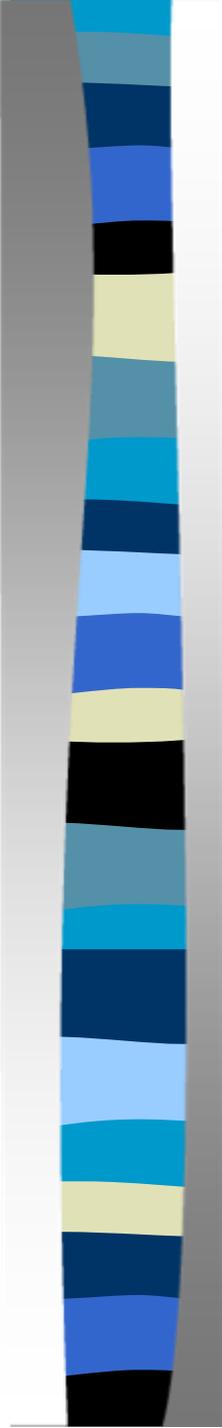


# An Introduction to Analytical Queuing Models (9)



## Reading:

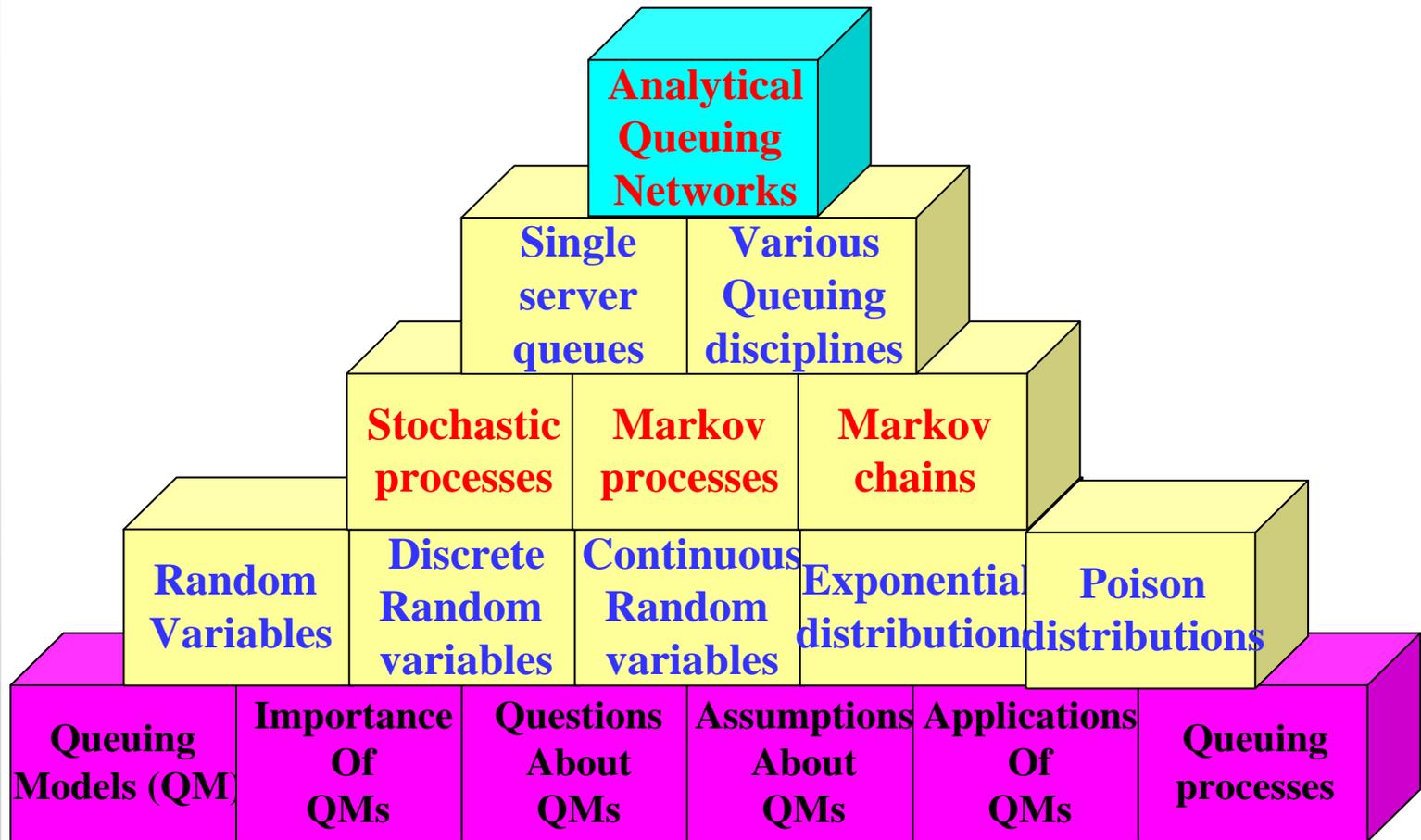
1. Cassady, R. and Nachlas, J. A. (2009), *Probability Models in Operational Research*, CRC Press, Taylor & Francis Group.
2. Askin, G. R. and Standridge, C. R. (1993), *Modelling and Analysis of Manufacturing Systems*, John Wiley & Sons Corp. (Chapter 11)

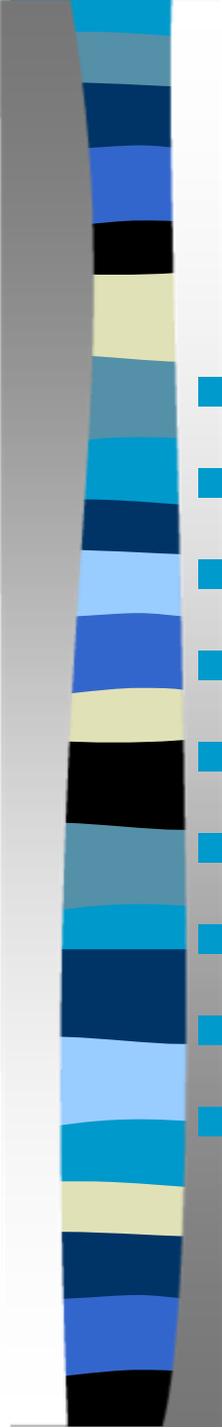


# Topic

1. The concept of queues and why they are important.
2. Become familiar with some of the key questions that are usually asked about queuing processes.
3. Understand the basics of probability distributions and Markov Chains
4. Understand and solve problems involving single server queuing models.
5. Be able to formulate and solve problems involving queuing networks
6. Understand the advantages of simulation over queuing models

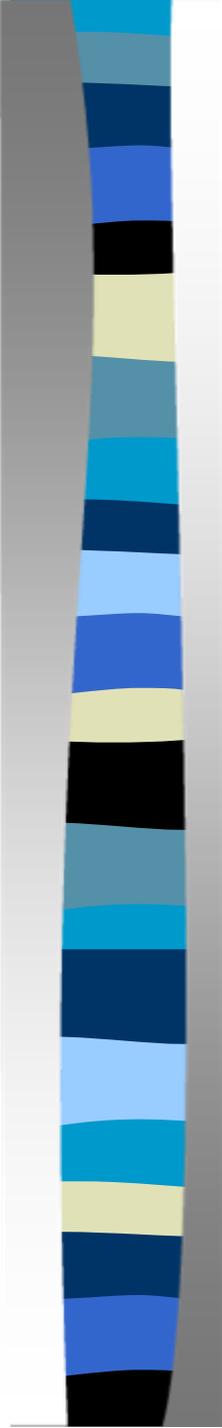
# Overview





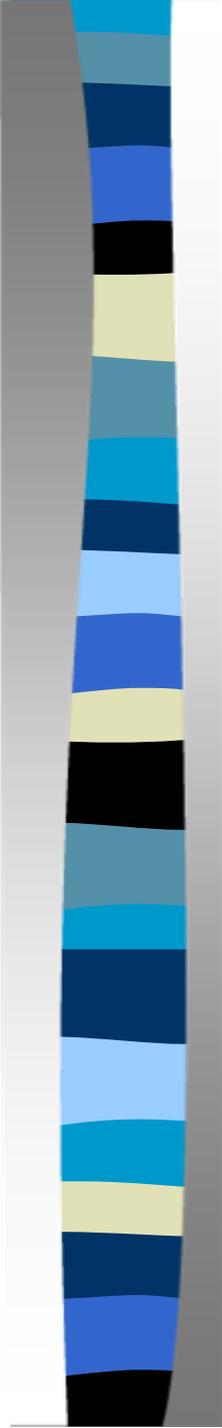
# Outline

- What are queues?
- Why study queuing processes?
- Some key questions about queuing processes.
- General assumptions behind queuing models.
- Essentials for understanding queuing models.
- A single server queuing model.
- Queuing networks.
- Simulation models and queuing models.
- Examples of queuing problems.



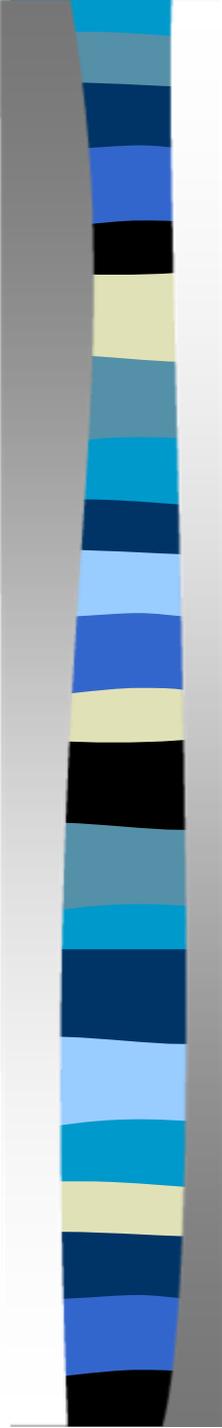
# What are queues?

- Queues result whenever entities (parts, customers, patients etc) have to compete for scarce resources (machines, cashiers, doctors etc).
- They occur in almost every system (manufacturing, banking, healthcare, etc)



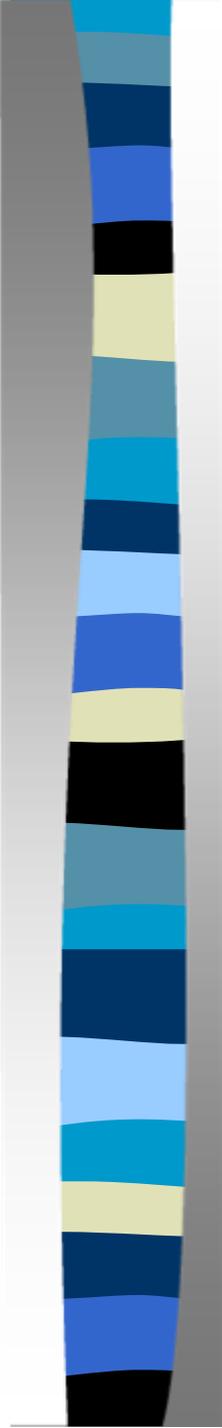
# Why study queuing processes?

- Resources in the real world are always limited.
- The need to always improve performance and satisfy the customer.
- The need to eliminate waste and reduce cost.



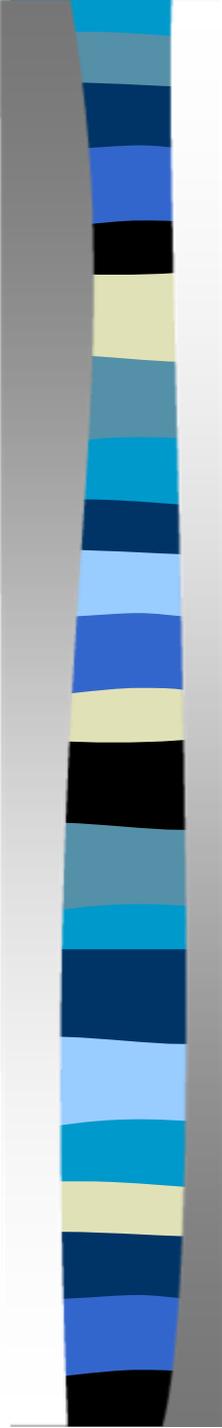
# Some key questions about queuing processes (1)

- How many servers are needed to reduce the proportion of customers waiting for more than 2hrs by 5%?
- What is the effect of the introduction of priorities on customers' waiting time?



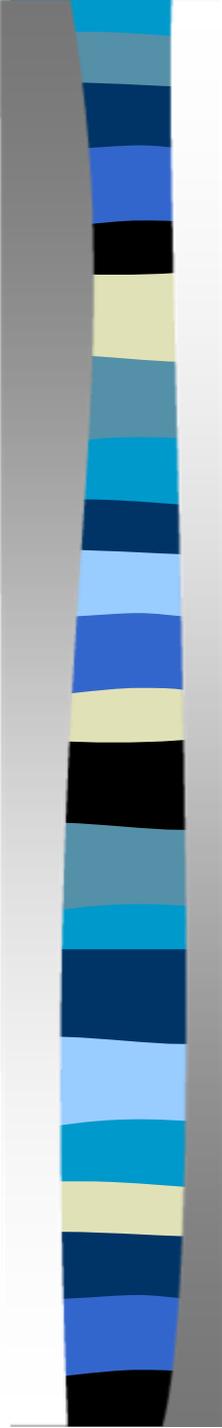
## Some key questions about queuing processes (2)

- How much must the waiting room be enlarged to reduce the proportion of customers turned away to less than 1 in 10?
- Would it be cheaper to employ another server or to increase the size of the waiting room?
- What is the utilization of current resources?



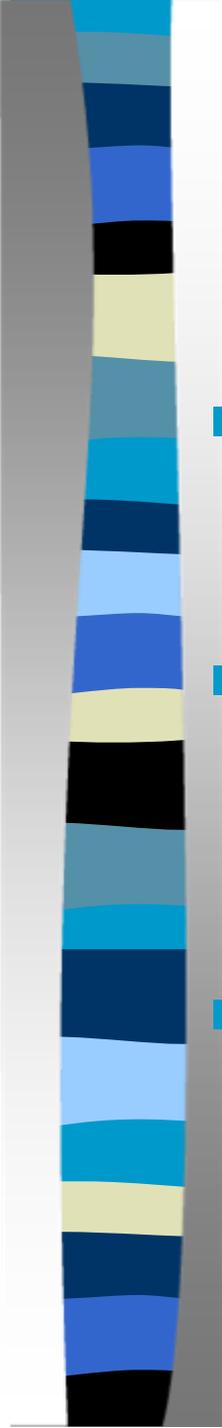
# General assumptions behind queuing models (1)

- Suitable for quickly evaluating average steady-state performance.
- Finding “rough-cut” or approximate results for the long-term average behaviour of static systems.
- **“Static” indicates that process parameters such as mean service time do not change over time.**



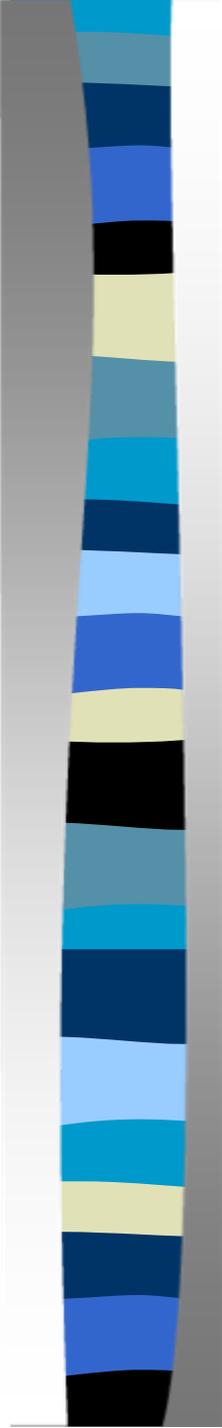
# General assumptions behind queuing models (2)

- **System stability is also assumed in the sense that capacity, measured by maximum production rate, exceeds average demand.**
- **System stability assures finite expected inventory levels and waiting times.**



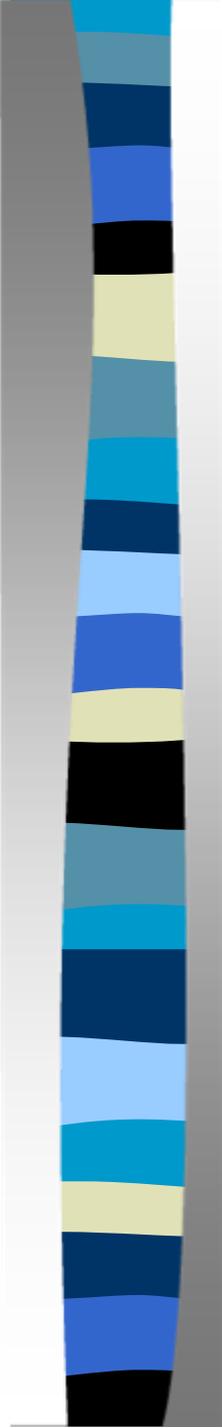
# Essentials for understanding queuing models (1)

- Systems conceptualisation (*the first step of the modelling process*)
- “Conceptualisation is an abstract, simplified view of the reference reality which is represented for some purpose” (Gunter et al, 2006).
- There are two forms of conceptualisation for the purpose of performance evaluation;
  - “Routed job flow” modelling paradigm (*structure oriented*)
  - “State – transition” modelling paradigm (*behaviour oriented*)



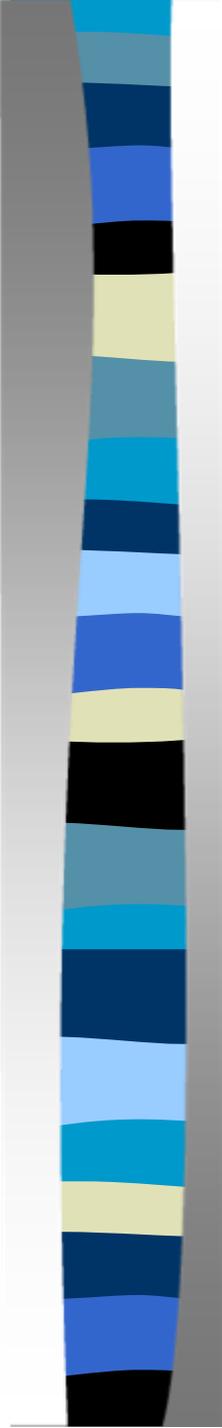
# Essentials for understanding queuing models (2)

- System Evaluation (*second step of the modelling process*)
- Evaluation is the deduction of performance measures by the application of appropriate solution methods depending on the conceptualisation chosen.



# Essentials for understanding queuing models (3)

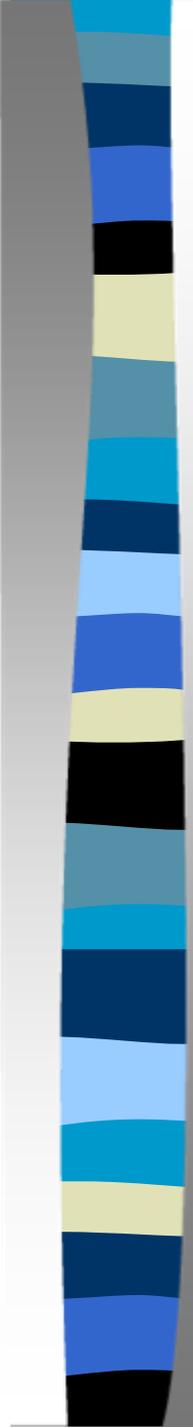
- Some solution methods;
  - **Analytical solutions**
    - *Closed-form solutions*
    - *Numerical solutions*
  - **Simulation solutions**
    - *For many complex systems, no analytical solution is feasible. Discrete Event Simulation (DES) is the most applicable.*
  - **Hybrid solutions**
    - *A combination of analytical and simulation solutions*



# Essentials for understanding queuing models (4)

## ■ Introduction to Probability distributions

- **Radom variables:** A random variable is a function that reflects the results of a random experiment (E.g. throwing a die, number of arrivals at a bank per hour, time between consecutive arrivals of jobs in a manufacturing system, etc)
  - *Discrete random variables*
  - *Continuous random variables*

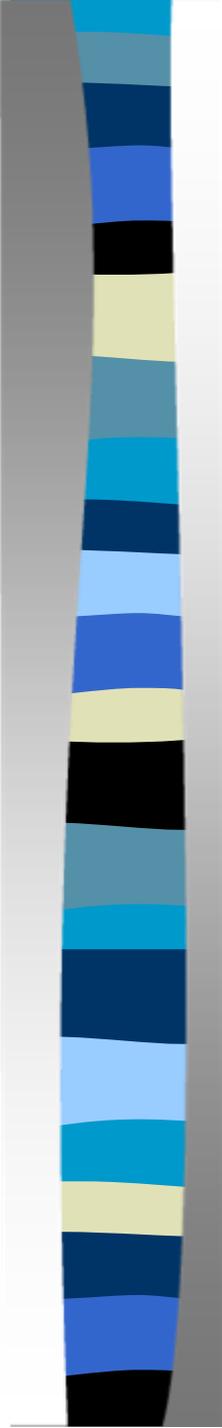


# Essentials for understanding queuing models (5)

- Discrete Random Variables: Can only assume discrete values and are often non-negative integers.
- Discrete Random Variables are described by;
  - Positive random values they can assume
  - Probability for each of these values. The set of these probabilities is called the **probability mass function** (pmf) of the random variable.
  - Thus if the possible values of a random variable  $X$  is given by the non-negative integers, then the **pmf** is given by the probabilities;

$$p_k = P(X = k), \text{ for } k = 0, 1, 2, \dots,$$

$$\text{Such that } P(X = k) \geq 0 \text{ and } \sum_{\text{all } k} P(X=k) = 1$$



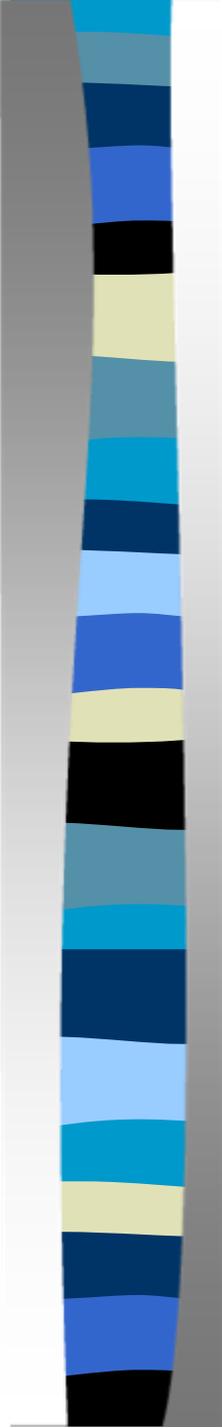
# Essentials for understanding queuing models (6)

- Examples of Discrete Random Variables are;
  - **Bernoulli random variables** (Experiment with only two possible outcomes e.g. tossing a coin,  $k = 0, 1$ )

$$P(X=0)=1-p \text{ and } P(X=1)=p, \text{ with } 0 < p < 1$$

- **Binomial random variable** (The experiment with two possible outcomes is carried out  $n$  times.  $X$  is the number of times the outcome  $k$  occurs).

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \dots, n$$



# Essentials for understanding queuing models (7)

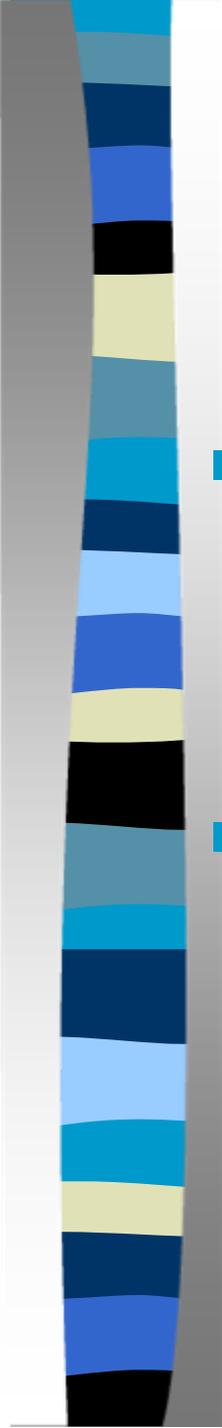
- **Geometric random variable** (The experiment with two possible outcomes is carried out several times.  $X$  is now the number of trials it takes for the outcome 1 to occur).

$$P(X = k) = p(1 - p)^{k-1}, \quad k = 1, 2, \dots$$

- **Poisson random variable** (Number of events in a period of time)

$$P(X = k) = \frac{(\alpha)^k}{k!} \cdot e^{-\alpha}, \quad k = 0, 1, 2, \dots; \alpha > 0$$

- Several important parameters can be derived from the **pmf** of a discrete random variable such as the expected value, variance, standard deviation, etc.



# Essentials for understanding queuing models (8)

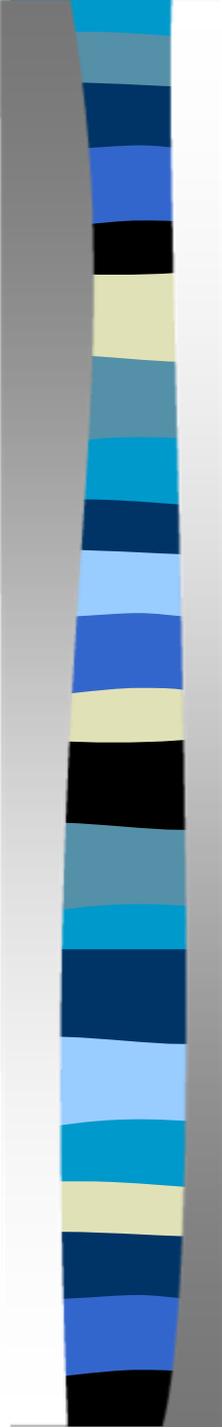
- **Continuous Random Variables:** Can assume all values in the interval  $[a,b]$ , where  $-\infty \leq a < b \leq +\infty$ 
  - They are usually measurements, e.g. height, weight, distance, time required for a task etc.

- **Continuous Random Variables are described by;**
  - Their ***Distribution Function*** also called ***Cumulative Distribution Function*** (CDF):

$$F_X(x) = P(X \leq x) \quad \text{for all } x$$

- The probability density function (pdf) is given by;

$$f_X(x) = \frac{dF_X(x)}{dx}$$



# Essentials for understanding queuing models (9)

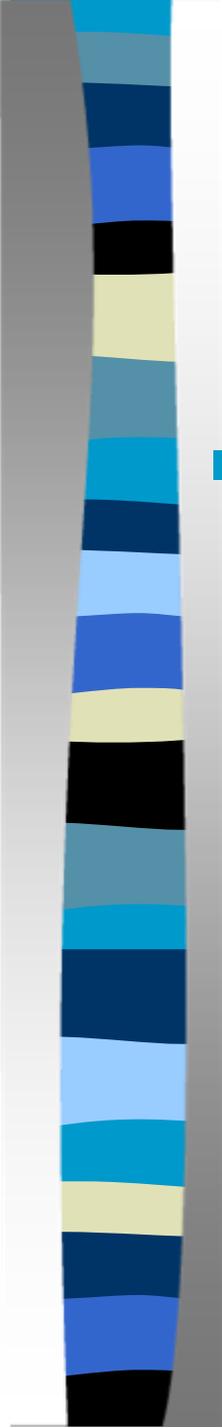
## ■ Examples of Continuous Random Variables

- **Normal Distribution:** The CDF of a normally distributed random variable  $X$  is given by;

$$F_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^x \exp\left(-\frac{(u - \bar{X})^2}{2\sigma_X^2}\right) du$$

and the pdf by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x - \bar{X})^2}{2\sigma_X^2}\right)$$



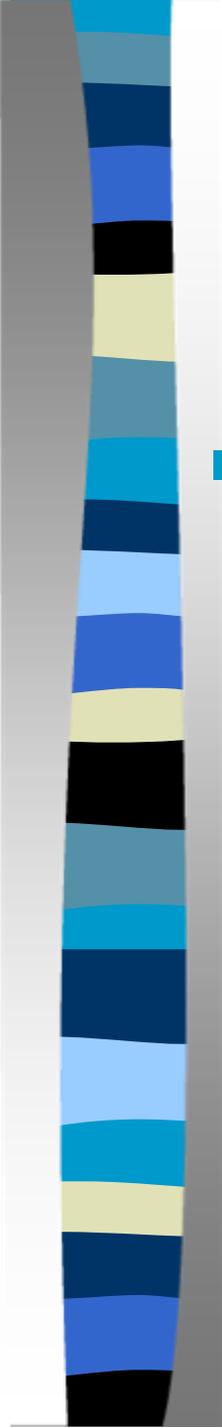
# Essentials for understanding queuing models (10)

## ■ Examples of Continuous Random Variables

- **Exponential Distribution:** This is the most important and easiest to use in queuing theory. The CDF of an exponentially distributed random variable  $X$  is given by;

$$F_X(x) = \begin{cases} 1 - \exp\left(-\frac{x}{\bar{X}}\right), & 0 \leq x < \infty, \\ 0, & \text{Otherwise} \end{cases}$$

$$\text{With } \bar{X} = \begin{cases} \frac{1}{\lambda}, & \text{if } X \text{ represents interarrival times,} \\ \frac{1}{\mu}, & \text{if } X \text{ represents service times.} \end{cases}$$



# Essentials for understanding queuing models (11)

- For an exponentially distributed random variable with parameter  $\lambda$ ,

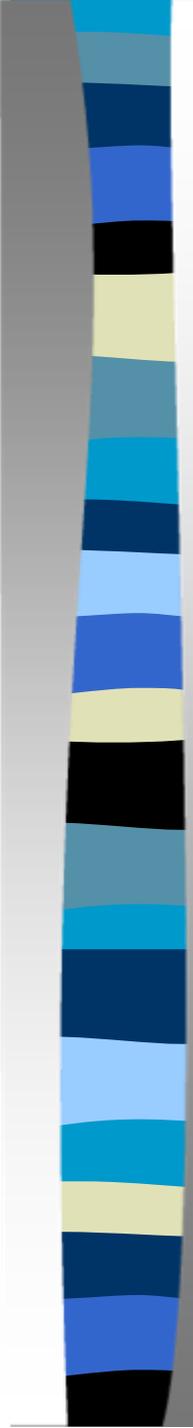
–pdf:  $f_X(x) = \lambda e^{-\lambda x}$ ,

–mean:  $\bar{X} = \frac{1}{\lambda}$ ,

–variance:  $\text{var}(X) = \frac{1}{\lambda^2}$ ,

–Coefficient of variation:  $c_X = 1$ .

**Thus the exponential distribution is completely determined by its mean.**



# Essentials for understanding queuing models (12)

- Important properties of the exponential distribution
  - It is completely determined by its mean as above.
  - It is memoryless (only continuous distribution that has this)

$$P(X \leq u + t | X > u) = 1 - \exp\left(-\frac{t}{X}\right) = P(X \leq t).$$

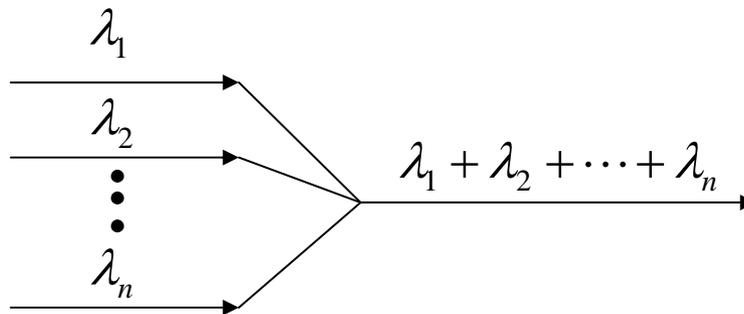
Example of bus arrivals at a bus stop.

- Special relation to the discrete Poisson random variable
  - *If interarrival times are exponential*
  - *If successive interarrival times are independent with identical mean  $\bar{X}$ .*
  - *Then number of arrivals in a fixed time interval has a poisson distribution with parameter  $\alpha = t/\bar{X}$ .*

# Essentials for understanding queuing models (13)

- Important properties of the exponential distribution

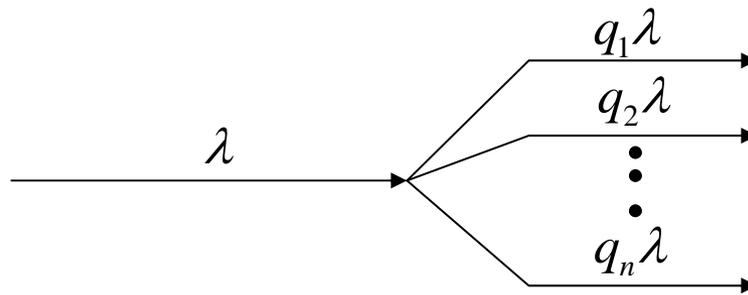
- **Merging of Poisson processes:** If we merge  $n$  Poisson processes with interarrival time distribution of  $1 - e^{-\lambda_i t}$ ,  $1 \leq i \leq n$ , into a single process, the result is a Poisson process for which the interarrival times have the distribution  $1 - e^{-\lambda t}$  with  $\lambda = \sum_{i=1}^n \lambda_i$



# Essentials for understanding queuing models (14)

## Important properties of the exponential distribution

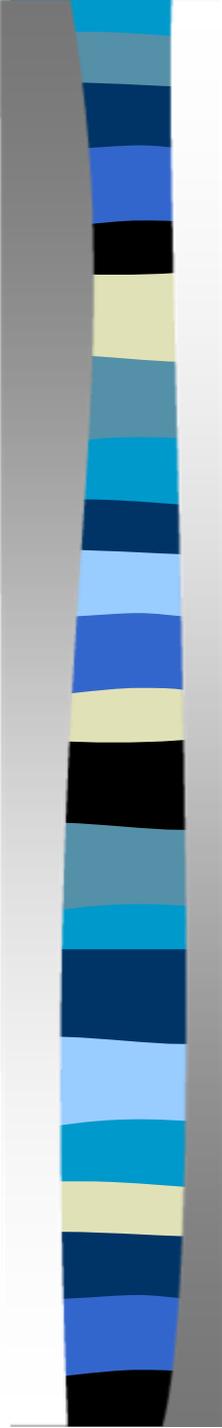
- **Splitting of Poisson processes:** A Poisson process can be split into  $n$  Poisson processes if the interarrival time distribution is exponential,  $1 - e^{-\lambda t}$ .



- The  $i$ th sub-process has an interarrival time distribution of

$$1 - e^{-q_i\lambda t}, \quad 1 \leq i \leq n.$$

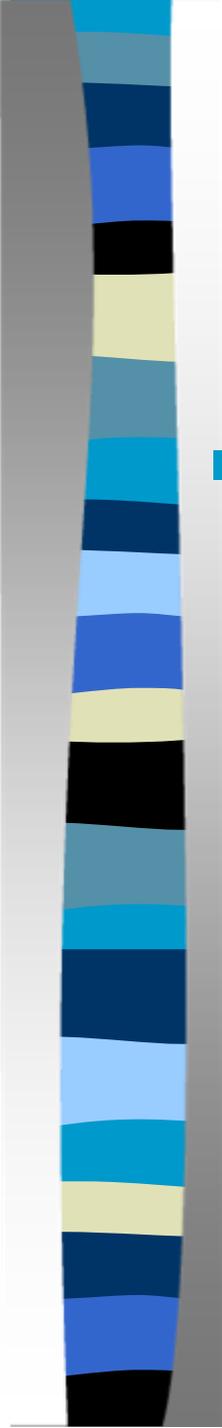
- Note that though the exponential distribution is very useful, it is not always a good approximation of observed distributions.



# Essentials for understanding queuing models (15)

## ■ Introduction to Markov Chains

- Markov processes provide a flexible and powerful means for the description and analysis of dynamic system properties.
- Markov processes constitute the fundamental theory underlying the concept of queuing systems.
- Each queuing system can be mapped onto an instance of a Markov process and then mathematically evaluated in terms of this process.



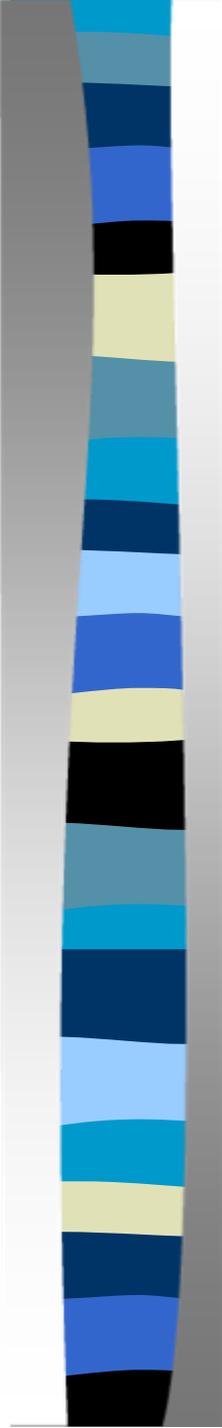
# Essentials for understanding queuing models (16)

## ■ Stochastic and Markov Processes

- A *stochastic process* is defined as a family of random variables  $\{X_t : t \in T\}$  where each random variable  $X_t$  is indexed by parameter  $t \in T$ , which is usually called the ***time parameter*** if

$$T \subseteq \mathbb{R}_+ = [0, \infty ]$$

- The set of all possible values of  $X_t$  (for each  $t \in T$ ) is known as the ***state space***  $S$  of the *stochastic process*.
- *Markov processes* constitute a special, perhaps the most important, subclass of stochastic processes.



# Essentials for understanding queuing models (17)

## ■ Types of stochastic processes

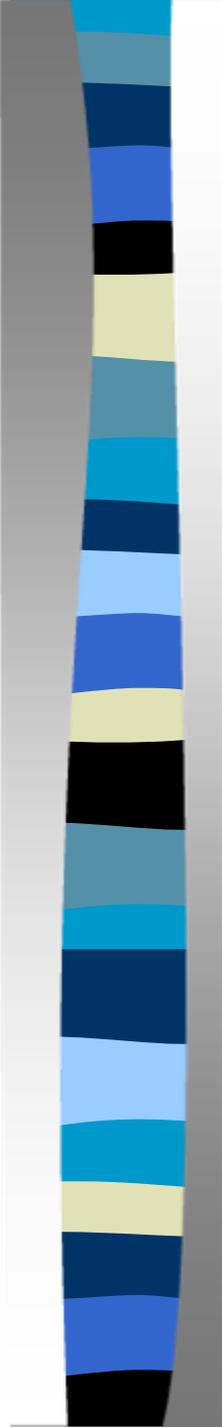
### – Discrete-Parameter processes

- *Involves a discrete parameter set  $T$  commonly represented by (a subset of)  $\mathbb{N}_0 = \{0, 1, \dots\}$ .*

### – Continuous-Parameter processes

- *Can be probabilistically characterised by the joint (cumulative) distribution function (CDF)  $F_X(s; t)$  for a given set of random variables  $\{X_{t_1}, X_{t_2}, \dots, X_{t_n}\}$ , parameter vector  $t = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$ , and state vector  $s = (s_1, s_2, \dots, s_n) \in \mathbb{R}^n$ , where  $t_1 < t_2 < \dots < t_n$  given by;*

$$F_X(s, t) = P(X_{t_1} \leq s_1, X_{t_2} \leq s_2, \dots, X_{t_n} \leq s_n).$$



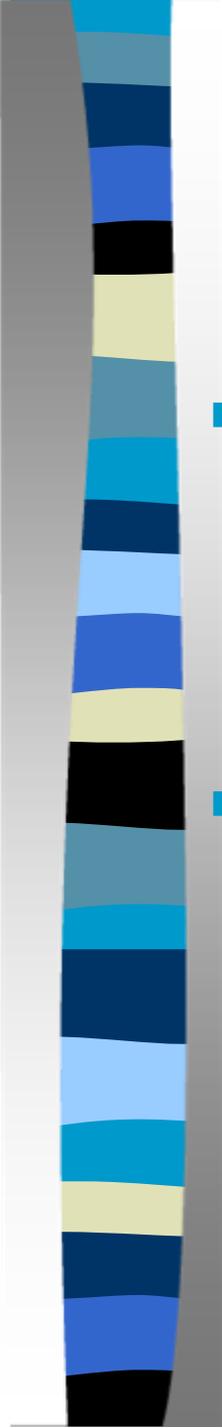
# Essentials for understanding queuing models (18)

## ■ The Markov property;

- A stochastic process  $\{X_t : t \in T\}$  constitutes a *Markov process* if for all  $0 = t_0 < t_1 < \dots < t_n < t_{n+1}$  and all  $s \in S$  the CDF of  $X_{t_{n+1}}$  depends only on the last previous value  $X_{t_n}$  and not on the earlier values  $X_{t_0}, X_{t_1}, \dots, X_{t_{n-1}}$  :

$$P\left(X_{t_{n+1}} \leq s_{n+1} \mid X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \dots, X_{t_0} = s_0\right) = P\left(X_{t_{n+1}} \leq s_{n+1} \mid X_{t_n} = s_n\right).$$

- Thus any stochastic process with the Markov property is a Markov process.
- **Example:** Card games verses dice games



# Essentials for understanding queuing models (19)

## ■ Homogeneity of Markov processes

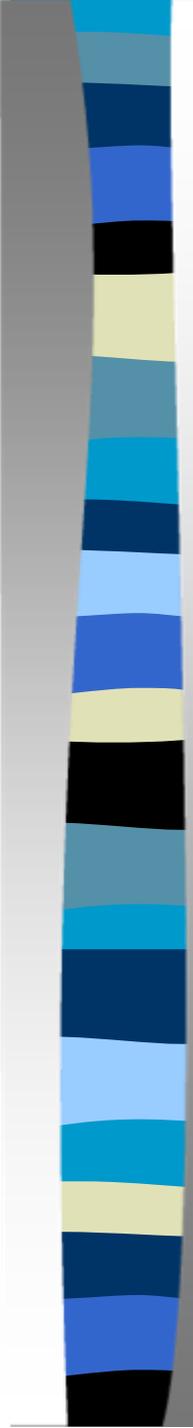
- Time homogeneous (or time independent)

$$P\left(X_{t_{n+1}} \leq s_{n+1} \mid X_{t_n} = s_n\right) = P\left(X_{t_{n+1}-t_n} \leq s_{n+1} \mid X_{t_{n-1}} = s_n\right)$$

- Time non-homogeneous (or time dependent)

## ■ Parameters and state spaces of Markov processes

- Discrete-parameter Markov processes (Set  $T$  is discrete)
  - *May have discrete or continuous state spaces*
- Continuous-parameter Markov processes (Set  $T$  is continuous)
  - *May have discrete or continuous state spaces*
- Markov processes with discrete state spaces are usually called **Markov Chains**.



# Essentials for understanding queuing models (20)

- Two main types of Markov Chains;

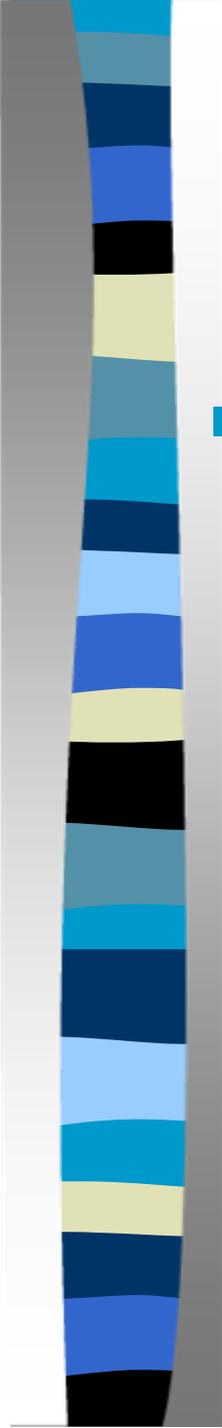
- **Discrete-Time Markov Chains (DTMC)**

A given stochastic process  $\{X_0, X_1, \dots, X_{n+1}, \dots\}$  at the consecutive points of observation  $0, 1, \dots, n+1$  constitutes a DTMC if the following relation on the *conditional pmf* holds;

$$P(X_{n+1} = s_{n+1} | X_n = s_n, X_{n-1} = s_{n-1}, \dots, X_0 = s_0) = P(X_{n+1} = s_{n+1} | X_n = s_n)$$

$P(X_{n+1} = s_{n+1} | X_n = s_n)$  is the conditional pmf of transitions from state  $s_n$  at time step  $n$  to state  $s_{n+1}$  at time step  $n+1$ . Or

$$p_{ij}^{(1)}(n) = P(X_{n+1} = s_{n+1} = j | X_n = s_n = i)$$



# Essentials for understanding queuing models (21)

## ■ Evolution of a DTMCs

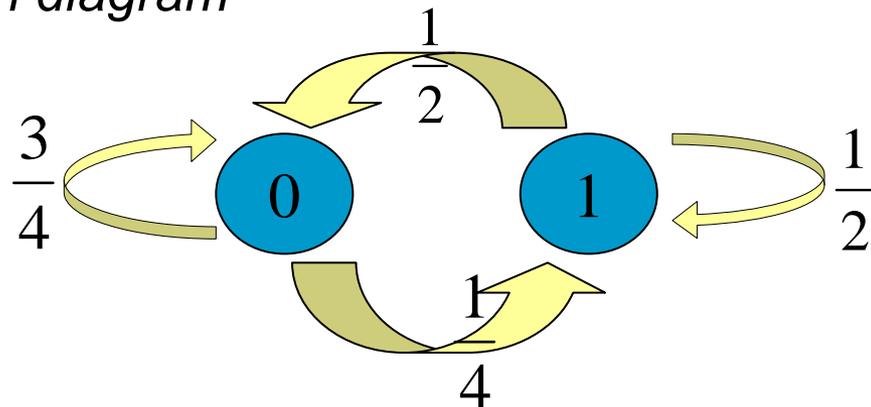
- Given an initial state  $s_0$ , a DTMC evolves over time, that is, step by step, according to *one-step transition probabilities*.
- The one-step transition probabilities are usually summarized in a non-negative, stochastic transition matrix  $P$ :

$$P = P^{(1)} = [p_{ij}] = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

- The elements of each row sum up to 1

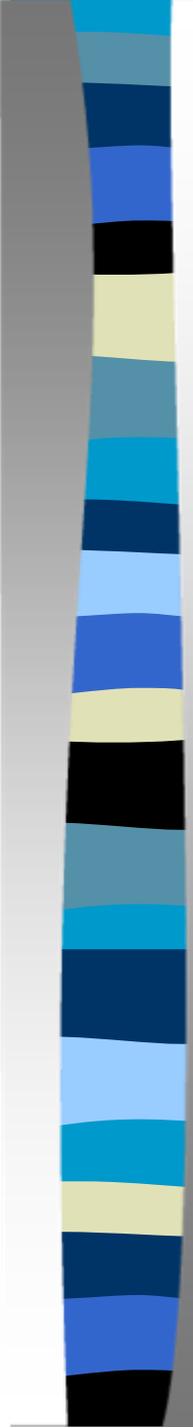
# Essentials for understanding queuing models (22)

- Graphically, a finite-state DTMC is represented by a *state transition diagram*



- The one-step transition probability matrix will be given by;

$$P^{(1)} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{pmatrix}$$



# Essentials for understanding queuing models (23)

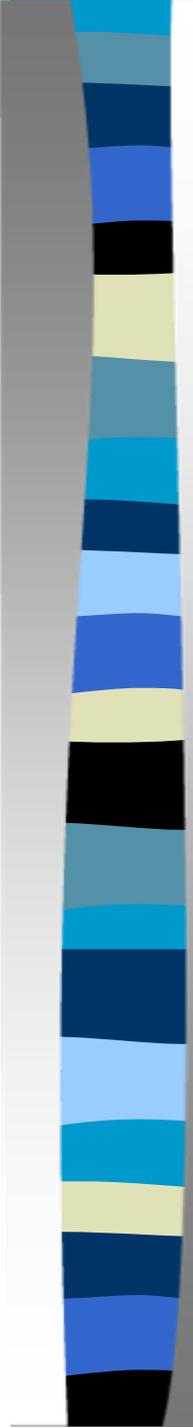
- A very simple weather model
  - The probabilities of weather conditions, given the weather on the preceding day, can be represented by the state transition matrix;

$$P^{(1)} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$$

*State 0 = Sunny*

*State 1 = Rainy*

- That is a sunny day is 90% likely to be followed by another sunny day, and a rainy day is 50% likely to be followed by another rainy day.



# Essentials for understanding queuing models (24)

- Predicting the weather

- If the weather on day 0 is known to be sunny, then it is considered 100% sunny and 0% rainy. This is represented by the vector;

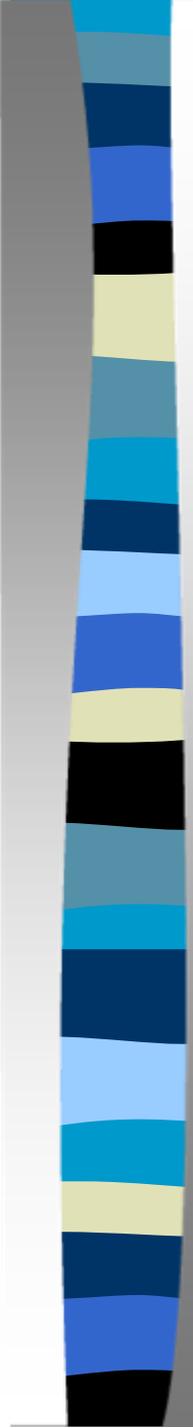
$$X^{(0)} = (1 \quad 0)$$

- The weather on day 1 can be predicted by;

$$X^{(1)} = X^{(0)}P = (1 \quad 0) \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} = (0.9 \quad 0.1)$$

- The weather on day 2 can also be predicted by;

$$X^{(2)} = X^{(1)}P = X^{(0)}P^2 = (0.9 \quad 0.1) \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} = (0.86 \quad 0.14)$$



# Essentials for understanding queuing models (25)

- General rules for day  $n$  are;

$$X^{(n)} = X^{(n+1)}\mathbf{P}$$

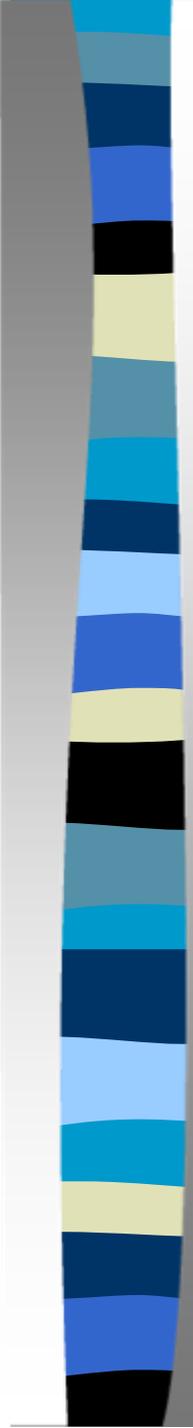
$$X^{(n)} = X^{(0)}\mathbf{P}^n$$

- Steady state of the weather

- In the steady state, number of days  $n$  goes to infinity and we have the solution

$$(p_0 \quad p_1) = (0.833 \quad 0.167)$$

- In conclusion, in the long term, 83% of days are sunny. 35

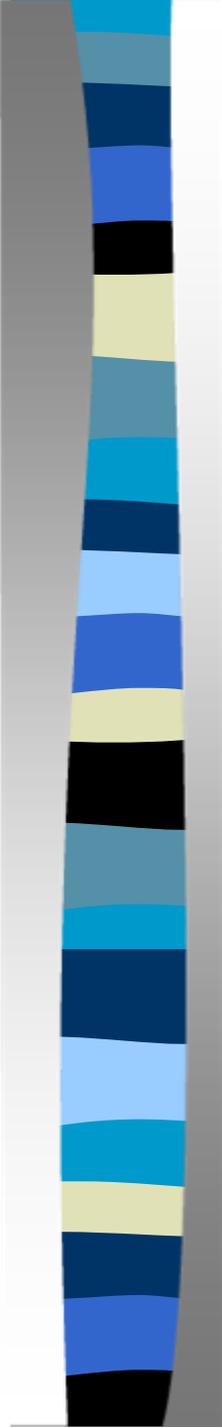


# Essentials for understanding queuing models (26)

- **Two main types of Markov Chains (cont.);**
  - **Continuous-Time Markov Chains (CTMC)**

A given stochastic process  $\{X_t : t \in T\}$  constitutes a CTMC if for all arbitrary  $t_i \in \mathbb{R}_0^+$ ,  $0 = t_0 < t_1 < \dots < t_n < t_{n+1}$ ,  $\forall n \in \mathbb{N}$ , and  $\forall s_i \in S = \mathbb{R}_0$  for the *conditional pmf* the following relation holds;

$$P\left(X_{t_{n+1}} = s_{n+1} \mid X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \dots, X_{t_0} = s_0\right) = P\left(X_{t_{n+1}} = s_{n+1} \mid X_{t_n} = s_n\right).$$



# A single server queuing model

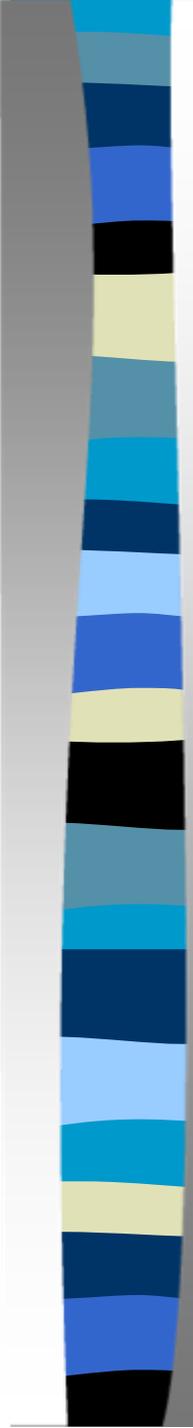
## ■ Kendall's notation

### – **A/B/m-Queuing discipline**

- *A* indicates the distribution of the inter-arrival times
- *B* denotes the distribution of the service times
- *m* is the number of servers ( $m \geq 1$ )

## ■ Markovian queues

- **M/M/1 Queue**
- **M/M/ $\infty$  Queues**
- **M/M/1/K Finite Capacity Queue**



# A single server queuing model (2)

- The M/M/c Queue

*Let  $\lambda$  be the average arrival rate*

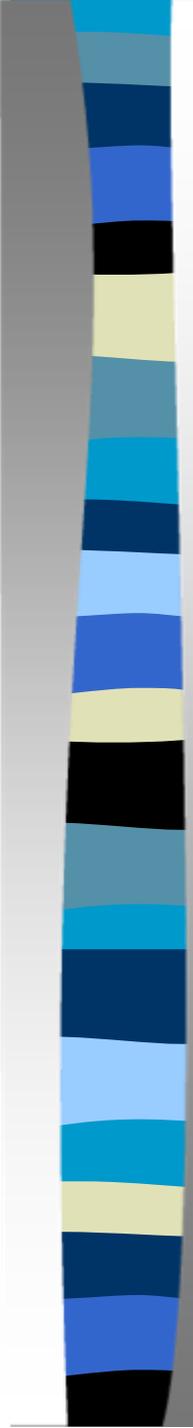
*$\mu$  the average service rate and*

*$\rho = \frac{\lambda}{c\mu}$ , the utilisation factor.*

*$L$  is the number of customers at the workstation*

*$W$  is the expected throughput time for an arbitrary customer.*

- The state of the system is the number of jobs at workstation,  $n$ .



# A single server queuing model (3)

- We denote the probability of  $n$  jobs at the workstation at time  $t$ , by

$$p_t(n)$$

- Thus for steady-state,

$$p_t(n) = p_{t+\delta t}(n)$$

- Hence the time index is dropped when referring to steady-state results.

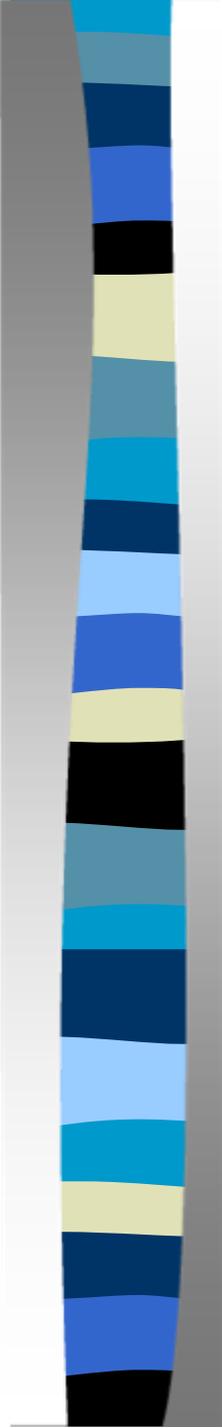
# A single server queuing model (4)

- Some M/M/1 Queuing results

$$\rho = \frac{\lambda}{c\mu} \qquad p(0) = 1 - \rho \qquad \text{Utilisation Factor}$$

$$L_q = \frac{\rho^2}{1 - \rho} \qquad L = \frac{\rho}{1 - \rho} \qquad \begin{array}{l} \text{No. Entities} \\ \text{Waiting} \end{array}$$

$$W_q = \frac{\rho}{\mu(1 - \rho)} \qquad W = \frac{1}{\mu(1 - \rho)} \qquad \begin{array}{l} \text{Expected} \\ \text{Throughput time} \end{array}$$



# A single server queuing model (5)

## ■ Example

- A manufacturing facility operates as a flow shop.
- 
- Arrival rate of orders,  $\lambda = \text{EXPO}(10)$  per week
- Service rate of orders,  $\mu = \text{EXPO}(12)$  per week
- Queuing discipline = FCFS
- Find the average time from order arrival to completion.

# A single server queuing model (6)

## ■ Solution

- We treat the facility as a single server workstation

$$\text{We have } \rho = \frac{\lambda}{c\mu} = \frac{10}{1 \times 12}, \quad p(0) = 1 - \rho = 1 - \frac{10}{12} = \frac{1}{6}$$

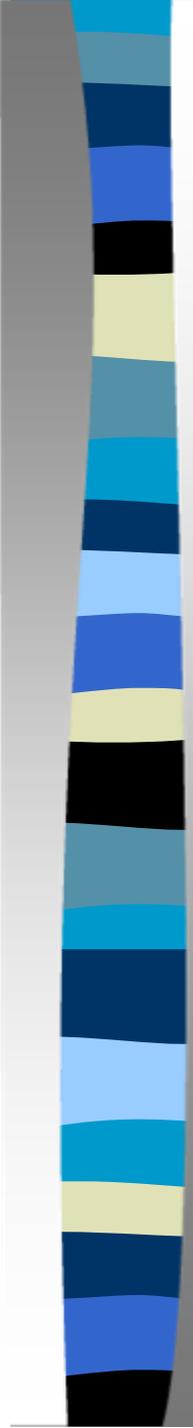
$\therefore$  the facility is idle  $\frac{1}{6}$  of the time.

The throughput time is

$$W = \frac{1}{\mu(1-\rho)} = \frac{1}{12\left(\frac{1}{6}\right)} = 0.5 \text{ weeks}$$

$$\text{Average processing time} = \frac{1}{12}$$

$$\therefore \text{Queuing time, } W_q = 0.5 - \frac{1}{12} = \frac{5}{12}$$



# Queuing networks

- A network has  $M$  workstations with jobs moving between workstations according to their processing sequence.
- Types of queuing networks
  - **Open**
  - **Closed**
  - **Hybrid**