

# Introduction to Queueing Theory

**Raj Jain**

Washington University in Saint Louis  
Jain@eecs.berkeley.edu or Jain@wustl.edu

A Mini-Course offered at UC Berkeley, Sept-Oct 2012

These slides and audio/video recordings are available on-line at:

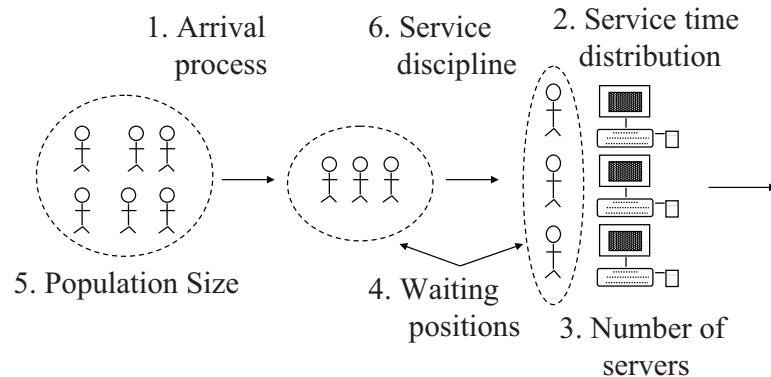
<http://amplab.cs.berkeley.edu/courses/queue>

and <http://www.cse.wustl.edu/~jain/queue>



- Queueing Notation
- Rules for All Queues
- Little's Law
- Types of Stochastic Processes

## Basic Components of a Queue



## Kendall Notation $A/S/m/B/K/SD$

- $A$ : Arrival process
- $S$ : Service time distribution
- $m$ : Number of servers
- $B$ : Number of buffers (system capacity)
- $K$ : Population size, and
- $SD$ : Service discipline

## Arrival Process

- ❑ Arrival times:  $t_1, t_2, \dots, t_j$
- ❑ Interarrival times:  $\tau_j = t_j - t_{j-1}$
- ❑  $\tau_j$  form a sequence of *Independent and Identically Distributed* (IID) random variables
- ❑ Notation:
  - M = Memoryless  $\Rightarrow$  Exponential
  - E = Erlang
  - H = Hyper-exponential
  - D = Deterministic  $\Rightarrow$  constant
  - G = General  $\Rightarrow$  Results valid for all distributions

## Service Time Distribution

- ❑ Time each student spends at the terminal.
- ❑ Service times are IID.
- ❑ Distribution: M, E, H, D, or G
- ❑ Device = Service center = Queue
- ❑ Buffer = Waiting positions

## Service Disciplines

- ❑ First-Come-First-Served (FCFS)
- ❑ Last-Come-First-Served (LCFS) = Stack (used in 9-1-1 calls)
- ❑ Last-Come-First-Served with Preempt and Resume (LCFS-PR)
- ❑ Round-Robin (RR) with a fixed quantum.
- ❑ Small Quantum  $\Rightarrow$  Processor Sharing (PS)
- ❑ Infinite Server: (IS) = fixed delay
- ❑ Shortest Processing Time first (SPT)
- ❑ Shortest Remaining Processing Time first (SRPT)
- ❑ Shortest Expected Processing Time first (SEPT)
- ❑ Shortest Expected Remaining Processing Time first (SERPT).
- ❑ Biggest-In-First-Served (BIFS)
- ❑ Loudest-Voice-First-Served (LVFS)

## Example *M/M/3/20/1500/FCFS*

- ❑ Time between successive arrivals is exponentially distributed.
- ❑ Service times are exponentially distributed.
- ❑ Three servers
- ❑ 20 Buffers = 3 service + 17 waiting
- ❑ After 20, all arriving jobs are lost
- ❑ Total of 1500 jobs that can be serviced.
- ❑ Service discipline is first-come-first-served.
- ❑ Defaults:
  - Infinite buffer capacity
  - Infinite population size
  - FCFS service discipline.
- ❑  $G/G/1 = G/G/1/\infty/\infty/FCFS$

## Quiz 30A

□ Key: A/S/m/B/K/SD

T F

- □ The number of servers in a M/M/1/3 queue is 3
- □ G/G/1/30/300/LCFS queue is like a stack
- □ M/D/3/30 queue has 30 buffers
- □ G/G/1 queue has  $\infty$  population size
- □ D/D/1 queue has FCFS discipline

## Exponential Distribution

□ Probability Density Function (pdf):

$$f(x) = \frac{1}{a} e^{-x/a}$$

□ Cumulative Distribution Function (cdf):

$$F(x) = P(X < x) = \int_0^x f(x) dx = 1 - e^{-x/a}$$

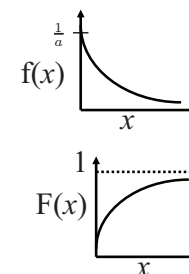
□ Mean:  $a$

□ Variance:  $a^2$


□ Coefficient of Variation = (Std Deviation)/mean = 1

□ Memoryless:

- > Expected time to the next arrival is always  $a$  regardless of the time since the last arrival
- > Remembering the past history does not help.



## Erlang Distribution

□ Sum of  $k$  exponential random variables   
Series of  $k$  servers with exponential service times

$$X = \sum_{i=1}^k x_i \text{ where } x_i \sim \text{exponential}$$

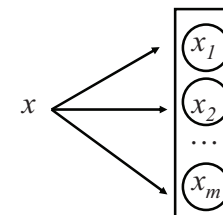
□ Probability Density Function (pdf):

$$f(x) = \frac{x^{k-1} e^{-x/a}}{(k-1)! a^k}$$

- Expected Value:  $ak$
- Variance:  $a^2 k$
- CoV:  $1/\sqrt{k}$

## Hyper-Exponential Distribution

□ The variable takes  $i^{\text{th}}$  value with probability  $p_i$



$x_i$  is exponentially distributed with mean  $a_i$

- Higher variance than exponential
- Coefficient of variation  $> 1$

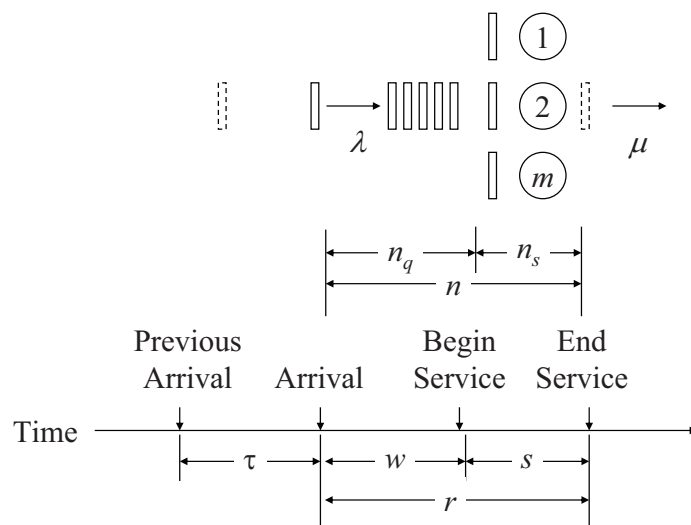
## Group Arrivals/Service

- ❑ Bulk arrivals/service
- ❑  $M^{[x]}$ :  $x$  represents the group size
- ❑  $G^{[x]}$ : a bulk arrival or service process with general inter-group times.
- ❑ Examples:
  - $M^{[x]}/M/1$  : Single server queue with bulk Poisson arrivals and exponential service times
  - $M/G^{[x]}/m$ : Poisson arrival process, bulk service with general service time distribution, and  $m$  servers.

## Quiz 30B

- ❑ Exponential distribution is denoted as \_\_\_\_
- ❑ \_\_\_\_\_ distribution represents a set of parallel exponential servers
- ❑ Erlang distribution  $E_k$  with  $k=1$  is same as \_\_\_\_\_ distribution

## Key Variables



## Key Variables (cont)

- ❑  $\tau$  = Inter-arrival time = time between two successive arrivals.
- ❑  $\lambda$  = Mean arrival rate =  $1/E[\tau]$   
May be a function of the state of the system, e.g., number of jobs already in the system.
- ❑  $s$  = Service time per job.
- ❑  $\mu$  = Mean service rate per server =  $1/E[s]$
- ❑ Total service rate for  $m$  servers is  $m\mu$
- ❑  $n$  = Number of jobs in the system.  
This is also called **queue length**.
- ❑ Note: Queue length includes jobs currently receiving service as well as those waiting in the queue.

## Key Variables (cont)

- $n_q$  = Number of jobs waiting
- $n_s$  = Number of jobs receiving service
- $r$  = Response time or the time in the system  
= time waiting + time receiving service
- $w$  = Waiting time  
= Time between arrival and beginning of service

## Rules for All Queues

Rules: The following apply to  $G/G/m$  queues

1. Stability Condition: Arrival rate must be less than service rate  
 $\lambda < m\mu$

Finite-population or finite-buffer systems are always stable.

Instability = infinite queue

Sufficient but not necessary. D/D/1 queue is stable at  $\lambda=\mu$

2. Number in System versus Number in Queue:

$$n = n_q + n_s$$

Notice that  $n$ ,  $n_q$ , and  $n_s$  are random variables.

$$E[n] = E[n_q] + E[n_s]$$

If the service rate is independent of the number in the queue,

$$\text{Cov}(n_q, n_s) = 0$$

$$\text{Var}[n] = \text{Var}[n_q] + \text{Var}[n_s]$$

## Rules for All Queues (cont)

3. Number versus Time:  
If jobs are not lost due to insufficient buffers,  
Mean number of jobs in the system  
= Arrival rate  $\times$  Mean response time
4. Similarly,  
Mean number of jobs in the queue  
= Arrival rate  $\times$  Mean waiting time

This is known as **Little's law**.

5. Time in System versus Time in Queue

$$r = w + s$$

$r$ ,  $w$ , and  $s$  are random variables.

$$E[r] = E[w] + E[s]$$

## Rules for All Queues(cont)

6. If the service rate is independent of the number of jobs in the queue,

$$\text{Cov}(w, s) = 0$$

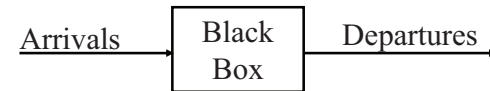
$$\text{Var}[r] = \text{Var}[w] + \text{Var}[s]$$

## Quiz 30C

- ❑ If a queue has 2 persons waiting for service, the number in system is \_\_\_\_\_
- ❑ If the arrival rate is 2 jobs/second, the mean inter-arrival time is \_\_\_\_\_ second.
- ❑ In a 3 server queue, the jobs arrive at the rate of 1 jobs/second, the service time should be less than \_\_\_\_\_ second/job for the queue to be stable.

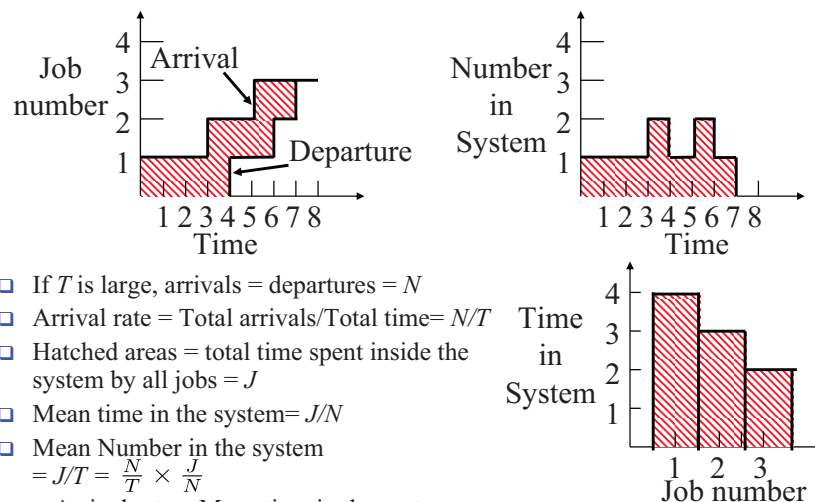
## Little's Law

- ❑ Mean number in the system  
= Arrival rate  $\times$  Mean response time
- ❑ This relationship applies to all systems or parts of systems in which the number of jobs entering the system is equal to those completing service.
- ❑ Named after Little (1961)
- ❑ Based on a black-box view of the system:



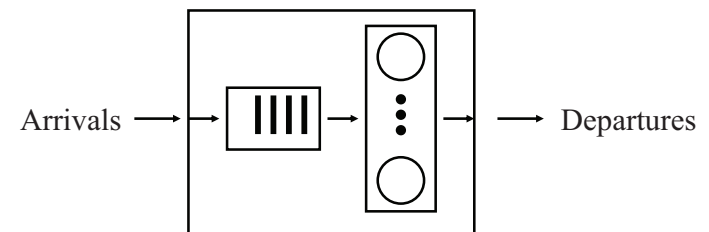
- ❑ In systems in which some jobs are lost due to finite buffers, the law can be applied to the part of the system consisting of the waiting and serving positions

## Proof of Little's Law



- ❑ If  $T$  is large, arrivals = departures =  $N$
- ❑ Arrival rate = Total arrivals/Total time =  $N/T$
- ❑ Hatched areas = total time spent inside the system by all jobs =  $J$
- ❑ Mean time in the system =  $J/N$
- ❑ Mean Number in the system  
=  $J/T = \frac{N}{T} \times \frac{J}{N}$   
= Arrival rate  $\times$  Mean time in the system

## Application of Little's Law



- ❑ Applying to just the waiting facility of a service center
- ❑ Mean number in the queue = Arrival rate  $\times$  Mean waiting time
- ❑ Similarly, for those currently receiving the service, we have:
- ❑ Mean number in service = Arrival rate  $\times$  Mean service time

## Example 30.3

- ❑ A monitor on a disk server showed that the average time to satisfy an I/O request was 100 milliseconds. The I/O rate was about 100 requests per second. What was the mean number of requests at the disk server?
- ❑ Using Little's law:  
Mean number in the disk server  
= Arrival rate  $\times$  Response time  
= 100 (requests/second)  $\times$  (0.1 seconds)  
= 10 requests

## Quiz 30D

- ❑ Key:  $n = \lambda R$
- ❑ During a 1 minute observation, a server received 120 requests. The mean response time was 1 second. The mean number of queries in the server is \_\_\_\_\_

## Stochastic Processes

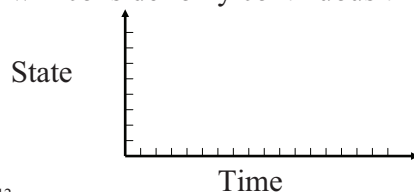
- ❑ **Process:** Function of time
- ❑ **Stochastic Process:** Random variables, which are functions of time
- ❑ **Example 1:**
  - $n(t)$  = number of jobs at the CPU of a computer system
  - Take several identical systems and observe  $n(t)$
  - The number  $n(t)$  is a random variable.
  - Can find the probability distribution functions for  $n(t)$  at each possible value of  $t$ .
- ❑ **Example 2:**
  - $w(t)$  = waiting time in a queue

## Types of Stochastic Processes

- ❑ Discrete or Continuous State Processes
- ❑ Markov Processes
- ❑ Birth-death Processes
- ❑ Poisson Processes

## Discrete/Continuous State Processes

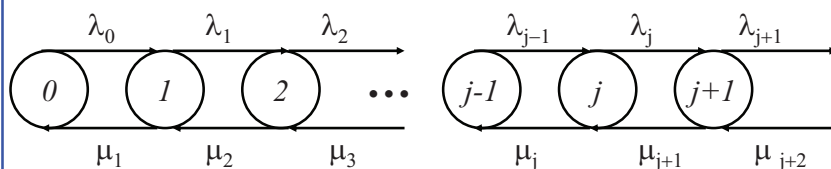
- ❑ Discrete = Finite or Countable
- ❑ Number of jobs in a system  $n(t) = 0, 1, 2, \dots$
- ❑  $n(t)$  is a discrete state process
- ❑ The waiting time  $w(t)$  is a continuous state process.
- ❑ **Stochastic Chain**: discrete state stochastic process
- ❑ Note: Time can also be discrete or continuous  
 $\Rightarrow$  Discrete/continuous **time** processes  
 Here we will consider only continuous time processes



## Markov Processes

- ❑ Future states are independent of the past and depend only on the present.
- ❑ Named after A. A. Markov who defined and analyzed them in 1907.
- ❑ **Markov Chain**: discrete state Markov process
- ❑ Markov  $\Rightarrow$  It is not necessary to history of the previous states of the process  $\Rightarrow$  Future depends upon the current state only
- ❑  $M/M/m$  queues can be modeled using Markov processes.
- ❑ The time spent by a job in such a queue is a Markov process and the number of jobs in the queue is a Markov chain.

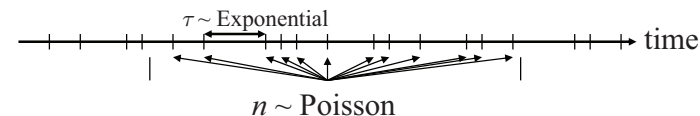
## Birth-Death Processes



- ❑ The discrete space Markov processes in which the transitions are restricted to neighboring states
- ❑ Process in state  $n$  can change only to state  $n+1$  or  $n-1$ .
- ❑ Example: the number of jobs in a queue with a single server and individual arrivals (not bulk arrivals)

## Poisson Distribution

- ❑ If the inter-arrival times are exponentially distributed, number of arrivals in any given interval are Poisson distributed



$$f(\tau) = \lambda e^{-\lambda\tau} \quad E[\tau] = \frac{1}{\lambda}$$

$$P(n \text{ arrivals in } t) = (\lambda t)^n \frac{e^{-\lambda t}}{n!} \quad E[n] = \lambda t$$

- ❑ M = Memoryless arrival = Poisson arrivals
- ❑ Example:  $\lambda=4 \Rightarrow$  4 jobs/sec or 0.25 sec between jobs on average

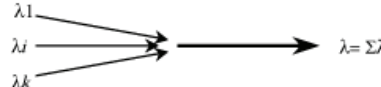


## Poisson Processes

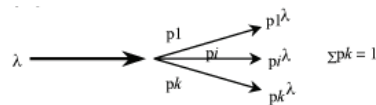
- Interarrival time  $s = \text{IID}$  and exponential  
 $\Rightarrow$  number of arrivals  $n$  over a given interval  $(t, t+x)$  has a Poisson distribution  
 $\Rightarrow$  arrival = Poisson process or Poisson stream

### Properties:

1. Merging:  $\lambda = \sum_{i=1}^k \lambda_i$



2. Splitting: If the probability of a job going to  $i$ th substream is  $p_i$ , each substream is also Poisson with a mean rate of  $p_i \lambda$



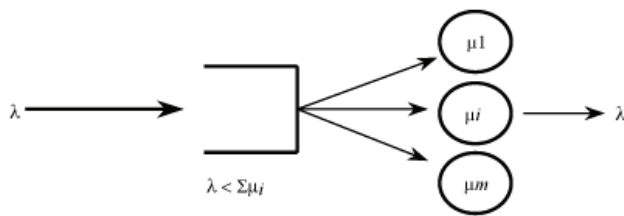
## Poisson Processes (Cont)

3. If the arrivals to a single server with exponential service time are Poisson with mean rate  $\lambda$ , the departures are also Poisson with the same rate  $\lambda$  provided  $\lambda < \mu$ .



## Poisson Process(cont)

4. If the arrivals to a service facility with  $m$  service centers are Poisson with a mean rate  $\lambda$ , the departures also constitute a Poisson stream with the same rate  $\lambda$ , provided  $\lambda < \sum_i \mu_i$ . Here, the servers are assumed to have exponentially distributed service times.

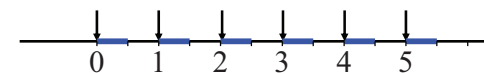


## PASTA Property



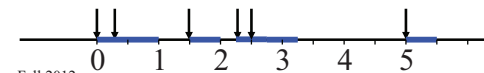
- Poisson Arrivals See Time Averages
- Poisson arrivals  $\Rightarrow$  Random arrivals from a large number of independent sources
- If an external observer samples a system at a random instant:  
 $P(\text{System state} = x) = P(\text{State as seen by a Poisson arrival is } x)$

Example: D/D/1 Queue: Arrivals = 1 job/sec, Service = 2 jobs/sec

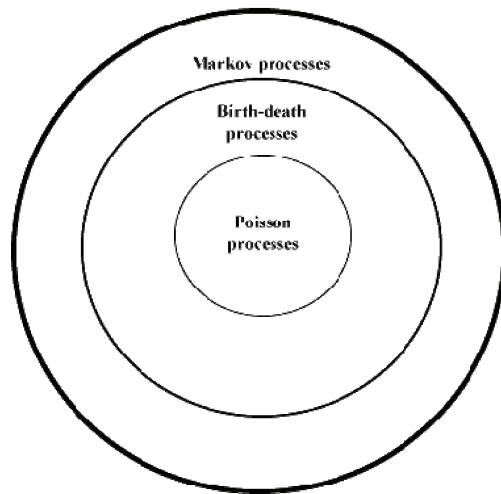


All customers see an empty system.

- M/D/1 Queue: Arrivals = 1 job/sec (avg), Service = 2 jobs/sec  
 Randomly sample the system. System is busy half of the time.



## Relationship Among Stochastic Processes



UC Berkeley, Fall 2012

©2012 Raj Jain

30-41

## Quiz 30E

- T F** Birth-death process can have bulk service
- Merger of Poisson processes results in a \_\_\_\_\_ Process
- The number of jobs in a M/M/1 queue is Markov \_\_\_\_\_
- T F** A discrete time process is also called a chain

UC Berkeley, Fall 2012

©2012 Raj Jain

30-42

## Summary



- Kendall Notation:  $A/S/m/B/k/SD$ , M/M/1
- Number in system, queue, waiting, service  
Service rate, arrival rate, response time, waiting time, service time
- Little's Law:  
Mean number in system = Arrival rate  $\times$  Mean time in system
- Processes: Markov  $\Rightarrow$  Only one state required,  
Birth-death  $\Rightarrow$  Adjacent states  
Poisson  $\Rightarrow$  IID and exponential inter-arrival

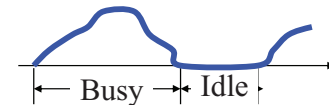
UC Berkeley, Fall 2012

©2012 Raj Jain

30-44

## Homework 30

- Form a team of 2 students. Select any system. Monitor the times of arrivals and departures of requests starting from an idle state and ending at the end of the next idle state with at least one idle state in the middle. (You can have more than one cycle)
  - Examples: System you are designing, network using a network monitor, traffic light, restaurant, grocery store, bank, ...
1. Compute:
    - $\triangleright$  Mean arrival rate
    - $\triangleright$  Mean service rate
    - $\triangleright$  Server utilization
    - $\triangleright$  Mean response time
    - $\triangleright$  Mean number in system
  2. What distribution (M, D, or G) according to you these arrivals and service should have and why? There is no need to analytically verify.  
Due: Via email to jain@eecs by noon next Monday.



UC Berkeley, Fall 2012

©2012 Raj Jain

30-45

## Reading List

- ❑ If you need to refresh your probability concepts, read chapter 12
- ❑ Read Chapter 30
- ❑ Refer to Chapter 29 for various distributions