

Communication Switching Systems

Course Code: ETU07402

Florent Morice Mtuka

Mobile: +255 764 281 463

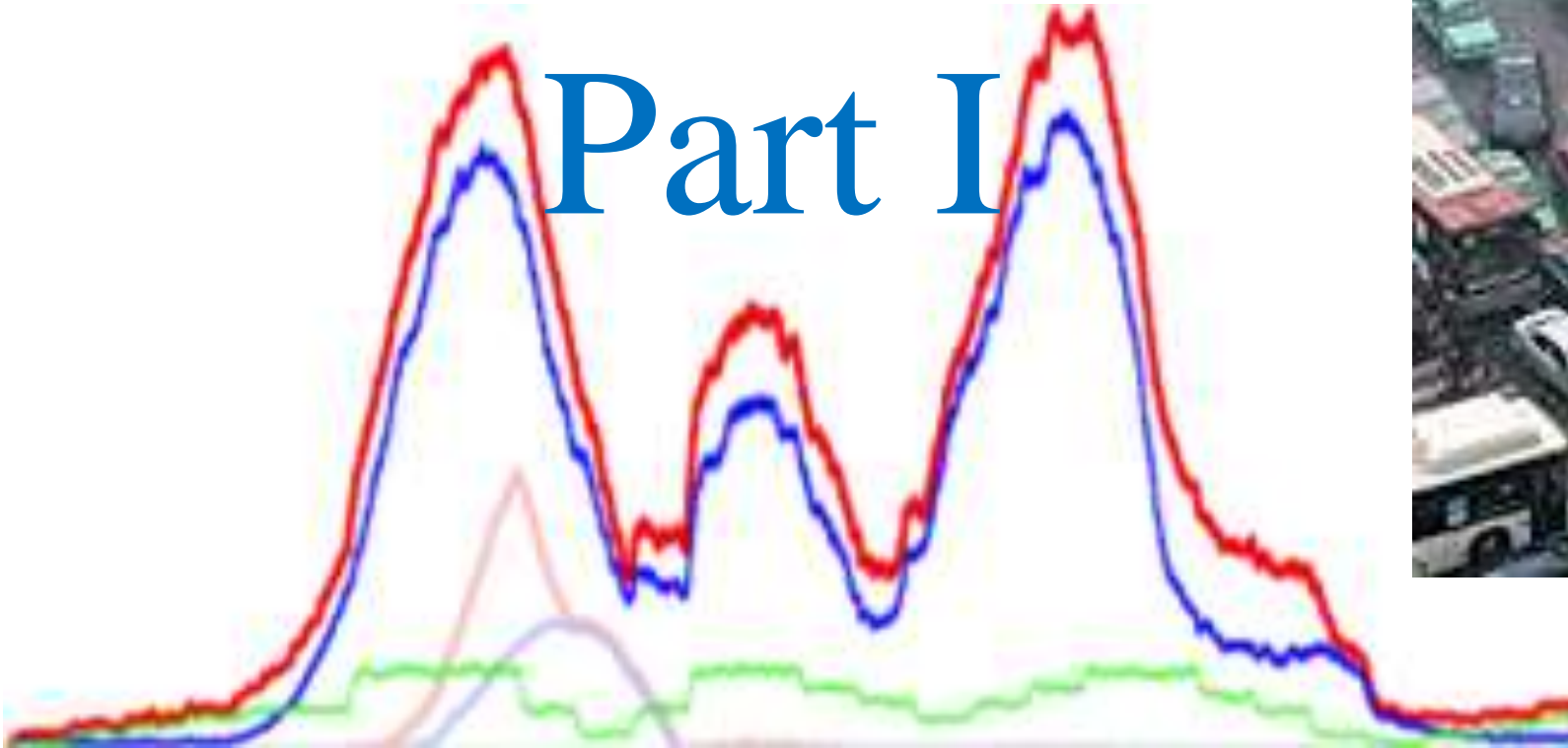
Email: Mtukaf@yahoo.Com

Electronics & Telecommunication Engineering



Traffic Engineering

Part I



Lecture 04

4.1 Traffic Engineering Introduction

- 4.1: Introduction to Traffic Engineering
- 4.2: Traffic Design Requirements
- 4.3. Modelling of Traffic
- 4.4. Loss Systems - Blocking
- 4.4. Delay System
- 4.5. Packet Level Model For Data Traffic

4.1: Introduction to Traffic Engineering

- The fundamental purpose of a Telco is to offer services (Services) in profitable manner. Voices and Data are the main source of Network traffic.
- Traffic is defined as the occupancy of the server, hence.
- **Traffic Engineering**; Determines
 - ➡ The conditions under which adequate service is provided to subscribers while making economical use of the resources providing the service.
 - ➡ The ability of a telecom network to carry a **given traffic** at a particular **loss probability**.
 - ➡ *Traffic theory* and *queuing theory* are used to estimate the probability of the occurrence of *call blocking*.

4.1. Introduction to Traffic Engineering

- **Traffic Engineering** provides the basis for analysis and design of telecommunication networks or model.
- Provides means to determine the major equipment required to provide a particular level of service for a given traffic pattern and volume.
- **Traffic model** predicts *accessibility* and *utilization of telephone lines, channel and trunks* and *cost effectiveness of various sizes and configuration of networks*.

4.1.1. Traffic pattern

Nature of telephone traffic and its distribution with respect to time (traffic load) which is normally 24 hours. It helps in determining the amount of lines/Servers required to serve the subscriber needs. The variations are not uniform and varies season to season, month to month, day to day and hour to hour.

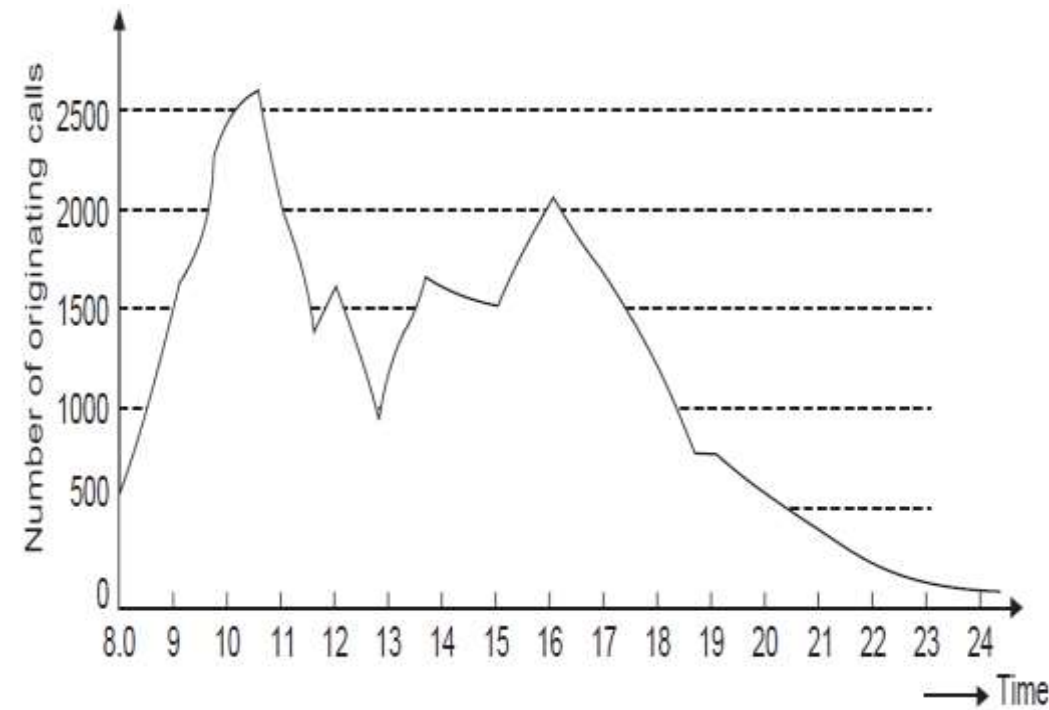


Fig. 4.1. Variations of call from 8 A.M. to midnight

4.1.1. Traffic pattern

- **Busy hour.** Continuous 60 minutes interval for which the traffic volume or the number of call attempts is greatest.
- **Peak busy hour.** It is the busy hour each day varies from day to day, over a number of days.
- **Time consistent busy hour.** The 1 hour period starting at the same time each day for which the average traffic volume or the number of call attempts is greatest over the days under consideration.

4.1.2. Traffic Theory

- **Homogeneous type:** Describe the classical telecommunication services based on voice transmission and switching
- **Heterogeneous type:** Includes integrated traffic streams from different sources (voice, audio, video, data) into a single network
- Covers specific types of random processes in telecommunications
 - ➡ Average connection duration
 - ➡ Average number of users
 - ➡ Busy time
 - ➡ Service time
 - ➡ Call arrival

4.1.3: Why Traffic Engineering

- Required in telecommunications network planning to ensure that network costs are minimized without compromising the quality of service delivered to the user of the network.
- ➡ It is based on probability theory and can be used to telecommunications networks.

4.1.4: Traffic Design Requirements

Due to peak hours, business hours, seasons, weekends, festival, location of exchange, University campus area, tourism area etc., the traffic is unpredictable and random in nature. Therefore, Traffic Pattern/characteristics of an exchange should be analyzed for the system design. The **Grade of Service** and the **Blocking Probability** are also important parameters for the traffic study.



4.1.5. Traffic statistics

The statistical descriptions of a traffic is important for the *analysis* and *design* of a *Switching* System.

- **Calling rate:** The average number of requests for connection that are made per unit time.
- **Holding Time:** the length of time that a resource is being held (e.g. duration of a phone call)
- **Service Rate** in calls per hour is given as $\mu = 1/h$

4.1.5. Traffic statistics ...

- **Distribution of destinations:** Probability of a call request being for particular destination (Note: *Number of calls receiving at a exchange may be destined to its own exchange or remote exchange or a foreign exchange*)
- **User behavior:** These behavior varies person to person and also depends on the situation. *The statistical properties of the switching system are a function of the behavior of users who encounter call blocking.*

4.1.5. Traffic statistics

- **Average occupancy:** Suppose, the average number of calls to and from a terminal during a period T seconds is 'n' and the average holding time is 'h' seconds, the average occupancy of the terminal is given by

$$A = nh/T = \lambda h = \lambda/\mu$$

- Therefore, *Average occupancy is the ratio of average arrival rate to the average service rate*
- **Traffic volume** - for an interval is the sum of all the traffic holding times for that interval

4.1.5. Traffic statistics

- **Erlangs** - describe traffic intensity in terms of the number of hours of resource time required per hour of elapsed time.
- ***Centum Call Seconds (CCS)*** - measures the exact same traffic intensity as the Erlangs but expresses it as the number of 100 second holding times required per hour. Traffic registers sample stations every 100 seconds per hour to check for busies. Since there are 36 sets of hundred seconds in an hour.
- $CCS = 36 \times \text{Erlangs}$

4.1.6. Traffic Measurement Units

■ **Erlangs:** (Traffic intensity) (named after of a Danish mathematician) is the average number of calls simultaneously in progress over a certain time. It is a dimensionless unit.

➡ **Erlang**

- *One hour of continuous use of one channel = 1 Erlang*
- *1 Erlang = 1 hour (60 minutes) of traffic*

➡ In data communications, an 1 E = 64 kbps of data

➡ In telephone, 1 Erlang = 60mins = 1 x 3600 call seconds

➡ **% of Occupancy**



A.K. Erlang, 1878-1929

4.1.6. Traffic Measurement Units

Example 4.1

A group of user made 30 calls in one hour, and each call had an average duration of 5 minutes, then the number of Erlangs this represents is worked out as follows:

■ **Minutes of traffic in the hour = number of calls x duration**

➡ *Minutes of traffic in the hour = 30×5*

➡ *Minutes of traffic in the hour = 150*

➡ *Hours of traffic in the hour = $150 / 60$*

➡ *Hours of traffic in the hour = 2.5*

➡ *Traffic figure = 2.5 Erlangs*

4.1.6. Traffic Measurement Units

Example 4.2

If a group of 20 trunk carries 10erlangs and the average call duration is 3mins, calculate (a) average number of calls in progress (b) total number of calls originating per hour:

Solution

*a) Traffic intensity per trunk = 10erlangs/20 = **0.5 erlangs/trunk***

Average no. of calls per trunk for 1erlang for 60 minutes = 20

For 0.5erlang, average no. of calls in progress = 10.

a) Traffic intensity = $A = nh/T = 10$ erlangs

total number of calls originating per hours

$$n = \frac{10 * 60}{3} = 200 \text{ calls}$$

4.1.6. Traffic Measurement Units

Example 4.3

Consider a group of 1200 subscribers who generate 600 calls during the busy hour. The average holding time is 2.2 minutes. What is the offered traffic in Erlang, CCS and call minutes (CM).

Solution

- ✓ *Traffic intensity in erlangs = $A = nh/T = \frac{600 * 2.2}{60} = 22 \text{ erlangs}$*
- ✓ *Traffic intensity in CCS = $36E = 36 * 22 = 792 \text{ CCS}$*
- ✓ *Traffic intensity in CS = $100 * \text{CCS} = 792 * 100 = 79200 \text{ CS}$*
- ✓ *Traffic intensity in CM = $\text{CS}/60 = 79200/60 = 1320 \text{ CM}$*

Waiting Line Systems

(Queueing Models)

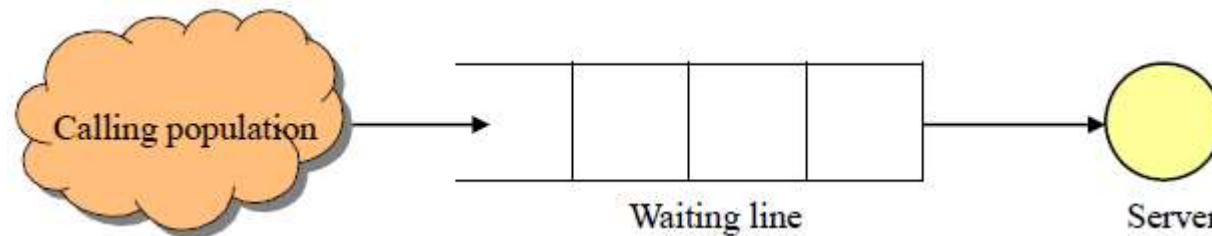
Part II

4.2. Waiting line systems

- 4.3.1. Characteristics of Queueing Systems
- 4.3.2. Queueing Notation – Kendall Notation
- 4.3.3. Long-run Measures of Performance of Queueing Systems
- Networks of Queues

4.2.1. Waiting line systems - Introduction

- **A waiting line system (or queuing system):** is defined by two elements: the population source of its customers (Calls) and the process or service system (Switch) itself.



- Queueing models provide the analyst with a powerful tool for designing and evaluating the performance of queueing systems.
 - Typical measures of system performance
 - Server utilization, length of waiting lines, and delays of customers

4.2.1. Waiting line systems - Introduction

- Used for analyzing network performance.
- In circuit switched networks want to know call blocking probability
 - How many circuits do we need to limit the blocking probability?
- In packet networks, events are random
 - Random packet arrivals
 - Random packet lengths
- While at the physical layer we were concerned with bit-error-rate, at the network layer we care about delays
 - How long does a packet spend waiting in buffers ?
 - How large are the buffers ?

4.2.2 Characteristics of Queueing Systems

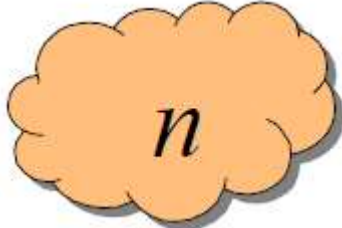
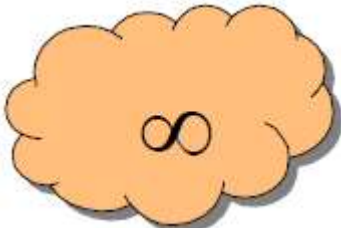
- **Arrivals or inputs to the system.** These have characteristics such as population size, behavior, and a statistical distribution.
- **Queue discipline, or the waiting line itself.** Characteristics of the queue include whether it is limited or unlimited in length and the discipline of people or items in it.
- **The service facility (Server)** Its characteristics include its design and the statistical distribution of service times.

4.2.2 Characteristics of Queueing Systems

- **Customer:** refers to anything that arrives at a facility and requires service, e.g., people, machines, trucks, emails, packets, frames.
- **Server:** refers to any resource that provides the requested service, e.g., repairpersons, machines, host, switch, router, disk drive, algorithm.

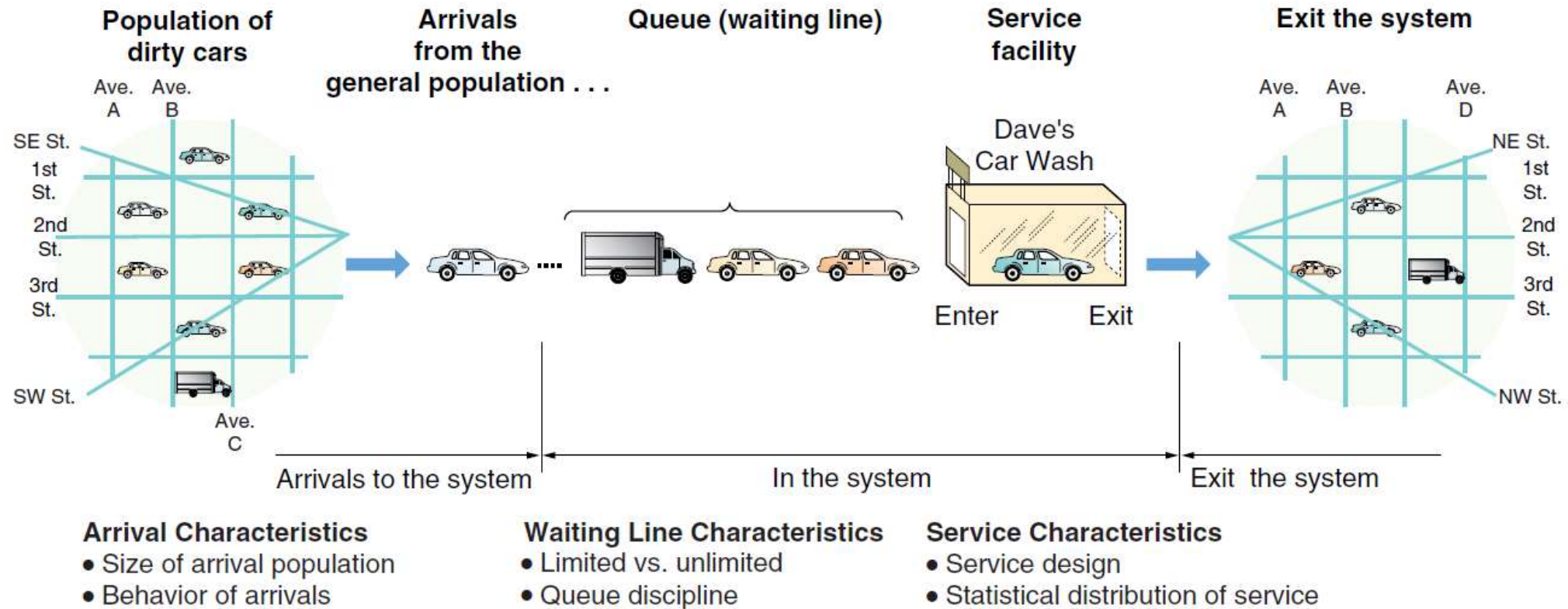
System	Customers	Server
Reception desk	People	Receptionist
Hospital	Patients	Nurses
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Computer	Jobs	CPU, disk, CD
Network	Packets	Router

4.2.2 Characteristics of Queueing Systems

- The input source that generates arrivals or customers for a service system has three major characteristics:
 - Size of the arrival population (Number of calls) are considered either unlimited (essentially infinite) or limited (finite)
 - Behavior of arrivals
 - Pattern of arrivals (statistical distribution).
- **Finite population model:** *if arrival rate depends on the number of calls/customers being served and waiting,*
- **Infinite population model:** *if arrival rate is not affected by the number of customers being served and waiting, e.g., systems with large population of potential customers.*

4.2.3 Pattern of Arrivals at the System

- Customers/Calls arrive at a server either according to known schedule or randomly.
- Random arrivals are independent of one another and their occurrence cannot be predicted exactly.



4.2.4. Arrival types

- **Random arrivals:** interarrival times usually characterized by a probability distribution.
 - *Most important model: Poisson arrival process (with rate λ), where a time represents the interarrival time between customer $n-1$ and customer n , and is exponentially distributed (with mean $1/\lambda$).*
- **Scheduled arrivals:** interarrival times can be constant or constant plus or minus a small random amount to represent early or late arrivals.
 - *Example: patients to a physician or scheduled airline flight arrivals to an airport*
- At least one customer is assumed to always be present, so the server is never idle, e.g., sufficient raw material for a machine.

4.2.4 Arrival types

- In queuing problems, the number of arrivals per unit of time can be estimated by a probability distribution known as the Poisson distribution
- **Poisson distribution:** A discrete probability distribution that often describes the arrival rate in queuing theory.

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } x = 0, 1, 2, 3, 4, \dots$$

where $P(x)$ = probability of x arrivals

x = number of arrivals per unit of time

λ = average arrival rate

$e = 2.7183$ (which is the base of the natural logarithms)

4.2.5 Queue Behavior and Queue Discipline

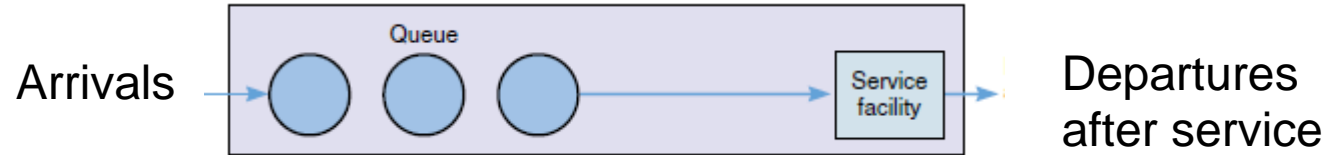
- **Queue behavior:** the actions of customers while in a queue waiting for service to begin, for example.
 - *Balk:* leave when they see that the line is too long
 - *Reneg:* leave after being in the line when its moving too slowly
 - *Jockey:* move from one line to a shorter line
- **Queue discipline:** the logical ordering of customers in a queue that determines which customer is chosen for service when a server becomes free, for example:
 - *First-in-first-out (FIFO)*
 - *Last-in-first-out (LIFO)*
 - *Service in random order (SIRO)*
 - *Shortest processing time first (SPT)*
 - *Service according to priority (PR)*

4.2.6 Service Times and Service Mechanism

- Service times of successive arrivals are denoted by S_1, S_2, S_3 .
 - *May be constant or random.*
 - *$\{S_1, S_2, S_3, \dots\}$ is usually characterized as a sequence of independent and identically distributed (IID) random variables, e.g., Exponential, Gamma, Lognormal, and Truncated normal distribution.*
- A queueing system consists of a number of service centers and interconnected queues.:
 - *Each service center consists of some number of servers (c) working in parallel, upon getting to the head of the line, a customer takes the 1st available server.*

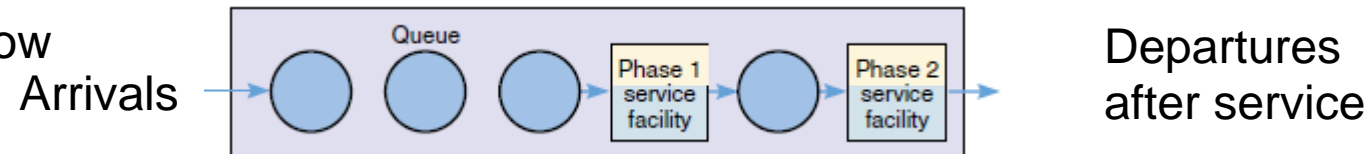
4.2.6 Service Mechanism - Examples

A family
dentist's office



Single-channel, single-phase system

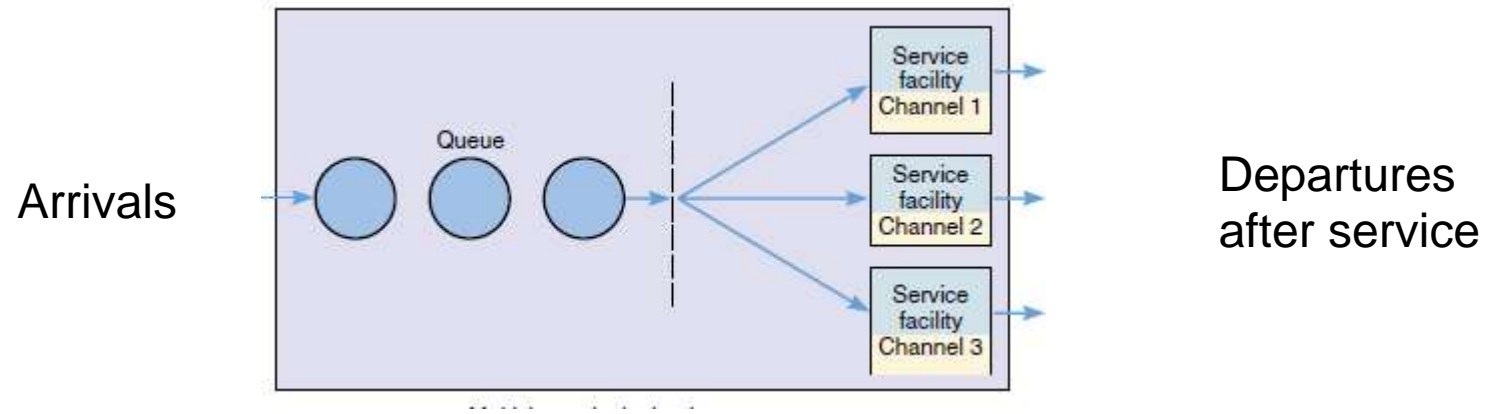
McDonald's dual-window
drive-through



Single-channel, multiphase system

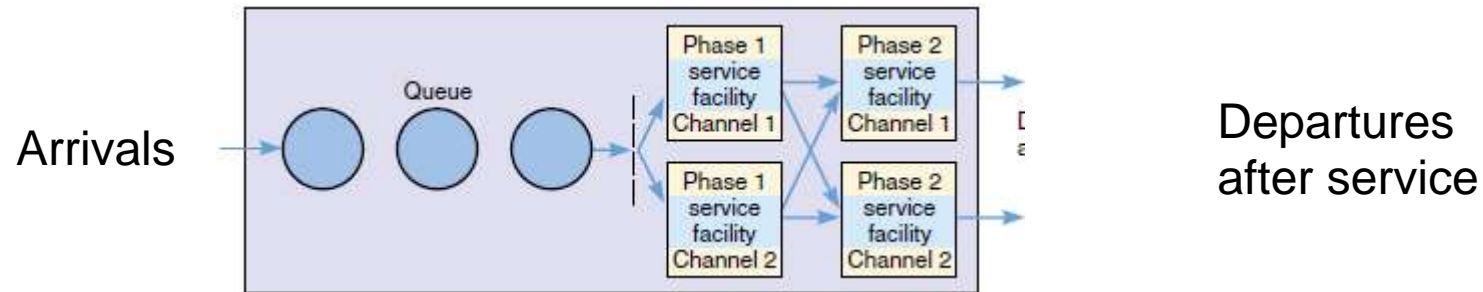
4.2.6 Service Mechanism - Examples

Most bank and post office service windows



Multichannel, single-phase system

Some college registrations



Multichannel, multiphase system

4.4 Queueing /The Kendall Notation

- A notation system for parallel server queues:

A/B/c/N/K

- **A** represents the interarrival-time distribution
- **B** represents the service-time distribution
- **c** represents the number of parallel servers
- **N** represents the system capacity
- **K** represents the size of the calling population
- **Take Note**

N, K are usually dropped, if they are infinity

4.2.7. Queueing /The Kendall Notation

■ Common symbols for A and B

- M Markov, exponential distribution
- D Constant, deterministic
- E_k Erlang distribution of order k
- H Hyperexponential distribution
- G General, arbitrary

■ Examples

- M/M/1/ ∞ / ∞ same as M/M/1: Single-server with unlimited capacity and call population. Interarrival and service times are exponentially distributed
- G/G/1/5/5: Single-server with capacity 5 and call-population 5.
- M/M/5/20/1500/FIFO: Five parallel server with capacity 20, call-population 1500, and service discipline FIFO

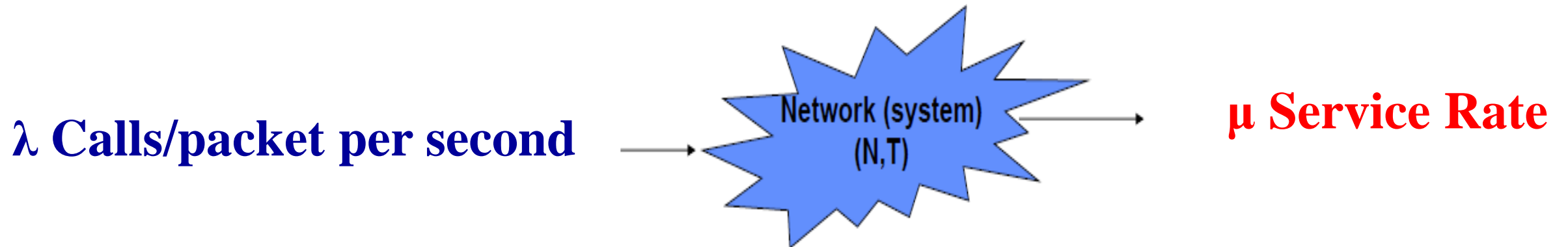
4.2.7 Queueing Notation

■ General performance measures of queueing systems:

- P_n *steady-state probability of having n customers in system*
- $P_n(t)$ *probability of n customers in system at time t*
- λ *arrival rate*
- λ_e *effective arrival rate*
- μ *service rate of one server*
- ρ *server utilization*
- A_n *interarrival time between customers $n-1$ and n*
- S_n *service time of the n -th arriving customer*
- W_n *total time spent in system by the n -th customer*

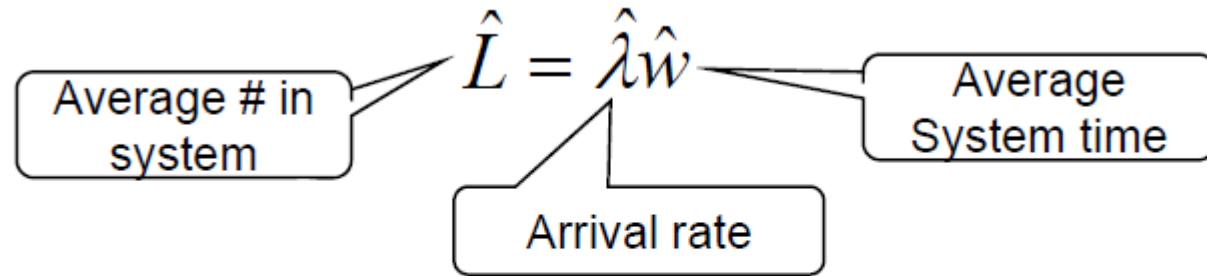
4.2.8. The Conservation Equation: Little's Law

- One of the most common theorems in queueing theory
- Mean number of customers in system
- Conservation equation (a.k.a. Little's law)



Average number in system = arrival rate \times average system time

4.2.8. The Conservation Equation: Little's Law



$$L = \lambda w \quad \text{as } T \rightarrow \infty \text{ and } N \rightarrow \infty$$

- *Holds for almost all queueing systems or subsystems (regardless of the number of servers, the queue discipline, or other special circumstances).*
- *G/G/1/N/K example (cont.): On average, one arrival every 4 time units and each arrival spends 4.6 time units in the system. Hence, at an arbitrary point in time, there are $(1/4)(4.6) = 1.15$ customers present on average.*

4.2.8. Little's theorem

Calling rate: Average number of requests for connection made per unit time; defined as

$$\lambda = \frac{n}{T} \quad \longrightarrow \quad n = \lambda T$$

n average number of calls to/from a terminal

λ calling rate

period T seconds

4.2.8. Single-server Waiting Line Model

- The customers are patient and come from a population that can be considered infinite.
- Customer arrivals are described by a *Poisson distribution* with a mean arrival rate of λ (lambda).
- This means that the time between successive customer arrivals follows an exponential distribution with an average of $1/\lambda$
- The customer service rate is described by a Poisson distribution with a mean service rate of μ (mu).
- This means that the service time for one customer follows an exponential distribution with an average of $1/\mu$
- The waiting line priority rule used is first-come, first-served.
- Using these assumptions, we can calculate the operating characteristics of a waiting line system using the following formulas:

4.2.9. Single-server Waiting Line Model

λ mean arrival rate of customers (average number of customers arriving per unit of time)

μ mean service rate of customers (average number of customers that can be served per unit of time)

$\rho = \frac{\lambda}{\mu}$ the average utilization of the system

$L = \frac{\lambda}{\mu - \lambda}$ the average number of customers in the service system

$L_Q = \rho L$ the average number of customers waiting in line

$W = \frac{1}{\mu - \lambda}$ the average time spent waiting in the system, including service

$W_Q = \rho W$ the average time spent waiting in line

$P_n = A = (1 - \rho)\rho^n$ the probability that n customers are in the service system at a given time

4.2.10. Grade of Service

- Is a measure of the call blocking in voice traffic, where resources allocation is deterministic (allocation and switching of channels)
- The ability to make call during the busiest time
- Is typically given as the likelihood that a call is blocked or the likelihood of a call experiencing a delay greater than a certain queuing time.
- Is determined by the available number of channels and used to estimate the total number of users that a network can support.

4.2.10. Grade of Service

For example, if $GOS = 0.05$, one call in 20 will be blocked during the busiest hour because of insufficient capacity

4.2.10. Grade of Service

$$GOS = \frac{\textit{Blocked Busy Hour calls}}{\textit{Offered Busy Hour calls}}$$

$$GOS = \frac{A - A_0}{A}$$

where

A_0 = carried traffic

A = offered traffic

$A - A_0$ = lost traffic.

4.2.10. Grade of Service

Example 4.4

During a busy hour, 1400 calls were offered to a group of trunks and 14 calls were lost. The average call duration has 3 minutes. Find (a) Traffic offered (b) Traffic carried (c) GOS and (d) The total duration of period of congestion

Solution:

a) During a busy hour $A = \frac{1400 \times 3}{60} = \mathbf{70E}$

b) a) Traffic carried $A_0 = \frac{1386 \times 3}{60} = \mathbf{69.3E}$

where $A - A_0 = 70 - 69.3 = 0.7 \text{ E}$ (lost traffic)

c) $GOS = \frac{0.7}{69.3} = \mathbf{0.01}$

(d) Total duration = $0.01 \times 3600 = 36 \text{ seconds}$

4.2.11. Traffic Intensity

Traffic Intensity: Is a measure of the average occupancy of a resource during a specified period of time, normally a busy hour.

$$A = \mu H \text{ Erlangs}$$

where

- ➡ H is the average holding time of a call
- ➡ μ is the average number of call requested/hour

For example, if $GOS = 0.05$, one call in 20 will be blocked during the busiest hour because of insufficient capacity

4.2.12. Offered Traffic

Example 4.5

Consider a PSTN which receives 240 calls/hr. Each call lasts an average of 5 minutes. What is the outgoing traffic intensity to the public network.

Solution:

$$A = \mu * H$$

$$\mu = 240 \text{ calls/hr and } H = 5 \text{ minutes}$$

$$A = (240 \text{ calls /hr}) \times (5 \text{ min/call}) = 1200 \text{ min/hr}$$

Erlang cannot have any unit so

$$A = 1200 \text{ min/hr} * (1 \text{ hour}/60 \text{ minutes}) = 20 \text{ Erlangs}$$

So 20 hours of circuit talk time is required for every hour of elapsed time. An average of T1 voice circuits busy at any time is 20. (Or 20 hours of continuous use of 20 channels.)

4.2.12. Traffic Congestion

- **Traffic Congestion** is the probability that the offered traffic load exceeds predefined value or capacity of a resource; hence, no new calls can be accepted.
- There are two ways of specifying congestion.
- **Time congestion** is the percentage of time that all servers in a group are busy.
- **Call Or Demand Congestion** is the proportion of calls arising that do not find a free server.

Assignment

The Electronic lab at DIT - ETE Department has technicians to assist students working on practical. The students patiently form a single line to wait for help. Students are served based on a first-come, first-served priority rule. On average, 15 students per hour arrive at the lab. Student arrivals are best described using a Poisson distribution. The technicians can help an average of 20 students per hour, with the service rate being described by an exponential distribution. Calculate the following operating characteristics of the service system.

- A. The average utilization of the technicians
- B. The average number of students in the system
- C. The average number of students waiting in line
- D. The average time a student spends in the system
- E. The average time a student spends waiting in line
- F. The probability of having more than 4 students in the system

End of Part 2

Traffic Engineering

Modelling of Traffic

Part III

4.3. Modelling of Traffic

- To analyze the statistical characteristics of a switching system, *traffic flow* and *service time*, it is 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 behavior of the quantity under consideration. Also the model provides an analytical solution to a teletraffic 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.

4.3.1. Loss Systems

Service rate of a Switching system 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 economical and not possible for number of subscriber equal number of lines.

- A Blocked Condition exist where the incoming calls finds all available lines busy.
- *Service rate* of depends on the number of lines. If available *lines equal to subscribers*, there is no question of *traffic analysis*. But it is not possible and it is uneconomical.

4.3.1. Loss Systems

- The Erlang loss system may be defined by the following specifications.
 - ➡ *The arrival process of calls is assumed to be Poisson with a rate of λ calls per hour.*
 - ➡ *The holding times are assumed to be mutually independent and identically distributed random variables following an exponential distribution with $1/\mu$ seconds.*
 - ➡ *Calls are served in the order of arrival.*

4.3.1. Loss Systems

- There are three models of loss systems.
 - *Lost calls cleared (LCC)*
 - *Lost calls returned (LCR)*
 - *Lost calls held (LCH)*
- In a loss system some calls are lost
 - A call is lost if all N channels are occupied when the call arrives
 - The term **Blocking** refers to this event

4.3.1. Loss Systems - Blocking

■ There are two different types of blocking quantities

➔ **Call blocking B_C** = *Probability that an arriving call finds all N channels occupied*
= *the fraction of calls that are lost*

➔ **Time blocking B_t** = probability that all N channels are occupied at an arbitrary time
= the fraction of time that all N channels are occupied

■ The two blocking quantities are not necessarily equal

➔ Example: your own mobile

➔ But if calls arrive according to a Poisson process, then $B_C = B_t$

➔ **Call blocking is a better measure for the quality of service experienced by the subscribers but, time blocking is easier to calculate**

4.3.1. Loss Systems - Call rates

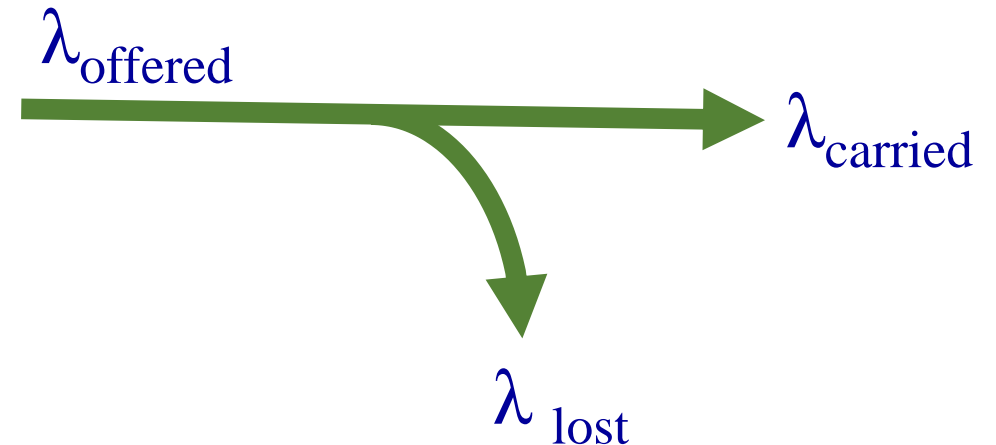
- In a loss system each call is either lost or carried. Thus, there are three types of call rates:

➡ λ_{offered} = arrival rate of all call attempts

➡ λ_{carried} = arrival rate of carried calls

➡ λ_{lost} = arrival rate of lost calls

$$\lambda_{\text{offered}} = \lambda_{\text{carried}} + \lambda_{\text{lost}} = \lambda$$



4.4. Loss Systems - Call rates

■ The three call rates lead to the following three traffic concepts:

➡ Traffic offered $A_{\text{offered}} = \lambda_{\text{offered}} * h$

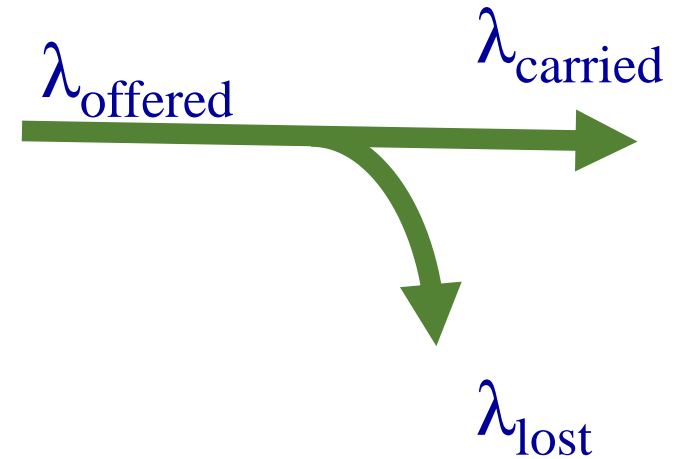
➡ Traffic carried $A_{\text{carried}} = \lambda_{\text{carried}} * h$

➡ Traffic lost $A_{\text{lost}} = \lambda_{\text{lost}} * h$

$$A_{\text{offered}} = A_{\text{carried}} + A_{\text{lost}} = A$$

$$A_{\text{carried}} = A * (1 - B_C)$$

$$A_{\text{Lost}} = A * B_C$$



4.3.1. Loss Systems

- **Erlang B Formula:** All blocked calls are cleared; The most common
- *Engset formula (probability of blocking in low density areas); used where B model fails.*
- **Extended Erlang B:** Similar to Erlang B, but takes into account that a percentage of calls are immediately represented to the system if they encounter blocking (a busy signal). The retry percentage can be specified.
- **Erlang C Formula:** Blocked calls delayed or held in queue Indefinitely
- *Poisson Formula: Blocked calls held in queue for a limited time only.*
- *Binomial Formula: Lost calls held*

4.3.1. Loss Systems

- Quality of service (QoS) is expressed in terms of blocking probability as:

$$P_B = (A * C) * B$$

Where

- ➡ $B = \text{Erlang} - B \text{ Formula}$
- ➡ $A = \text{The traffic intensity}$
- ➡ $C=N = \text{No of channels (lines)}$

4.3.1. Lost calls cleared (LCC) Model

- 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 model also known as Erlang B Model*
- 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*

4.3.1. Lost calls cleared (LCC) Model

$$P_B(C, A) = \frac{\frac{A^C}{C!}}{\sum_{k=0}^C \frac{A^k}{k!}}$$

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

where

- ➡ *A is the traffic intensity*
- ➡ *C=N is the number of channels*

4.3.1. Lost calls cleared (LCC) Model

Erlang B Model – Characteristics

- Provides the probability of blockage at the switch due to congestion.

Assumptions:

- ➡ No waiting is allowed (lost calls are cleared) (*i.e. they disappear from the system. This assumption is valid for systems that can overflow blocked calls onto another trunk (e.g. a high usage trunk)*)

4.3.1. Lost calls cleared (LCC) Model

- Traffic originated from an infinite numbers of sources
- Limited No. of trunk (or serving channels)
- Memory-less, channel requests at any time
- The probability of a user occupying a channel is based on exponential distribution
- Calls arrival rate at the network = Poisson process (the holding time or duration of the call has exponentially distribution)

4.3.1. Lost calls cleared (LCC) Model

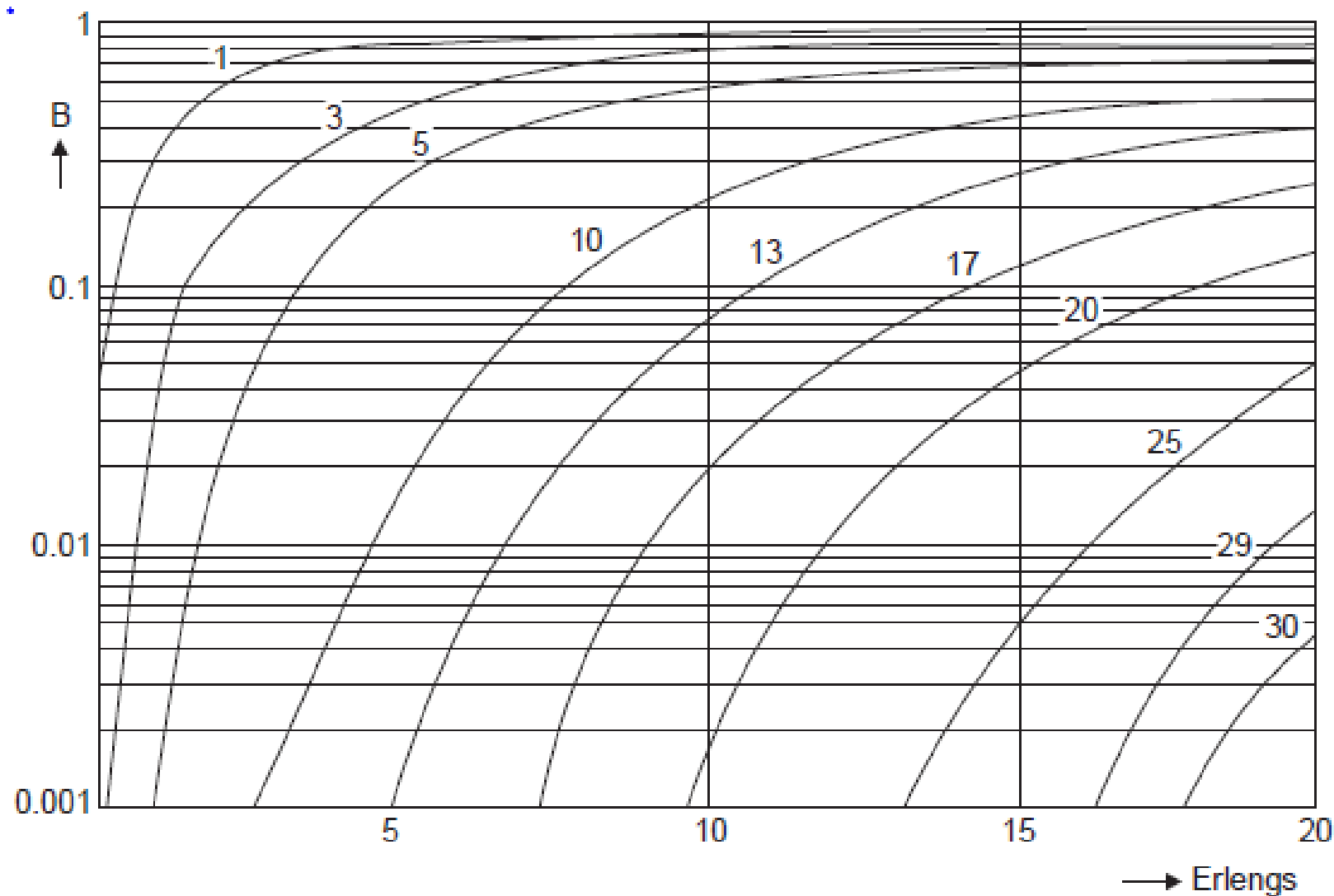


Fig. 4.5: Erlang B Chart

Assignment

- By applying Erlang B formula develop Erlang B Chart

4.3.1. Lost calls cleared (LCC) Model

- The carried traffic is

$$A_{ca} = A [1 - P_B]$$

- The efficiency of the channel usage is

$$\eta = \frac{A_{ca}}{C}$$

- ➡ The start-up systems usually begins with a GOS of 0.02 (2% of the blocking probability) rising up to 0.5 as the system grows.
- ➡ If more subscribers are allowed in the system the blocking probability may reach unacceptable values.

4.3.1. Lost calls cleared (LCC) Model

Example 4.6

Consider a trunk group with an offered load 4.5E and a blocking probability of 0.01. If the offered traffic increased to 13E, to keep same blocking probability, find the number of trunks needed. Also calculate the trunk occupancies.

Solution:

Given $A = 4.5$, $B = 0.01$ From Erlang Table No. of trunks $(N/C) = 10$

For the increase in load of 13E, from Erlang table $N/C = 21$ for same $B = 0.01$ required

The trunk occupancies calculated as

$$\eta = \frac{A_{ca}}{C} = \frac{A [1 - P_B]}{C}$$

For $N = 10$, $A = 4.5$ $\eta = 4.5(1 - 0.001)/10 = 0.4455$

For $N = 21$, $A = 13$ $\eta = 13(1 - 0.001)/10 = 0.613$

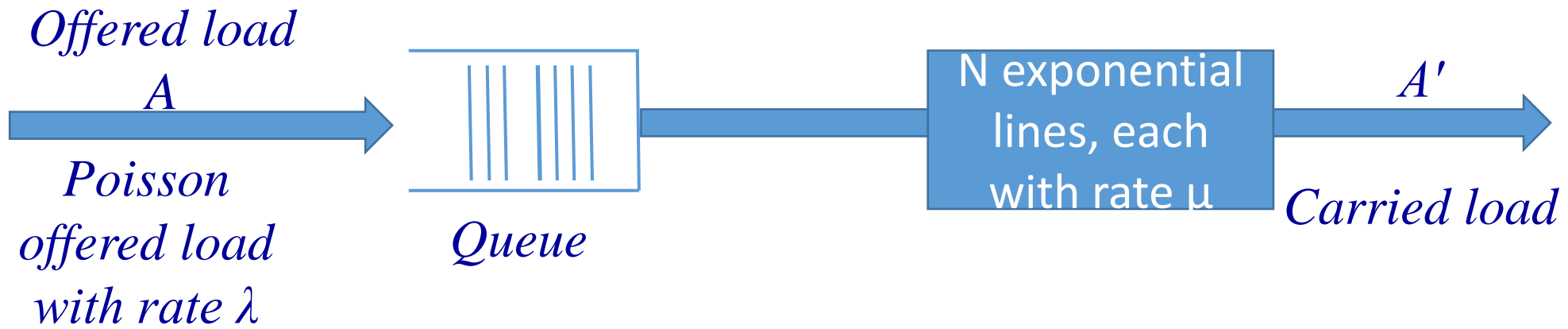
4.3.2. Lost Calls Held (LCH) System

- 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.
- LCH systems generally arise in real time applications in which the sources are continuously in need of service, whether or not the
- Normally, telephone network does not operate in a lost call held manner

Modelling of a Delay System

Part IV

4.4. Delay System



4.4. Delay System

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 facilities is available. These are referred to as delay, waiting-call and Queuing systems.

- The delay systems are analyzed using queuing theory which is sometimes known as waiting line theory.
- Note: *The basic purpose of the investigation of delay system is to determine the probability distribution of waiting times.*

4.4. Delay System

- The service discipline of the queue involves two important factors.
 - *Waiting calls are selected on of FCFS) or FIFO service*
 - *The blocking probability or delay probability in the system is based on the queue size in comparison with number of effective sources.*

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

where

B = Blocking probability for a LCC system

N = Number of servers

A = Offered load (Erlangs)

These Equations are referred as Erlang second, Erlang delay or C formula

4.4. Delay System

- For single server systems ($N = 1$), the Prob. of delay reduces to % utilization (ρ) of the resource, i.e. the output utilization or traffic carried
- The distribution of waiting times for random arrivals, random service times, and a FIFO service discipline is

$$(P > t) = P(> 0) e^{-(N-A)t/h}$$

Where

$P(> 0)$ = probability of delay

h = average service time of negative exponential service time distribution.

4.4. Delay System

- The average waiting time for all arrivals can be determined

$$W(t)_{avg} = \frac{C(N, A)h}{N - A}$$

$W(t)_{avg}$ is the expected delay for all arrivals.

- The average delay of only those arrivals that get delayed is denoted as

$$T_w = \frac{h}{N - A}$$

4.4. Delay System

Example 4.7

A message switching network is to be designed for 90% utilization of its transmission link. Assuming exponentially distributed message lengths and an arrivals rate of 10 messages per min. What is the average waiting time and what is the probability that the waiting time exceeds 3 minutes ?

4.4. Delay System

Example 4.7: Solution

Given $A = \rho = 90\% = 0.9$. $\lambda = 10$ messages/minute. Assume $N = 1$

For $N = 1$, prob (delay) = $P(> 0) = \rho = 0.9$. Also $A = \rho = 0.9$.

The average service time $h = \text{Prob. (delay)}/\lambda = 0.9/10 = 0.09$

Average waiting time $W(t)_{avg} = P(>0)h/(N - A) = 0.9 \times 0.09/(1 - 0.9) = 0.81$ min.

Prob. of the waiting time exceeding 3 minutes

$$= P(> 3) = P(> 0) e^{-(N-A)t/h} = 0.9 \times e^{-(1-0.9)3/0.09} = 0.032$$

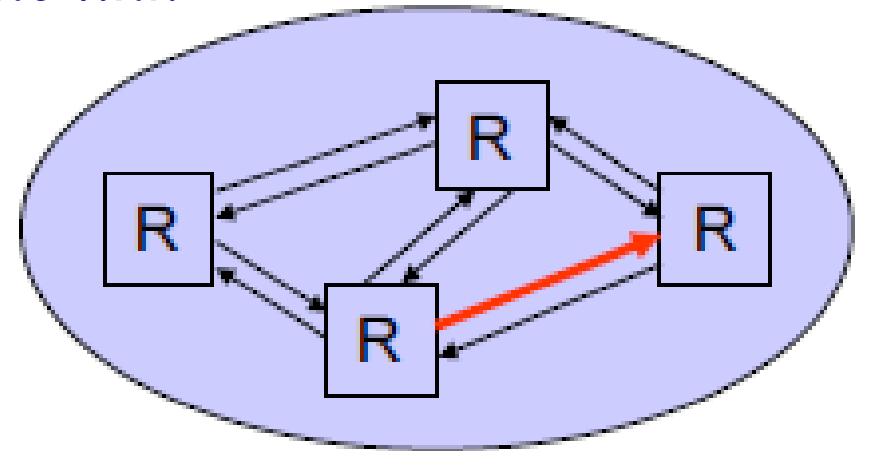
Thus 3.2% of the message experience queuing delay of more than 3 minutes.

Packet Level Model For Data Traffic

Part V

4.5. Packet Level Model For Data Traffic

- Queuing models are suitable for describing (packet-switched) data traffic at packet level
 - ➔ *Pioneering work made by many people in 60's and 70's related to ARPANET, in particular L. Kleinrock (<http://www.lk.cs.ucla.edu/>)*
- Consider a link between two packet routers
 - ➔ *Traffic consists of data packets transmitted along the link*



4.5. Packet Level Model For Data Traffic

This can be modelled as a pure queuing system with a single server ($n = 1$) and an infinite buffer ($m = \infty$)

■ customer = packet

➔ λ = packet arrival rate (packets per time unit)

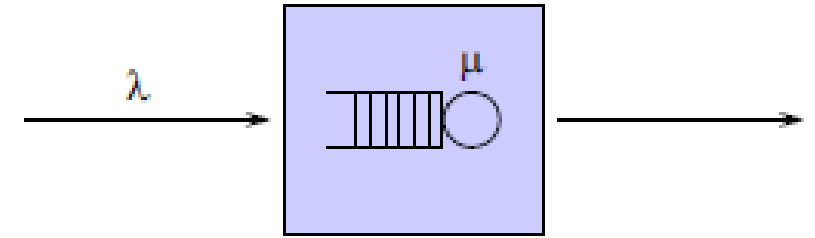
➔ L = average packet length (data units)

■ server = link, waiting places = buffer

➔ C = link speed (data units per time unit)

■ Service time = packet transmission time

➔ $1/\mu = L/C$ = average packet transmission time (time units)



4.5. Packet Level Model For Data Traffic

- The *strength of the offered traffic* is described by the traffic load ρ
- By definition, the traffic load ρ is the ratio between the arrival rate λ and the service rate $\mu = C/L$:

$$\rho = \frac{\lambda}{\mu} = \frac{\lambda L}{C}$$

- System capacity
 - ➔ C = link speed in kbps
- Traffic load
 - ➔ λ = packet arrival rate in pps (considered here as a variable)
 - ➔ L = average packet length in Kbits (assumed here to be constant 1kbit)

4.5. Packet Level Model For Data Traffic

Example 4.8

Consider a link between DIT & MUST packet routers. Assume that,

- ➡ *on average, 50,000 new packets arrive in a second,*
- ➡ *the mean packet length is 1500 bytes, and*
- ➡ *the link speed is 1 Gbps.*

Then the traffic load (as well as, the utilization) is

Solution:

$$\rho = 50,000 * 1500 * 8 / 1,000,000,000 = 0.60 = 60\%$$

4.5.1. Delay in Traffic Data

- In a queuing system, some packets have to wait before getting served
 - ➡ An arriving packet is buffered, if the link is busy upon the arrival
- Delay of a packet consists of
 - The waiting time, which depends on the state of the system upon the arrival, and
 - ➡ The transmission time, which depends on the length of the packet and the capacity of the link

4.5.1. Delay in Traffic Data

Example 4.9

Packet length = 1500bytes; Link speed = 1Gbps

*Transmission time = $1500 * 8 / 1,000,000,000 = 0.000012 \text{ s} = 12\mu\text{s}$*

Take Note:

This is from the Operator/ISP point of view

4.5.1. Delay in Traffic Data

- Quality of service (*From the users' point of view*)

- ➡ P_z = probability that a packet has to wait “too long”, i.e. longer than a given reference value z (assumed here to be constant $z = 0.00001\text{ s} = 10\mu\text{s}$)

- Assume an M/M/1 queuing system:

- ➡ Packets arrive according to a Poisson process (with rate λ)

- ➡ Packet lengths are independent and identically distributed according to the exponential distribution with mean L

4.5.1. Delay in Traffic Data

- Then the quantitative relation between the three factors (*system, traffic, and quality of service*) is given by the following formula:

$$P_z = \text{Wait}(C, \lambda, L, Z) = \begin{cases} \frac{\lambda L}{C} \exp\left(-\left(\frac{C}{L} - \lambda\right)Z\right) = \rho \exp(-\mu(1 - \rho)Z) & \text{if } \lambda L < C (\rho < 1) \\ 1, & \text{if } \lambda L \geq C (\rho \geq 1) \end{cases}$$

Take Note:

- ➡ *The system is stable only in the former case ($\rho < 1$). Otherwise the number of packets in the buffer grows without limits.*

4.6.1. Delay in Traffic Data

Example 4.10

A router installed at DIT Server Room; assume that packets arrive at rate $\lambda = 600,000$ (pps = 0.6 packets/ μ s) and the link speed is $C = 1.0$ Gbps = 1.0 kbit/ μ s. (a) Is the system Stable? (b) The probability P_z that an arriving packet has to wait too long (i.e. longer than $z = 10\mu$ s)

Solution:

*a) The system is stable since $\rho = 50,000 * 1500 * 8 / 1,000,000,000 = 0.60 = 60\%$*

*b) $P_z = \text{Wait}(1.0, 0.6; 1, 10) = 0.6 \exp(-(\frac{1 \text{Kbit}/\mu\text{s}}{1 \text{Kbit}} - 0.6 \text{p}/\mu\text{s}) * 10\mu\text{s}) = 0.6 \exp(-0.4) \approx 0.4\%$*

4.5.2. Throughput

- In a sharing system the service capacity is shared among all active flows. It follows that all flows get delayed (unless there is only a single active flow)
- Throughput: The ratio between the *average flow size S* and the *average total delay D* of a flow

$$\Theta = S/D$$

Example 4.11

- ➡ $S = 1 \text{ Mbit}$
- ➡ $D = 5 \text{ s}$
- ➡ $\Theta = S/D = 0.2 \text{ Mbps}$

4.5.2. Throughput

- System capacity

- C = link speed in Mbps

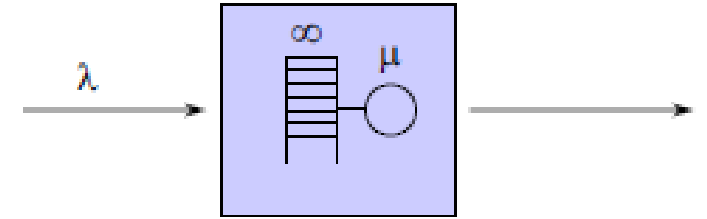
- Traffic load

- λ = flow arrival rate in flows per second (considered here as a variable)

- – S = average flow size in Kbits (assumed here to be constant 1Mbit)

- Quality of service (from the users' point of view)

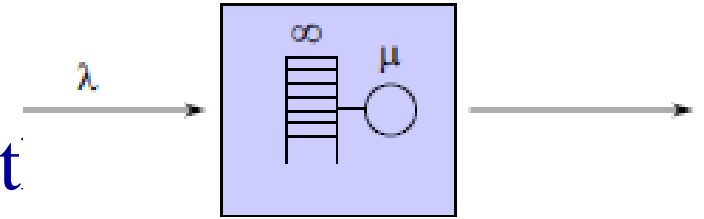
- Θ = throughput



4.5.2. Throughput

■ Assume an **M/G/1-PS** sharing system:

- Flows arrive according to a Poisson process (with rate λ)
- Flow sizes are independent and identically distributed according to any distribution with mean S



4.6.2. Throughput

- Then the quantitative relation between the three factors (system, traffic, and quality of service) is given by the following formula:

$$\vartheta = X_{put}(C, \lambda, S) = \begin{cases} C - \lambda S = \rho(1 - \rho), & \text{if } \lambda S < C (\rho < 1) \\ 0, & \text{if } \lambda S \geq C (\rho \geq 1) \end{cases}$$

- Take Note:

The system is stable only in the former case ($\rho < 1$). Otherwise the number of flows as well as the average delay grows without limits.

In other words, the throughput of a flow goes to zero

4.6.2. Throughput

Example 4.12

Assume that flows arrive at rate $\lambda = 600$ flows per second and the link speed is $C = 1000\text{Mbps} = 1.0\text{Gbps}$. (a) Is the system stable? (b) The Throughput

Solution:

*a) The system is stable since $\rho = \lambda S / C = 50,000 * 1500 * 8 / 1,000,000,000 = 0.60 = 60\%$*

b) $\Theta = X_{\text{put}}(1000, 600; 1) = 1000 - 600 = 400\text{Mbps} = 0.4\text{Gps}$

Reference

- 1.Manav T. V., (2015) "Telecommunication Switching Systems and Networks" PHI Learning private Limited. ISBN 9788120350830
- 2.Rubin M, Haller C. E., (1966) "Communication Switching Systems" Chapman et Hall. ISBN: 0882752324,
- 3.Hobbs M. (1974) "Modern Communications Switching Systems" G/L Tab Books, ISBN:0830646787,
V. S. Bagad, (2014) "Telecommunication Switching Systems and Networks" Technical Publication.
ISBN-10: 9350993724, ISBN-13: 978-9350993729
- 4.Lawrence V. B, Ahamed S. V., (1997) "Design and Engineering of Intelligent Communication Systems" Kluwer Academic Publishers. ISBN 9781461562917
- 5.Gnanasivam P. (2010) "Telecommunication Switching and Networks" New Age International Pvt Ltd Publishers. ISBN-10: 812241950X,

Any Questions ?

End of Lecture 04

Thank you Class for your Attention