**Networking**

**Explanation**

**On Queuing Models**

**Satwik 2018UCO1647**

**Rajnish 2018UCO1653**

**Sachin 2018UCO1663**

**Rahul 2018UCO1665**

**Ritvik 2018UCO1667**

**Geetansh 2018UCO1668**

**Before learning about queuing models, let’s get the basics right.**

**What is a MODEL?**

A model is an abstraction of a system: an attempt to distill, from the mass of details that is the system itself, exactly those aspects that are essential to the system’s behavior. Once a model has been defined through this abstraction process, it can be parameterized to reflect any of the alternatives under study, and then evaluated to determine its behavior under this alternative. Using a model to investigate system behavior is less laborious and more flexible than experimentation, because the model is an abstraction that avoids unnecessary detail.

**What is a Queueing Network Model?**

Queueing network modelling, the specific subject of this book, is a par- ticular approach to computer system modelling in which the computer system is represented as a network of queues which is evaluated analytically. A network of queues is a collection of service centers, which represent system resources, and customers, which represent users or transactions. Analytic evaluation involves using software to solve efficiently a set of equations induced by the network of queues and its parameters.

**Need for Queuing Models**

The queuing networks are widely used to analyze the performance of complex systems involving service. The queuing network is the primary methodological framework for analyzing network delay. In communication networks, minimizing delay from entry to exit is the major concern of users. In user optimal routing, each user selects a path with minimum delay and without loss of the data from entry to exit. This network problem can be analyzed using a queuing model.

In a communication network when too many data packets arrive from many input lines and all need the same line to move out, a queue will build up. The data has to wait in the queue for transmission to its destination. However, as traffic increases the nodes are no longer able to cope and they begin losing data. At very high traffic, performance collapses completely and almost no packets are delivered. Therefore, congestion prevention is an important problem of packet switching network management

**Few Important Terms**

**Workload intensity,** which in this case is the rate at which customers arrive

**Service demand,** which is the average service requirement of a customer

**Utilisation,** the proportion of time the server is busy

**Residence time,** the average time spent at the service center by a customer, both queueing and receiving service

**Queue length,** the average number of customers at the service center, both waiting and receiving service

**Throughput,** the rate at which customers pass through the service center

**Waiting time in queue**

Time spent by a customer in the queue before being

served.

**Waiting time in the system**

It is the total time spent by a customer in the system. It can be calculated as follows: Waiting time in the system = Waiting time in queue + Service time

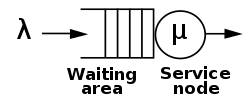
**Queue length**

Number of persons in the system at any time

**Average length of line**

The number of customers in the queue per unit of time

**QUEUING SYSTEM**

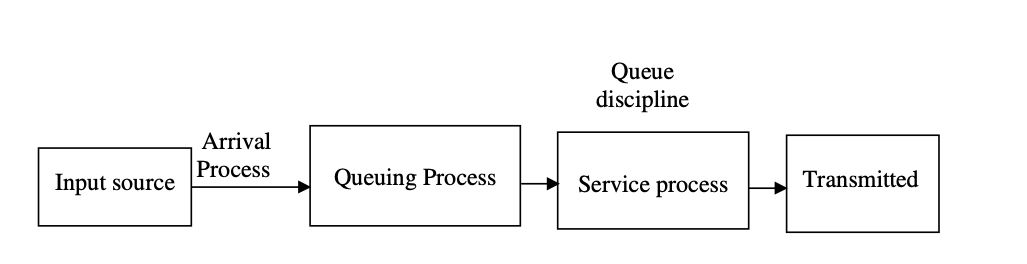


From the above figure we can clearly see that customers first have to wait in a queue before they can be served.

Since the rate at which the customers come and leave the centre is different the length of the queue can increase over time.

The essential features of a queuing system consists of

* Input source
* Queuing process
* Queue discipline
* Service process.



**Input Source Characteristics**

Input source is characterized by size, behaviour of the arrival of data and pattern of arrival of data at the system size as the data is either finite or infinite. Data on arriving at the service system stays in the system until served no matter how much the data has to wait for service. The rate, either constant or random at which data arrive at the service facility is determined by the pattern of arrival process.

The arrival process of data to the service system is classified into two categories - static and dynamic. In the static arrival process, the control depends on the nature of arrival rate as random or constant. Random arrivals are either at a constant rate or vary with time. To analyze the queuing system, it is necessary to describe the probability of distribution of arrivals. Dynamic arrival process is controlled by both service facility and data.

The arrival time distribution are

• Poisson distribution

• Exponential distribution

• Erlang distribution.

**POISSON PROCESS**

It is a probabilistic phenomenon where the number of arrivals in an interval of length t follows a Poisson distribution with parameter t, which is the rate of arrival.

The Poisson distribution is a discrete probability distribution of the number of data packets arriving in some time interval. Exponential distribution is the expected or average time between arrivals.

The Poisson distribution is a discrete probability distribution of the number of data arriving in some time interval. Considering a Poisson process involving the number of arrivals n over a time period t. If λ is the average number of arrivals per unit time, then expected number of arrivals during a time interval t will be λ t.

Then Poisson probability mass function is

P(x=n/Pn =λt)=(λt)n e-λt /n!,n=0,1,2 (4.1)

Then in the time interval from 0 to t, the probability of no arrival is given by t

P(x=0/Pn =λt)=(λt)0e-λt /0!=e-λt (4.2)

T can be defined as a random variable, as time between successive arrivals. Since a data can arrive at any time, T must be a continuous random variable. The probability of no arrival in the time interval from 0 to t will be equal to the probability that exceeds t,

by

P(T>t)=P(x=0/Pn =λt)=e-λt (4.3)

The probability that there is an arrival interval from 0 to t is given P ( T <= t ) =1 - P ( T > t ) = 1- e- λ t ; t >= 0

**MARKOV CHAIN**

Markov Chain. Let (Xn)n∈IN be a sequence of random variables taking values in a countable space E (Xn)n∈IN is a Markov chain if and only if

P(Xn = in|Xn−1 = in−1, Xn−2 = in−2, .., X0 = i0) = P(Xn = in|Xn−1 = in−1)

Thus for a Markov chain, the state of the chain at a given time contains all the information about the past evolution which is of use in predicting its future behavior. We can also say that given the present state, the past and the future are independent.

**Kendall’s Notation**

Single queueing nodes are usually described using Kendall’s Notation in the form A/S/*c* where *A* describes the distribution of durations between each arrival to the queue, *S* the distribution of service times for jobs and *c* the number of servers at the node. For an example of the notation, the M/M/1 queue is a simple model where a single server serves jobs that arrive according to a Poisson process (inter-arrival durations are exponentially distributed) and have exponentially distributed service times. In an M/G/1 queue, the G stands for general and indicates an arbitrary probability distribution for service times.

**Single Server Queuing Model**

A typical communication network as in Figure 4.6 is selected for modeling and analysis. Data are transmitted from node 1 to node 7 through node 4. At the same time node 2, and node 3 need to send data to 7 through 4. A queue will be formed at node 4. The selected model for this is a single queue single server model. Performance measure of this model is based on certain assumptions about the queuing system.

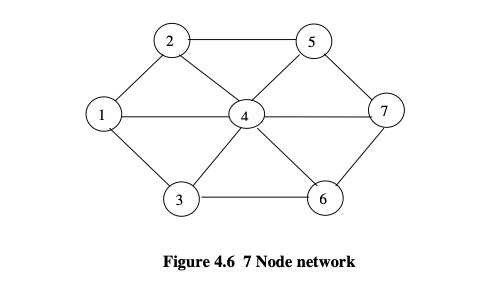
i) Exponential or Poisson distribution of arrivals as data.

ii) Single waiting line with no restriction on length of queue that

is infinite.

iii) Queue discipline is first-come, first served (FCFS).

iv) Single server with exponential distribution of service time.



**Little’s Law**

is a theorem by John Little which states that the long-term average number L of customers in a stationary system is equal to the long-term average effective arrival rate λ multiplied by the average time W that a customer spends in the system. Expressed algebraically the law is

Although it looks intuitively easy, it is quite a remarkable result, as the relationship is "not influenced by the arrival process distribution, the service distribution, the service order, or practically anything else."

The result applies to any system, and particularly, it applies to systems within systems. So in a bank, the customer line might be one subsystem, and each of the tellers another subsystem, and Little's result could be applied to each one, as well as the whole thing. The only requirements are that the system be stable and non-preemptive; this rules out transition states such as initial startup or shutdown.

# M/M/1 queue

In queueing theory, a discipline within the mathematical theory of probability, an M/M/1 queue represents the queue length in a system having a single server, where arrivals are determined by a Poisson process and job service times have an exponential distribution. The model name is written in Kendall's notation

The model is the most elementary of queueing modelsand an attractive object of study as closed-form expressions can be obtained for many metrics of interest in this model. An extension of this model with more than one server is the M/M/c queue.

## Model definition

An M/M/1 queue is a stochastic process whose state space is the set {0,1,2,3,...} where the value corresponds to the number of customers in the system, including any currently in service.

* Arrivals occur at rate λ according to a Poisson process and move the process from state i to i + 1.
* Service times have an exponential distribution with rate parameter μ in the M/M/1 queue, where 1/μ is the mean service time.
* A single server serves customers one at a time from the front of the queue, according to a first-come, first-served discipline. When the service is complete the customer leaves the queue and the number of customers in the system reduces by one.
* The buffer is of infinite size, so there is no limit on the number of customers it can contain.

The model is considered stable only if λ < μ. If, on average, arrivals happen faster than service completions the queue will grow indefinitely long and the system will not have a stationary distribution. The stationary distribution is the limiting distribution for large values of t.

Various performance measures can be computed explicitly for the M/M/1 queue. We write ρ = λ/μ for the utilization of the buffer and require ρ < 1 for the queue to be stable. ρ represents the average proportion of time which the server is occupied.

### Average number of customers in the system

The probability that the stationary process is in state i (contains i customers, including those in service) is



We see that the number of customers in the system is geometrically distributed with parameter 1 − ρ. Thus the average number of customers in the system is ρ/(1 − ρ) and the variance of number of customers in the system is ρ/(1 − ρ)2. This result holds for any work conserving service regime, such as processor sharing.

### Response time

The average response time or sojourn time (total time a customer spends in the system) does not depend on scheduling discipline and can be computed using Little's law as 1/(μ − λ). The average time spent waiting is 1/(μ − λ) − 1/μ = ρ/(μ − λ). The distribution of response times experienced does depend on scheduling discipline.

# M/G/1 Queue

An M/G/1 queue is a queue model where arrivals are Markovian (modulated by a Poisson process), service times have a General distribution and there is a single server. The model name is written in Kendall's notation, and is an extension of the M/M/1 queue, where service times must be exponentially distributed. The classic application of the M/G/1 queue is to model performance of a fixed head hard disk.

## Model definition

A queue represented by a M/G/1 queue is a stochastic process whose state space is the set {0,1,2,3...}, where the value corresponds to the number of customers in the queue, including any being served. Transitions from state i to i + 1 represent the arrival of a new customer: the times between such arrivals have an exponential distribution with parameter λ. Transitions from state i to i − 1 represent a customer who has been served, finishing being served and departing: the length of time required for serving an individual customer has a general distribution function. The lengths of times between arrivals and of service periods are random variables which are assumed to be statistically independent.

## Queue length

### Pollaczek–Khinchine method

The probability generating function of the stationary queue length distribution is given by the Pollaczek–Khinchine transform equation



where g(s) is the Laplace transform of the service time probability density function. In the case of an M/M/1 queue where service times are exponentially distributed with parameter μ, g(s) = μ/(μ + s).

This can be solved for individual state probabilities either by direct computation or using the method of supplementary variables. The Pollaczek–Khinchine formula gives the mean queue length and mean waiting time in the system.

## Waiting/response time

Writing W\*(s) for the Laplace–Stieltjes transform of the waiting time distribution, is given by the Pollaczek–Khintchine transform as



where g(s) is the Laplace–Stieltjes transform of service time probability density function.