CHAPTER 4

QUEUING MODEL FOR NETWORK CONGESTION IN COMMUNICATION NETWORK

4.1 INTRODUCTION

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 queuing model.

4.2 NETWORK CONGESTION

In 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 (Lotfi 1993, Fendick 1992, Narvaez 1998).
4.3 CAUSES FOR NETWORK CONGESTION

Congestion can be brought about by several factors. If all of a sudden streams of packets begin arriving on three or four input lines and all need the same output line, a queue will be formed. If there is insufficient memory to hold all of them, packets will be lost. Adding more memory may help but it has been discovered that if routers have an infinite amount of memory (Nagle 1987), congestion gets worse, not better, because by the time packets get to the front of the queue, they have already timed out, and duplicates have been sent.

All these packets will be forwarded to the next router, increasing the load all the way to the destination. Slow processors with low bandwidth lines can also cause congestion. If a router has no free buffers, congestion can worsen.

4.4 PRINCIPLES OF CONGESTION CONTROL

Many problems in complex systems, such as computer networks, can be viewed from a control theory, which leads to dividing all solutions into two groups.

- Open loop congestion control
- Closed loop congestion control

Open loop congestion control solutions attempt to solve the problem by good design, in essence to make sure it does not occur in the first place. Once the system is up and running, midcourse corrections are not made. Tools for doing open-loop control include deciding when to discard packets and which ones and making scheduling decisions at various points in
the network. All of these have in common the fact that they make decisions without regards to the current state of the network.

Closed loop congestion control solutions are based on the concept of a feedback loop. Closed loop congestion control has three functions:

- Monitor the system to detect when and where congestion occurs.
- Pass this information on to places where action can be taken.
- Adjust system operation to correct the problem.

Various metrics can be used to monitor the subnet for congestion. Chief among these are the percentage of all packets discarded due to lack of buffer space, the average queue lengths, the number of packets that time out and are retransmitted, the average packet delay and the standard deviation of packet delay. In all cases, rising numbers indicate growing congestion.

Feedback loop transfers the information about the congestion from the point where it is detected to the point where something can be done about it. The obvious way is for the router which detects the congestion to send a packet to the traffic sources, announcing the problem. Extra packets increase the load at precisely the moment that more load is not needed, namely when the subnet is congested. There are many polices which are used to prevent congestion.

4.4.1 Congestion Prevention Polices

Congestion prevention polices are mainly classified as Transport Layer Polices, Network Layer Polices and Data link Layer Polices.
Transport Layer Policies:

- Retransmission policy
- Out-of-order caching policy
- Acknowledgement policy
- Flow control policy
- Timeout determination

Network Layer Policies:

- Virtual circuits versus datagram inside the subnet
- Packet queuing and service policy
- Packet discard policy
- Routing algorithm
- Packet lifetime management

Data link Layer Policies:

- Retransmission policy
- Out-of-order caching policy
- Acknowledgement policy
- Flow control policy

In open loop control congestion, the systems are designed to minimize congestion in the first place, rather than letting it happen and reaching after the congestion. The systems try to achieve the goal by using appropriate policies at various levels. Different data link, network, and transport policies are also affected by congestion.
Finally, routing algorithm can help avoid congestion by spreading the traffic over all the lines, where as a bad one can send too much traffic over already congested lines. Packet life time management deals with how long a packet may live before being discarded. If it is too long or if it is too short, packets may some times time out before reaching their destination, thus inducing retransmissions. This place queue will be set up by the data packets. Analysis of network congestion in queuing model has been a fundamental research area in the field of data communication network.

4.5 QUEUING MODEL

Queuing model is one of the tools to analyze network congestion. The queuing theory is an important method to analyze the characteristics of the network. Many networks, especially data networks, are commonly modeled on single server queuing model (Cohen 1997, Sharma 2003). Analysis of network congestion using single queue, single server queuing model is presented.

Network queuing theory

The network queuing theory is a particular approach to computer system modeling in which the computer is represented as a network of queues, which is evaluated analytically. The network of queues is a collection of service nodes, which represent system resources and packets. The analytic evaluation involves usage of software to solve a set of equations induced by the network of queues and its parameters.

The queuing network model is viewed as a small subset of the techniques of queuing theory, which is selected and specialized for modeling computer systems. Much of the queuing theory is oriented towards modeling
a complex system using a single node with complex characteristics. The mathematical techniques are employed to analyze these models.

General networks of queues, which obviate many of these assumptions, are evaluated analytically. But the algorithms require time and space that grow prohibitively quickly with the size of the network. They are useful in certain specialized circumstances, but not for the direct analysis of realistic computer systems. Each delay consists of four components - processing delay, queuing delay, transmission delay, and propagation delay (Chhabra 1979).

**Network delay parameters**

In many cases the packet arrival and service rates are not sufficient to determine the delay characteristics of the system. It mainly focuses on the packet delay within the communication subnet. This delay is the sum of delays on each subnet link traversed by the packet. Each delay in turn consists of four components.

- **Processing delay:** It is the time the packet is correctly received at the head node of the link and at the time when the packet is assigned to outgoing link queue for transmission.

- **Queuing delay:** It is defined as the time the packet is assigned to a queue for transmission and the time when it starts to transmit. During this time the packet waits while other packets in the transmission queue are transmitted.

- **Transmission delay:** It is the time between the first and last bits of the packet that are transmitted.
• **Propagation delay**: It is described as the time between the transmission of the last bit at the heads node of the link and the time when the last bit is received at the tail node.

4.6 QUEUING SYSTEM

The essential features of a queuing system consists of

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

![Diagram of Queuing System](image)

**Figure 4.1 Queuing system**

4.6.1 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 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 queueing 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.

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 \( \lambda \) is the average number of arrivals per unit time, then expected number of arrivals during a time interval \( t \) will be \( \lambda t \).

Then Poisson probability mass function is

\[
P (x = n / P = \lambda t ) = (\lambda t)^n e^{-\lambda t} / n! , \ n = 0,1,2 
\]

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

\[
P (x = 0 / P = \lambda t ) = (\lambda t)^0 e^{-\lambda t} / 0! = e^{-\lambda t}
\]
$T$ can be defined as 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 $t$,

$$P(T > t) = P(x = 0 / P_n = \lambda , t) = e^{-\lambda t}$$  \hspace{1cm} (4.3)

The probability that there is an arrival interval from $0$ to $t$ is given by

$$P(T \leq t) = 1 - P(T > t) = 1 - e^{-\lambda t}; \ t \geq 0$$  \hspace{1cm} (4.4)

### 4.6.2 Queuing Process

The queuing process refers to the number of queues, and respective lengths. The number of queues depend upon the layout of a service system. There may be a single queue or multiple queues. Certain service systems adopt a number of policies to avoid queue formation. The length or size of the queue depends upon the operational situation such as physical space, legal restrictions, etc.

In certain cases, a service system is unable to accommodate more than the required number of data at a time. No former data are allowed to enter until space becomes available to accommodate new data. Thus it is referred to as finite or limited source queue. If a service system is able to accommodate any number of customers at a time, then it is referred to as infinite or unlimited source queue (Hamdy 2002).
4.6.3 Queue Discipline

The queue discipline is the order or manner in which data packets from one queue are selected for service. There are a number of ways in which data packets in the queue are served.

1. Static queue disciplines are based on the individual data packets status in the queue.
   i) If the data are served in the order of their arrival, then this is known as the first-come first-served (FCFS) service discipline.
   ii) The other discipline in use is last-come, first-served (LCFS).

2. Dynamic queue disciplines are based on the individual data attributes in the queue.
   i) Service in Random order: Data are selected for service at random irrespective of their arrivals in the service system.
   ii) Priority service: Data are grouped in priority classes on the basis of some attributes such as service time or urgency.
   iii) Pre-emptive priority: Under this rule, the highest priority customer is allowed to enter into the service immediately after entering into the system even if a data with lower priority is already in service.
   iv) Non-pre-emptive priority: In this case, highest priority data goes ahead in the queue, but service is started immediately on completion of the current service.
4.6.4 Service Process

The service process is concerned with the manner in which data are serviced and leave the service system.

It is characterized by

- The arrangement or capacity of server
- The distribution of service times
- Servers behaviour
- Management policies.

The server may be single channel in series or in parallel or mixed. Queue may be single queue or multiple queue system. The framework of various typical queuing systems is shown depicted in Figures 4.2 to 4.4.

![Figure 4.2: Single queue single server model](image)

Figure 4.2 Single queue single server model

Figure 4.2 shows the arrangement as single queue and single server model.
Figure 4.3 Multiple queues, multiple servers

Figure 4.4 Single queue multiple servers in series

Figure 4.5 Single queue multiple servers in parallel
Analytical model

Analytical models are mathematical models that have a closed form solution, that is the solution to the equations used to describe changes in a system can be expressed as a mathematical analytic function. Analytical technique provides exact solution. Thus, analytical model fails, if it is required to study the system in detail. After developing a conceptual model of a physical system it is natural to develop a mathematical model that will allow one to estimate the quantitative behavior of the system. Quantitative results from mathematical models can be compared with observational data to identify a strengths and weaknesses of a model. Mathematical models are an important component of the final complete model of a system.

Simulation model

In operations research, simulation is a problem solving technique which uses a computer aided experimental approach to study problems that cannot be analyzed using direct and formal analytical methods. Simulation is suitable to analyze large and complex real life problems which cannot be solved by usual quantitative methods. Simulations are done with the model, not on the system itself. Simulations model does not produce answer by itself. The building of a simulation model requires a long time, if no ready simulation software is available.

4.7 QUEUE MODEL FOR NETWORK CONGESTION

Data transmission with service rate in single queue, single server queuing model is analyzed (Cohen 1997). In this thesis, the model proposed is single server queuing model at the network level. It focuses on long term average performance summarizing the complexities of transient congestion through the arrival and service rate distribution of data. The performance
analysis of single server queuing model is proposed to reduce the network congestion.

In an infinite buffer single queue single server with Poisson distribution of arrivals and with exponential distribution of service time, packets are processed on first come first service order (Sharma 2003, Sundaresan 1998). Different models in queuing theory are classified by using standard notations described initially by D.G. Kendall in 1953 in the form \(a/b/c\). Later Am. Lee in 1966 added the symbols \(d\) and \(e\) to Kendall notation.

In the Literature of queuing theory, the standard format used to describe the main characteristics is \((a/b/c) : (d/e)\), where

\[
\begin{align*}
    a & \quad \text{arrival of data distribution} \\
    b & \quad \text{service of data distribution} \\
    c & \quad \text{number of servers} \\
    d & \quad \text{maximum of data allowed in the system} \\
    e & \quad \text{queue service}
\end{align*}
\]

Single server model represented as \((a/b/1) : (\infty / FCFS)\).

The various performance measures of single queue, single server model of network congestion are as follows:

\[
\begin{align*}
    n & \quad \text{number of data in the system.} \\
    P_n & \quad \text{probability of } n \text{ data in the system.} \\
    \lambda & \quad \text{average number of arrival of data per unit of time in the queue system.} \\
    \mu & \quad \text{average number of data served per unit of time in the queue system.} \\
    \rho & = \frac{\lambda}{\mu} \quad \text{Utilization factor or Traffic Intensity.}
\end{align*}
\]
4.8 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 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 sever with exponential distribution of service time.

Figure 4.6 7 Node network

Taking a small interval of time, just before time \( t \), it is assumed that the system is in state \( n \) (where \( n \) is number of data) at time \( t \).
1. The system is in state $n$ (number of data) and there is no arrival and no waiting time for a data in the queue departure, leaving the total to $n$ data.

2. The system is in state $n+1$ (number of data) and there is no arrival and one departure, reducing the total to $n$ data.

3. The system is in state $n-1$ (number as data) and has one arrival and no departure, bringing the total to $n$ customers.

\[ \lambda \rightarrow \text{expected data of arrival rate per unit of time in the queuing system} \]

\[ \mu = \text{Average expected service rate of data served per unit time at the place of service.} \]

Figure 4.7 shows the process of determining $P_n$, probability of $n$ data in the system by considering each possible number of customers either waiting or receiving service at each state which may be entered by the arrival of a new data or left by the completion of the loading server (Sharma 2003).

\[ P_n = p^n (1-p) \text{ where } p = \frac{\lambda}{\mu} < 1, \ n = 0, 1, 2, \ldots \] (4.1)

**Figure 4.7  Single server queuing system states**
This expression 4.1 gives the required probability distribution of exactly \( n \) data in the queuing system. Single server queuing model is represented as

\[(m/m/1); (\alpha/FCFS)] \tag{4.2}\]

### 4.9 MATHEMATICAL QUEUING MODEL

The expected number of data (Sharma 2003) in the system is,

\[
Q_1 = \sum_{n=0}^{\infty} n P_n \tag{4.3}
\]

\[
= \sum_{n=0}^{\infty} n (1-\rho)^n \rho^n, \; 0 < \rho < 1 \tag{4.4}
\]

\[
Q_1 = \frac{\lambda}{\mu - \lambda} \tag{4.5}
\]

The queue length, that is expected number of data waiting in the queue is,

\[
Q_2 = \sum_{n=1}^{\infty} n (n-1) P_n \tag{4.6}
\]

\[
Q_2 = \frac{\lambda^2}{\mu (\mu - \lambda)} \tag{4.7}
\]

The expected waiting time for a data in the queue is,

\[
Q_3 = \frac{\lambda}{\mu (\mu - \lambda)} \tag{4.8}
\]
The probability of an arrival during the service time when system contains no data

\[ Q_4 = \frac{\mu}{(\mu + \lambda)} \]  \hspace{1cm} (4.9)

The Variance of queue length of the data is

\[ Q_5 = \sum_{n=1}^{\alpha} n^2 P_n - \left( \sum_{n=1}^{\alpha} n P_n \right)^2 \]  \hspace{1cm} (4.10)

\[ Q_5 = \frac{\lambda \mu}{(\mu - \lambda)^2} \]  \hspace{1cm} (4.11)

The traffic Intensity is,

\[ Q_6 = \frac{\lambda}{\mu} \]  \hspace{1cm} (4.12)

The total delay of the data is,

\[ Q_7 = \frac{\lambda}{\mu (\mu - \lambda)} + \frac{1}{\mu} \]  \hspace{1cm} (4.13)

The expected length of non-empty queue is,

\[ Q_8 = \frac{\mu}{(\mu - \lambda)} \]  \hspace{1cm} (4.14)

4.10 PERFORMANCE MEASURES OF QUEUING MODEL

The various performance measures of the single queue single server model of network congestion using analytical model are depicted in Figures 4.8 to 4.20.
Figure 4.8 Performance measure of expected number of data in the system

Figure 4.9 Performance measure of queue length
Figure 4.10  Comparison of waiting time of data in the queue

Figure 4.11  Measure of probability of arrival during service time
Figure 4.12 Measure of variance of queue length

Figure 4.13 Performance measure of utilization factor
Figure 4.14 Measure of total delay of the data

Figure 4.15 Expected length of non empty queue
Comparative Analysis of Analytical model and Simulation model

Figure.4.16 Comparison of total delay

Figure 4.17 Comparison of queue length
Figure 4.18 Comparison of utilization factor

Figure 4.19 Utilization factor
4.11 CONCLUSION

The performance measure of the single server single queue model using analytical is illustrated in Figures 4.8 to 4.15. The arrival rates of data are taken as 1.5 Mbps (AR 1), 1 Mbps (AR 2) and service rate up to 5 Mbps. Figures 4.13 and 4.19 shows the utilization factor reduced with the arrival rate of data, where arrival rates are taken as 1Mbps, 1.5Mbps, 3Mbps, 5Mbps and 10Mbps and service rates are taken up to 20Mbps. From Figure 4.14 for the selected time of 1.6 micro seconds for the data, the service rate of 2.5 Mbps and the arrival rate of 1.5 Mbps, the data are entered and serviced and transmitted with out loss. If the total delay of the data is reduced to 1.5 micro seconds and service rate to 2.5 Mbps, the arrival rate is 1 Mbps, loss of data arriving with in 0.1 micro seconds.

If the selected total delay is 0.4 micro seconds and service rate at 13 Mbps the arrival rate is 10 Mbps, data arriving with in 0.2 micro second will be lost. If the total delay of the data is taken as 0.6 micro seconds for the
data, the service rates of 13 Mbps and the arrival rate of 10 Mbps are entered and serviced and transmitted without loss as shown in Figures 4.19 and 4.20.

If selected time to leave the data in transmission is within 1 micro second, the service rate at 2.5 Mbps and arrival rate at 1 Mbps, the actual total delay is 1.6 micro seconds, data with in the queue 0.6 micro seconds will be lost due to congestion. For the same delay, if the service rate is increased to 3 Mbps for the same arrival rate, all the data will be transmitted without loss for the maximum utilization factor 0.5 as shown in Figures 4.13 and 4.14.

The results are validated using computer simulation model. The simulation is done by taking the arrival time distribution approximated by probability distribution as Poisson distribution. The Poisson process involves the number of arrivals \( n \) over a time period \( t \). The number of arrivals in \( t \) is taken as 40. Comparative analysis of delay of the data, queue length, utilization factor in transmission by analytical model and simulation model, are very close to each other as shown in Figures 4.16 to 4.18. The minor differences may be due to the following considerations in an analytical model as listed.

1. Single waiting line with no restriction on length of queue that is of infinite capacity
2. Analytical model is a mathematical model, which provides an exact solution
3. Analytical model is time independent model.

This model provides optimization of service rate of data transmission in communication network for various arrival rate of data with the optimized value maximum throughput of the system is obtained with full
utilization of link capacity and traffic intensity. At the same time, minimum delay in data transmission without loss of data is also obtained.

4.12 SUMMARY

This chapter addresses network congestion, various aspects of queuing model, analysis of single queue single server model for network congestion and analysis of network congestion, without loss of data with maximum throughput.