Simulation and Modeling (CSc 302)

By
Surya Bam
Faculty B.Sc CSIT
Academia International College
Gwarko, Lalitpur

Chapter 1

Introduction to Simulation

1. What is simulation?

Simulation is the imitation of the operation of a real-world process or system over time. Simulation involves the generation of an artificial history of the system, and the observation of that artificial history to draw inferences concerning the operating characteristics of the real system that is represented.

Simulation is the numerical technique for conducting experiments on digital computer, which involves logical and mathematical relationships that interact to describe the behavior and the structure of a complex real world system over extended period of time.

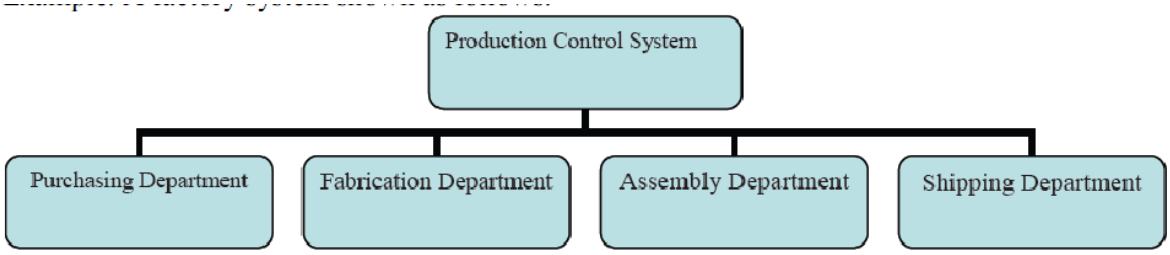
The process of designing a model of a real system, implementing the model as a computer program, and conducting experiments with the model for the purpose of understanding the behavior of the system, or evaluating strategies for the operation of the system.

2. System Concepts

A system is defined as a group f objects that are joined together in some regular intraction or interdependence for the accomplishment of some task. For example: *Production system for manufacturing automobiles.*

A system is usually considered as a set of inter–related factors, which are described as entities activities and have properties or attributes. Processes that cause system changes are called activities. The state of a system is a description of all entities, attributes and the activities at any time.

Example: A factory system shown as follows:



Components of system

2.1 Entity, attribute and activities

An **entity** represents an object that requires explicit definition. An entity can be dynamic in that it moves through the system, or it can be static in that it serves other entities. In the example, the customer is a dynamic entity, whereas the bank teller is a static entity.

An entity may have **attributes** that pertain to that entity alone. Thus, attributes should be considered as local values. In the example, an attribute of the entity could be the time of arrival. Attributes of interest in one investigation may not be of interest in another investigation. Thus, if red parts and blue parts are being manufactured, the color could be an attribute.

Processes that cause system changes are called **activities or events.**
Example

In the bank example, events include the arrival of a customer for service at the bank, the beginning of service for a customer, and the completion of a service.

System	Entities	Attributes	Activities
Traffic	Cars, bus, pedestrian	Speed, model	Driving, walking
Bank	Customer	Balance	Depositing, Arrival of customer, …..
Supermarket	Customers	Shopping List	Checking out…

There are both internal and external events, also called endogenous and exogenous events, respectively. For instance, an endogenous event in the example is the beginning of service of the customer since that is within the system being simulated. An exogenous event is the arrival of a customer for service since that occurrence is outside of the simulation.

2.2 State variables

The state of a system is defined to be that collection of variables necessary to describe the system at any time, relative to the objectives of the study. In the study of a bank, possible state variables are the number of busy tellers, the number of customers waiting in line or being served, and the arrival time of the next customer.
So the system state variables are the collection of all information needed to define what is happening within the system to a sufficient level (i.e., to attain the desired output) at a given point in time.

Example

System	*Entities*	*Attributes*	*Activities*	*Events*	*State Variables*
Banking	Customers	Checking account balance	Making deposits	Arrival; departure	Number of busy tellers; number of customers waiting
Rapid rail	Riders	Origination; destination	Traveling	Arrival at station; arrival at destination	Number of riders waiting at each station; number of riders in transit
Production	Machines	Speed; capacity; breakdown rate	Welding; stamping	Breakdown	Status of machines (busy, idle, or down)
Communications	Messages	Length; destination	Transmitting	Arrival at destination	Number waiting to be transmitted
Inventory	Warehouse	Capacity	Withdrawing	Demand	Levels of inventory; backlogged demands

2.3 Open System/Close System

A system with exogenous activities is considered as open system and a system with strict endogenous activities is called a closed system.

2.4 System Environment

The external components which interact with the system and produce necessary changes are said to constitute the system environment. In modeling systems, it is necessary to decide on the boundary between the system and its environment. This decision may depend on the purpose of the study.

Example: In a factory system, the factors controlling arrival of orders may be considered to be outside the factory but yet a part of the system environment. When, we consider the demand and supply of goods, there is certainly a relationship between the factory output and arrival of orders. This relationship is considered as an activity of the system.

Endogenous System

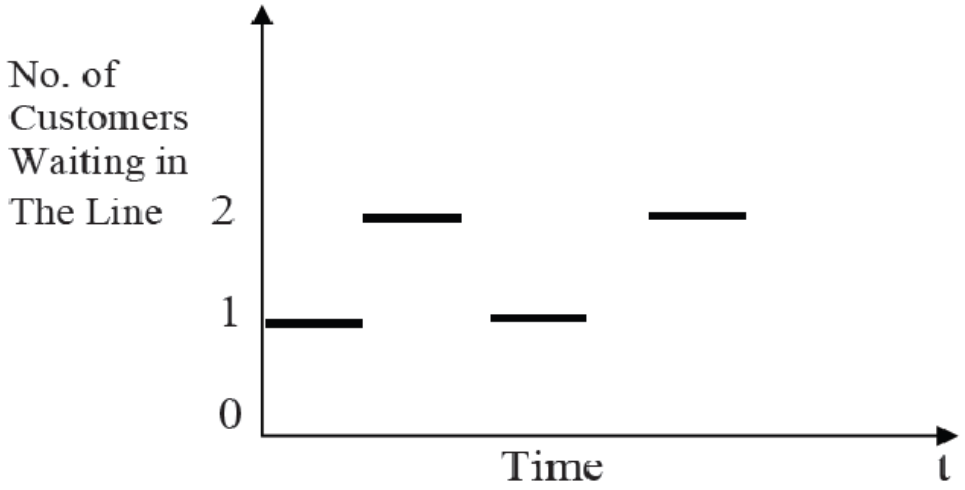
The term endogenous is used to describe activities and events occurring within a system. Example: Drawing cash in a bank.

Exogenous System

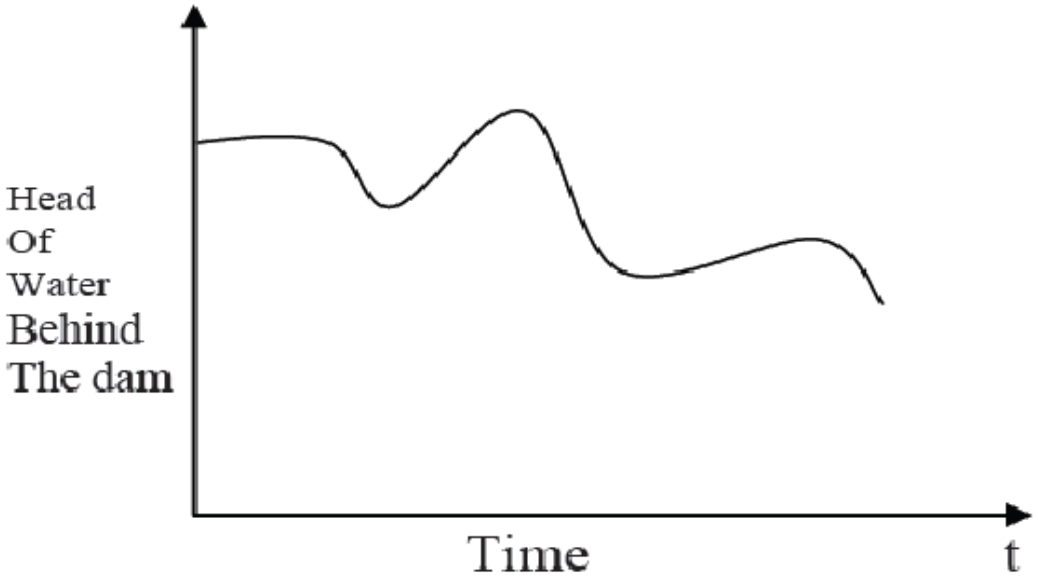
The term exogenous is used to describe activities and events in the environment that affect the system. Example: Arrival of customers.

3. Discrete and continuous system

Discrete system is one in which the state variables changes only at a discrete set of time. For example: banking system in which no of customers (state variable) changes only when a customer arrives or service provided to customer i.e customer depart form system. The figure below show how no of customer changes only at discrete points in time



Continuous system is one in which the state variables change continuously over time. For example, during winter seasons level of which water decreases gradually and during rainy season level of water increase gradually. The change in water level is continuous. The figure below shows the change of water level over time.



4. System Modeling

A model is defined as a representation of a system for the purpose of studying the system. It is necessary to consider only those aspects of the system that affect the problem under investigation. These aspects are represented in a model, and by definition it is a simplification of the system. The aspect of system that affect the problem under investigation, are represented in a model of the system. Therefore model is the simplification of the real system. There is no unique

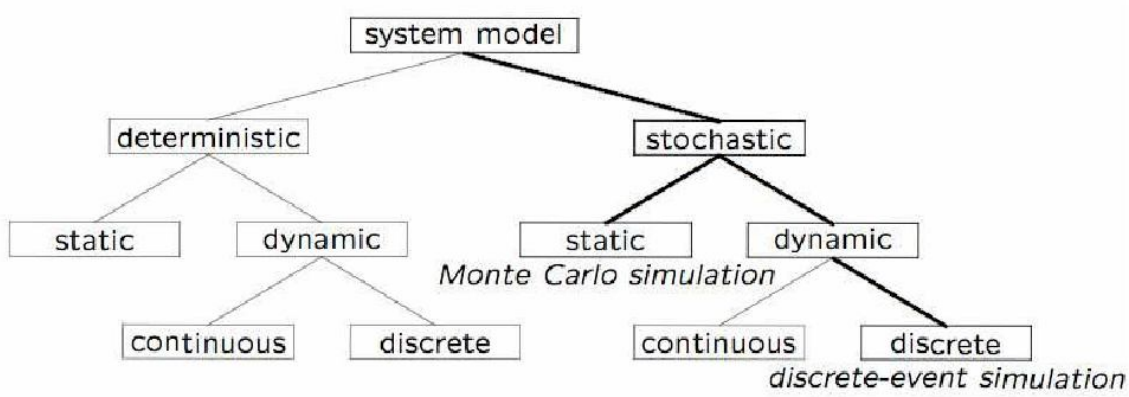
model of a system. Different models of the same system will be produced by different system analysts who are interested in different aspect of system. The task of deriving a model of a system may be divided broadly into two subtasks: Establishing model parameter and supplying data. Establishing model structure determines system boundary and identifies the entities, attributes, activities and events of a system. Supplying data provides value contained an attribute and define relationships involved in the activities.

Types of Model

The various types of models are shown in figure below.

- Mathematical and Physical Model
- Static Model
- Dynamic Model
- Deterministic Model
- Stochastic Model
- Discrete Model
- Continuous Model

The chart to represent different model in a hierarchy is as shown below



Physical model

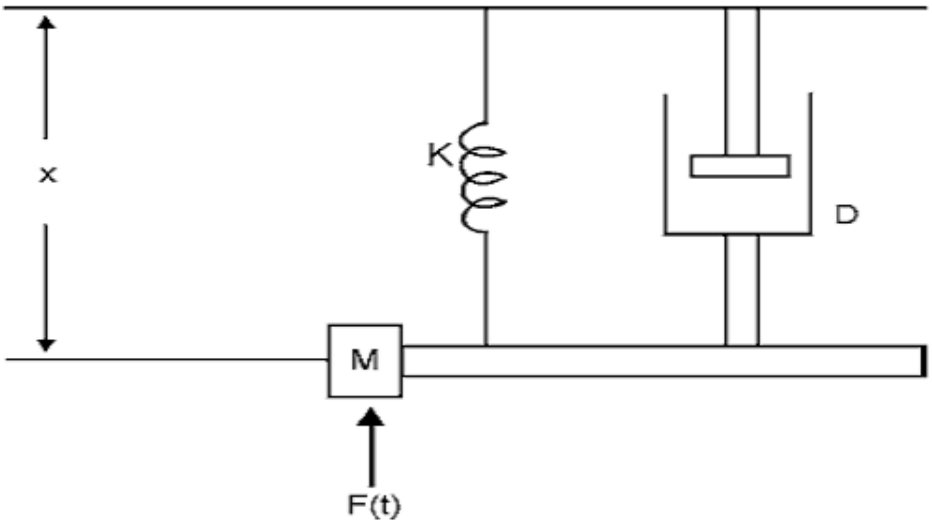
These models are based on some analogy between mechanical and electrical system The system attributes are represented by physical measures such as voltage. The system activities are represented by physical laws.

Physical models are of two types, static and dynamic. Static physical model is a scaled down model of a system which does not change with time. An architect before constructing a building makes a scaled down model of the building, which reflects all it rooms, outer design and other important features. This is an example of static physical model. Similarly for conducting trials in water, we make small water tanks, which are replica of sea, and fire small scaled down shells in them. This tank can be treated as a static physical model of ocean. Dynamic physical models are ones which change with time or which are function of time. In wind tunnel, small aircraft models (static models) are kept and air is blown over them with different velocities and pressure profiles

are measured with the help of transducers embedded in the model. Here wind velocity changes with time and is an example of dynamic physical model.

Let us take an example of hanging wheel of a stationary truck and analyze its motion under various forces. Consider a wheel of mass *M*, suspended in vertical direction, a force *F*(*t*), which varies with time, is acting on it. Mass is connected with a spring of stiffness *K*, and a piston with damping factor *D*. When force *F* (*t*), is applied, mass *M* oscillates under the action of these three forces.

This model can be used to study the oscillations in a motor wheel. Figure 1.2 shows such a system. This is a discrete physical static model. Discrete in a sense, that one can give discrete values *F* and observe the oscillations of wheel with some measuring equipment. When force is applied on it, which is a function of time, this discrete physical static model becomes dynamic model. Parameters *K* and *D* can also be adjusted in order to get controlled oscillations of the wheel. This type of system is called spring-mass system or wheel suspension. Load on the beams of a building can be studied by the combination of spring-mass system.



Mathematical Model

It uses symbolic notation and mathematical equation to represent system. The system attributes are represented by variables and the activities are represented by mathematical function.

Example: f(x) = mx+c is a mathematical model of a line.

Static Model

Static models can only show the values that the system attributes value does not change over time. Example: Scientist has used models in which sphere represents atom, sheet of metal to connect the sphere to represent atomic bonds. Graphs are used to model the various system based on network. A map is also a kind of graph. These models are sometimes said to be iconic models and are of kind static physical models.

Dynamic Model

Dynamic models follow the changes over time that result from system activities. The mechanical and electrical systems are the example of dynamic system. Generally, dynamic models involve the computation of variable value over time and hence they are represented by differential equations.

Analytical Models:

In mathematical model, we can differentiate the model on the basis of solution technique used to solve the model. Analytical technique means using deductive reasoning of mathematical theory to solve a model. Such models are known as analytical model.

Numerical models

Numerical models involve applying computational process to solve equations. For example: we may solve differential equation numerically when the specific limit of variable is given. The analytical methods to produce solution may take situation numerical methods are preferred.

Deterministic Model

Contains no random variables. They have a known set of inputs which will result in a unique set of outputs. Ex: Arrival of patients to the Dentist at the scheduled appointment time.

Stochastic Model

Has one or more random variable as inputs. Random inputs leads to random outputs. Ex: Simulation of a bank involves random inter-arrival and service times.

Principles used in Modeling

Guidelines used in modeling

- It is not possible provide rule by which models are built. But a number of guidelines can be stated.
- The different viewpoints from which we can judge whether certain info. Should be included as excluded in models are:

1. **Block –Building:** The description of system should be organized as a sequence of blocks. It simplifies the interaction between block within system. Then it will be easy to describe the whole system in terms of interaction between the block and can be represented graphically as simple block diagram. For example:-the block of factory system.

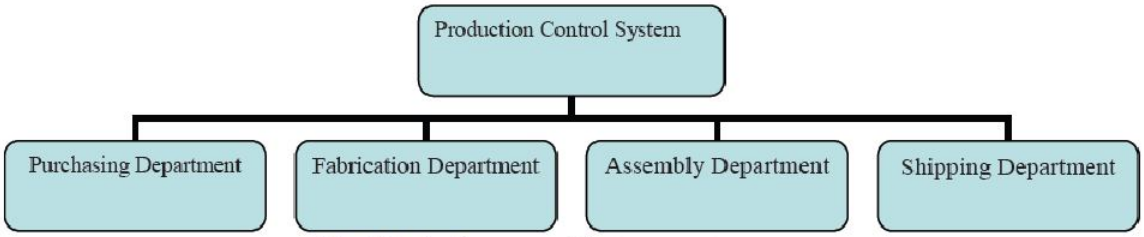


Fig: Block diagram of factory system

2. **Relevance:** The model should only include relevant information. For example, if the factory system study aims to compare the efficient of different operating rules efficiency it is not relevant to consider the mining of employee as an activity. Irrelevant information should not include despite of being no harm because it increases the complexity of model and takes more time and effort to solve model.

3. **Accuracy:** The gathered information should be accurate as well. For example in aircraft system the accuracy as movement of the aircraft depends upon the representations of airframe such as a rigid body.

4. **Aggregation:** It should be considered that to which numbers of individual entities can be grouped into a block. For example in factory system, different department are grouped together handled by production manger.

Distributed lag model

Models that have the property of changing only at fixed interval of time and based on current values of variables on other current values of variables are called distributed lag model. In economic studies some economic data are collected over uniform time interval such as a month or year. This model consists of linear algebraic equations that represent continuous system but data are available at fixed points in time.

For example: Mathematical model of national economy

Let

C=consumption

I=investment

T=Taxes

G=government expenditures

Y=national income Then

C=20+0.7(Y-T)

I=2+0.1Y

T=0.2Y

Y=C+I+G

All the equation are expressed in billions of rupees. This is static model and can be made dynamic by lagging all the variables as follows $C=20+0.7(Y_{-1}-T)$
$I=2+0.1Y_{-1}$
$T=0.2Y_{-1}$
$Y=C_{-1}+I_{-1}+G_{-1}$
Any variable that can be expressed in the form of its current value and one or more previous value is called lagging variable. And hence this model is given the name distributed lag model. The variable in a previous interval is denoted by attaching –n suffix to the variable. Where -n indicate the nth interval.

Advantages of distributed lag model

 Simple to understand and can be computed by hand, computers are extensively used to run them.
 there is no need for special programming language to organize simulation task.

When simulation is appropriate Tool?

The availability of special-purpose simulation languages, massive computing capabilities at a decreasing cost per operation, and advances in simulation methodologies have made simulation one of the most widely used and accepted tools in operations research and systems analysis. Simulation can be used for the following purposes:
1. Simulation enables the study of, and experimentation with, the internal interactions of a complex system, or of a subsystem within a complex system.
2. Informational, organizational, and environmental changes can be simulated, and the effect of these alterations on the model's behavior can be observed.
3. The knowledge gained in designing a simulation model may be of great value toward suggesting improvement in the system under investigation.
4. By changing simulation inputs and observing the resulting outputs, valuable insight may be obtained into which variables are most important and how variables interact.
5. Simulation can be used to experiment with new designs or policies prior to implementation, so as to prepare for what may happen.
6. Simulation can be used to verify analytic solutions.
7. By simulating different capabilities for a machine, requirements can be determined.
8. Simulation models designed for training allow learning without the cost and disruption of on-the-job learning.

When the simulation is not appropriate?
1. Simulation should not be used when the problem can be solved using common sense.
2. Simulation should not be used if the problem can be solved analytically.
3. Simulation should not be used if it is easier to perform direct experiments.etc
4. Simulation should be used when the problem cannot be solved using common sense.
5. Simulation should not be used, if the costs exceed savings.

6. Simulation should not be performed, if the resources or time are not available.
7. If no data is available, not even estimate simulation is not advised.
8. If there is not enough time or the persons are not available, simulation is not appropriate.
9. If managers have unreasonable expectation say, too much soon – or the power of simulation is over estimated, simulation may not be appropriate.
10. If system behavior is too complex or cannot be defined, simulation is not appropriate.

Advantages of simulation
1. Simulation can also be used to study systems in the design stage.
2. Simulation models are run rather than solver.
3. New policies, operating procedures, decision rules, information flow, etc can be explored without disrupting the ongoing operations of the real system.
4. New hardware designs, physical layouts, transportation systems can be tested without committing resources for their acquisition.
5. Hypotheses about how or why certain phenomena occur can be tested for feasibility.
6. Time can be compressed or expanded allowing for a speedup or slowdown of the phenomena under investigation.
7. Insight can be obtained about the interaction of variables.
8. Insight can be obtained about the importance of variables to the performance of the system.
9. Bottleneck analysis can be performed indication where work-in process, information materials and so on are being excessively delayed.
10. A simulation study can help in understanding how the system operates rather than how individuals think the system operates.
11. "what-if" questions can be answered. So it is useful in the design of new systems.

Disadvantage of simulation
1. Model building requires special training.
2. Simulation results may be difficult to interpret.
3. Simulation modeling and analysis can be time consuming and expensive.
4. Simulation is used in some cases when an analytical solution is possible or even preferable.

Applications of Simulation

Manufacturing Applications
1. Analysis of electronics assembly operations
2. Design and evaluation of a selective assembly station for high precision scroll compressor shells.
3. Comparison of dispatching rules for semiconductor manufacturing using large facility models.

4. Evaluation of cluster tool throughput for thin-film head production.
5. Determining optimal lot size for a semiconductor backend factory.
6. Optimization of cycle time and utilization in semiconductor test manufacturing.
7. Analysis of storage and retrieval strategies in a warehouse.
8. Investigation of dynamics in a service oriented supply chain.
9. Model for an Army chemical munitions disposal facility.

Semiconductor Manufacturing

1. Comparison of dispatching rules using large-facility models.
2. The corrupting influence of variability.
3. A new lot-release rule for wafer fabs.
4. Assessment of potential gains in productivity due to proactive retied management.
5. Comparison of a 200 mm and 300 mm X-ray lithography cell.
6. Capacity planning with time constraints between operations.

Military Applications

1. Modeling leadership effects and recruit type in a Army recruiting station.
2. Design and test of an intelligent controller for autonomous underwater vehicles.
3. Modeling military requirements for non war fighting operations.
4. Multi trajectory performance for varying scenario sizes.
5. Using adaptive agents in U.S. Air Force retention.

Steps on simulation Study

1. Problem formulation

Every study begins with a statement of the problem, provided by policy makers. Analyst ensures it's clearly understood. If it is developed by analyst policy makers should understand and agree with it.

2. Setting of objectives and overall project plan

The objectives indicate the questions to be answered by simulation. At this point a determination should be made concerning whether simulation is the appropriate methodology. Assuming it is appropriate, the overall project plan should include

- A statement of the alternative systems
- A method for evaluating the effectiveness of these alternatives
- Plans for the study in terms of the number of people involved
- Cost of the study
- The number of days required to accomplish each phase of the work with the anticipated results.

Model conceptualization

The construction of a model of a system is probably as much art as science. The art of modeling is enhanced by ability:

- To abstract the essential features of a problem
- To select and modify basic assumptions that characterizes the system
- To enrich and elaborate the model until a useful approximation results

Thus, it is best to start with a simple model and build toward greater complexity. Model conceptualization enhances the quality of the resulting model and increases the confidence of the model user in the application of the model.

Data collection
There is a constant interplay between the construction of model and the collection of needed input data. It is done in the early stages. Objective kind of data are collected.

Model translation
Real-world systems result in models that require a great deal of information storage and computation. It can be programmed by using simulation languages or special purpose simulation software. Simulation languages are powerful and flexible. Simulation software models development time can be reduced.

Verified
It pertains to the computer program and checking the performance. If the input parameters and logical structure are correctly represented, verification is completed.

Validated
It is the determination that a model is an accurate representation of the real system. It is achieved through calibration of the model. The calibration of model is an iterative process of comparing the model to actual system behavior and the discrepancies between the two.

Experimental Design
The alternatives that are to be simulated must be determined. Which alternatives to simulate may be a function of runs? For each system design, decisions need to be made concerning
· Length of the initialization period
· Length of simulation runs
· Number of replication to be made of each run

Production runs and analysis
They are used to estimate measures of performance for the system designs that are being simulated.

More runs
Based on the analysis of runs that have been completed, the analyst determines if additional runs are needed and what design those additional experiments should follow.

Documentation and reporting
Two types of documentation.
· Program documentation
· Process documentation

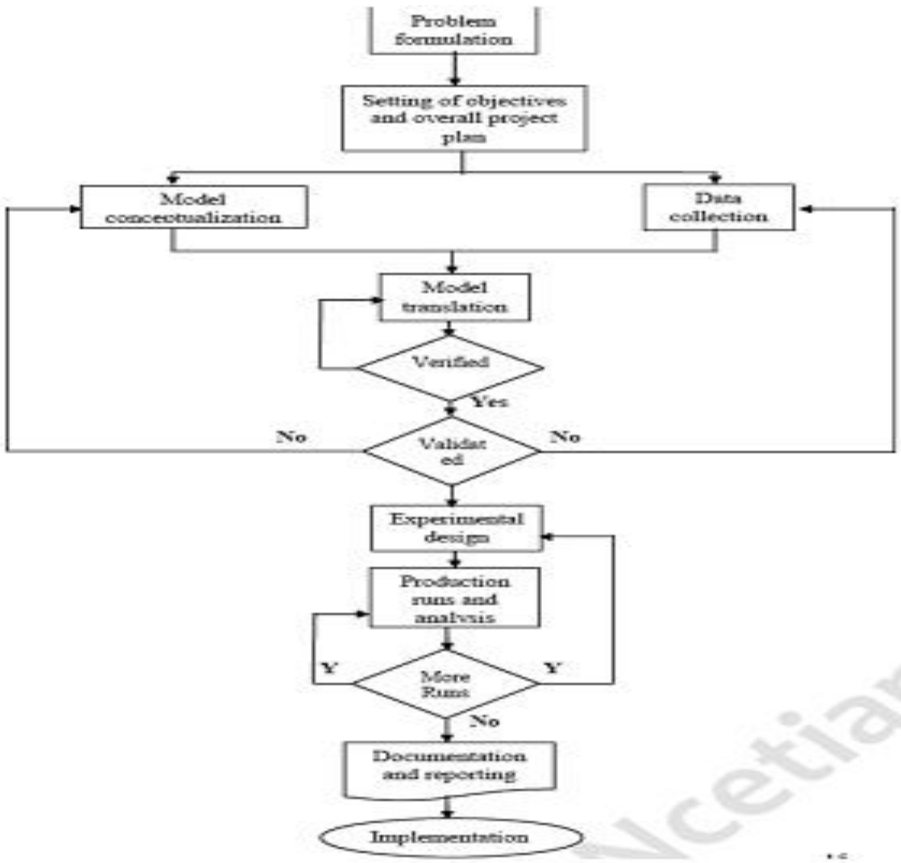
Program documentation
Can be used again by the same or different analysts to understand how the program operates. Further modification will be easier. Model users can change the input parameters for better performance.

Process documentation
Gives the history of a simulation project. The result of all analysis should be reported clearly and concisely in a final report. This enables to review the final formulation and alternatives, results of the experiments and the recommended solution to the problem. The final report provides a vehicle of certification.

Implementation
Success depends on the previous steps. If the model user has been thoroughly involved and understands the nature of the model and its outputs, likelihood of a vigorous implementation is enhanced. The simulation model building can be broken into 4 phases.

Phase of Simulation Study
I Phase
· Consists of steps 1 and 2
· It is period of discovery/orientation
· The analyst may have to restart the process if it is not fine-tuned
· Recalibrations and clarifications may occur in this phase or another phase.
II Phase
· Consists of steps 3,4,5,6 and 7
· A continuing interplay is required among the steps
· Exclusion of model user results in implications during implementation
III Phase
· Consists of steps 8,9 and 10
· Conceives a thorough plan for experimenting
· Discrete-event stochastic is a statistical experiment
· The output variables are estimates that contain random error and therefore proper statistical analysis is required.

IV Phase
· Consists of steps 11 and 12
· Successful implementation depends on the involvement of user and every steps successful completion.



Hybrid Simulation: For most studies, the system under study is clearly either of continuous or discrete nature and it is the determining factor in deciding whether to use an analog or digital computer for system simulation.

If the system being simulated is an interconnection of continuous and discrete subsystem, then such system simulation is known as hybrid simulation. Such hybrid system can be digital computer being linked together

Hybrid simulation required high speed converters to transform signals from analog to digital from and vice –versa.

Real time simulation: In real time simulation, actual device (which are part of a system) are used in conjunction with either digital computer or hybrid computer. It provides the simulation of the points of systems that do not exist or that cannot be easily used in an experiment.

i.e. the basic idea of real time simulation is „uses the actual part if they are appropriate to use in experiment otherwise use the simulation of the points of the system".

A well-known examples is "simulation to train pilots". It uses the devices for training pilots by giving them the impression that is at the control of an aircraft.

It requires real time simulator of the plane its control system, the weather and other environmental conditions. Sometimes, real time simulation also refers to a computer model of a physical system that can execute at the same rate as actual system can. For example: if a machine takes 10 minutes to fill a tank in real world, the simulation also would take 10 minutes.

Real time simulation of an engineering system becomes possible when replace physical device with virtual device are.

(Note : please read about the mathematical and physical model form the Gorden book with example of wheel suspension and electric capacity)

Questions: Differentiate between dynamic physical models and static physical models with example.

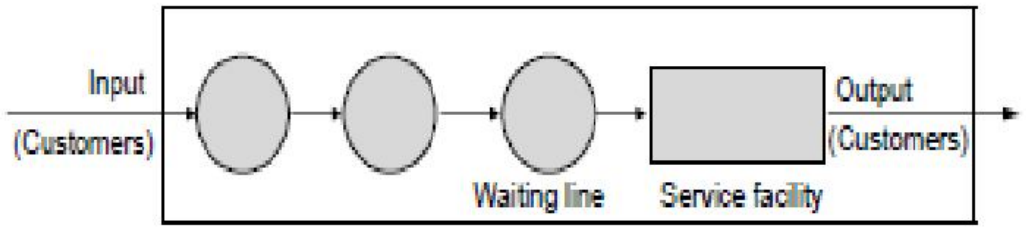
Chapter 2
Queuing system and Markov Chains

1. Queuing system: Introduction

Most systems of interest in a simulation study contain a process in which there is a demand for services. The system can service entities at a rate which is greater than the rate at which entities arrives. The entities are then said to join waiting line. The line where the entities or customers wait is generally known as queue.

The combination of all entities in system being served and being waiting for services will be called a **queuing system.**

The general diagram of queuing system can be shown as



A queuing system involves customers arriving at a constant or variable time rate for service at a service station. Customers can be students waiting for registration in college, aero -plane queuing for landing at airfield, or jobs waiting in machines shop.

If the customer after arriving can enter the service center, it is good, otherwise they have to wait for the service and form a queue i.e. **waiting line**. They remain in queue till they are provided the service. Sometimes queue being too long, they will leave the queue and go, it results a loss of customer. Customers are to be serviced at a constant or variable rate before they leave the service station.

3. Characteristics or elements of queuing system

In order to model queuing systems, we first need to be a bit more precise about what constitutes a queuing system. The three basic elements common to all queuing systems are:

1. Arrival Process or patterns
2. Service process or patterns
3. Queuing discipline

a) Arrival Process or patterns

Any queuing system must work on something − customers, parts, patients, orders, etc. We generally called them as **entities or customers**. Before entities can be processed or subjected to waiting, they must first enter the system. Depending on the environment, entities can arrive

smoothly or in an unpredictable fashion. They can arrive one at a time or in clumps (e.g., bus loads or batches). They can arrive independently or according to some kind of correlation.
A special arrival process, which is highly useful for modeling purposes, is the **Markov** arrival process. Both of these names refer to the situation where entities arrive one at a time and the times between arrivals are **exponential** random variables. This type of arrival process is *memoryless*, which means that the likelihood of an arrival within the next *t* minutes is the same no matter how long it has been since the last arrival.
Examples where this occurs are phone calls arriving at an exchange, customers arriving at a fast food restaurant, hits on a web site, and many others.

b) Service Process
Once entities have entered the system they must be served. The physical meaning of "service" depends on the system. Customers may go through the checkout process. Parts may go through machining. Patients may go through medical treatment. Orders may be filled. And so on. From a modeling standpoint, the operational characteristics of service matter more than the physical characteristics. Specifically, we care about whether service times are long or short, and whether they are regular or highly variable. We care about whether entities are processed in first-come-first-serve (FCFS) order or according to some kind of priority rule. We care about whether entities are serviced by a single server or by multiple servers working in parallel etc.

Markov Service Process
A special service process is the **Markov** service process, in which entities are processed one at a time in FCFS order and service times are independent and **exponential**. As with the case of Markov arrivals, a Markov service process is memoryless, which means that the expected time until an entity is finished remains constant regardless of how long it has been in service.

For example, in the Marcrohard example, a Markov service process would imply that the additional time required resolving a caller's problem is 15 minutes, no matter how long the technician has already spent talking to the customer. While this may seem unlikely, it does occur when the distribution of service times looks like the case shown in Figure 1. This depicts a case where the average service time is 15 minutes, but many customers require calls much shorter than 15 minutes (e.g., to be reminded of a password or basic procedures) while a few customers require significantly more than 15 minutes (e.g., to perform complex diagnostics or problem resolution). Simply knowing how long a customer has been in service doesn't tell us enough about what kind of problem the customer has to predict how much more time will be required.

c) Queuing Discipline:

The third required component of a queuing system is a queue, in which entities wait for service.
The number of customer can wait in a line is called **system capacity**.
The simplest case is an unlimited queue which can accommodate any number of customers. It is called system with unlimited capacity.
But many systems (e.g., phone exchanges, web servers, call centers), have limits on the number of entities that can be in queue at any given time.
Arrivals that come when the queue is full are rejected (e.g., customers get a busy signal when trying to dial into a call center). Even if the system doesn't have a strict limit on the queue size,

The logical ordering of customer in a waiting line is called Queing discipline and it determines which customer will be chosen for service. We may say that queuing discipline is a rule to choose the customer for service from the waiting line.
The queuing discipline includes:

a) **FIFO (First in First out):** According to this rule, Service is offered on the basis of arrival time of customer. The customer who comes first will get the service first. So in other word the customer who get the service next will be determine on the basis of longest waiting time.

b) **Last in First Out (LIFO):** It is usually abbreviated as LIFO, occurs when service is next offered to the customer that arrived recently or which have waiting time least. In the crowded train the passenger getting in or out from the train is an example of LIFO.

c) **Service in Random order (SIRO):** it means that a random choice is made between all waiting customers at the time service is offered. i.e a customer is picked up randomly from the waiting queue for the service.

d) **Shortest processing time First(SPT)**: it means that the customer with shortest service time will be chosen first for the service. i.e. the shortest service time customer will get the priority in the selection process.
e) **Priority**: a special number is assigned to each customer in the waiting line and it is called priority. Then according to this number, the customer is chosen for service.

Queuing Behavior
Customers may balk at joining the queue when it is too long (e.g., cars pass up a drive through restaurant if there are too many cars already waiting). It is called balking.
Customer may also exit the system due to impatience (e.g., customers kept waiting too long at a bank decide to leave without service) or perishability (e.g., samples waiting for testing at a lab spoil after some time period). It is called *reneging*.
When there is more than one line forming for the same service or server, the action of moving customer from one line to another line because they think that they have chosen slow line. It is called Jockeying.

3) Queuing Notations (or KENDALL'S NOTATION)

We will be frequently using notation for queuing system, called Kendall's notation,
i.e A/B/c/N/K,
where, A, B, c, N, K respectively indicate arrival pattern, service pattern, number of servers, system capacity, and Calling population.
The symbols used for the probability distribution for inter arrival time, and service time are, D for deterministic, M for exponential (or Markov) and E_k for Erlang.
If the capacity is not specified, it is taken as infinity, and if calling population is not specified, it is assumed unlimited or infinite
For example
a) M/D/2/5/∞ stands for a queuing system having exponential arrival times, deterministic service time, 2 servers, capacity of 5 customers, and infinite population.
b) If notation is given as M/D/2 means exponential arrival time, deterministic service time, 2 servers, infinite service capacity, and infinite population.

4) Single server queuing system

For the case of simplicity, we will assume for the time being, that there is single queue and only one server serving the customers. We make the following assumptions.
• First-in, First-out (FIFO): Service is provided on the first come, first served basis.
• Random: Arrivals of customers is completely random but at a certain arrival rate.
• Steady state: The queuing system is at a steady state condition.
The above conditions are very ideal conditions for any queuing system and assumptions are made to model the situation mathematically.
First condition only means irrespective of customer, one who comes first is attended first and no priority is given to anyone.

Poison arrival Patterns
Second condition says that **arrival of a customer** is completely random. This means that an arrival can occur at any time and the time of next arrival is independent of the previous arrival. With this assumption it is possible to show that the distribution of the inter-arrival time is exponential. This is equivalent to saying that the number of arrivals per unit time is a random variable with a Poisson's distribution. This distribution is used when chances of occurrence of an event out of a large sample is small.
i.e. if X = number of arrivals per unit time, then, probability distribution function of arrival is given as

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad \begin{cases} x = 0, 1, 2, \ldots \\ \lambda > 0 \end{cases}$$

$$E(X) = \lambda$$

where λ. is the average number of arrivals per unit time ($1/\tau$), and x is the number of customers per unit time. This pattern of arrival is called **Poisson's arrival pattern**.

Illustrative example
In a single pump service station, vehicles arrive for fueling with an average of 5 minutes between arrivals. If an hour is taken as unit of time, cars arrive according to Poison's process with an average of $\lambda = 12$ cars/hr.
The distribution of the number of arrivals per hour is,

$$f(x) = \Pr(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-12}12^{x}}{x!}, \quad \begin{cases} x = 0, 1, 2, \ldots \\ \lambda > 0 \end{cases}$$

$$E(X) = 12 \text{ cars/hr}$$

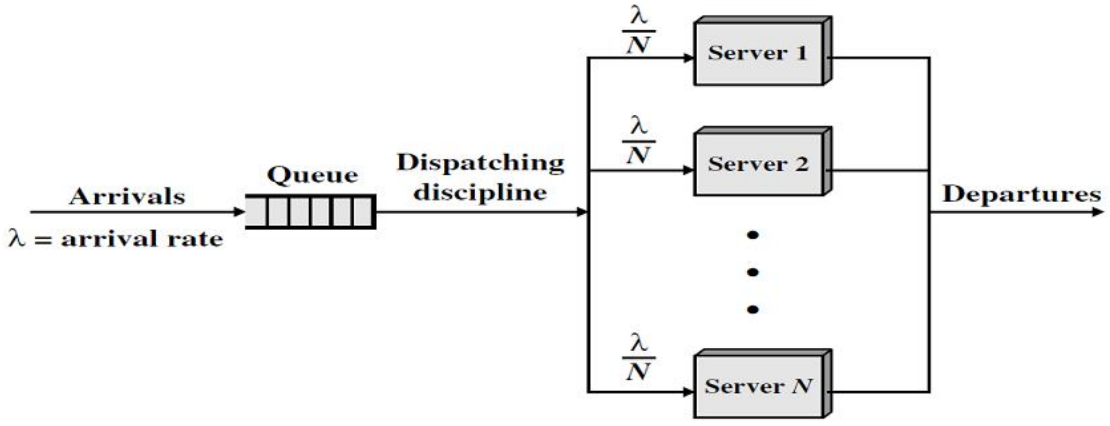
5) Measure of Queues

We have already defined the mean inter arrival time Ta and the mean service time Ts and the corresponding rates;
Arrival rate λ=1/Ta
Service rate μ=1/Ts
The following measures are used in the analysis of queue system
Traffic intensity: the ratio of the mean service time to the mean inter arrival time is called traffic intensity.
I.e. u= λ‟Ts or u=Ts/Ta
If there is any balking or reneging, not all arriving entities get served. It is necessary therefore to distinguish between actual arrival rate and the arrival rate of entities that get served.
Here λ‟ denoted the all arrivals including balking or reneging.
Server utilization: It consists of only the arrival that gets served. It is denoted by and defined as
= λTs= λ/ μ (server utilization for single server).
This is also the average number of customers in the service facility.
Thus probability of finding service counter free is
(1 – ρ)

That is there is zero customers in the service facility.

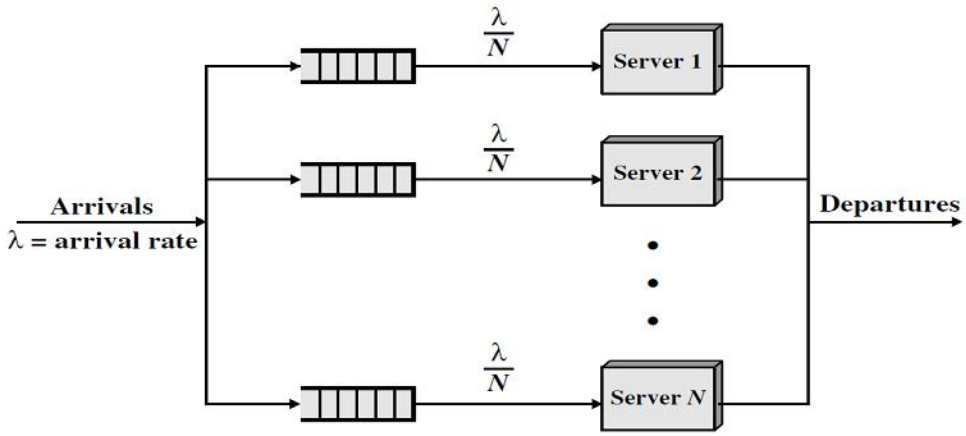
6) Concept of Multi-server Queue

Figure 3a shows a generalization of the simple model we have been discussing for multiple servers, all sharing a common queue. If an item arrives and at least one server is available, then the item is immediately dispatched to that server. It is assumed that all servers are identical; thus, if more than one server is available, it makes no difference which server is chosen for the item. If all servers are busy, a queue begins to form. As soon as one server becomes free, an item is dispatched from the queue using the dispatching discipline in force.
The key characteristics typically chosen for the multi-server queue correspond to those for the single-server queue. That is, we assume an infinite population and an infinite queue size, with a single infinite queue shared among all servers. Unless otherwise stated, the dispatching discipline is FIFO. For the multi-server case, if all servers are assumed identical, the selection of a particular server for a waiting item has no effect on service time.



The total server utilization in case of Multi-server queue for N server system is
Where μ is the service rate and lemda is the arrival rate.

There is another concept which is called multiple single server queue system as shown below



7) Some notation or Formula used to Measure the different parameter of queue

Two principal measures of queing system are;

a) The mean number of customers waiting and

b) The mean time the spend waiting

Bothe these quantities may refer to the total number of entities in the system, those waiting and those being served or they may refer only to customer in the waiting line.

Average number of customers in the queue L_Q is same as expected number in the system – the expected number in the service facility:

$$\overline{L}_Q = \overline{L}_S - \rho = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{(1 - \rho)}$$

Average time a customer spends in the system is denoted by WS , and is equal to expected number of customers in the system at time t, divided by number of customers arrived in unit time i.e.,

$$\overline{W}_S = \frac{\lambda}{\mu - \lambda} \cdot \frac{1}{\lambda} = \frac{1}{(\mu - \lambda)}$$

Average time a customer spends in the queue (W_Q) is same as average time a customer spends in the system – average time a customer spends in the server i.e.,

$$\overline{W}_Q = \overline{W}_S - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}$$

Example
At the ticket counter of football stadium, people come in queue and purchase tickets. Arrival rate of customers is 1/min. It takes at the average 20 seconds to purchase the ticket.

(a) If a sport fan arrives 2 minutes before the game starts and if he takes exactly 1.5 minutes to reach the correct seat after he purchases a ticket, can the sport fan expects to be seated for the tip-off ?
Solution:
(a) A minute is used as unit of time. Since ticket is disbursed in 20 seconds, this means, three customers enter the stadium per minute, that is service rate is 3 per minute.
Therefore,
λ = 1 arrival/min
μ = 3 arrivals/min
W_S = waiting time in the system=1/(μ- λ)=0.5
The average time to get the ticket and the time to reach the correct seat is 2 minutes exactly, so the sports fan can expect to be seated for the tip-off.

Example2
Customers arrive in a bank according to a Poisson‟s process with mean inter arrival time of 10 minutes. Customers spend an average of 5 minutes on the single available counter, and leave. Discuss
(a) What is the probability that a customer will not have to wait at the counter?
(b)What is the expected number of customers in the bank?
(c) How much time can a customer expect to spend in the bank?

Solution: We will take an hour as the unit of time. Thus
λ = 6 customers/hour,
μ = 12 customers/hour.

The customer will not have to wait if there are no customers in the bank. Thus
P0 = 1 – λ/μ= 1− 6/12 = 0.5
Expected numbers of customers in the bank are given by
$L_S = \lambda /(\mu - \lambda)$=6/6=1
Expected time to be spent in the bank is given by
W_S=1/(μ – λ)= 1/(12-6) = 1/6 hour = 10 minutes.

8) Markov Chains and its applications

a) Markov chains and Markov Process
Important classes of stochastic processes are Markov chains and Markov processes. A Markov chain is a discrete-time process for which the future behavior, given the past and the present,

only depends on the present and not on the past. A Markov process is the continuous-time version of a Markov chain. Many queuing models are in fact Markov processes. This chapter gives a short introduction to Markov chains and Markov processes focusing on those characteristics that are needed for the modeling and analysis of queuing problems.

A Markov chain

A Markov chain, named after Andrey Markov, is a mathematical system that undergoes transitions from one state to another, between a finite or countable number of possible states. It is a random process characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it. This specific kind of "memorylessness" is called the Markov property. Markov chains have many applications as statistical models of real-world processes.

Formally

A Markov chain is a sequence of random variables X1, X2, X3, ... with the Markov property, namely that, given the present state, the future and past states are independent. i.e

$$\Pr(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x | X_n = x_n).$$

Example; A simple whether model

The probabilities of weather conditions (modeled as either rainy or sunny), given the weather on the preceding day, can be represented by a transition matrix:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

The matrix P represents the weather model in which 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. The columns can be labelled "sunny" and "rainy" respectively, and the rows can be labeled in the same order.

(P)i j is the probability that, if a given day is of type i, it will be followed by a day of type j.

Notice that the rows of P sum to 1: This is because P is a stochastic matrix.

The weather on day 0 is known to be sunny. This is represented by a vector in which the "sunny" entry is 100%, and the "rainy" entry is 0%:

$$\mathbf{x}^{(0)} = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

The weather on day 1 can be predicted by:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} P = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix}$$

The weather on day 2 can be predicted in the same way:

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} P = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.86 & 0.14 \end{bmatrix}$$

Or

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} P = \mathbf{x}^{(0)} P^2 = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}^2 = \begin{bmatrix} 0.86 & 0.14 \end{bmatrix}$$

In general

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} P = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.86 & 0.14 \end{bmatrix}$$

General rules for day *n* are:

$$\mathbf{x}^{(n)} = \mathbf{x}^{(n-1)} P$$
$$\mathbf{x}^{(n)} = \mathbf{x}^{(0)} P^n$$

b) Markov chain and its Applications

Physics

Markovian systems appear extensively in thermodynamics and
statistical mechanics
, whenever probabilities are used to represent unknown or unmodelled details of the system, if it can be ssumed that the dynamics are time-invariant, and that no relevant history need be considered which is not already included in the state description.

Queueing theory

Markov chains are the basis for the analytical treatment of queues (queueing theory). Agner Krarup Erlang initiated the subject in 1917. This makes them critical for optimizing the performance of telecommunications networks, where messages must often compete for limited resources (such as bandwidth).

Internet applications

The Page Rank of a webpage as used by Google is defined by a Markov chain. It is the probability to be at page i in the stationary distribution on the following Markov chain on all (known) web pages

Statistics
Markov chain methods have also become very important for generating sequences of random numbers to accurately reflect very complicated desired probability distributions, via a process called Markov chain Monte Carlo (MCMC) And many more.

9) Differential and partial deferential equations

Continuous model
When continuous system is modeled mathematically, the variables of model representing the attribute of system are controlled by continuous functions. The distributed lag model is an example of a continuous model. Since in continuous system, the relationship between variables describe the rate at which the value of variable change, these system consist of differential equations.
Continuous system simulation uses the notation of differential equation to represent the change on the basic parameter of the system with respect to time. Hence the Mathematical model for continuous system simulation is usually represented by differential and partial differential equations.

Differential Equations
An example of a linear differential equation with constant coefficients to describe the wheel suspension system of an automobile can be given as

$$M\ddot{x} + D\dot{x} + Kx = KF(t)$$

Here the dependent variable x appears together with first and second derivatives single dot and double dot respectively.
The simple differential equation can model the simplest continuous system and they an have one or more linear differential equation with constant coefficients. It is then often possible to solve the model without using simulation technique i.e. we can solve such equations using analytical methods as (we have done in Numerical methods)
However when non linearity involves into the model, it may be impossible or at least very difficult to solve such model without simulation.

Partial Differential Equations
When more than one independent variable occurs in a differential equation the equation is sait to be partial differential equations. It can involve the derivatives of the same dependent variable with respect to each of the independent variable.

Differential equations both linear and nonlinear occur frequently in scientific and engineering studies. The reason for this is that most physical and chemical process involves rates of change, which require differential equation to represent their mathematical descriptions.
Since partial deferential equation can also represent a growth rate, continuous model can also be applied to the problems of a social or economic nature.

Analog Computer
Before the invention of digital computer, there existed devices whose behavior is equivalent of mathematical operation such as addition or subtraction ot integration. Putting together these device in a manner specified by a mathematical model or equation of a system, allowed us to simulate the system.
Some devices have been created for simulation continuous system and called analog computer or differential analyzer.

Digital analog simulators
To avoid the disadvantages of analog computers, many digital computer programming language have been written to produce digital-analog simulators. They allow or facilitate a continuous model to be programmed on a digital computer in essentially the same way as it is solved on analog computer. The language contains micro instructions that carry the action of addition, integration and sign changer. A program is written to link together these micro instructions in the same way as operational amplifiers are connected in analog computer.
Since more powerful digital computer and programming language have been developed for this purpose of simulating continuous system on digital computer, the digital-analog simulators are now in extensive use.

Exercise
(Please find the answer from above notes and prepare it on your copy)
Long questions
1) What do you mean by Queuing system? Explain the characteristics of Queuing system with example.
OR
Define the queuing system. Explain the elements of queuing system with example.
2) Explain about the Poison arrival process and Service process with example.

Short questions
1) Define a Markov chains and its application
OR
What are the key features of Markov chains?
2) Explain about the server utilization and Traffic intensity.
3) What do you mean by multi server queues?
4) What are the kendall notations of queuing system?
OR
What do you mean by Queuing notation? Explain with example
5) Explain about the Queuing Discipline and behaviors.
6) Explain about the uses of differential equations in simulations.