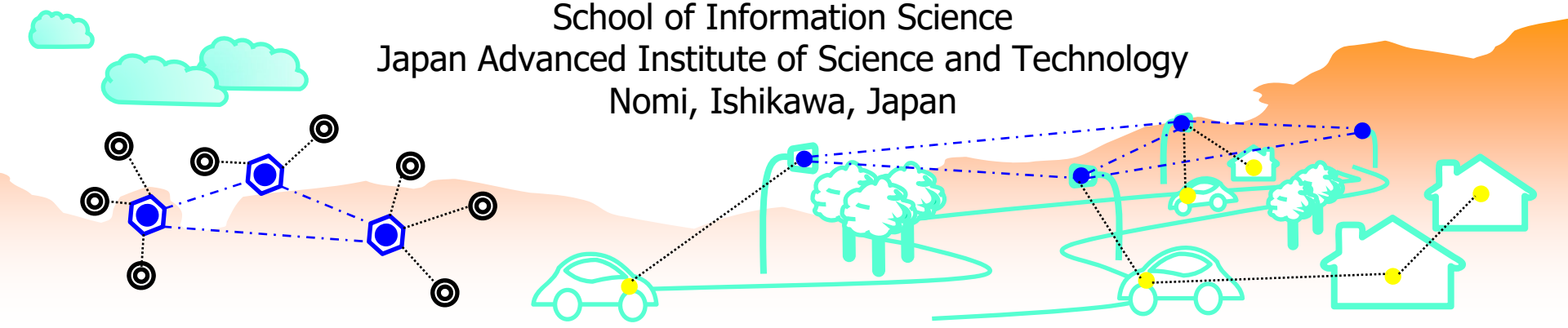
I226
Computer Networks

Chapter 10

Traffic and Communication Engineering

Assoc. Prof. Yuto Lim

School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan



Objectives of this Chapter

- Give an understanding what is the traffic theory
- Show a few examples of traffic intensity using Erlang B
- Offer the different between the QoS and GoS
- Provide the knowledge of queuing theory
- Also explain how the define the system works properly. Dependability?

Outline

- Traffic Theory
  - Traffic Engineering
  - Traffic Intensity
  - Erlang B Formula
  - Quality of Service (QoS) and Grade of Service (GoS)
- Queuing Theory
  - Little's Law
  - Kendall Notation
  - Analysis of M/M/1 queue
- Dependability
  - Failure in Time (FIT)

Traffic Theory

- Depends on the type of traffic in the network
  - Circuit switched network with homogenous/heterogeneous traffic
  - Packet switched network with homogenous/heterogeneous traffic
- Homogeneous type: describe the classical telecommunication services based on voice transmission and switching
- Heterogeneous type: includes integrated traffic streams from different sources (voice, text, image, video) 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

Traffic Engineering

- Requires network planning to ensure that network costs are minimized without compromising the quality of service (QoS) delivered to the user of the network
- Mobile radio networks have traffic issues that do not arise in the fixed line PSTN
  - Mobile handset that is moving in a cell, receives a signal with varying strength
  - This signal strength is subject to slow fading, fast fading, interference from other signals
  - Thus resulting in degradation of the carrier-to-interference (C/I) ratio

Traffic Engineering (cont.)

- Traffic engineering balances the following factors based on given amount of traffic
  - Resources (e.g., trunk channels)
  - Grade of Service (GoS)
- 2 types of systems implemented to provide voice communications
  - Blocking
    - Voice or data is blocked (by a busy signal) if network resource (e.g., trunk channel) is not available
    - GoS = blocking probability
  - Delay system
    - Voice or data is queued until network resource is available
    - GoS = queuing probability and average time in queue

Terms of Traffic Engineering

- Traffic intensity: traffic volume / time interval which is a measure of demand
- Traffic volume for an interval: the sum of all the traffic holding times for that interval
- Holding time: the length of time that a resource is being held (e.g., the duration of a phone call)
- Erlangs: describe traffic intensity in terms of the no. 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 100-second holding times required per hour
  - CCS = 1/36 x Erlangs

* Centum = hundred

Traffic Measurement Unit

- Erlangs
  - Traffic intensity (named after of a Danish mathematician, Agner Krarup Erlang) is the average number of calls simultaneously in progress over a certain time
  - Erlang is a dimensionless unit
  - 1 Erlang = 1 hour of continuous use of one channel or 1 hour of traffic
  - In telephone, 1 Erlang = 1 x 3600 call seconds
  - In data communications, 1 Erlang = 64 kbps of data
- Example: a group of user made 30 calls in 1 hour, and each call had an average call duration of 5 minutes, What is the number of Erlangs?
  - Minutes of traffic in one hour = number of calls x duration
= 30 x 5
= 150 minutes
  - Hours of traffic in one hour = 150 / 60
= 2.5
  - Traffic figure is 2.5 Erlangs

Example 1

A call established at 1:00 am between a mobile and MSC. Assuming a continuous connection and data transfer rate at 30 kbit/s, determine the traffic intensity if the call is terminated at 1:50 am.

Solution:

Traffic intensity = 1 call x 50 minutes x (1 hour/60 minute)

= 0.833 Erlangs

Notes:

- *Traffic intensity has nothing to do with the data rate, only the holding time is taken into account*
- *If the traffic intensity > 1 Erlang, then the incoming call rate exceeds the outgoing calls, thus resulting in queuing delay which will grow without bound (when the traffic intensity stays the same)*
- *If the traffic intensity < 1 Erlang, then the network can handle more average traffic*

Traffic Intensity

- Traffic intensity is a measure of the average occupancy of a resource during a specified period of time, normally a busy hour
- Traffic intensity offered by each user is
  - $A = \mu H$ [Erlangs]
  - μ is the average no. of call requested in one hour
  - H is the average holding time of a call
- U users in an unspecified no. of channels, total offered traffic intensity is
  - $A_T = UA$ [Erlangs]
- In a trunks system of C channels and equally. *Distributed* traffic among the channels, traffic intensity per channel is
  - $A_C = UA/C$ [Erlangs/channels]
- Traffic volume is a measure of the total work done by a resource or facility, normally over 24 hours
  - $V_T = AT$ [Erlangs-hours]

Example 2

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

Solution:

Traffic intensity, A = μ x H, μ = 240 calls/hour and H = 5 minutes/call

A = (240 calls/hour) x (5 minutes/call)

= 1200 minutes/hour

Because Erlang cannot have any unit, thus

A = 1200 minutes/hour x (1 hour/60 minutes)

= 20 Erlangs

Note that 20 hours of circuit talk time is required for every hour of elapsed time. An average number of T1 voice circuits busy at any time is 20. (or one hour of continuous use of 20 channels)

Offered Traffic

- Offered traffic: volume of traffic offered to a switch that are all processed
  - Total offered traffic intensity (A_T) = carried traffic + overflow traffic
  - Carried traffic (A_{ca}): the actual traffic carried by a switch
  - Overflow (blocked) traffic (A_{ov}): portion of the traffic not processed
- Busy hour call attempts (BHCA)
  - Used to evaluate and plan the capacity for telephone networks
  - Is the no. of telephone calls made at the peak hour
  - Higher the BHCA, higher the stress on the network processors
  - Not to be confused with busy hour call completion (BHCC), which truly measures the throughput capacity of the network

Traffic Intensity Models

- Erlang B formula: all blocked calls are cleared
  - Engset formula (probability of blocking in low density areas, i.e., M/M/C/C/N queue) is used where Erlang B formula fails
- Extended Erlang B formula: 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 is also 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

Erlang B Formula Characteristics

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

$$P_B = \frac{\frac{A^N}{N!}}{\sum_{x=0}^{N} \frac{A^x}{x!}}$$

P_B = Erlang B loss probability
N = Number of trunks in full availability group
A = Traffic offered to group in Erlangs

- 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)
  - Traffic originated from an infinite no. of sources
  - Limited no. of trunk (or serving channels)
  - Memory-less, channel requests at any time
  - 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)

Erlang B Formula Table

Number of servers (N)	Capacity (erlangs) for grade of service of: $P = 0.02$ (1/50)	$P = 0.01$ (1/100)	$P = 0.005$ (1/200)	$P = 0.002$ (1/500)	$P = 0.001$ (1/1000)
1	0.02	0.01	0.005	0.002	0.001
4	1.09	0.87	0.7	0.53	0.43
5	1.66	1.36	1.13	0.9	0.76
10	5.08	4.46	3.96	3.43	3.09
20	13.19	12.03	11.10	10.07	9.41
24	16.64	15.27	14.21	13.01	12.24
40	31.0	29.0	27.3	25.7	24.5
70	59.13	56.1	53.7	51.0	49.2
100	87.97	84.1	80.9	77.4	75.2

Erlang B Chart

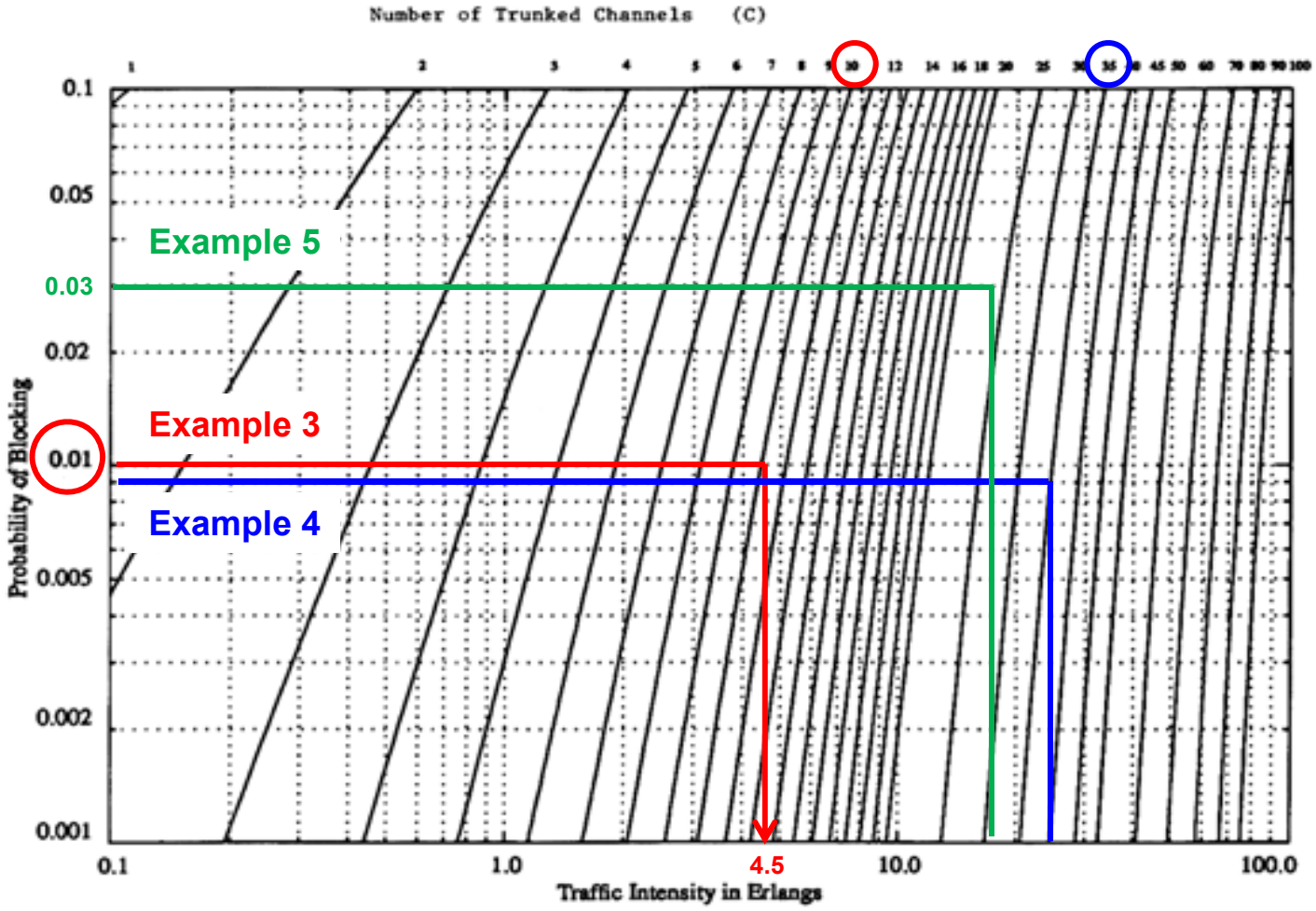


Figure: The Erlang B chart showing the probability of blocking as functions of the number of channels and traffic intensity in Erlangs

Blocking Probability, P_B

- GoS is expressed in terms of blocking probability as $P_B = A_T \times C$ where A_T is the total offered traffic intensity and C is the no. of channels
- Lost calls cleared
  - Assume that blocked calls are cleared (i.e., lost from the system. This assumption is valid for systems that can overflow blocked calls onto another trunk, e.g., a high usage trunk)
  - $A_T = A_{ca} / (1 - P_B)$
- Lost calls returning
  - Assume that blocked calls are re-tried until they are successfully carried. This assumption is valid for PBXs* and corporate lines
  - $A_T \geq A_{ca}$
- Efficiency of the channel usage, $\eta = A_{ca} / C$
  - 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

* PBXs= Private Branch eXchanges

Example 3

A single GSM service provider support 10 digital speech channels. Assume the probability of blocking is 1.0%. From the Erlang B chart find the traffic intensity. How many 3 minutes of calls does this represent?

Solution:

From the Erlang B chart, the traffic intensity, A is about 4.5 Erlangs.

A = 4.5 Erlangs / (1 hour/60 minutes) = 270 minutes/hour.

$\Rightarrow A = \mu \times H$

$\Rightarrow \mu = A / H$

$= (270 \text{ minutes/hours}) / (3 \text{ minutes/call})$

$= \underline{90 \text{ calls/hour}}$

Example 4

A telephone switching board at the UNN* can handle 120 phones. Assuming the followings, on average 5 calls/hour per phone, average call duration time = 4 minutes, 60% of all calls made are external, and GoS = 0.9%. Determine the outgoing traffic intensity and the number of channels.

Solution:

$\Rightarrow A_T = U \times A$

$\Rightarrow A_T = U \times \mu \times H$

$A_T = (60\%/100\%) \times (120 \text{ calls} \times 5 \text{ calls/hour}) \times (4 \text{ minutes/call})$

$= 1440 \text{ minutes/hour}$

Therefore, $A_T = 1440 \text{ minutes/hour} \times (1 \text{ hour}/60 \text{ minutes})$

$= \underline{24 \text{ Erlangs}}$

This means that 24 hours of circuit talk time is required for every hour of elapsed time. No. of channels, C from Erlang B chart is about <u>35</u>

* UNN= UNallocated Number

Example 5

Consider a telephone switching board with 150 phones. Assuming the number of calls is 3/hour/line, the average call duration is 4 minutes, and 58% of all call are made external via a T1* trunk (24 channels) to the PSTN. Determine carried traffic and channel usage.

Solution:

Total offered traffic is

A_T = U x μ x H
= (58%/100%) x (150 calls x 3 calls/hour) x (4 minutes/call)
= 1044 minutes/hour

A_T = (1044 minutes/hour) x (1 hour/60 minutes) = 17.4 Erlangs

Blocking probability is

P_B = 0.03 by looking at 17.4 Erlangs and 24 channels from Erlang B chart

Carried traffic is

A_{ca} = A_T x (1 – P_B) = 17.4 Erlangs x (1 – 0.03) = 16.9 Erlangs

Channel usage

η = A_{ca} / C = 16.9 / 24 = 0.7 or 70%

How many call are blocks?

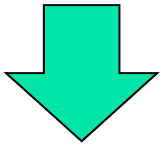
Answer: P_B = 0.03 means that 1 call in 33.3 calls will be blocked and results 0.5 Erlangs

Note that 16.9 Erlangs of traffic attempts to go across the T1 trunk and 0.5 Erlangs (17.4 – 16.9) is blocked

* See chapter 9 at slide number 31

Traffic Capacity

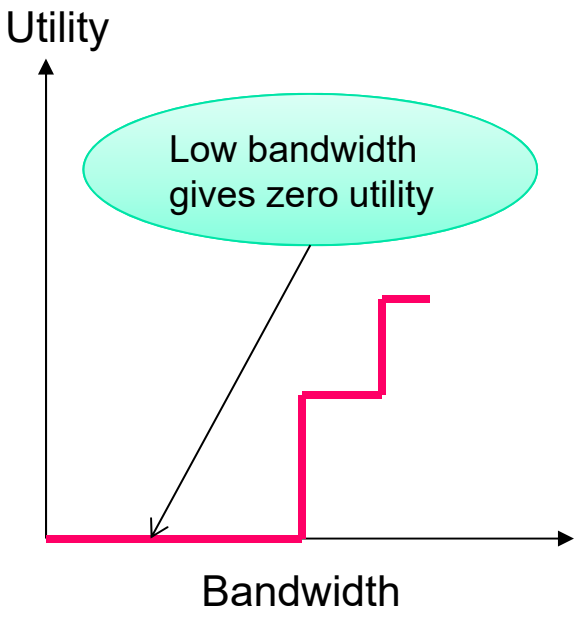
- Traffic capacity measurements:
  - Traffic congestion and blocking
  - Probability of waiting before a call is connected
  - Dominant coverage area
  - C/I ratio
  - Dropped call rate
  - Handover failure rate
  - Overall call success rate



All these measurement can be explained by Quality of Service (QoS)

What is QoS?

- Many applications are sensitive to the effects of delay (+ jitter*) and packet loss
  - May have bandwidth is not enough, in which leads to utility drops to zero
- Existing Internet architecture provides a best effort service
  - All traffic is treated equally (generally, FIFO queuing)
  - No mechanism for distinguishing between delay sensitive traffic and best effort traffic
  - Original IPv4 has Type of Service (ToS) field in packet header. This ToS specifies a datagram's priority and request a route for low-delay, high-throughput, or highly-reliable service. However, ToS never use in practice



* Jitter is defined as a variation in the delay of received packets

Factors Affecting QoS

- Standard metrics of QoS to the user that can be measured to rate the QoS are
  - Coverage: the strength of the measured signal is used to estimate the size of the cell
  - Accessibility that includes GoS: is about determining the ability of the network to handle successful calls from mobile-to-fixed networks and from mobile-to-mobile networks
    - Connection duration of call is in tens of seconds or minutes
    - Packet transmission or serving measured in milliseconds or even microseconds
    - User movement measured in seconds, minutes or hours
  - Audio quality: monitoring a successful call for a period of time for the clarity of the communication channel

Grade of Service (GoS)

- GoS is a mechanism for controlling the performance, reliability and usability of a telecommunications service
- GoS is a measure of the call blocking in voice traffic, where resources allocation is deterministic or is a measure of the ability to make call during the busiest time
- GoS 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
- GoS is determined by the available no. of channels and used to estimate the total no. of users that a network can support
- Example: GoS = 0.05 means that 1 call in 20 calls will be blocked during the busiest hour

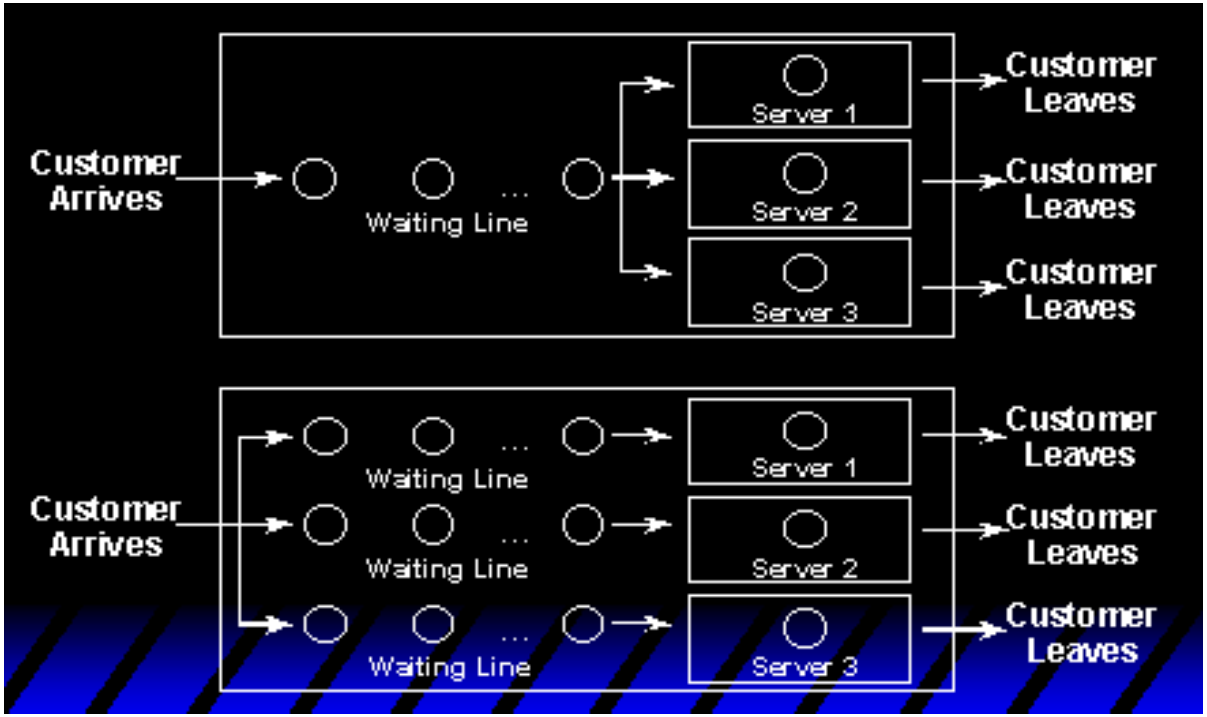
GoS in Cellular Networks

- In general, GoS is measured by
  - Looking at traffic carried
  - Traffic offered
  - Calculating the traffic blocked and lost
  - The proportion of lost calls is the measure of GoS
    GoS = no. of lost calls / no. of offered calls
- In cellular, GoS acceptable = 0.02 at busy period means that 2 users out of 100 users will encounter a call refusal
- GoS is calculated using the Erlang-B formula, as a function of the no. of channels required for the offered traffic intensity
- Trade-off between GoS and channel utilization

Queuing Theory Definitions

- (Bose) The basic phenomenon of queuing arises whenever a shared facility needs to be accessed for service by a large number of jobs or customers
- (Wolff) The primary tool for studying these problems [of congestions] is known as queuing theory
- (Kleinrock) We study the phenomena of standing, waiting, and serving, and we call this study queuing theory. Any system in which arrivals place demands upon a finite capacity resource may be termed a queuing system
- (Mathworld) The study of the waiting times, lengths, and other properties of queues

Applications of Queuing Theory



- Applications
  - Telecommunications, traffic control, determining the sequence of computer operations, predicting computer performance, health services (e.g., control of hospital bed assignments), airport traffic, airline ticket sales, layout of manufacturing systems

Queuing Theory for Networks

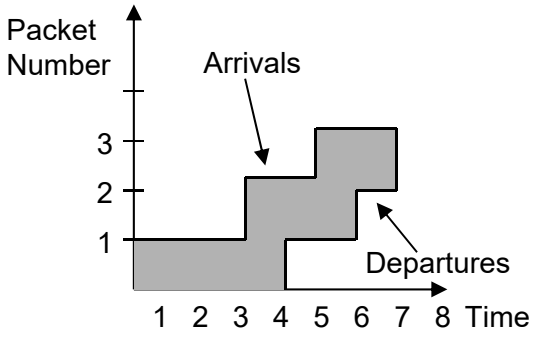
- View network as collections of queues
  - FIFO data-structures
- Queuing theory provides probabilistic analysis of these queues
- Examples
  - Average length
  - Average waiting time
  - Probability queue is at a certain length
  - Probability a packet will be lost

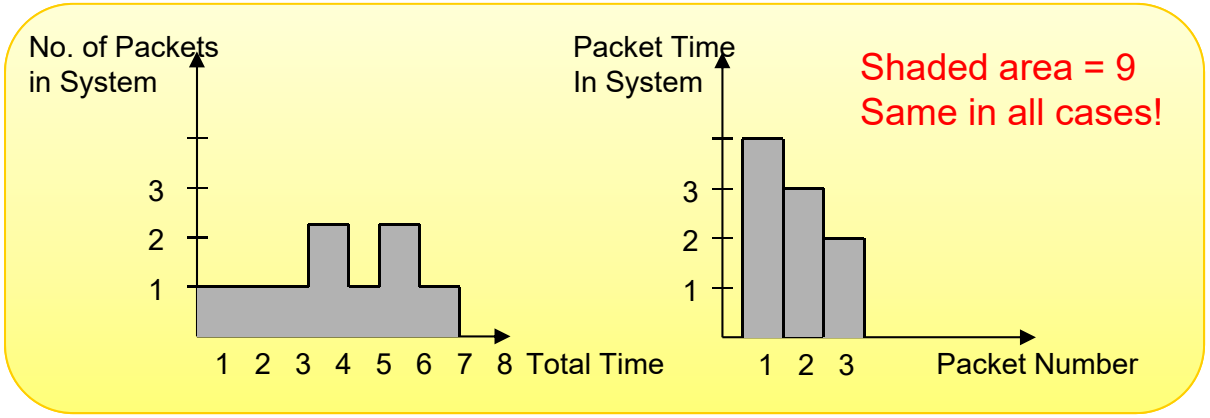
Little's Law

- Little's Law in a system
  - mean number of tasks = mean arrival rate X mean response time

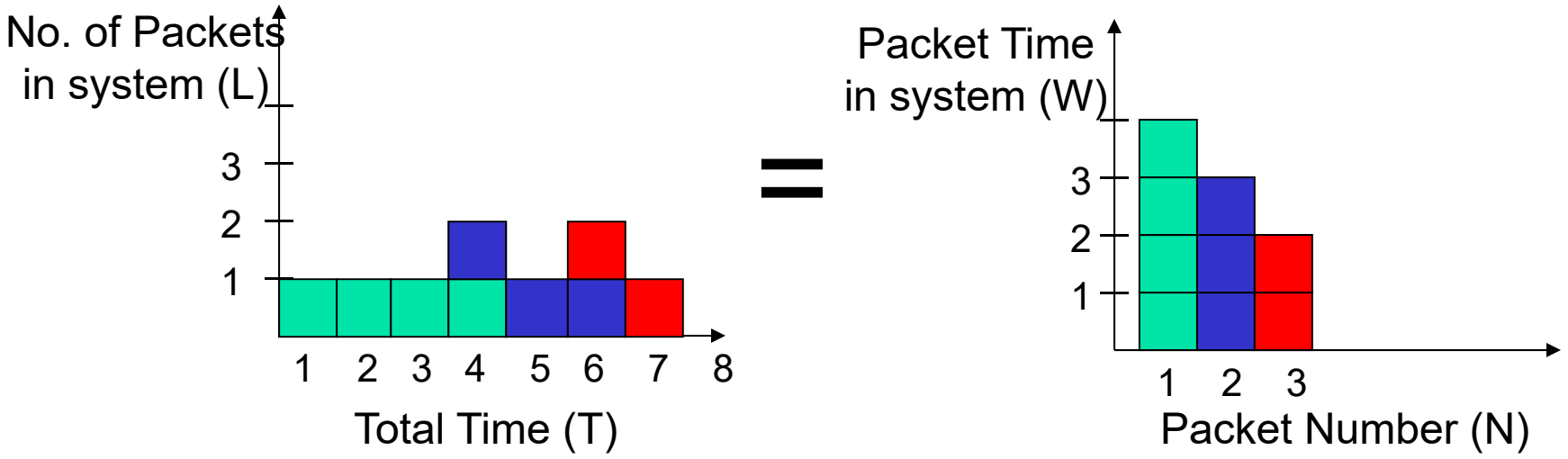
$$L = \lambda W$$

- Applies to any system in equilibrium, as long as nothing in black box is creating or destroying tasks





Proving Little's Law



Definitions:

- J: "Area" from previous slide
- N: Number of jobs (packets)
- T: Total time
- λ: Average arrival rate = N/T
- W: Average job time in the system = J/N
- L: Average number of jobs in the system = J/T

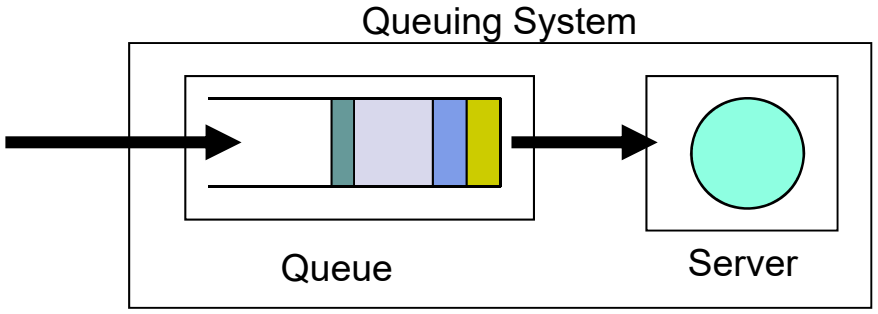
Proof:

$$J = TL = NW$$
$$L = (\frac{N}{T})W$$
$$L = (\lambda)W$$
$$L = (\frac{N}{T})W$$
$$\frac{J}{T} = (\frac{N}{T})(\frac{J}{N})$$
$$\frac{J}{T} = \frac{J}{T}$$

Tautology

* Tautology a logical argument by repeating the same concept using different terminology to result a valid evidence

Model Queuing System



- Use queuing model to
  - Describe the behavior of queuing systems
  - Evaluate system performance
- Characteristics of queuing system
  - Arrival Process: the distribution that determines how the tasks arrives in the system
  - Service Process: the distribution that determines the task processing time
  - Number of Servers: total number of servers available to process the tasks

Kendall Notation

A / B / S / K / N / D

- 6 parameters in shorthand
  - First three typically used, unless specified
- Legend:
  - A: Arrival Distribution
  - B: Service Distribution
  - S: Number of Servers
  - K: Total Capacity (infinite if not specified)
  - N: Population Size (infinite if not specified)
  - D: Service Discipline
- Distributions
  - M: "Markovian", implying exponential distribution for service times or inter-arrival times
  - D: Deterministic (e.g. fixed constant)
  - Ek: Erlang with parameter k
  - Hk: Hyper exponential with parameter k
  - G: General (anything)

Symbol for A

M	Markovian	Poisson process (or random) arrival process.
M^X	batch Markov	Poisson process with a random variable *X* for the number of arrivals at one time.
MAP	Markovian arrival process	Generalization of the Poisson process.
BMAP	Batch Markovian arrival process	Generalization of the *MAP* with multiple arrivals
MMPP	Markov modulated Poisson process	Poisson process where arrivals are in "clusters".
D	Degenerate distribution	A deterministic or fixed inter-arrival time.
Ek	Erlang distribution	An Erlang distribution with *k* as the shape parameter.
G	General distribution	Although *G* usually refers to independent arrivals, some authors prefer to use *GI* to be explicit.
PH	Phase-type distribution	Some of the above distributions are special cases of the phase-type, often used in place of a general distribution.

Symbol for D

FIFO/FCFS	First In First Out/First Come First Served	The customers are served in the order they arrived in.
LIFO/LCFS	Last in First Out/Last Come First Served	The customers are served in the reverse order to the order they arrived in.
SIRO	Service In Random Order	The customers are served in a random order with no regard to arrival order.
PNPN	Priority service	Priority service, including preemptive and non-preemptive. (see Priority queue)
PS	Processor Sharing	

Symbol for B

M	Markovian	Exponential service time.
D	Degenerate distribution	A deterministic or fixed service time.
Ek	Erlang distribution	An Erlang distribution with *k* as the shape parameter.
G	General distribution	Although *G* usually refers to independent service time, some authors prefer to use *GI* to be explicit.
PH	Phase-type distribution	Some of the above distributions are special cases of the phase-type, often used in place of a general distribution.

Examples

- M/M/1 - is the simplest 'realistic' queue
  - Poisson arrivals and exponential service
  - 1 server
  - infinite capacity
  - Population
  - FCFS (FIFO)
- M/M/m
  - Same, but m servers
- G/G/3/20/1500/SPF
  - General arrival and service distributions
  - 3 servers
  - 17 queue slots (20 – 3)
  - 1500 total jobs
  - Shortest Packet First

Analysis of M/M/1 Queue

- Given:
  - λ: Arrival rate of jobs (packets on input link)
  - μ: Service rate of a server (packets on output link)
- Solve:
  - L_s: Average number in the queuing system
  - L_q: Average number in the queue
  - W_s: Average waiting time in the whole system
  - W_q: Average waiting time in the queue

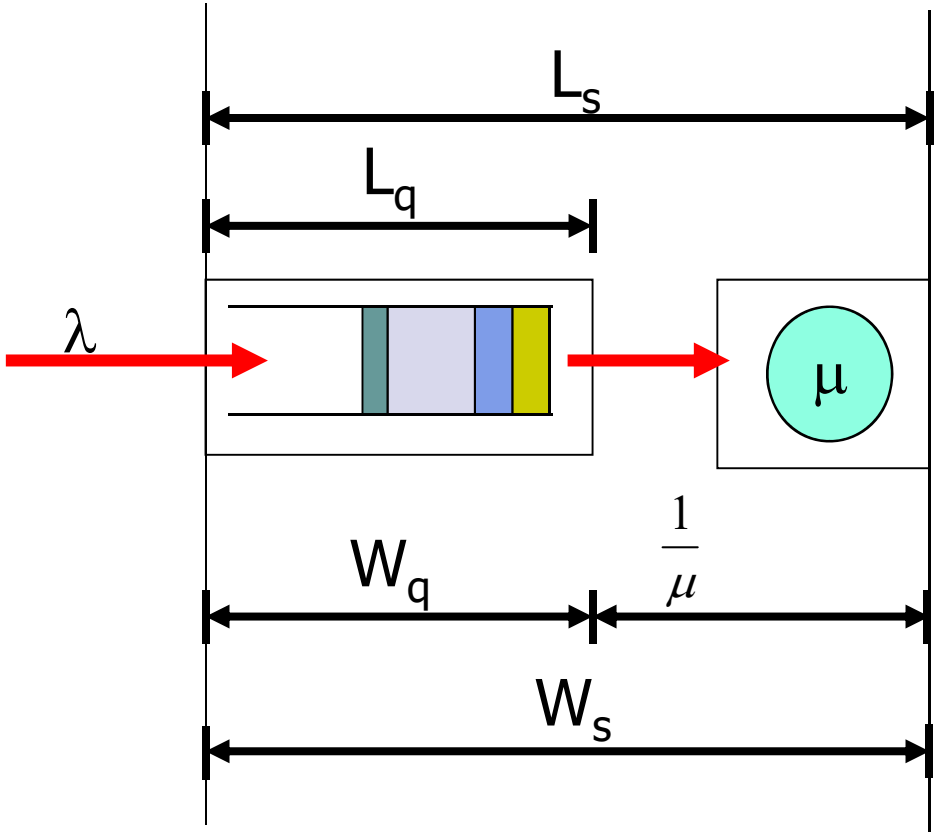


Figure: M/M/1 queue model

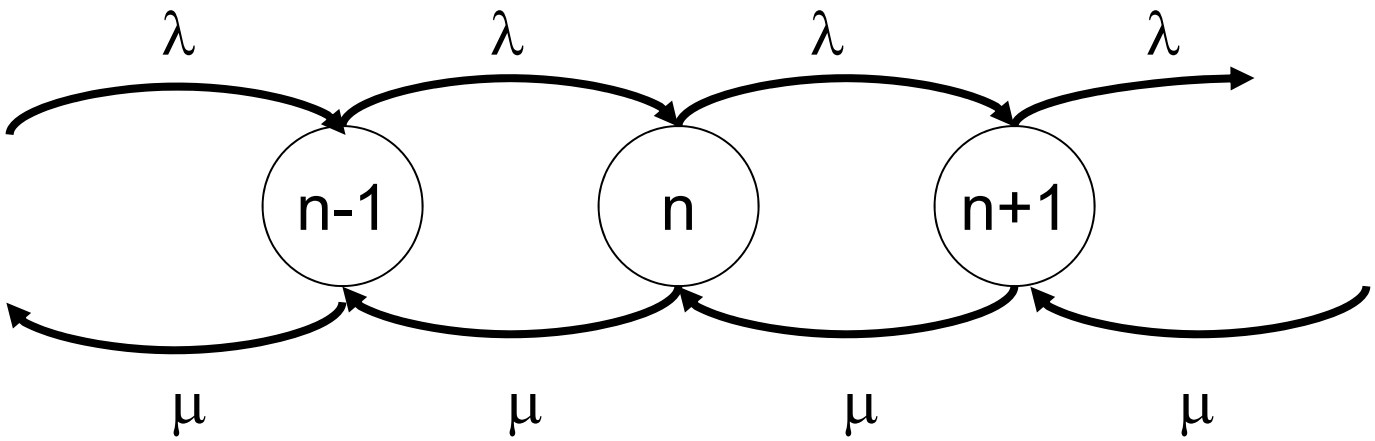
Solving M/M/1 Queue

- 4 unknowns: L_s, L_q, W_s, W_q
- Relationships:
  - $L_s = \lambda W_s$
  - $L_q = \lambda W_q$ (steady-state argument)
  - $W_s = W_q + (1/\mu)$
- If we know any one of them, we can find the others
- Finding L_s is hard or easy depending on the type of system. In general,

$$L_s = \sum_{n=0}^{\infty} nP_n$$

- Goal: a closed form expression of the probability of the number of jobs in the queue (P_i) given only λ and μ

Equilibrium Conditions



Define $P_n(t)$ to be the probability of having n tasks in the system at time t

$$P_0(t+\Delta t) = P_0(t)[(1-\mu\Delta t)(1-\lambda\Delta t) + \mu\Delta t\lambda\Delta t] + P_1(t)[(\mu\Delta t)(1-\lambda\Delta t)]$$

$$P_n(t+\Delta t) = P_n(t)[(1-\mu\Delta t)(1-\lambda\Delta t) + \mu\Delta t\lambda\Delta t] + P_{n+1}(t)[(\mu\Delta t)(1-\lambda\Delta t)] + P_{n-1}(t)[(\lambda\Delta t)(1-\mu\Delta t)]$$

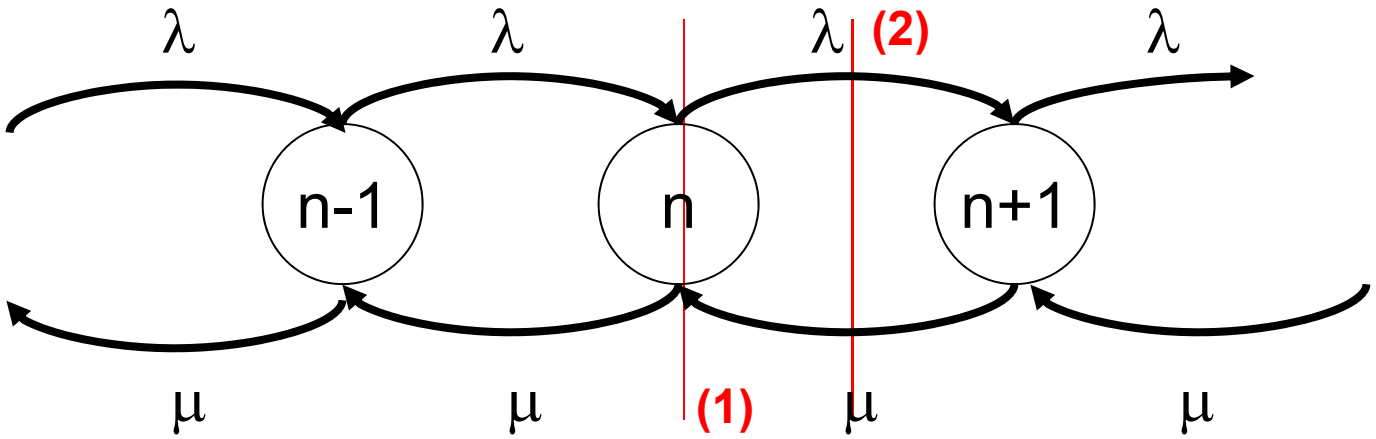
$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t) + \mu P_1(t)$$

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \lambda P_{n-1}(t) - (\lambda+\mu)P_n(t) + \mu P_{n+1}(t)$$

Stablize when $\lambda \le \mu$, $\lim_{t\to\infty} P_n(t) = P_n$, $\lim_{t\to\infty} \frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = 0$

Proofing the stability condition

Equilibrium Conditions (cont.)



In the equilibrium, in flow = out flow

$$\lambda P_n + \mu P_n = \lambda P_{n-1} + \mu P_{n+1} \quad \text{...(1)}$$

$$\lambda P_n = \mu P_{n+1} \quad \text{...(2)}$$

For stability

$$\lambda \le \mu,\ \rho = \frac{\lambda}{\mu},\ \rho \le 1$$

Solving for P_0 and P_n

Use eqn. (2) in slide number 37

Step 1

$$P_1 = \frac{\lambda}{\mu} P_0, \quad P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_0, \quad P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

First, get P_0
Then, get P_n

Step 2

$$\sum_{n=0}^{\infty} P_n = 1, \quad then \quad P_0 \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = 1, \quad \Rightarrow \quad P_0 = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n}$$

Step 3

$$\rho = \frac{\lambda}{\mu}, \; then \; \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = \sum_{n=0}^{\infty} \rho^n = \frac{1-\rho^{\infty}}{1-\rho} = \frac{1}{1-\rho}, \; \text{for } \rho < 1$$

Step 4

$$P_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1-\rho \quad \text{and} \quad P_n = \rho^n (1-\rho)$$

Solving for L_s

$$L_s = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1}$$

$$(1-\rho)\rho\frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right) = (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$$(1-\rho)\rho\left(\frac{1}{(1-\rho)^2}\right) = \frac{\rho}{(1-\rho)} = \boxed{\frac{\lambda}{\mu-\lambda}}$$

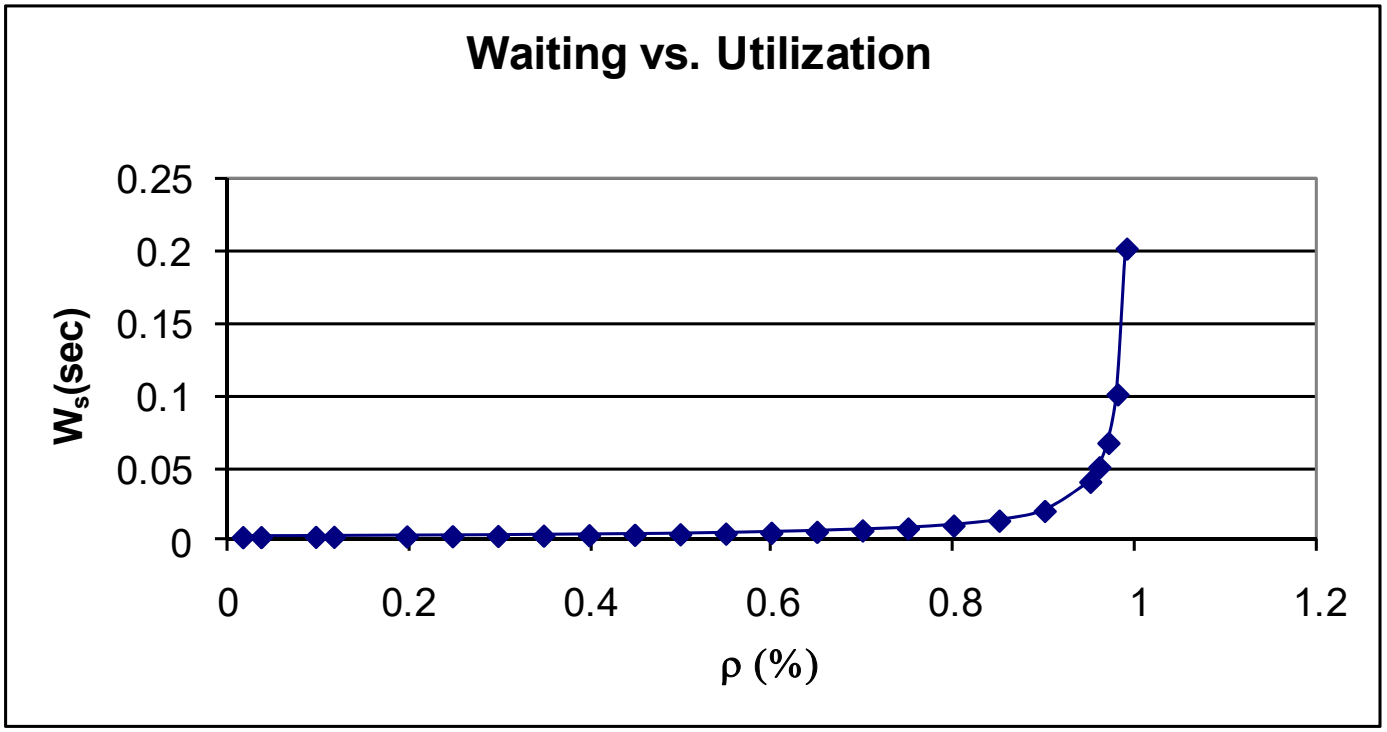
Solving for W_s, W_q, L_q

$$W_s = \frac{L}{\lambda} = \left(\frac{\lambda}{\mu-\lambda}\right)\left(\frac{1}{\lambda}\right) = \boxed{\frac{1}{\mu-\lambda}}$$

$$W_q = W_s - \frac{1}{\mu} = \left(\frac{\lambda}{\mu-\lambda}\right) - \left(\frac{1}{\mu}\right) = \boxed{\frac{\lambda}{\mu(\mu-\lambda)}}$$

$$L_q = \lambda W_q = \lambda \frac{\lambda}{\mu(\mu-\lambda)} = \boxed{\frac{\lambda^2}{\mu(\mu-\lambda)}}$$

Response Time vs. Arrivals



$$W_s = \frac{1}{\mu - \lambda}$$

Example 6

On a network gateway, measurements show that the packets arrive at a mean rate of 125 packets per second (pps) and the gateway takes about 2 ms to forward them. Assuming an M/M/1 model,

- What is the gateway's utilization?
- What is the probability of n packets in the gateway?
- What is the mean number of packets in the gateway?
- What is the probability of buffer overflow if the gateway had only 13 buffers?

Solution:

- Arrival rate λ = 125 pps and service rate μ = 1/0.002 = 500 pps
  Therefore, gateway utilization, ρ = λ/μ = 0.25
- Probability of n packets in gateway is $(1-\rho)\rho^n = 0.75(0.25)^n$
- Mean number of packets in gateway is $\frac{\rho}{1-\rho} = \frac{0.25}{0.75} = 0.33$
- Probability of buffer overflow = P(more than 13 packets in gateway)
  $= \rho^{13} = 0.25^{13} = 1.49 \times 10^{-8}$
  = 15 packets per billion packets

Example 7

Assume a drive-up window at a fast food restaurant, customers arrive at the rate of 20 per hours. The employee can serve one customer every two minutes. Assume Poisson arrival and exponential service rate, Determine the following questions.

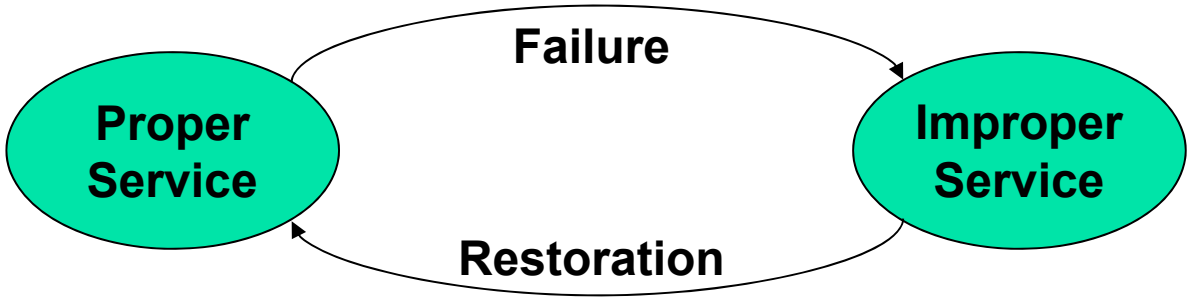
- What is the average utilization of the employee?
- What is the average number of customers in line?
- What is the average number of customers in the system?
- What is the average waiting time in line?
- What is the average waiting time in the system?

Solution:

- Arrival rate λ = 20 customers/hour
  service rate μ = 1 customer/2 minutes = 30 customers/hour
  Therefore, employee utilization, ρ = λ/μ = 0.6667
- Average number of customers in line, $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = 1.3333$
- Average number of customers in the system, $L_s = \frac{\lambda}{\mu-\lambda} = 2$
- Average waiting time in line, $W_q = \frac{L_q}{\lambda} = 0.066\ hours = 4\ \text{min}$
- Average waiting time in the system, $W_s = \frac{L_s}{\lambda} = 0.1\ hours = 6\ \text{min}$

Define and Quantify Dependability

- How to decide when system is operating properly?
- *Dependability* is the ability of a system to deliver a specified service
- Infrastructure providers now offer Service Level Agreements (SLA) to guarantee that their networking or power service will be dependable
- Systems alternate between 2 states of service with respect to an SLA
  - Service accomplishment, where service is delivered as specified in SLA
  - Service interruption, where delivered service is different from the SLA
- Failure = a transition from proper to improper service
- Restoration = a transition from improper to proper service

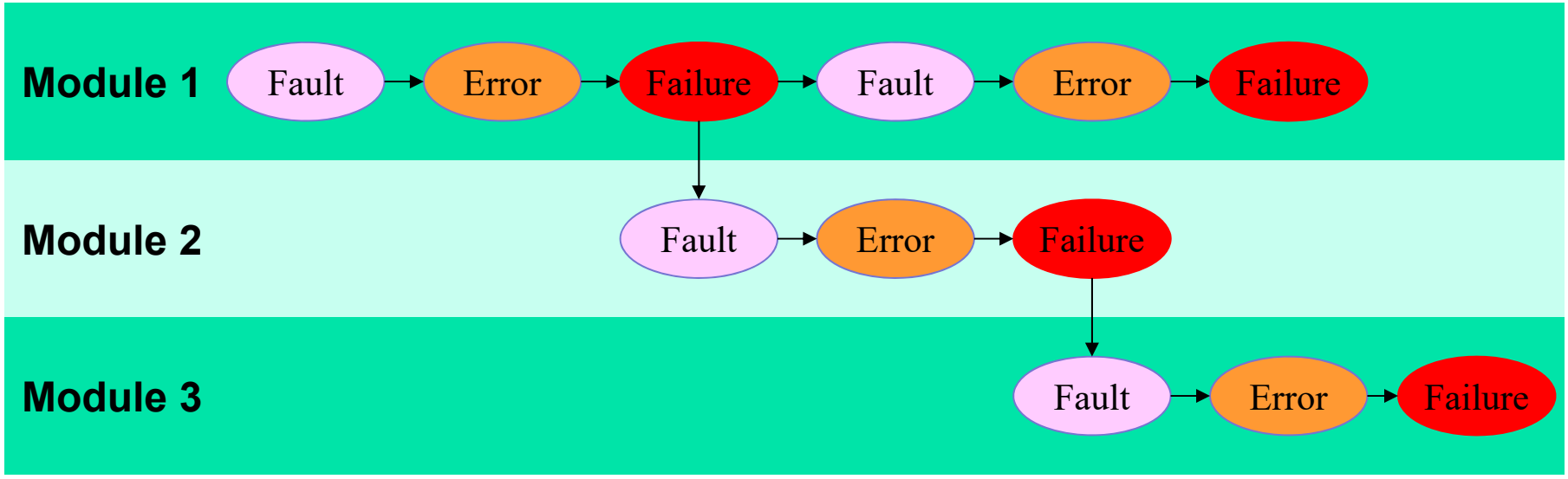


Dependability Concepts

- Measures – properties expected from a dependable system
  - Availability
  - Reliability
  - Safety
  - Confidentiality
  - Integrity
  - Maintainability
  - Coverage
- Means – methods to achieve dependability
  - Fault avoidance
  - Fault tolerance
  - Fault removal
  - Dependability assessment
- Impairments – causes of undependable operation
  - Faults
  - Errors
  - Failures

Faults, Errors, Failures Can Cause Improper Service

- Failure – a transition from proper to improper service
- Error – that part of system state that is liable to lead to subsequent failure
- Fault – the hypothesized cause of error(s)



Reliability and Availability

- Reliability measures of the continuous delivery of service (or time to failure)
  - Time to failure measures of the time to failure from last restoration. It is also referred to as Mean time to failure (MTTF)
  - Failures in time (FIT) = 1/MTTF, the rate of failures (conventionally it reported as failures per billion hours of operation)
- Maintainability measures of the time to restoration from last experienced failure
  - Mean time to repair (MTTR) measures service interruption
- Availability measures service as alternation between deliveries of proper and improper services (number between 0 and 1, e.g. 0.9)
  - Availability = MTTF / (MTTF + MTTR)

Example 8

If modules have exponentially distributed lifetimes (age of module does not affect probability of failure), overall failure rate is sum of failure rates of individual modules. Calculate FIT and MTTF for 10 disks (1M hour MTTF per disk), 1 disk controller (0.5M hour MTTF), and 1 power supply (0.2M hour MTTF).

Solution:

$$FIT = 10 \times (1/1{,}000{,}000) + 1/500{,}000 + 1/200{,}000$$
$$= (10 + 2 + 5)/1{,}000{,}000$$
$$= 17/1{,}000{,}000$$
$$= \underline{17{,}000\,FIT}$$
$$MTTF = 1{,}000{,}000{,}000/17{,}000$$
$$\approx \underline{59{,}000\,hours}$$

Announcement

- Next is Chapter 11 Design of Network Equipments and Protocols
- 10:50 ~ 12:30 on 16 November (Wednesday)