Lecture – 8 (Dt. 12th April 2020)

Electronic Switching (EC- 8th Sem)

Traffic Engineering- Part-2

- **Blocking Probabilities &Congestion**
- **Modelling of Traffic**

References :

1) S. Agrawal : Lecture Notes (VSSUT)

2) Telecommunication Switching Systems & Networks, Thiagrajan

3) Telecommunication System Engineering, R.L. Freeman

4) Telecommunication Switching and Networks, By, P. Gnanasivam

5) Internet sources

Acknowledgment :

Special thanks to Prof. S. Agrawal (VSSUT) for providing online lecture notes.

Blocking Probability and Congestion

The value of the blocking probability is one aspect of the telephone company's grade of service. The basic difference between GOS and blocking probability is that GOS is a measure from subscriber point of view whereas the blocking probability is a measure from the network or switching point of view. Based on the number of rejected calls, GOS is calculated, whereas by observing the busy servers in the switching system, blocking probability will be calculated. The blocking probabilities can be evaluated by using various techniques. Lee graphs and Jacobaeus methods are popular and accurate methods. The blocking probability B is defined as the probability that all the servers in a system are busy. Congestion theory deals with the probability that the offered traffic load exceeds some value. Thus, during congestion, no new calls can be accepted. There are two ways of specifying congestion. They are time congestion and call congestion. Time congestion is the percentage of time that all servers in a group are busy. The call or demand congestion is the proportion of calls arising that do not find a free server. In general GOS is called call congestion or loss probability and the blocking probability is called time congestion. If the number of sources is equal to the number of servers, the time congestion is finite, but the call congestion is zero. When the number of sources is large, the probability of a new call arising is independent of the number already in progress and therefore the call congestion is equal to time congestion.

MODELLING OF TRAFFIC

To analyse the statistical characteristics of a switching system, traffic flow and service time, itis necessary to have a mathematical model of the traffic offered to telecommunication systems. The model is a mathematical expression of physical quantity to represents the behaviour of the quantity under consideration. Also the model provides an analytical solutions to a tele traffic problems. As the switching system may be represented in different ways, different models are possible. Depending on the particular system and particular circumstance,

a suitable model can be selected. In practice, the facilities of the switching systems are shared by many users. This arrangement may introduce the possibility of call setup inability due to lack of available facilities. Also in data transfer, a system has to buffer message while waiting for transmission. Here size of the buffer depends on traffic flow. As serving the number of subscribers subject to fluctuation (due to random generation of subscriber calls, variations in holding time, location of the exchange, limitation in servers etc.), modelling of traffic is studied using the concepts and methods of the theory of probability. If a subscriber finds no available server for his call attempt, he will wait in a line (queue) or leave immediately. This phenomenon may be regarded as a queueing system. The mathematical description of the queueing system characteristics is called a queueing model.

The random process may be discrete or continuous. Similarly the time index of random variables can be discrete or continuous. Thus, there are four different types of process namely (*a*) continuous time continuous state (*b*) continuous time discrete state (*c*) discrete time continuous state and (*d*) discrete time discrete state. In telecommunication switching system, our interest is discrete random process and therefore for modelling a switching system, we use discrete state stochastic process. A discrete state stochastic process is often called a chain. A statistical properties of a random process may be obtained in two ways:

(*i*) Observing the behaviour of the system to be modelled over a period of time repeatedly. The data obtained is called a single sample. The average determined by measurements on a single sample function at successive times will yield a **time average**.

(*ii*) Simultaneous measurements of the output of a large number of statistically identical random sources. Such a collection of sources is called an **ensemble** and the individual noise waveforms is called the **sample function**. The statistical average made at some fixed time $t = t1$ on all the sample functions of the ensample is the **ensemble average.**

The above two ways are analogous to obtaining the statistics from tossing a die repeatedly (large number) or tossing one time the large number of dice. In general, time average and ensemble average are not the same due to various reasons. When the statistical characteristics of the sample functions do not change with time, the random process is described as being **stationary**. The random process which have identical time and ensemble average are known as **ergodic processes.** An ergodic process is stationary, but a stationary process is not necessarily ergodic.

Telephone traffic is nonstationary. But the traffic obtained during busy hour may be considered as stationary (which is important for modelling) as modelling non-stationary is difficult.

Pure Chance Traffic

Here, the call arrivals and call terminations are independent random events. If call arrivals are independent random events, their occurrence is not affected by previous calls. This traffic is therefore sometimes called **memory less traffic.** A.A. Markov in 1907, defined properties and proposed a simple and highly useful form of dependency. This class of processes is of great interest to our modelling of switching systems. A discrete time Markov chain *i.e.* discrete time discrete state Markov process is defined as one which has the following property.

$$P\,[\{X(t_{n+1}) = x_{n+1}\}/\{X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \ldots\ldots X(t_1) = x_1\}]$$
$$= P\,[\{X(t_{n+1}) = x_{n+1}\}/\{X(t_n) = x_n\}]$$

where $t_1 < t_2 \ldots\ldots < t_n < t_{n+1}$ and x_i is the ith discrete state space value.

Equation states that the probability that the random variable X takes on the value x_{n+1}

at time step $n + 1$ is entirely determined by its state value in the previous time step n and is independent of its state values in earlier time steps ; $n - 1$, $n-2$, $n-3$ etc.

The Birth and Death Process

The birth and death process is a special case of the discrete state continuous time Markov process, which is often called a continuous-time Markov chain. The number of calls in progress is always between 0 and N. It thus has N + 1 states. If the Markov chain can occur only to adjacent states (*i.e.* probability change from each state to the one above and one below it) the process is known as birth-death (B–D) process. The basic feature of the method of Markov chains is the kolmogorov differential-difference equation, for the limiting case, can provide a solution to the state probability distribution for the Erlang systems and Engest systems.

Let N(t) be a random variable specifying the size of the population at time t. For a complete description of a birth and death process, we assume that N(t) is in state k at time t and has the following properties:

1. P(k) is the probability of state k and P($k + 1$) is the probability of state $k + 1$.

2. The probability of transition from state k to state $k + 1$ in short duration Δt is $\lambda\,\Delta t$, where λ is called the birth rate in state k.

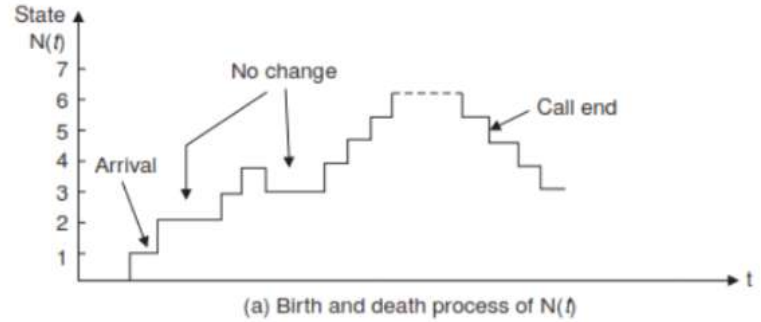
3. The probability of transition from state k to state $k-1$ in the time interval Δt is $\mu k\,\Delta t$, where μ is called the death rate in state k.

4. The probability of no change of state in the time interval Δt is equal to $1-(\lambda k)\,\Delta t$.

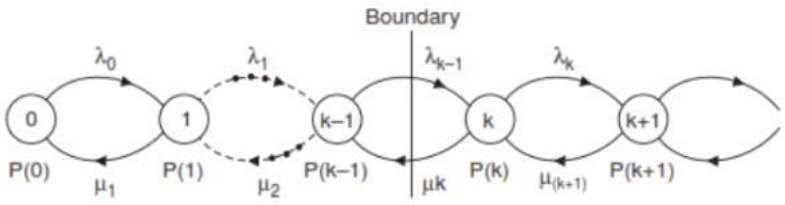
5. The probability in Δt, from state k to a state other than $k+1$ or $k-1$ is zero. Based on the above properties, birth and death process of N(t) and state transition rate diagram are shown in Fig. At statistical equilibrium (*i.e.* stationary), let P_{jk} is the conditional probability, that is the probability of state increases from j to k. Similarly P_{kj} is the probability of state decrease from k to j.

The probabilities P(0), P(1), ... P(N) are called the **state probabilities** and the conditional probabilities P_{jk}, P_{kj} are called **transition probabilities.** The transition probabilities satisfy the following condition:

$$P_{jk}(t) \geq 0, \sum_{k=0}^{\infty} P_{jk}(t) = 1$$



(a) Birth and death process of N(t)



(b) State transition diagram

Numbered circles denote the state of the process at t

Markov theorem states that for any Markov process characterized by the transition probability Pjk, the limit

$$\lim_{t \to \infty} P_{jk} = P(k)$$

exist and does not depend on j and the probability P(k).

According to Markov's,

$$\frac{d}{dt} P(k) = -(\lambda_k + \mu_k) P(k) + \lambda_{k-1} P(k-1) + \mu_{k+1} P(k+1)$$

$$k = 0, 1, 2, \ldots, \text{with } \lambda_{-1} = \mu_0 = P_{-1} = 0$$

This set of differential-difference equations represents the dynamic behaviour of the birth and death process. As $t \to \infty$,

$$-(\lambda_k + \mu_k) P(k) + \lambda_{k-1} P(k-1) + \mu_{k+1} P(k+1) = 0$$

with $\lambda - 1 = \mu_0 = P_{-1} = 0$.

This set of equations together with the normalization condition uniquely determines the required

$$\sum_{k=0}^{\infty} P(k) = 1$$

and state probabilities $P(k)$ as $P(k) = \frac{\lambda_{k-1}}{\mu_k} P(k-1)$

The probability P(0) can be determined by the equation

$$P(k) = \frac{\lambda_0 \lambda_1 \ldots\ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots\ldots \mu_k} P(0), k = 1, 2, 3, \ldots\ldots \qquad P(0) = \left[1 + \sum_{k=1}^{\infty} \frac{\lambda_0 \lambda_1 \ldots\ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots\ldots \mu_k}\right]^{-1}$$

$$P(k) = \frac{\lambda_{k-1} P(k-1)}{\mu_k}$$

LOSS SYSTEMS

The service of incoming calls depends on the number of lines. If number of lines equal to the number of subscribers, there is no question of traffic analysis. But it is not only uneconomical but not possible also. So, if the incoming calls finds all available lines busy, the call is said to be **blocked.** The blocked calls can be handled in two ways. The type of system by which a blocked call is simply refused and is lost is called **loss system.** Most notably, traditional analog telephone systems simply block calls from entering the system, if no line available. Modern telephone networks can statistically multiplex calls or even packetize for lower blocking at the cost of delay. In the case of data networks, if dedicated buffer and lines are not available, they block calls from entering the system. In the second type of system, a blocked call remains in the system and waits for a free line. This type of system is known as **delay system.** These two types differs in network, way of obtaining solution for the problem and GOS.

For loss system, the GOS is probability of blocking. For delay system, GOS is the probability of waiting.

Erlang determined the GOS of loss systems having N trunks, with offered traffic A, with the following assumptions. (*a*) Pure chance traffic (*b*) Statistical equilibrium (*c*) Full availability and (*d*) Calls which encounter congestion are lost. The first two are explained in previous section. A system with a collection of lines is said to be a fully-accessible system, if all the lines are equally accessible to all in arriving calls. For example, the trunk lines for interoffice calls are fully accessible lines. The lost call assumption implies that any attempted call which encounters congestion is immediately cleared from the system. In such a case, the user may try again and it may cause more traffic during busy hour. The Erlang loss system may be defined by the following specifications.

1. The arrival process of calls is assumed to be Poisson with a rate of λ calls per hour.

2. The holding times are assumed to be mutually independent and identically distributed random variables following an exponential distribution with $1/\mu$ seconds.

3. Calls are served in the order of arrival.

There are three models of loss systems. They are:

1. Lost calls cleared (LCC)

2. Lost calls returned (LCR)

3. Lost calls held (LCH)

Lost Calls Cleared (LCC) System

The LCC model assumes that, the subscriber who does not avail the service, hangs up the call, and tries later. The next attempt is assumed as a new call. Hence, the call is said to be cleared. This also referred as blocked calls lost assumption. The first person to account fully and accurately for the effect of cleared calls in the calculation of blocking probabilities was A.K.Erlang in 1917. Consider the Erlang loss system with N fully accessible lines and exponential holding times. The Erlang loss system can be modelled by birth and death process with birth and death rate as follows.

$$\lambda_k = \begin{cases} \lambda, k = 0, 1, \ldots\ldots N-1 \\ 0 \quad k \geq N \end{cases}$$

$$\mu_k = \begin{cases} k\mu, k = 0, 1, \ldots\ldots N \\ 0, \quad k > N \end{cases}$$

$$P(k) = \frac{\lambda_0 \lambda_1 \ldots\ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots\ldots \mu_k} P(0), k = 1, 2, 3$$

Substituting in the above equation, we get

$$P(k) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k P(0), k = 1, 2, 3, \ldots\ldots N$$

From equation the offered traffic is

$$A = \frac{\lambda}{\mu}$$

Substituting we get

$$P(k) = \frac{1}{k!} (A)^k P(0), k = 1, 2, 3, \ldots\ldots N$$

The probability P(0) is determined by the normalization condition

$$\sum_{k=0}^{N} P(k) = P(0) \sum_{k=0}^{N} \frac{A^k}{k!} = 1$$

$$P(0) = \frac{1}{\sum_{k=0}^{N} \frac{A^k}{k!}}$$

Substituting we get

$$P(k) = \frac{A^k / k!}{\sum_{k=0}^{N} \frac{A^k}{k!}}$$

The probability distribution is called the truncated **Poisson distribution** or **Erlang's loss distribution.** In particular when $k = N$, the probability of loss is given by

$$P(N) = B(N, A) = \frac{A^N}{N! \sum_{k=0}^{N} \left(\frac{A^k}{k!}\right)}$$

where $A = \lambda/\mu$.

This result is variously referred to as **Erlang's formula of the first kind, the Erlang's-B formula or Erlang's loss formula.**

Equation specifies the probability of blocking for a system with random arrivals from an infinite source and arbitrary holding time distributions. The Erlang B formula gives the time congestion of the system and relates the probability of blocking to the offered traffic and the number of trunk lines. Values from B(N, A) obtained from equation have been plotted against the offered traffic 'A' Erlang's for different values of the number of N lines in Fig. In design problems, it is necessary to find the number of trunk lines needed for a given offered traffic and a specified grade of service.

$$A' = A\,[1 - B\,(N, A)]$$

Thus, the carried load is the position of the offered load that is not lost from the system. The carried load per line is known as the trunk occupancy.

$$\rho = \frac{A'}{N} = \frac{A(1-B)}{N}$$

The trunk occupancy ρ is a measure of the degree of utilization of a group of lines and is sometimes called the utilization factor.

Example 8.7. *A group of 7 trunks is offered 4E of traffic, find (a) the grade of service (b) the probability that only one trunk is busy (c) the probability that only one trunk is free and (d) the probability that at least one trunk is free.*

Sol. Given data : N = 7, A = 4E

From equation 8.54,

(*a*)
$$B(7, 4) = \frac{4^7}{7!\left[1+\frac{4}{1}+\frac{4^2}{2!}+\frac{4^3}{3!}+\frac{4^4}{4!}+\frac{4^5}{5!}+\frac{4^6}{6!}+\frac{4^7}{7!}\right]}$$

$$= \frac{16384}{5040\,[1+4+8+21.3+10.6+8.5+5.7+3.25]}$$

$$B = 0.052 = \text{GOS}.$$

(*b*) The probability of only one trunk is busy

$$P(k) = \frac{A^k / k!}{\sum_{k=0}^{N} (A^k / k!)}$$

For $k = 1$
$$P(1) = \frac{4/1!}{62.35} = 0.064$$

(*c*) The probability that only one trunk is free

$$P(6) = \frac{4^6/6!}{62.35} = \frac{5.68}{62.35} = 0.0912$$

(*d*) The probability that at least one trunk is free

$$P(k < 7) = 1 - P(7) = 1 - B = 1 - 0.052 = 0.948.$$

Lost Calls Returned (LCR) System

In LCC system, it is assumed that unserviceable requests leave the system and never return. This assumption is appropriate where traffic overflow occurs and the other routes are in other calls service. If the repeated calls not exist, LCC system is used. But in many cases, blocked calls return to the system in the form of retries. Some examples are subscriber concentrator systems, corporate tie lines and PBX trunks, calls to busy telephone numbers and access to WATS lines. Including the retried calls, the offered traffic now comprise two components *viz.*, new traffic and retry traffic. The model used for this analysis is known as lost calls returned

1. All blocked calls return to the system and eventually get serviced, even if multiple retries are required.

2. Time between call blocking and regeneration is random statistically independent of each other. This assumption avoid complications arising when retries are correlated to each other and tend to cause recurring traffic peaks at a particular waiting time interval.

3. Time between call blocking and retry is somewhat longer than average holding time of a connection. If retries are immediate, congestion may occur or the network operation becomes delay system.

Consider a system with first attempt call arrival ratio of λ (say 100). If a percentage B (say 8%) of the calls blocked, B time's λ retries (*i.e.* 8 calls retries). Of these retries, however a percentage B will be blocked again.

Hence by infinite series, total arrival rate λ' is given as

$$\lambda' = \lambda + B\lambda + B^2\lambda + B^3\lambda + ...$$

$$\lambda' = \frac{\lambda}{1 - B}$$

where B is the blocking probability from a lost calls cleared (LCC) analysis.

The effect of returning traffic is insignificant when operating at low blocking probabilities. At high blocking probabilities, it is necessary to incorporate the effects of the returning traffic into analysis.

Lost Calls Held (LCH) System

In a lost calls held system, blocked calls are held by the system and serviced when the necessary facilities become available. The total time spend by a call is the sum of waiting time and the service time. Each arrival requires service for a continuous period of time and terminates its request independently of its being serviced or not. If number of calls blocked, a portion of it is lost until a server becomes free to service a call. An example of LCH system is the time assigned speech interpolation (TASI) system.

LCH systems generally arise in real time applications in which the sources are continuously in need of service, whether or not the facilities are available. Normally, telephone network does not operate in a lost call held manner. The LCH analysis produces a conservative design that

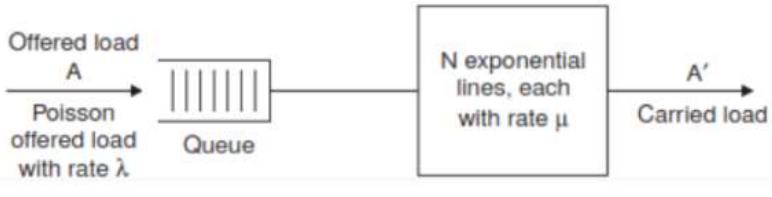
system concentrates some number of voice sources onto a smaller number of transmission channels. A source receives service only when it is active. If a source becomes active when all channels are busy, it is blocked and speech clipping occurs. Each speech segment starts and stops independently of whether it is served or not. Digital circuit multiplication (DCM) systems in contrast with original TASI, can delay speech for a small amount of time, when necessary to minimize the clipping. LCH are easily analysed to determine the probability of the total number of calls in the system at any one time. The number of active calls in the system at any time is identical to the number of active sources in a system capable of carrying all traffic as it arises. Thus the distribution of the number in the system is the Poisson distribution. The Poisson distribution is given as

$$P(x) = \frac{\mu^x}{x!} e^{-\mu}.$$

The probability that k sources requesting service are being blocked is simply the probability that $k + N$ sources are active when N is the number of servers.

DELAY SYSTEMS

The delay system places the call or message arrivals in a queue if it finds all N servers (or lines) occupied. This system delays non-serviceable requests until the necessary facilities become available. These systems are variously referred to as delay system, waiting-call systems and queueing systems. The delay systems are analysed using queueing theory which is sometimes known as waiting line theory. This delay system have wide applications outside the telecommunications. Some of the more common applications are data processing, supermarket checkout counters, aircraft landings, inventory control and various forms of services. Consider that there are k calls (in service and waiting) in the system and N lines to serve the calls. If $k = N$, k lines are occupied and no calls are waiting. If $k > N$, all N lines are occupied and $k - N$ calls waiting. Hence a delay operation allows for greater utilization of servers than does a loss system. Even though arrivals to the system are random, the servers see a somewhat regular arrival pattern. A queueing model for the Erlang delay system is shown in Fig.



The basic purpose of the investigation of delay system is to determine the probability distribution of waiting times. From this, the average waiting time W as random variable can be easily determined. The waiting times are dependent on the following factors:

1. Number of sources

2. Number of servers

3. Intensity and probabilistic nature of the offered traffic

4. Distribution of service times

5. Service discipline of the queue.

In a delay system, there may be a finite number of sources in a physical sense but an infinite number of sources in an operational sense because each source may have an arbitrary number of requests outstanding. If the offered traffic intensity is less than the servers, no statistical limit exists on the arrival of calls in a short period of time. In practice, only finite queue can be realised. There are two service time distributions. They are constant service times and exponential service times. With constant service times, the service time is deterministic and with exponential, it is random. The service discipline of the que involves two important factors.

1. Waiting calls are selected on of first-come, first served (FCFS) or first-in-first-out (FIFO) service.

2. The second aspect of the service discipline is the length of the queue. Under heavy loads, blocking occurs. The blocking probability or delay probability in the system is based on the queue size in comparison with number of effective sources. We can model the Erlang delay system by the birth and death process with the following birth and death rates respectively.

$\lambda_k = \lambda,\ k = 0, 1, \ldots\ldots$ and

$$\mu_k = \begin{cases} k\mu\ , k = 0, 1, \ldots\ldots \mathrm{N} - 1 \\ \mathrm{N}\mu\ k \geq \mathrm{S} \end{cases}$$

Under equilibrium conditions, the state probability distribution P(k) can be obtained by substituting these birth rates into the following equation

$$P(k) = \frac{\lambda_0 \lambda_1 \ldots\ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots\ldots \mu_k} P(0)\ k = 1, 2, \ldots\ldots$$

we set

$$P(k) = \begin{cases} \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k P(0) & 0 \le k \le S \\ \frac{\left(\frac{\lambda}{\mu}\right)^k}{N!N^{k-N}} P(0) & k \ge N. \end{cases}$$

As $A = \frac{\lambda}{\mu}$, we get

$$p(k) = \begin{cases} \frac{A^k}{K!} P(0) & 0 \le k \le S \\ \frac{A^k}{N!N^{k-N}} P(0); & k > N \end{cases}$$

Under normalised condition,

$$\sum_{k=0}^{\infty} P(k) = 1 \qquad \text{or} \qquad \sum_{k=0}^{N-1} \frac{A^k}{k!} P(0) + \sum_{k=N}^{\infty} \frac{A^k}{N!N^{k-N}} P(0) = 1$$

$$\frac{1}{P(0)} = \sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{N^N}{N!} \sum_{k=N}^{\infty} \left(\frac{A}{N}\right)^k$$

$$= \sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{N^N}{N!}\left[\left(\frac{A}{N}\right)^N + \left(\frac{A}{N}\right)^{N+1} + \left(\frac{A}{N}\right)^{N+2} + \ldots\ldots\right]$$

$$= \sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{N^N}{N!}\left(\frac{A}{N}\right)^N \left[1 + \frac{A}{N} + \left(\frac{A}{N}\right)^2 + \ldots\ldots\right]$$

$$= \sum_{k=0}^{N-1} \frac{A^k}{k!} + \frac{A^N}{N!}\left[\frac{1}{1 - A/N}\right] = \sum_{k=0}^{N-1} \frac{A^k}{k!} + \left[\frac{A^N}{N!} + \frac{A^N}{N!}\left(\frac{A}{N-A}\right)\right]$$

$$\frac{1}{P(0)} = \sum_{k=0}^{N} \frac{A^k}{k!} + \frac{A^N}{N!}\left(\frac{A}{N-A}\right)$$

$$\frac{1}{P(0)} = \sum_{k=0}^{N} \frac{A^k}{k!} + \frac{A^N}{N!}\left(\frac{A}{N-A}\right)$$

$$P(0) = \frac{1}{\sum_{k=0}^{N} \frac{A^k}{k!} + \frac{A^N}{N!}\left(\frac{A}{N-A}\right)}$$

We know

$$P(k) = \frac{A^k}{k!} P(0),\ k = 1, 2, \ldots\ldots, N$$

$$C(N, A) = \frac{A^N / N!}{\sum_{k=0}^{N} \frac{A^k}{k!} + \frac{A^N}{N!}\left(\frac{A}{N-A}\right)}$$

$$\frac{1}{C(N, A)} = \frac{\sum_{k=0}^{N} \frac{A^k}{k!}}{\frac{A^N}{N!}} + \frac{\frac{A^N}{N!}\left(\frac{A}{N-A}\right)}{\frac{A^N}{N!}!}$$

$$\frac{1}{C(N, A)} = \frac{1}{B} + \frac{A}{N-A} .$$

$$\text{Prob. (delay)} = P(>0)\, C(N, A) = \frac{BN}{N - A(1-B)}$$

where B = Blocking probability for a LCC system

N = Number of servers

A = Offered load (Erlangs)

Equation above are referred as **Erlang's second formula, Erlang's delay formula or Erlang's C formula.**

For single server systems (N = 1), the probability of delay reduces to ρ, which is simply the output utilization or traffic carried by the server. Thus the probability of delay for a single server system is also.