cannot reduce the time required to clear a given backlog (= amount of unfinished work).

Denote the backlog by $V(t)$. Then $\{V(t)\}$ is a continuous-time stochastic process (in fact, a Markov process for $M/G/1$ systems) which has the form:



By considering the mean value of $\{V(t)\}$, it can be shown (e.g. Kleinrock (1970)) that

$$\boxed{\sum_{k=1}^{m} \rho_k E[W_k] = \left(\frac{\rho}{1-\rho}\right) E[R]} \qquad (17)$$

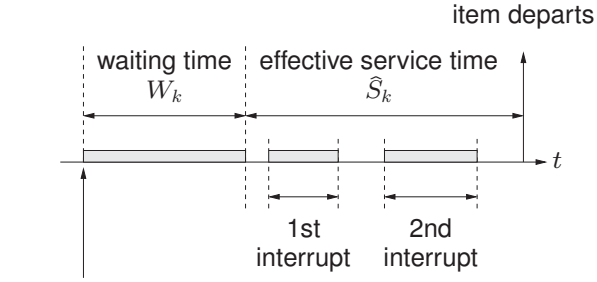
for any work-conserving priority queue discipline.

$M/G/1$ Queuing System with Preemptive Priority

The easier case to analyse is preemptive resume.

We must now consider the *total* time spent in the system by a class k item, i.e. the transit time, T_k.

The general picture is now:



and

$$\boxed{T_k = W_k + \hat{S}_k}$$

Computation of $E[W_k]$

Class k items can now *interrupt* any lower-class service in progress. So the only contributions to $E[R]$ will be class k and above:

$$E[R] = \left(\frac{1}{2}\sum_{i=1}^{k} \lambda_i E[S_i^2]\right) \Leftarrow \boxed{\text{call this } E[R_k]}$$

Apart from this modification, $E[W_k]$ will be given by equation (14)

$$E[W_k] = \frac{E[R_k]}{(1-\sigma_{k-1})(1-\sigma_k)} \qquad (18)$$

Computation of $E[\hat{S}_k]$

Let

V_k = work brought into the system, during $\hat{S}_k$, by higher-priority arrivals

Then

$$\begin{aligned} E[V_k] &= \sum_{i=1}^{k-1} \left(\lambda_i E[\hat{S}_k]\right) E[S_i] \\ &= \left(\sum_{i=1}^{k-1} \rho_i\right) E[\hat{S}_k] \quad \text{using } \rho_i = \lambda_i E[S_i] \\ &= \sigma_{k-1} E[\hat{S}_k] \end{aligned}$$

Then

$$\begin{aligned} E[\hat{S}_k] &= E[\text{true service time}] + E[\text{interrupt time}] \\ &= E[S_k] + E[V_k] \\ &= E[S_k] + \sigma_{k-1} E[\hat{S}_k] \end{aligned}$$

$$\Rightarrow E[\hat{S}_k] = \frac{E[S_k]}{1-\sigma_{k-1}}$$

And finally,

$$E[T_k] = E[W_k] + E[\hat{S}_k]$$

so that

$$\boxed{E[T_k] = \left[\frac{E[R_k]}{(1-\sigma_{k-1})(1-\sigma_k)}\right] + \left[\frac{E[S_k]}{1-\sigma_{k-1}}\right]} \quad (19)$$

Note

Case of preemptive restart results in a longer effective service time, $\hat{S}_k$.

Traffic Theory & Queueing Systems

E4.05 - S07: Part 3

Javier A. Barria
j.barria@imperial.ac.uk
1009a DEEE

LECTURE ON
Traffic Characterisation B-ISDN

Packet Voice Model

- Two state process. N multiplexed Independent voice sources
- (N+1)-state B-D Model.

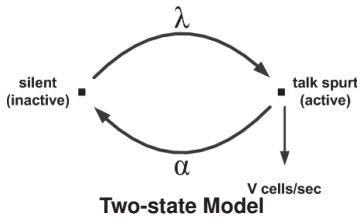
$$\pi_i = \binom{N}{i}\left(\frac{\lambda}{\lambda+\alpha}\right)^i\left(\frac{\alpha}{\lambda+\alpha}\right)^{N-i} \quad (1)$$

Probability of i out of N, two-state sources, one active

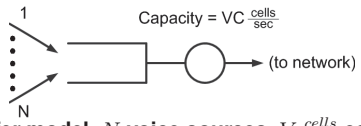
- Also: Setting up the balance equations at each state

$\pi M = 0$

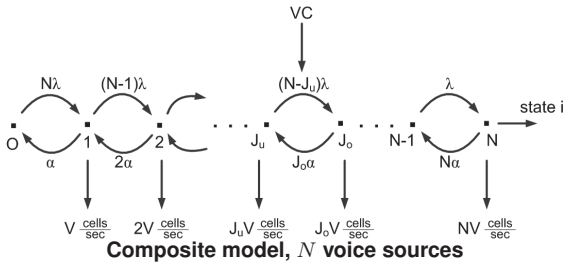
$$M = \begin{pmatrix} -N\lambda & N\lambda & 0 & \dots \\ \alpha & -[\alpha+(N-1)\lambda] & (N-1)\lambda & \dots \\ 0 & 2\alpha & -(N-2)\lambda-2\alpha & \dots \\ \vdots & & & \end{pmatrix}$$



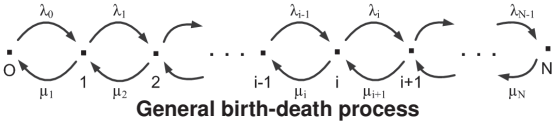
Two-state Model



Access buffer model, N voice sources, V $\frac{cells}{sec}$ each, when in talk spurt

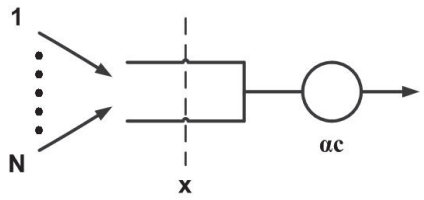


Composite model, N voice sources



General birth-death process

Fluid flow modelling of packet voice



- N_R of calls so large it appears like a continuous flow fluid
- Buffer capacity becomes a continuous R.V.

$F_i(t,x)$ = Probability distribution function at time t, with system in state x and i sources in talk spurt.

$\Delta t \to 0$, Taylor series expansion

$$\begin{aligned}\frac{\partial F_i(t,x)}{\partial t} &= [N-(i-1)]\lambda F_{i-1}(t,x) \\ &+ (i+1)\alpha F_{i+1}(t,x) \\ &- [(N-i)\lambda + i\lambda]F_i(t,x) \\ &- (i-c)\alpha\frac{\partial F_i(t,x)}{\partial x}\end{aligned}$$

Statistical Equilibrium

$$\begin{aligned}\frac{\partial F_i(t,x)}{\partial t} &= 0, F_i(t,x) \to F_i(x) \\ (i-c)\alpha\frac{\partial F_i(x)}{\partial x} &= [N-(i-1)]\lambda F_{i-1}(x) \\ &- [(N-i)\lambda + i\alpha]F_i(x) \\ &+ (i+1)\alpha F_{i+1}(x) \\ \frac{dF(x)}{dx}D &= F(x)M \\ F(x) &\equiv [F_0(x), F_1(x), \ldots, F_N(x)] \\ D &\equiv diag[-C\alpha, (1-C)\alpha, \ldots, (N-C)\alpha]\end{aligned}$$

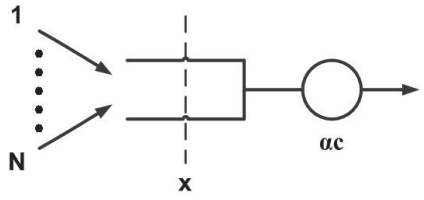
First order linear differential equation

$$F(x) = \sum_{i=0}^{N} a_i F_i e^{z_i x} \qquad (2)$$

(z_i, ϕ_i) = eigenvalue, eigenvector of: $z_j\phi_j D = \phi_j M$

Definition:
$G(x) = 1 - F(x)$ = Probability buffer occupancy exceeds x.

Example:



$\frac{1}{\alpha}$ = average length of talk spurt

$$\begin{aligned}\frac{dF(x)}{dx}D &= F(x)M \\ \frac{dF(x)}{dx} &= F(x)M', \qquad M' = MD'\end{aligned}$$

$N = 1$ (one voice source)

$$M = \begin{bmatrix}-\lambda & \lambda \\ \alpha & -\alpha\end{bmatrix} = \begin{bmatrix}-\gamma & \gamma \\ 1 & -1\end{bmatrix}$$

$$\gamma \equiv \frac{\lambda}{\alpha}$$

$$D = \begin{bmatrix}-C\alpha & 0 \\ 0 & (1-C)\alpha\end{bmatrix}$$

$$D' = \begin{bmatrix}\frac{1}{-C\alpha} & 0 \\ 0 & \frac{1}{(1-C)\alpha}\end{bmatrix}$$

$$M = MD' = \begin{bmatrix}\frac{\gamma}{C} & \frac{\gamma}{1-C} \\ -\frac{1}{C} & -\frac{1}{1-C}\end{bmatrix}_{|\alpha=1}$$

Eigen values given by:

$$zI - M' = 0$$
$$z = 0$$
$$z = \frac{\gamma}{C} - \frac{1}{1-C}$$

Stable system $\Rightarrow \rho \equiv (\frac{\gamma}{1+\gamma})\frac{1}{C} < 1$

$$z = -\frac{(1-\rho)(1+\gamma)}{1-C} \qquad (z<0), (N=1)$$

Asymptotic Result G(x) [ANI82]

- Probability distribution is given by the sum of negative exponentials
- The exponential with the smallest negative eigenvalue R will dominate

$$R = \frac{(1-\rho)(1+\gamma)}{1-\frac{C}{N}} \tag{3}$$

for N multiplexed sources

$$\gamma = \frac{\lambda}{\alpha} \tag{4}$$
$$\rho \equiv \frac{\gamma}{1+\gamma}\frac{N}{C} < 1 \tag{5}$$
$$G(x) \sim A_N \rho^N e^{-Rx} \tag{6}$$

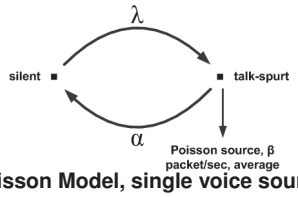
Example 1. Service Process: MMPP arrival

- Cell (packet) length distribution:
  Exponential $\frac{1}{\nu}$
- Each source: $\beta\frac{cells}{sec}$(when active)
- N active sources:
  Average Arrival Rate $= N\beta\frac{\lambda}{\alpha+\lambda}\left\{\frac{cells}{sec}\right\}$

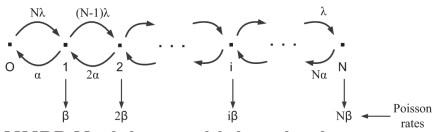
Stable system if:

$$N_\beta(\frac{\lambda}{\alpha+\lambda}) < \nu \tag{7}$$

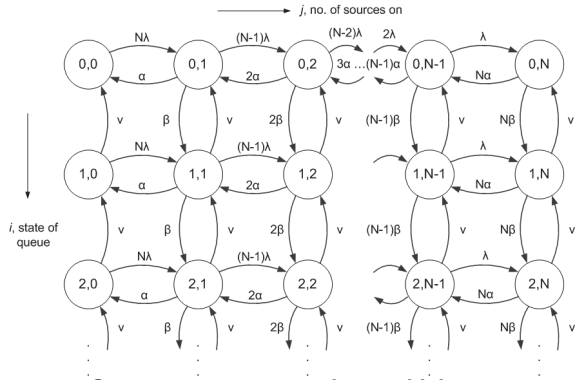
$F_j(x) =$ Probability buffer occupancy $\leq x$, j sources on.
$P_{ij} =$ P[Queue length $= i$, sources on $= j$].
$P_i = \sum_{j=0}^{N} P_{ij}$
$(i,j) \leftarrow$ defines a two-dimensional state space.



Poisson Model, single voice source



MMPP Model, N multiplexed voice sources



State space representation, multiplexer.

Example 2. MMPP Model for Video Traffic

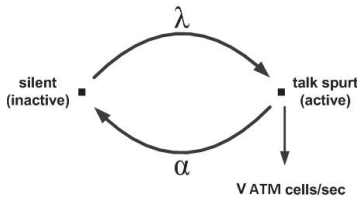
- Video source approximated by its quantised equivalent.
- Parameters estimation: Measuring transition probabilities from the actual sequence.
  - Calculate steady state π_j
  - Autocovariance function
  - Estimate P $\rightarrow$ Generate M

Note: If frame-length internal is long compared to the rate of cell arrivals then, the access buffer occupancy distribution reaches steady-state in each frame interval. To obtain $P(n)$:

1. Obtain $P(n|\lambda = \lambda_i)$ (Queueing theory)
2. $P(n) = \sum P(n|\lambda = \lambda_i)\pi_i$

Example 3. Packet Voice Modelling

- A single voice source can be represented by a two-state process, alternating between:
- Active periods (talk spurts)
- Silence Period



Composite Model, N voice sources

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \alpha & -\alpha \end{pmatrix}$$

$$\pi = \left(\frac{\alpha}{\lambda+\alpha}, \ \frac{\lambda}{\lambda+\alpha} \right)$$

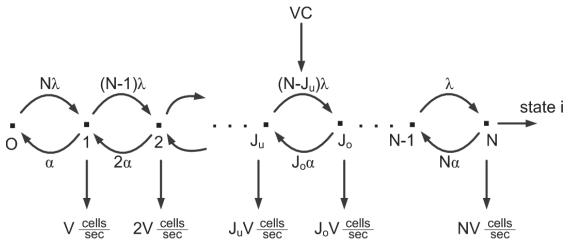
N-Multiplexed Independent Voices sources

- Steady state probability
- If N-sources are independent; the probability of having *i* sources on is:

$$\pi_i = \begin{pmatrix} N \\ i \end{pmatrix} P^i (1-P)^{N-i}$$

$$P = \frac{\lambda}{\alpha + \lambda}$$

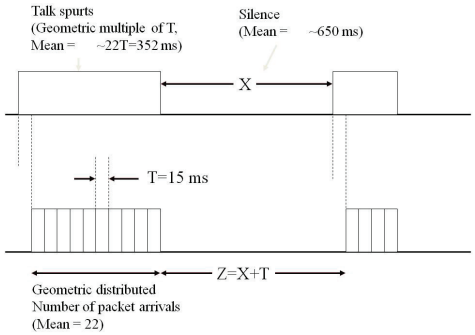
$$1 - P = \frac{\alpha}{\alpha + \lambda}$$



Composite Model, N voice sources

One Voice Source

- Active (talk spurt): arrival of packets at fixed intervals (T[ms])
  Talk spurt: is of random length NT
  assumption: the number of packet is geometrically distributed
- Silence: No packet arrivals
  assumption: the length of the silence period is a random variable x exponentially distributed



Arrival process of one voice source

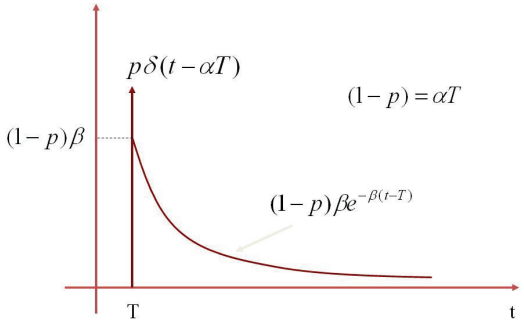
- Each packet interarrival time is of length T with probability

$$p = 21/22$$

- of length X+T with probability

$$1\text{-}p = 1/22$$

$$\alpha T = 1 - p$$



Probability density function of packet interarrival time for one voice source.

Packet arrival time distribution

$$F(T) = [(1- \alpha\ T) + \alpha\ T\ (1- e^{-\beta(t-T)})]\ U(t - T)$$

$$f(s) = \int_0^\infty e^{-st} dF(t) = [1 - \alpha T + \frac{\alpha T \beta}{(S+\beta)}] e^{-sT}$$

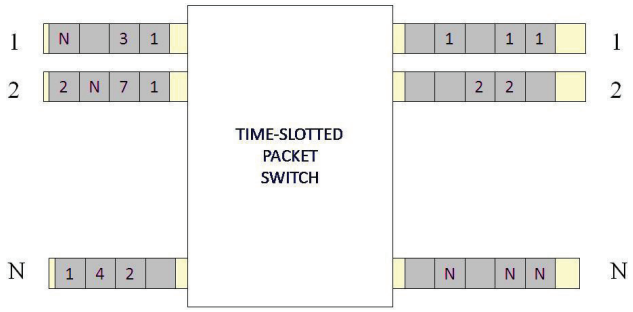
$$m_1 = T\left(1 + \frac{\alpha}{\beta}\right)$$

$$m_2 = T^2\left(1 + \frac{2\alpha}{\beta^2 T} + 2\frac{\alpha}{\beta}\right)$$

$$m_3 = T^3\left(1 + \frac{3\alpha}{\beta} + \frac{6\alpha}{\beta^2 T} + \frac{6\alpha}{\beta^3 T^2}\right)$$

Example 4. ATM switch Fabric

- Fixed-length cell switching.
- NxN switch functions
  Route incoming cells to its destination,
  Resolve contention is two or more cells arrive to the same destination.
- Note: if no buffer, carried load is less than offered load.



Generic time-slotted packet switch

Analysis of lost packet performance (no-buffers)

- Arrival totally un-correlated in time and across inputs.
- *p*: probability that a given tie slot contains an active cell.
- *p/N*: probability that a given input contains a cell destined for a particular output.

- The number of cells k having the same destination (arriving within the same time slot) is a random variable with a:

Binomial Probability Distribution

$$P(K=k) = \binom{N}{k} \left(\frac{P}{N}\right)^k \left(1 - \frac{P}{N}\right)^{N-k}, \quad k = 0,1,\ldots N$$

Average number of lost packets intended for a given output:

$$L = \sum_{k=2}^{N} (k-1) \binom{N}{k} \left(\frac{p}{N}\right)^k \left(1-\frac{p}{N}\right)^{N-k}$$

$$L = \sum_{k=0}^{N} (k-1) \binom{N}{k} \left(\frac{p}{N}\right)^k \left(1-\frac{p}{N}\right)^{N-k} + \left(1-\frac{p}{N}\right)^N$$

$$L = N\left(\frac{p}{N}\right) - 1 + \left(1-\frac{p}{N}\right)^N = p + \left(1-\frac{p}{N}\right)^N - 1$$

Offered load for any output: *F = p - L*

LECTURE ON
Admission and Access Control

- Given VP: How many VC $\sim$ QoS
- Given active VC: Admit a new call?

Admission Control
-ON-OFF Traffic Source:

- R_p Peak Rate
- $\frac{1}{\beta}$ Average burst length
- $\frac{1}{\alpha}$ Average silence length

Single-class Admission Control

- VP with capacity $C_L \sim QoS$

$p = \frac{\alpha}{\alpha+\beta}$: probability source in ON.

- Lower bound number of calls accepted
$NpR_p = N\frac{\alpha}{\alpha+\beta}R_p$: Average rate of transmission for N sources

$\rho = \frac{N_p R_p}{C_L}$ (Average bandwidth assignment method) may be unacceptable in terms of cell cost.

- Upper band number of calls accepted.
$NR_p = C_L$ (Peak bandwidth assignment)

Equivalent capacity

$$C_L = (m + k\sigma)R_p$$

$mR_p =$ Mean
$\sigma R_p =$ Standard deviation
$k = k(Qos)$
$m = N_p$ ON-OFF Source Model

$$\sigma^2 = N_p(1-p) = m(1-p)$$

$$C = \frac{C_L}{R_p} \rightarrow C = m + k\sigma$$

$$C = N_p + k\sqrt{N_p(1-p)}$$

$k(Qos) \sim P_L$

i) $P_L = \sum_{i=J_o}^{N} \frac{(i-C)\pi i}{m}$
ii) $E = \sum_{i=J_o}^{N} \pi_i$

A. Large number of sources multiplexed
$N \gg 1, P \ll 1$
$\pi_i = \binom{N}{i} P^i (1-P)^{N-i}$ (Binomial)

is approximated quite closely by the normal distribution ($m = N_p, \sigma^2 = N_p(1-p)$)

$$P_L = \frac{1}{m} \int_{J_0}^{\infty} \frac{e^{-\frac{(x-m)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} (x-c) dx$$

$$E = \int_{J_0}^{\infty} \frac{e^{-\frac{(x-m)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx$$

if $(C-m) > 3\sqrt{2}\sigma$

$$E = \frac{\sigma e^{-\frac{(C-m)^2}{2\sigma^2}}}{\sqrt{2\pi}(C-m)}; P_L = \frac{1-P}{C-m} E$$

Appling ln:

$$ln(\sqrt{2\pi}E) = ln(\frac{\sigma}{C-m}) - \frac{(C-m)^2}{2\sigma^2}$$

$$C_{LS} = mR_p + \sigma \underbrace{\sqrt{-ln(2\pi) - 2ln(E)}}_{k} R_p$$

B. Effect of the Access Buffer

$G(x) \sim A_N \rho^N e^{-\frac{\beta R_x}{R_p}}$ (probability buffer occupancy $> x$)

$$R = \frac{(1-\rho)(1+\frac{\alpha}{\beta})}{1 - \frac{C_L}{NR_p}}$$

$$\rho = \frac{N_p R_p}{C_L}$$

Use fluid flow approximation. If $\rho \sim 1, A_N \rho^N \sim 1$

$$P_L = e^{-\frac{\beta R_x}{R_p}}$$

Applying ln:

$$\frac{\beta R_x}{R_p} = -ln(P_L)$$

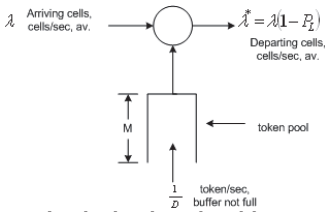
$$\frac{C_L}{R_p N} = \frac{1-k}{2} + \sqrt{\left(\frac{1-k}{2}\right)^2 + kp}$$

$$C_{LF} = R_p N \left(\frac{1-k}{2}\right) + R_p N \sqrt{\left(\frac{1-k}{2}\right)^2 + kp}$$

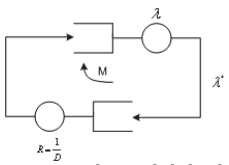
$$C_L = min[C_{LS}, C_{LF}]$$

LECTURE ON
Access Control: Leaky Bucket

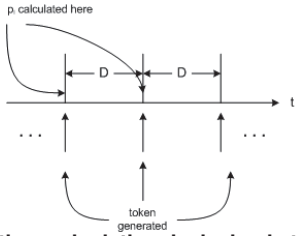
- Open loop access control
- UPC $\Leftrightarrow$ General rate algorithm (An equivalent representation is: leaky bucket technique)
- Token pool buffer
- Tokens are generated one per D sec
  λ^* = average throughput differs from offered load λ (because of possible cell loss)
  P_L represents the cell loss probability
  Closed queueing network model
- Cells generated at Poisson rate λ (only if upper queue has token). This queue increases at an average rate $R = D^{-1}$
- M tokens circulating in it (i.e. at most M cells served in succession)



Leaky bucket algorithm



Closed queueing network model, leaky bucket, M/M/1 approximation



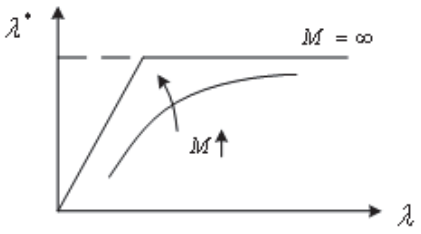
Discrete-time calculation, leaky bucket algorithm

$$\lambda^* = \lambda(1 - P_L) = \lambda(1 - P_0)$$

$P_L =$ Probability upper queue empty
Probability lower queue full (using M/M/1/M)

$$P_L = \frac{\rho^M(1-\rho)}{1-\rho^{M+1}}$$
$$\rho = \frac{\lambda}{R} = \lambda D$$
$$\lambda^* = \lambda[\frac{1-\rho^M}{1-\rho^{M+1}}]$$

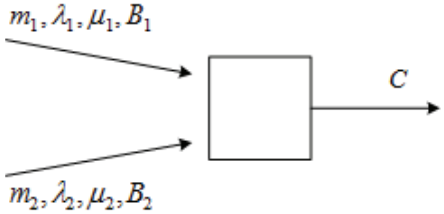
as $M \uparrow \rightarrow \lambda^*$ to ideal, i.e.



Single Line Model (Non-linear equivalent capacity

$G(n) =$ equivalent capacity function

Product-form solution



Product-form solution

Call admitted if: available link capacity is enough

Available link capacity

- Necessary to consider the connections in progress: (2-dimensional vectors) $(N_1(t), N_2(t))$
- $S = \{(M_1, M_2) \in I^2 | G_1(M_1) + G_2(M_2) \leq C\}$
  S= call admission region of the single link model

Call admitted if: Total consumed link capacity of all connections including new connection $\leq C$.
i.e.

$$G_1(M_1 + 1) + G_2(M_2) \leq C$$
$$G_1(M_1) + G_2(M_2 + 1) \leq C$$

$\{(N_1(t), N_2(t))\}$ steady state joint probability

$$\pi(M_1, M_2) = \frac{1}{K(s)} \frac{\left(\frac{\lambda_1}{\mu_1}\right)^{M_1} \left(\frac{\lambda_2}{\mu_2}\right)^{M_2}}{M_1! M_2!}$$
$$K(S) = \sum_{(n_1,n_2)\in S} \frac{\left(\frac{\lambda_1}{\mu_1}\right)^{M_1} \left(\frac{\lambda_2}{\mu_2}\right)^{M_2}}{M_1! M_2!}$$

LECTURE ON
Blocking Probabilities

$$B_1 = 1 - \frac{K(S_1)}{K(S)}$$
$$B_2 = 1 - \frac{K(S_2)}{K(S)}$$

$$S_1 = \{(M_1, M_2) \in S | (m_1 + 1, m_2) \in S\}$$
$$S_2 = \{(M_1, M_2) \in S | (m_1, m_2 + 1) \in S\}$$

Summing all the corresponding probabilities of state occupacy

LECTURE ON
ATM Multiplexer

Peak rate admission

$$S = \{(M_1, \ldots, M_k) : \sum_{R=1}^{K} b_K M_K \leq C\} \quad (8)$$

$b_K =$ peak rate for service K

Effective bandwidth admission

$$S = \{(M_1, \ldots, M_k) : \sum_{R=1}^{K} b_K^e M_K \leq C\} \quad (9)$$

$b_K^e =$ effective bandwidth for service K (e.g. $b_K^e = b_K$)

Service separation

$$S = \{(M_1, \ldots, M_k) : \beta_1(M_1) + \cdots + \beta_K(M_K) \leq C\} \quad (10)$$

$\beta_K(n) =$ minimum amount of transmission capacity for Qos to be met by VCs.
(e.g. $\beta_K(n) = b_K^e n_K$)

Stochastic Knapsack

- C resource units
- Arrival from K classes
  - Poisson arrival: λ_k
  - Exponential holding time: $\frac{1}{\mu_k}$
  - Hold b_k resource units
- Pure loss systems:
  $m = (M_1, \ldots, M_k)$ (State System)
  $b = (b_1, \ldots, b_k)$

Admit class-k arrival iff

$$b_k < C - b \cdot n \quad (11)$$

Dynamics Knapsack problem

$$S \equiv \{n \epsilon I^k : b \cdot n \leq C\}$$
$$X(t) = (X_1(t), \ldots, X_k(t)) \text{ state at } t$$
$$\{X(t)\} = \text{ Associated stochastic process}$$

(Aperiodic and irreductible Markov process over S)

Steady State Distribution: $\pi(n)$

$$\pi(n) = \frac{1}{G} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!} \quad (12)$$

$$G = \sum_{n \in S} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!} \quad (13)$$

1) $C = \infty$ $\{Y_t(t)\}$
- K independent B-D process
$Birth = \lambda_k$
$Death = \mu_k n_k$

$$\widehat{\pi}(n) = \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!} e^{-\rho_k} \quad (14)$$

2) $C < \infty$
- IDEM: Truncated Distribution

Blocking Probability of Class-k

$$S_k = \{n \in S : b \cdot n \leq C - b_k\} \quad (15)$$

(Subset Knapsack admits class-k).

Since arrivals are Poisson;

$$B_K = 1 - \sum_{n \in S_k} \pi(n) \quad (16)$$

Blocking Probability and throughput

$$TH_k = \lambda_k (a - B_K) \quad (17)$$

If $x = (x_1, \ldots, x_k)$

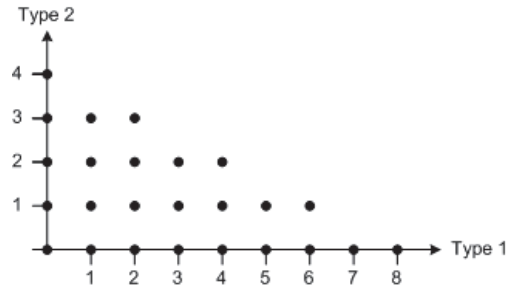
$$U = \sum b_i x_i$$
$$UTIL = E[U]$$

(Knapsack average utilisation)

Erlang Loss

- Stochastic Knapsack: Multidimensional generalisation of Erlang Model.
- Erlang Loss: Only one class and all object size =1.

Example:
$C = 8; b_1 = 1; b_2 = 2$



$S =$ all •
$S = \{n \in S : bn \leq C - b_k\}$

LECTURE ON
Performability of communication systems

Content

- Introduction to the Performability concept
- Performability model construction
- Examples of Performability models
- Designs of communication networks considering congestion and reliability constraints

Performance/Reliability Issues in Communication Systems

- Performability is a measure that quantifies a system's ability to perform in the presence of faults.
- A Performability model is obtained by:
  - Modelling the process X (the base model) that represents the behaviour of the system and,
  - Defining the performance variable Y for each state of the system.
- Lets define
  - S = Object system
  - T = observation period (or mission time)
  - Y = a random variable associated with the performance of the system

- and
  - A = Accomplishment (performance) levels of variable Y
- Performability (B) = Probability that S performs at a level in B
  - The determination of this probability is based on the underlying stochastic process $X(t)$ (the base model).

Reward based Perfomability model construction

- System behaviour (**Base model X**)
  - e.g. Markov chains
- Performability variable (**Variable Y**)
  - e.g. Link throughput

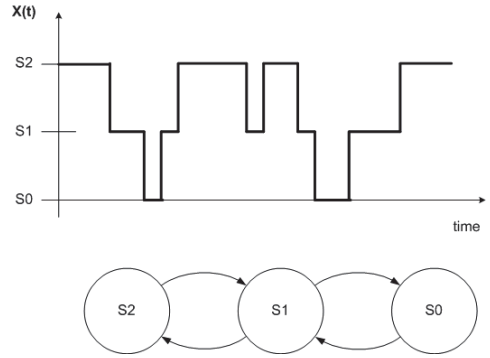
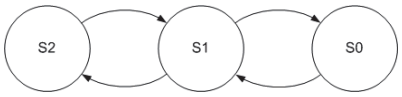


Figure 1a. Markovian model of a three level system performance



$$R = [r_{S2}, r_{S1}, r_{S0}]$$

$$Y(t) = \int_0^t r_{X(\tau)} d\tau$$

Accumulated reward up to time t.

Figure 1b. Markovian model of a three level system performance

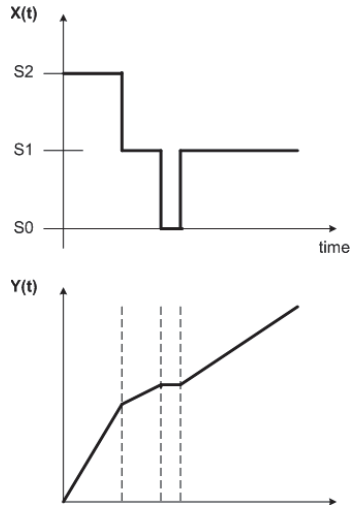
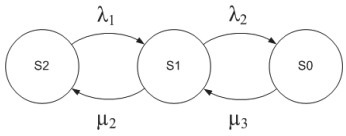


Figure 1c. Markovian model of a three level system performance

Markov Reward Models

- If $X(t)$ (the base model) can be represented by a continuous-time finite-state Markov chain (CTMC), and
- This base model $X(t)$ is extended assigning rewards rates to it states we are defining a:
- Rate-base Markov Reward Model (MRM)



$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 \\ \mu_2 & -(\mu_2 + \lambda_2) & \lambda_2 \\ 0 & \mu_3 & -\mu_3 \end{pmatrix}$$

$$Y(t) = \int_0^t r_{X(\tau)} d\tau = \sum_{i=0}^{N} r_i \tau_i; \quad t = \sum_{i=1}^{N} \tau_i$$

$$P_i(t, y) = P[Y(t) \leq y | X(0) = s_i]$$

Steady state solution behaviour of the system

$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 \\ \mu_2 & -(\mu_2 + \lambda_2) & \lambda_2 \\ 0 & \mu_3 & -\mu_3 \end{pmatrix}$$

$$Q^T \pi = 0, \qquad \pi e = 1$$

Some MRM measures of interest

Expected Values of Reward Rates

- E.g. It can represents the averaged rate at which the system completes the work

$$E[Z(t)] = \sum_{i \in S} r_i \pi_i(t)$$

Expected Accumulated Rewards

- E.g. It can represents the expected number of jobs processed in a given time interval

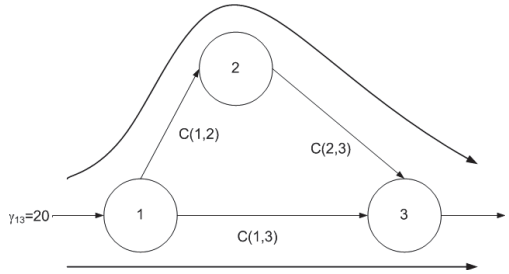
$$\lim_{t\to\infty} E[W(t)] = W(\infty) = \sum_{i=0}^{N} r_i \pi_i$$

Performability (MRM) evaluation

- Construction of a Performability model for the system and measure is question (i.e. Q and R)
- Evaluation of the measure via solution of the model, e.g.

$$E[Z(t)] = \sum_{i\in S} r_i \pi_i(t)$$

Example Performability (MRM) modelling

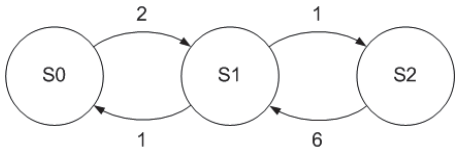


- State $S0$: failure-free state
  - Failures can occur in respect to link $(1,2)$ at a rate of 2 failures/unit time.
- State $S1$: is the state in which link $(1,2)$ is in failed condition
  - Link $(1,2)$ can be repaired as a rate of 1 repair/unit time or,
  - Link $(2,3)$ can fail at a rate of 1 failure/unit time.

- State $S2$: is the state in which link $(1,2)$ and link $(2,3)$ are in failed condition
  - Link $(2,3)$ can be repaired at a rate of 6 repairs/unit time.

(i) The transition matrix Q of the base model X is:

$$Q = \begin{pmatrix} -2 & 2 & 0 \\ 1 & -2 & 1 \\ 0 & 6 & -6 \end{pmatrix}$$



(ii)

$$T = \frac{1}{\gamma} \sum_{(i,j)} \frac{F_{ij}}{(C_{ij} - F_{ij})} \text{ Mean network delay}$$

$$T_0 = \frac{1}{20}\frac{10}{(30-10)} + \frac{1}{20}\frac{10}{(30-10)} + \frac{1}{20}\frac{10}{(30-10)}$$
$$T_1 = \frac{1}{20}\frac{20}{(30-20)}$$
$$T_2 = \frac{1}{20}\frac{20}{(30-20)}$$

Consider $R = \left\{ r_i = \frac{T_0}{T_i} \right\}$

$$r_0 = \frac{T_0}{T_0} = 1$$
$$r_1 = \frac{T_0}{T_1} = 0.75$$
$$r_2 = \frac{T_0}{T_2} = 0.75$$

(iii) Evaluate $\lim_{t\to\infty} E(W(t)); W(t) = \frac{Y(t)}{t}$

$$Q^T\pi = 0, \pi e = 1$$

$$\begin{pmatrix} -2 & 2 & 0 \\ 1 & -2 & 1 \\ 0 & 6 & -6 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

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

Sol.: $\pi^T = [0.3, 0.6, 0.1]$

$$\lim_{t\to\infty} E(W(t)) = W(\infty) = \frac{3}{10} + \frac{6}{10}\frac{3}{4} + \frac{1}{10}\frac{3}{4} = 0.825$$

Other MRM example

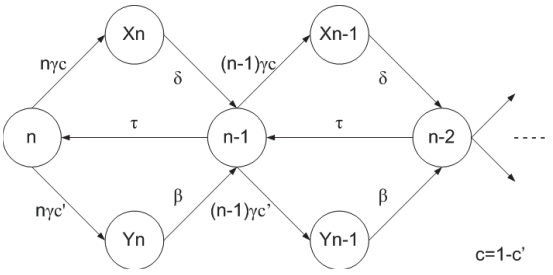


Figure 3a. Multiprocessor system

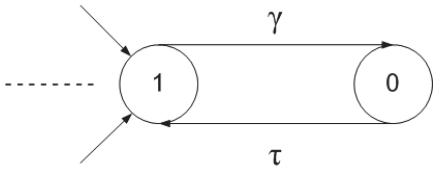


Figure 3b. Multiprocessor system