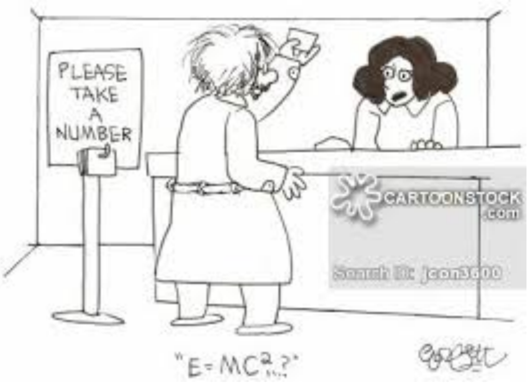
Part I Syllabus

Lecture	Date	Subject
1	10/08/2016	Introduction
2	10/08/2016	Layered network architecture & Physical resilience
3	17/08/2016	Data link layer – flow control
4	17/08/2016	Data link layer – error control
5	24/08/2016	Data link layer – HDLC
6	24/08/2016	Local area network – introduction
7	31/08/2016	Local area network – MAC
8	31/08/2016	Local area network – Ethernet
9	07/09/2016	Local area network – WLAN
10	07/09/2016	Packet switch network - Introduction
11	**14/09/2016**	**Packet switch network – queue analysis**
12	14/09/2016	Review and examples


PLEASE TAKE A NUMBER
CARTOONSTOCK .com
Search ID: jcon3600
"E=MC2...?"

CE3005/CPE302 Computer Networks

Lecture 10
Introduction to Queuing Theory

Contents

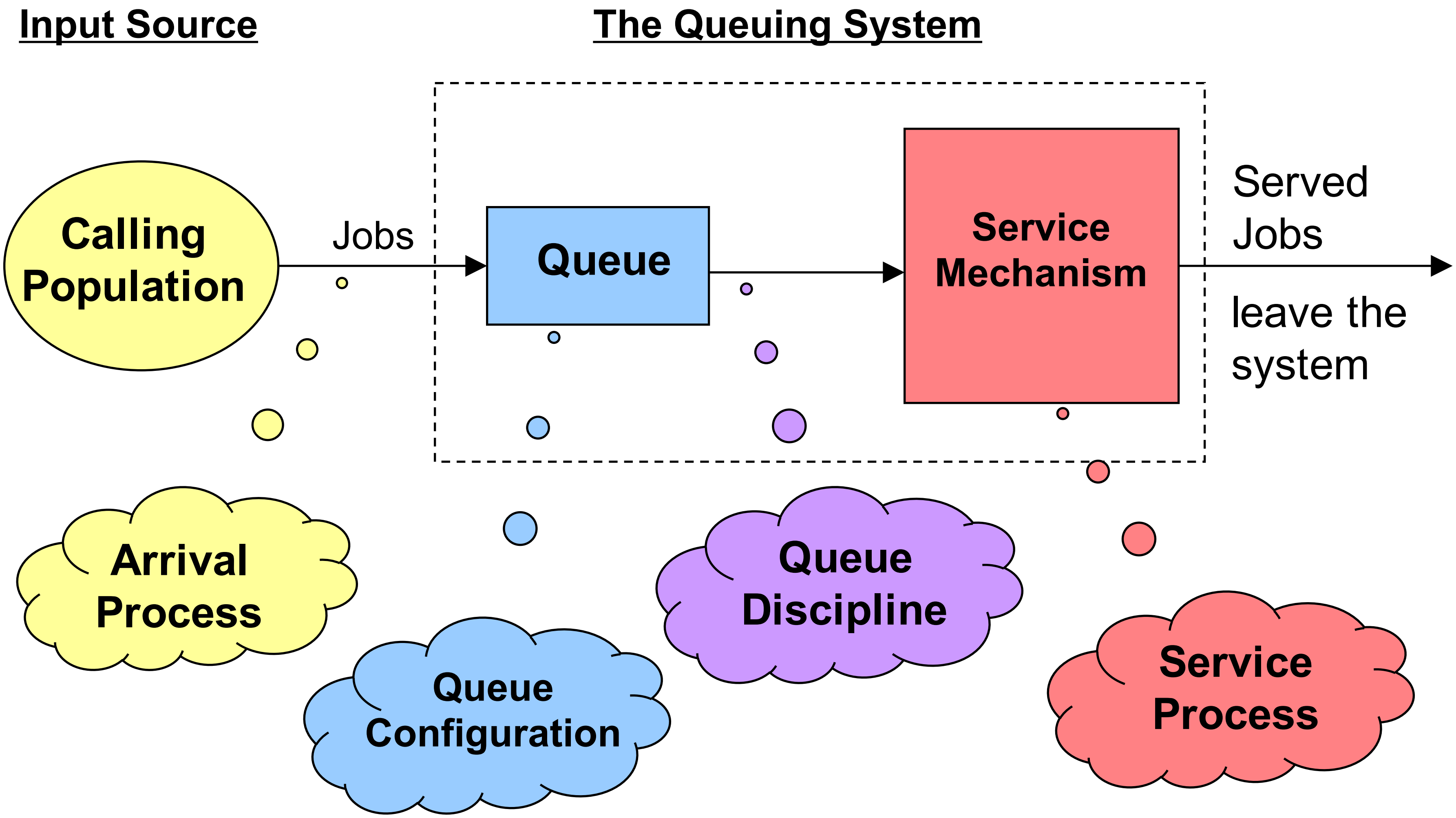
- **Introduction to Queuing System**
  - Queuing system definition and examples
  - Components of a queuing system
- **Probability Laws for Queuing Systems**
  - Exponential random variable
  - Poisson process
  - Arrival Process and Service Time
  - Kendall notation for queuing system
- **M/M/1 Queue Analysis**
  - Assumptions
  - Delay analysis and queue occupancy
- **Applications to Packet Switched Networks**
  - Computer network modeling
  - Delay analysis via M/M/1 queue
  - M/M/1 queue properties: Poisson Process and Feedback Queue

Introduction to Queuing System

Queuing System Definition

- **Queuing Systems**
  - Model processes in which customers arrive
  - Wait their turn for service
  - Serviced and then leave
- **Examples**
  - Supermarket checkouts stands
  - Bank reception counter
  - JP movie theater ticket booths
  - Doctors waiting rooms

Components of Queuing System



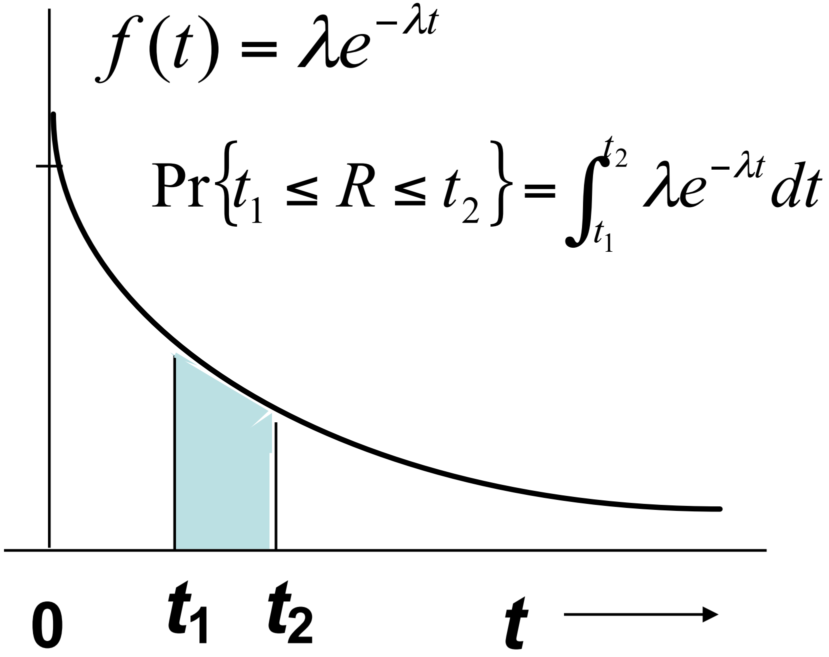
Characteristics of Queuing System

- A queuing system is characterized by
  - The inter-arrival time probability distribution function
  - The service-time probability distribution function
  - The number of servers
  - The queuing discipline
  - The amount of buffer space in the queues

Probability Laws for Queuing System

Exponential Random Variable (RV)

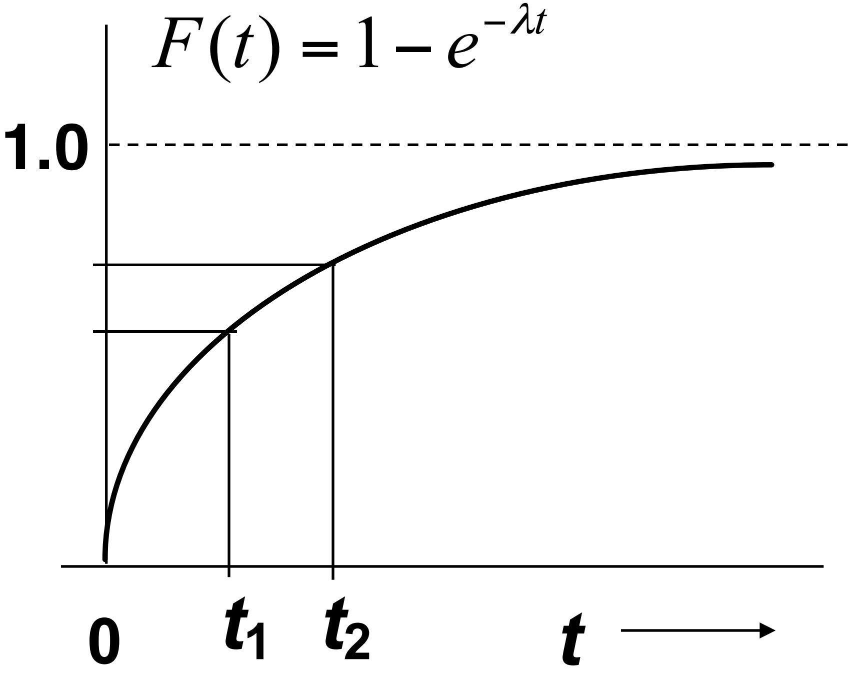
Density Function



$$f(t) = \lambda e^{-\lambda t}$$

$$\Pr\{t_1 \le R \le t_2\} = \int_{t_1}^{t_2} \lambda e^{-\lambda t}\,dt$$

Distribution Function



$$F(t) = 1 - e^{-\lambda t}$$

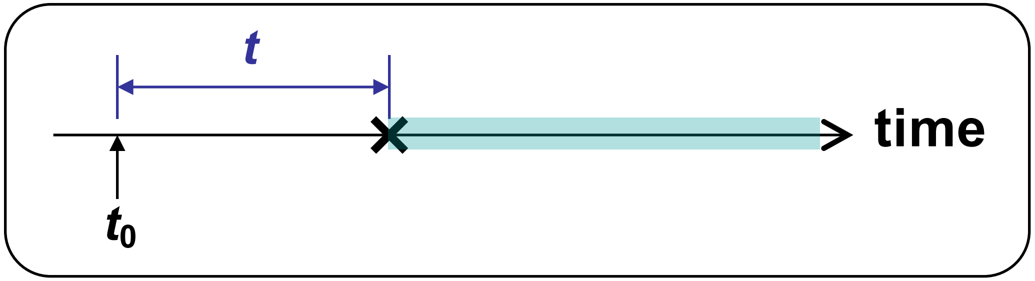
Example: $\lambda = 10$ packets per second (Average Rate)
Unit of *t* and *R* is second.

Memoryless Property for Exp. R.V.

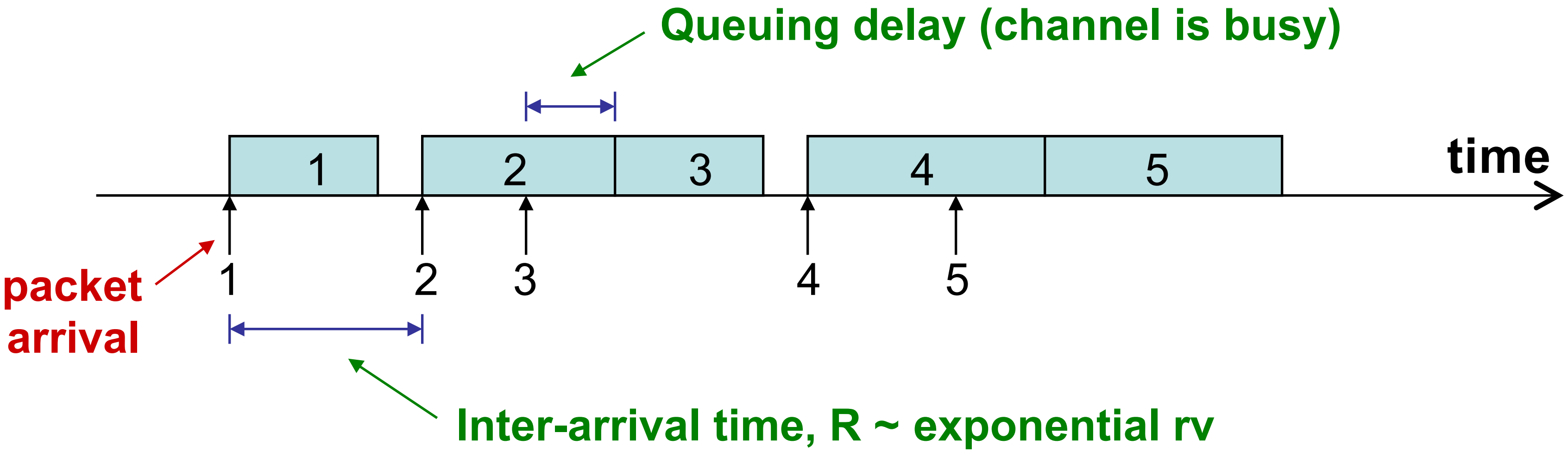
- **Exponential distribution is memoryless**
  - *R* be the waiting time for an event to occur
  - Given *R* is larger than t_0, it does not change the distribution of the remaining waiting time

$$\Pr\left\{R > t + t_0 \middle| R > t_0\right\} = \frac{\Pr\left\{R > t + t_0, R > t_0\right\}}{\Pr\left\{R > t_0\right\}}$$

$$= \frac{\Pr\left\{R > t + t_0\right\}}{\Pr\left\{R > t_0\right\}} = \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t_0}} = e^{-\lambda t} = \Pr\left\{R > t\right\}$$



Inter-Arrival Time


Queuing delay (channel is busy)
1
2
3
4
5
time
packet arrival
1
2
3
4
5
Inter-arrival time, R ~ exponential rv

$$\Pr\left\{R \le t_1\right\} = A(t_1) = 1 - e^{-\lambda t_1}$$

λ = Average packet arrival rate (packets/second)

E[*R*] = 1/λ = Average inter-arrival time (sec/packet)

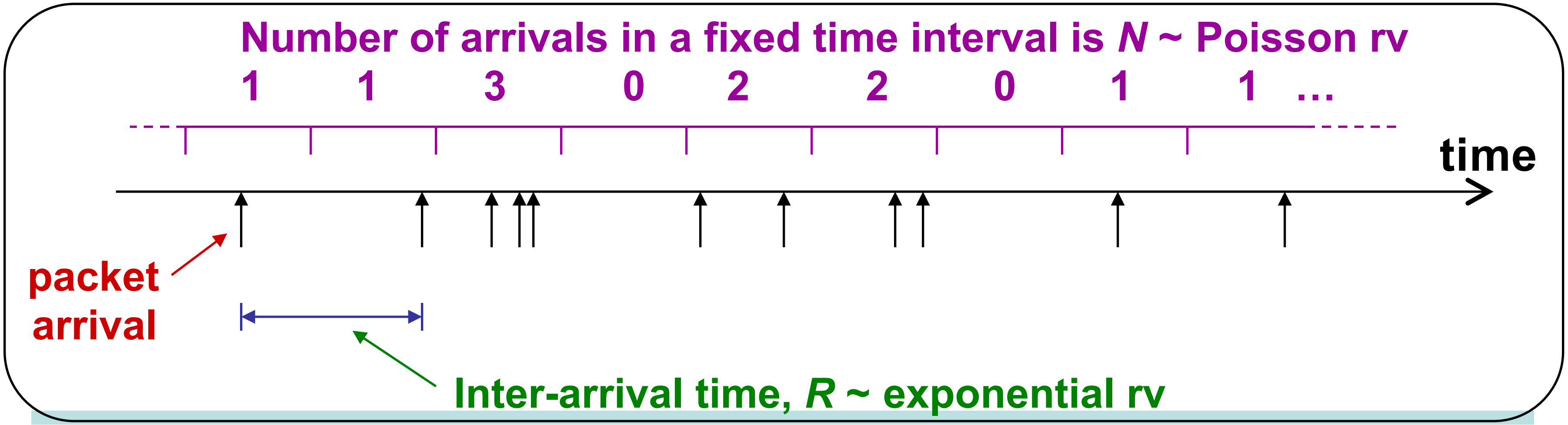
(Poisson) Arrival Process

- It describes the behavior of packet generation.
- If the inter-arrival time follows exponential r.v., then the arrival process is a Poisson Process.
- The # of arrivals in a fixed time interval follows Poisson Distribution for Poisson process

$$\Pr\{N = n\} = P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

λ = Average Packet Arrival Rate (packets/sec)

t = The fixed time interval


Number of arrivals in a fixed time interval is *N* ~ Poisson rv
1 1 3 0 2 2 0 1 1 …
time
packet arrival
Inter-arrival time, *R* ~ exponential rv

Poisson Arrival & Inter-Arrival Time

The following shows the relationship between Poisson Arrival & Exponential Inter-arrival Time

$$\Pr\{T > t\} = \Pr\{\text{no arrival in time } t\}$$

$$= P_0(t) = e^{-\lambda t}$$

According to Poisson Process

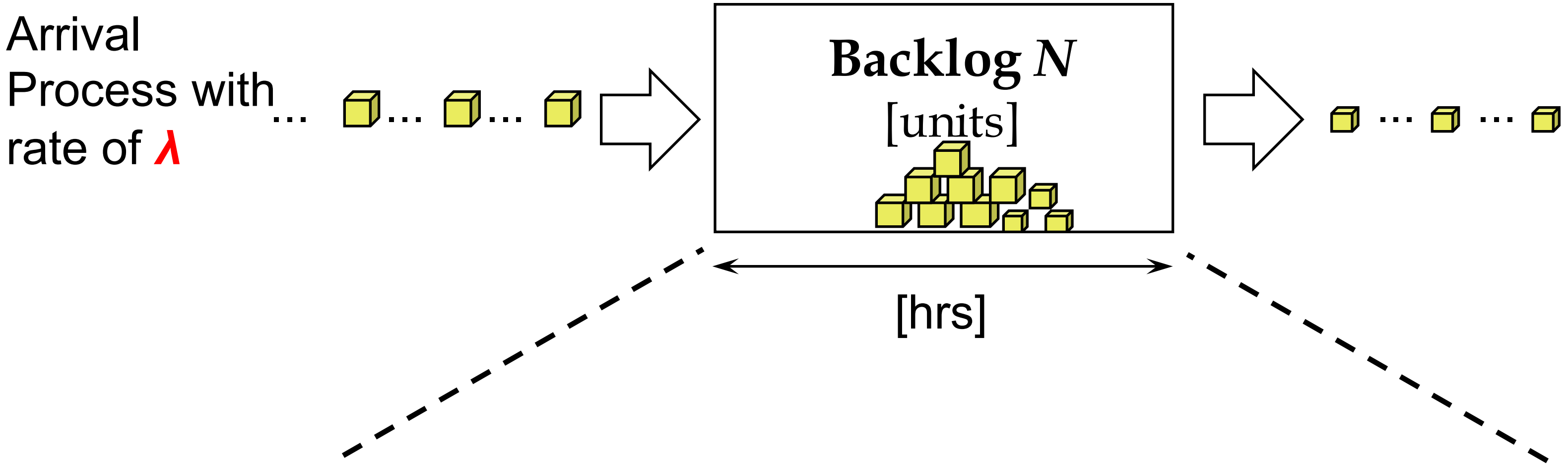
that is

$$\underline{\Pr\{T \le t\} = 1 - e^{-\lambda t}}$$

The time the next arrival appears follows Exponential distribution

Little's Law

- **Little’s Law**
  - Proved by John Little @ MIT in 1961
  - Independent of arrival process, service time, service order, or practically anything else



Arrival Process with rate of λ

Backlog = Arrival Rate x Response Time

$$N = \lambda \times T$$

Example: Arrival Process

Assume that students at NTU go to McDonald in Canteen A with a Possion arrival. On average, 1 student will arrive per minute.

- **What is the arrival rate?**

$$\lambda = 1 \text{ student/min}$$

- **What is the average inter-arrival time?**

$$1/\lambda = 1 \text{ min}$$

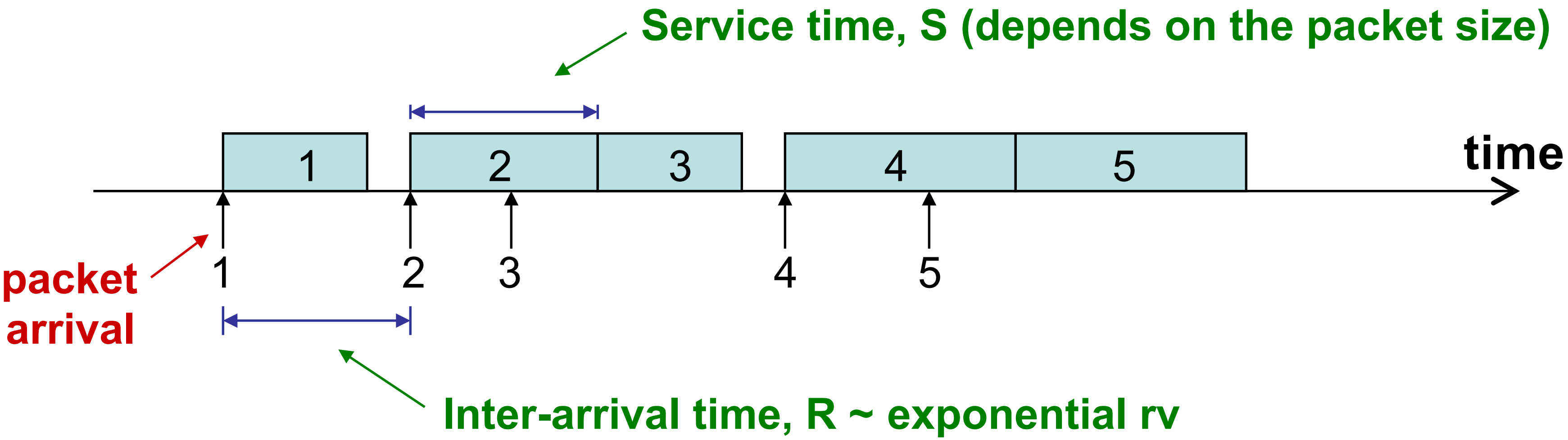
- **Between 12pm-1pm, how many students are expected to show up?**

$$N = \lambda T = 60 \text{ students}$$

- **Between 12-1pm, what is the probability that no student show up?**

$$\Pr\{N = n\}\big|_{n=0} = \frac{(\lambda t)^n}{n!} e^{-\lambda t} = e^{-60}$$

Service Time


Service time, S (depends on the packet size)
1
2
3
4
5
time
packet arrival
1
2
3
4
5
Inter-arrival time, R ~ exponential rv

For exponential service time: $\Pr\left\{S \leq t_1\right\} = 1 - e^{-\mu t_1}$

μ = **Average service rate (packets/second)**

E[*S*] = 1/μ = Average service time (sec/packet)

Example: Service Time

Assume that at McDonald in Canteen A, the time to serve each student is an exponential random variable. One registrar can serve 2 students within one minute.

- What is the service rate?

$$\mu = 2 \text{ students/min}$$

- What is the average service time?

$$1/\mu = 0.5 \text{ min}$$

- Between 12pm-1pm, how many students can be served potentially?

$$M = \mu T = 120 \text{ students}$$

- Given that the student arrives in 1 person per minute, what is the average occupancy of a registrar?

$$\rho = \frac{\lambda}{\mu} = 0.5$$

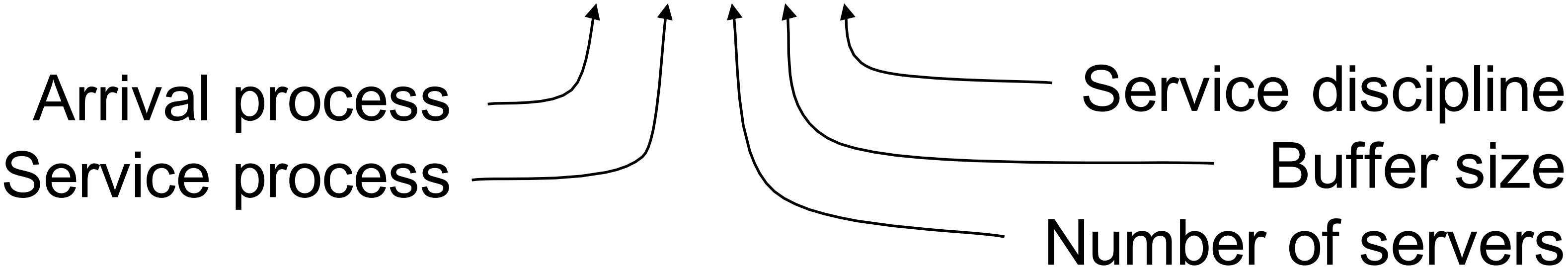
NANYANG TECHNOLOGICAL UNIVERSITY

Parameters for Servers

- **Number of servers:**
  - Single Server: eg a router with one output link **← our focus**
  - Multiple Servers: eg a bank with multiple tellers
- **Service Discipline (order in which packets are served):**
  - First Come First Serve **← our focus**
  - Higher priority first
  - others…
- **Buffer size**
  - Finite
  - Infinite **← our focus**

Queuing System Notation

- **Kendall notation: A/B/x/y/z**



Arrival process

Service process

Service discipline

Buffer size

Number of servers

Possible arrival and service processes:

M, D, E_k, H_k, PH, G, ...

Possible service discipline:

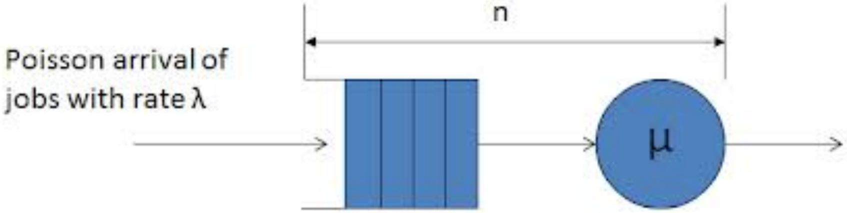
FCFS, LCFS, PR, GD, ...

We will only study **M/M/1**, ie. Markovian arrival/service, single server with unlimited buffer.

M/M/1 Queue Analysis

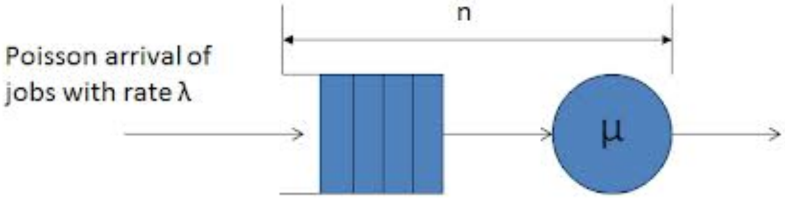
M/M/1 Queue Assumptions

- **M/M/1 Queue Assumptions**
  - ✓ Infinite Calling Populations
    - Independence between arrivals
  - ✓ Poisson arrival with an expected arrival rate λ
  - ✓ Single queue with possibly infinite length
  - ✓ Queue discipline is FIFO
  - ✓ Single server with exponentially distributed service times
    - μ = expected service rate when the server is busy



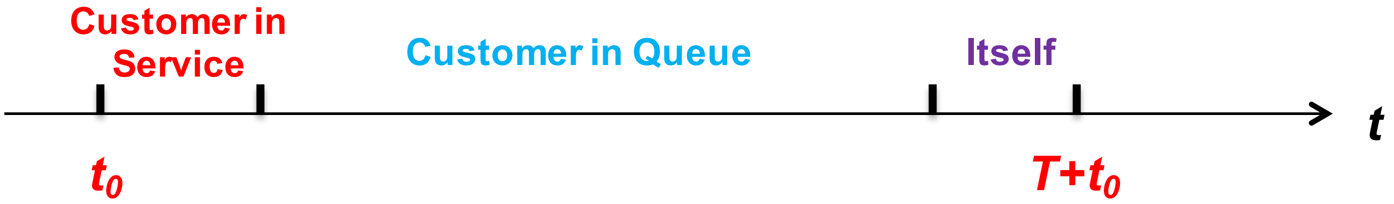
M/M/1 Queue Results

- **M/M/1 Parameters**
  - λ be the mean arrival rate (packets/sec)
  - μ be the mean service rate (packets/sec)
  - ρ be the system utilization, prob. that server is busy
  - N be the mean # of packets in the buffer
  - T be the delay
  - T_q be the queuing delay only



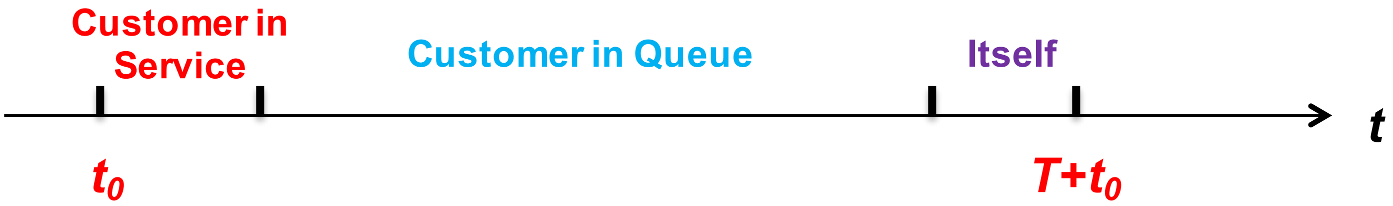
$\rho = \dfrac{\lambda}{\mu} < 1$	$T = \dfrac{1}{\mu - \lambda}$	$T_q = \dfrac{\rho}{\mu - \lambda}$	$N = \dfrac{\rho}{(1-\rho)}$

"Tagged Customer Method" (I): Three Sources of Delays



- **Residual service time of the customer in service, if the queue is busy at the arrival instant**
  - Probability of server being busy: ρ
  - Residual service time: $s_R = 1/\mu$ (memoryless property)
- **Service time for previous customers that are waiting in the queue but not being serviced at the moment**
  - Number of waiting customers: $N-\rho$
  - Average service time per customer: $s_A = 1/\mu$
- **Service time for the tagged customer**
  - Average service time: $s_T = 1/\mu$

"Tagged Customer Method" (II): Total Delay Calculation


Customer in Service

Customer in Queue

Itself

t_0

$T+t_0$

t

- Total Delay: $T = \rho s_R + (N-\rho)s_A + s_T$
- Little's Law: $N = \lambda T$
- Server Occupancy: $\rho = \lambda/\mu$

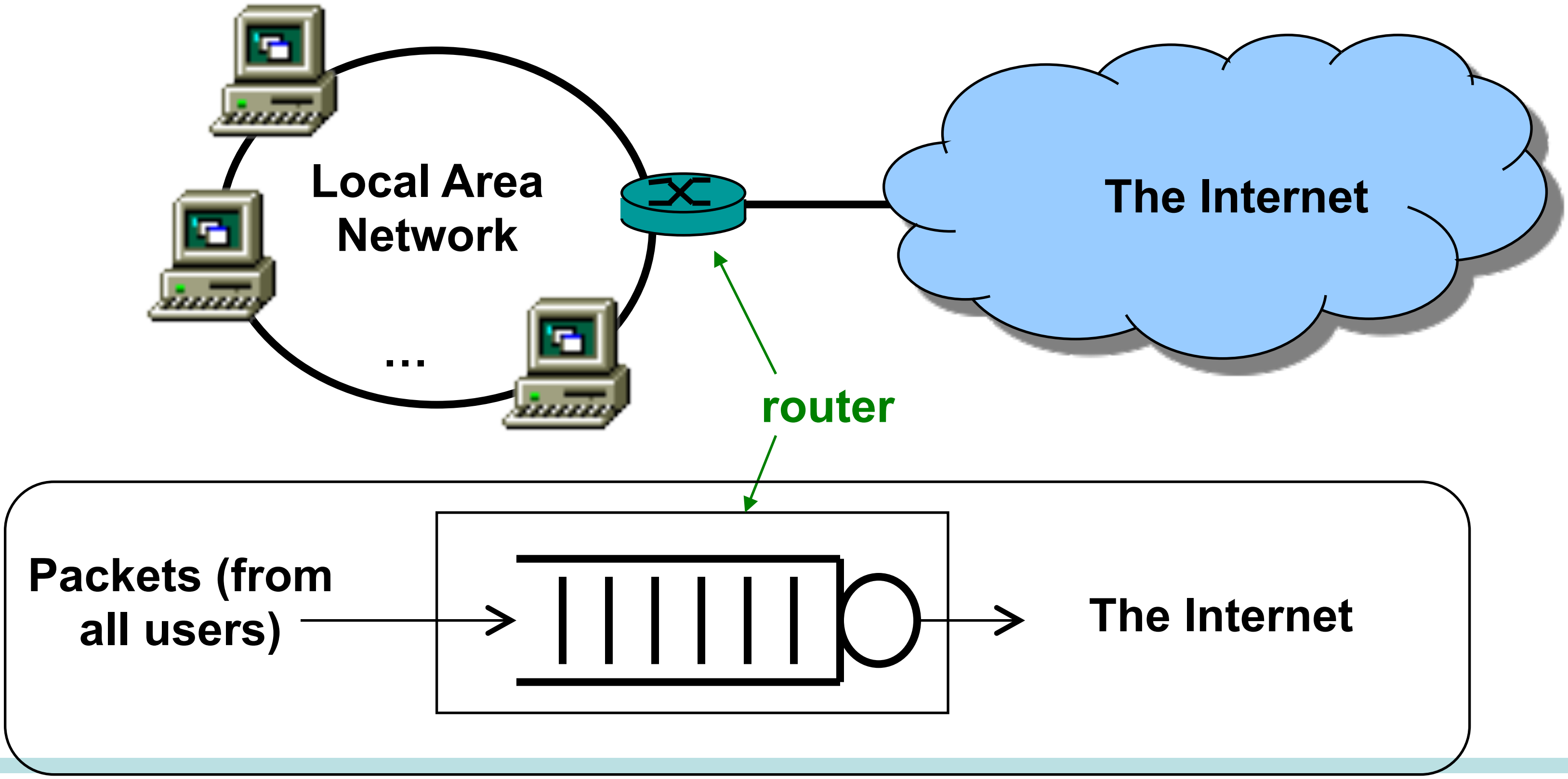
$$\boxed{T = \frac{\lambda/\mu}{\mu} + \frac{\lambda T - \lambda/\mu}{\mu} + \frac{1}{\mu}}$$

$$\boxed{T = \frac{1}{\mu - \lambda}} \quad \boxed{N = \lambda T = \frac{\lambda}{\mu - \lambda}} \quad \boxed{T_q = T - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}}$$

Applications in Packet Switched Networks

Computer Network Modeling

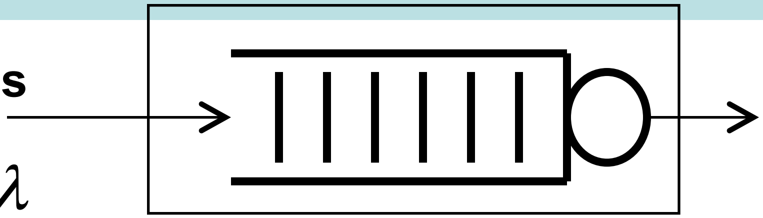
We model a network into a single server queue (eg. *M*/*M*/1) to study the performance



M/M/1 Summary

Arrival Process

arrival rate : λ



Departure Process

departure rate : λ

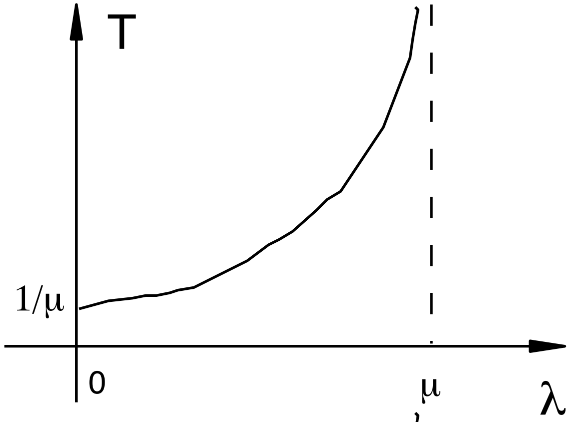
service rate : μ

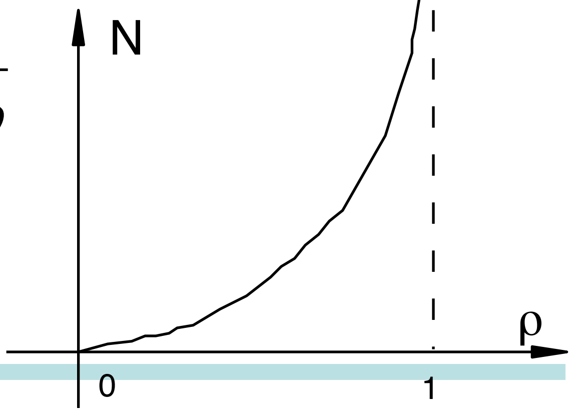
system utilization : $\rho = \dfrac{\lambda}{\mu}$

delay : $T = \dfrac{1}{\mu - \lambda}$

# of packets in queue : $N = \lambda T = \dfrac{\lambda}{\mu - \lambda} = \dfrac{\rho}{1-\rho}$

waiting time : $T_q = T - \dfrac{1}{\mu} = \dfrac{1}{\mu - \lambda} - \dfrac{1}{\mu} = \dfrac{\rho}{\mu - \lambda}$





M/M/1: Application (I)

At a link, packets arrive at an average Poisson rate of 8 packets/sec and the service time is exponentially distributed with an average rate of 10 packets/sec. Find the average # of packets in the system, and the average delay for a packet in the system , and average waiting time in the queue.

Context: Given $\lambda = 8$, $\mu = 10$, $\rho = 8/10 = 0.8$. Find N, T.

$$N = \frac{\rho}{(1-\rho)} = \frac{0.8}{1-0.8} = \frac{0.8}{0.2} = 4 \text{ packets}$$

$$T = \frac{1}{\mu - \lambda} = \frac{1}{10-8} = 0.5 \text{ sec}$$

Each newly arriving packet finds (on average) 4 packets in the system. Since, each packet's average service time is 0.1 sec, it waits for 0.4 sec and then it is served (another 0.1 sec). Thus total average delay is 0.5 sec.

M/M/1: Application (II)

A packet switching node with one outgoing link receives packets at an average Poisson rate of 10 packets per second. Packets are buffered in the queue (unlimited space) if the outgoing link is busy. Given that

- The outgoing link has bandwidth of 1.2 Mbps
- Packet sizes are exponentially distribution with mean size of 1500 bytes

Find the average number of packets in the system.

Given that

- Arrival rate (packets/sec) = 10
- Service rate (packets/sec) = 100, since
  - Channel data rate = 1.2 Mbps (constant rate)
  - Mean packet size = 1500 bytes (or 12000 bits)
- Using:

$$\boxed{\rho = \frac{\lambda}{\mu} < 1} \qquad \boxed{N = \frac{\rho}{(1-\rho)}}$$

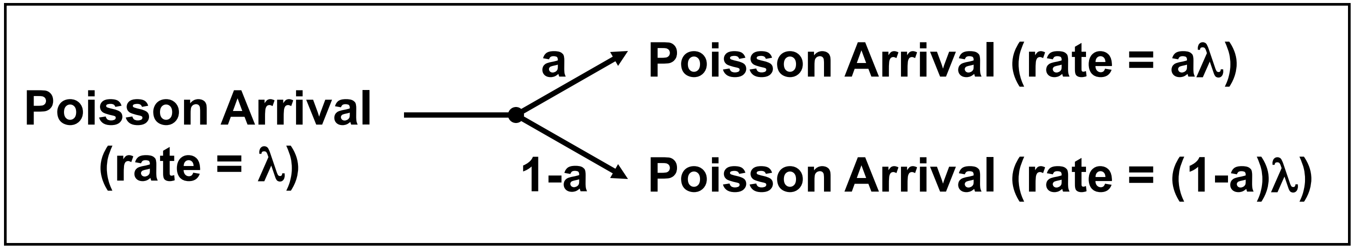
We get

ρ = 10/100 = 0.1

N = 0.1/0.9 = 1/9

M/M/1 Properties (I)

- **Splitting and Aggregating of Poisson Arrivals**



Poisson Arrival (rate = λ)

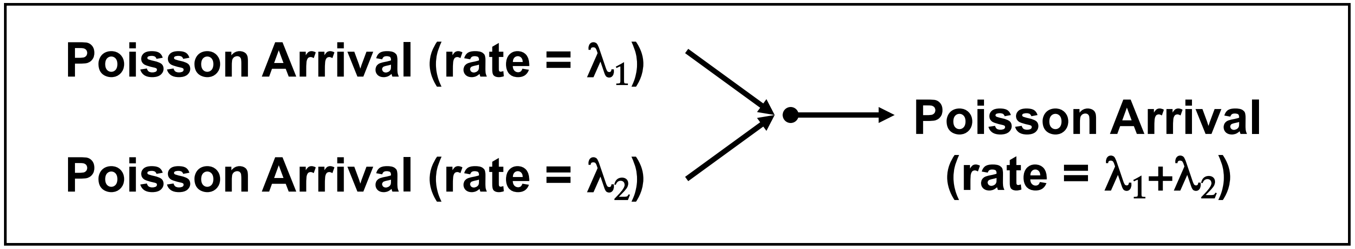
a

Poisson Arrival (rate = $a\lambda$)

1-a

Poisson Arrival (rate = $(1-a)\lambda$)

Example: at McDonld, there are two lines. When students arrive, they randomly with equal probability choose one line to join in. If the student arrival rate is 6 students per minutes. What is the average arrival rate for each line?



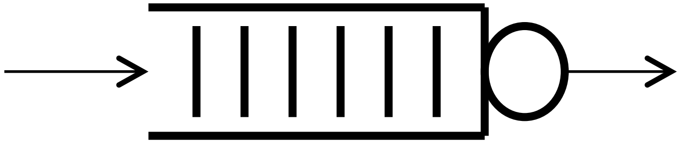
Poisson Arrival (rate = λ_1)

Poisson Arrival (rate = λ_2)

Poisson Arrival (rate = $\lambda_1+\lambda_2$)

M/M/1 Properties (II)

- **M/M/1 departure process**

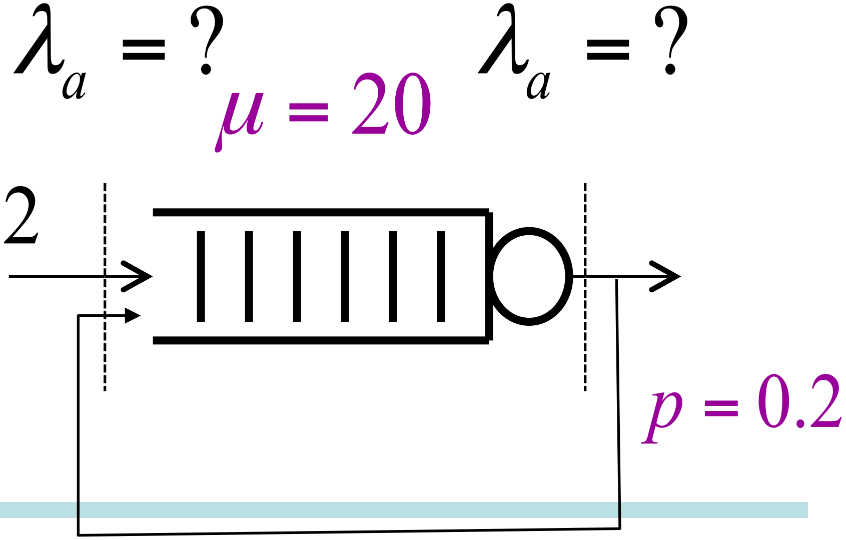


Example: in a freefood event, students arrive at 12 people per minute. The service rate is 20 students/minute. After picking up their food, 20% of students decide to line up again for another round. What is the average delay?

$$\lambda_a = \lambda + 0.2\lambda_a$$

$$\Rightarrow \lambda_a = \frac{1}{1-0.2}\lambda = 15$$

$$T = \frac{1}{\mu - \lambda_a} = \frac{1}{20-15} = 0.2\text{min}$$



Learning Objectives

- **Introduction to Queuing System**
  - Understand five components of a queuing system
- **Probability Laws for Queuing Systems**
  - Understand Exponential random variable
  - Understand Poisson process
  - Quantitative analysis in arrival Process and Service Time
- **M/M/1 Queue Analysis**
  - Understand assumptions
  - Understand Delay analysis and queue occupancy
- **Applications to Packet Switched Networks**
  - Analyze system performance in PSN with M/M/1 queue
  - Understand aggregating and splitting Poisson process
  - Analyze M/M/1 queue with feedback
  - Optimize system performance with proper traffic assignment

Reading Materials

Queue Components (I)

- **Calling Population**
  - The population from which customers/jobs originate
  - The size can be finite or infinite (the latter is most common)
  - Can be homogeneous (only one type of customers/ jobs) or heterogeneous (several different kinds of customers/jobs)
- **Arrival Process**
  - Determines how, when and where customer/jobs arrive to the system
  - Important characteristic is the customers'/jobs' inter-arrival times
  - To correctly specify the arrival process requires data collection of inter-arrival times and statistical analysis.
- **Queue Configuration**
  - Specifies the number of queues
    - Single or multiple lines to a number of service stations
  - Their location
  - Their maximum size (# of jobs the queue can buffer)
    - Distinction between infinite and finite capacity

Queue Components (II)

- **Service Mechanism**
  - Can involve one or several service facilities with one or several parallel service channels (**servers**) - Specification is required
  - The service provided by a server is characterized by its service time
    - Specification is required and typically involves data gathering and statistical analysis.
    - Most analytical queuing models are based on the assumption of exponentially distributed service times, with some generalizations.
- **Queue Discipline**
  - Specifies the order by which jobs in the queue are being served.
  - Most commonly used principle is FIFO.
  - Other rules are, for example, LIFO, SPT, EDD…
  - Can entail prioritization based on customer type.