LECTURE 5
QUEUEING THEORY

Instructor: Lu Wang

College of Business

Shanghai University of Finance and Economics

EVERYONE HAS THE EXPERIENCE OF WAITING IN LINE

“The average person spends 5 years waiting in line!”

“The other line always moves faster!”

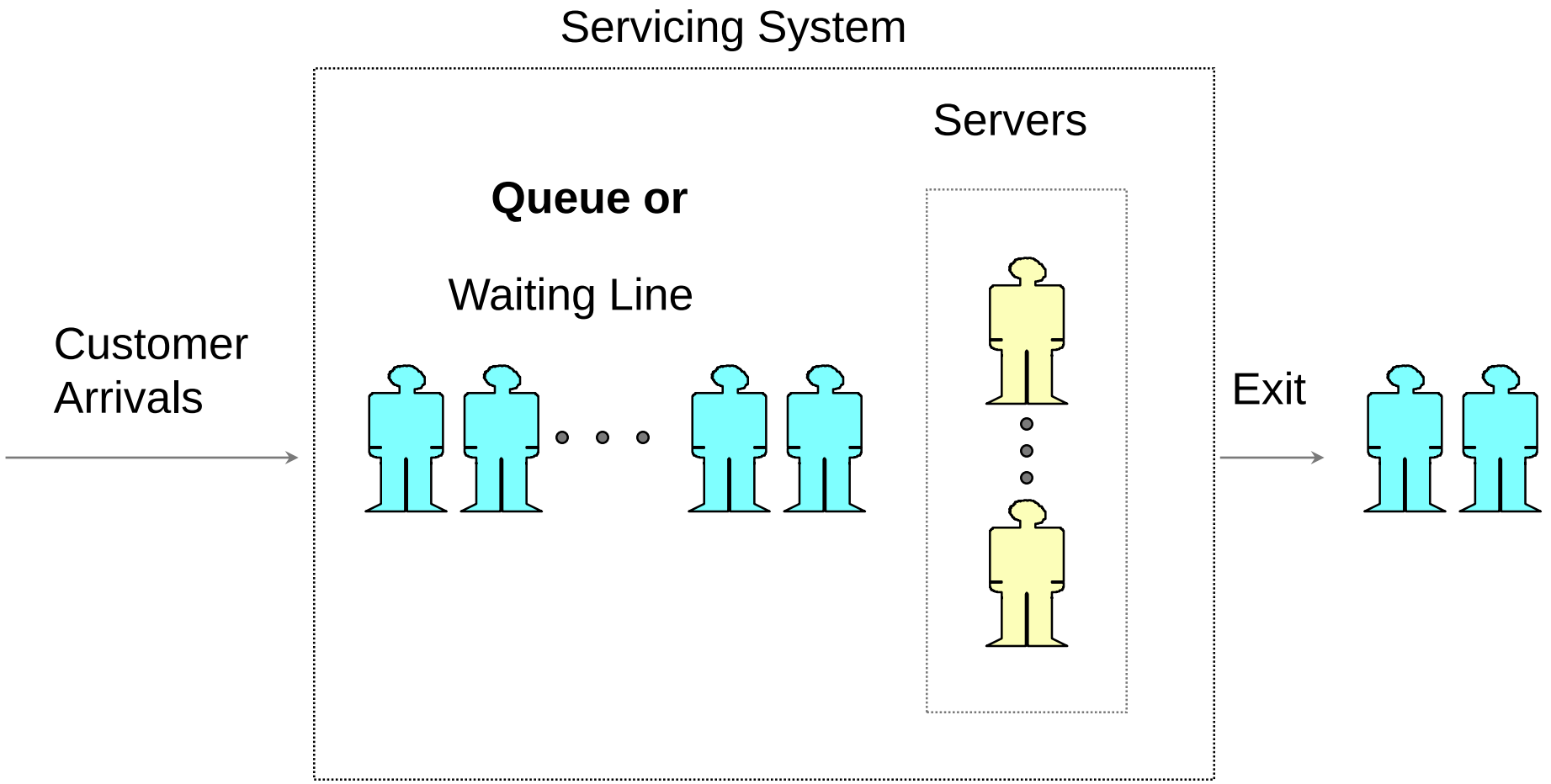
QUEUEING EXAMPLES

- Customers go to the bank, wait in line and make a deposit
- Patients go to the hospital, wait in the waiting room and get treatment
- Cars run into the entrance of a toll road, waiting to pay and enter the highway
- Parts run through the assembly line and get to a work station, wait for the workers to perform operations

QUEUEING THEORY

- **Is one of the core theories in operations research**
- **A large body of research**
- **From the management point of view:**
  - We can always increase capacity to decrease the customer waiting time
  - Balance cost of providing good service with cost of customers waiting.

COMPONENTS OF QUEUES



TERMINOLOGIES

- **Arrival: 1 person, machine, etc., that arrives and demands service**
- **Queue: the waiting line**
- **Server: people or machines that provide service to the arrivals**
- **Queueing discipline: Rules for determining the order that arrivals receive service (FIFO/LIFO/priority based)**
- **Channels: parallel servers**
- **Phases: sequential stages in service.**

ARRIVAL CHARACTERISTICS

Input source (population size)

- Infinite: Number in service does not affect the probability of a new arrival. A very large population is approximately infinite
- Finite: Number in service affects probability of a new arrival (Example: population = 10 aircrafts that may need repair)

Characterization of Arrivals:

- Random:
  - Mean arrival rate = λ (10 arrivals per hour)
  - Mean inter-arrival time = $1/\lambda$ (6 minutes between two arrivals)
- Deterministic (Non-random): Appointments

BEHAVIOR OF ARRIVALS

- **Patient**
  - Arrivals will wait in line for service
- **Impatient**
  - Balking: Arrival leaves before entering line
  - Reneging: Arrival leaves after waiting for a while

QUEUEING CHARACTERISTICS

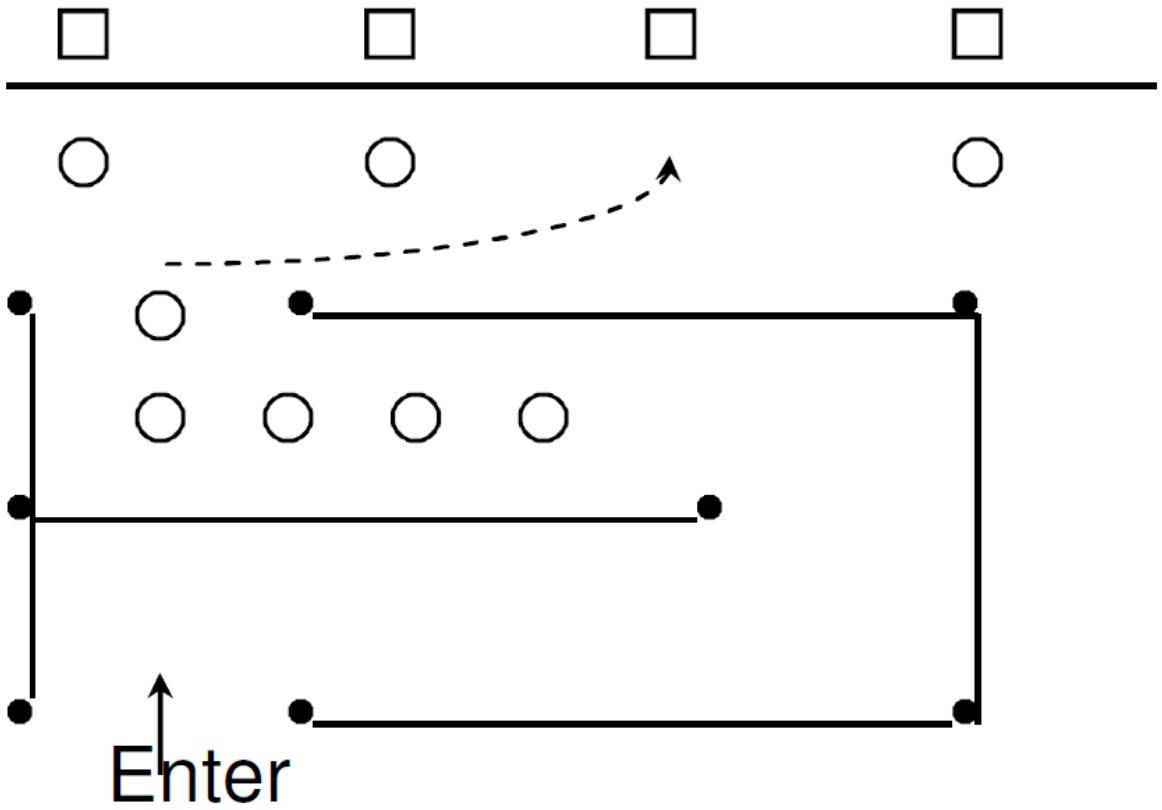
- **Queue capacity:**
  - Limited: Maximum number waiting is limited. (Limited space for waiting)
  - Unlimited: No limit on number waiting
- **Queue discipline:**
  - FIFO (FCFS): First in first out. (First come first serve)
  - LIFO (LCFS): Last in first out. (Last come first serve)
  - Random: Select arrival to serve randomly from the queue
  - Priority based: Give some arrivals priority

QUEUE CONFIGURATION

- **Single channel, single phase**

QUEUE CONFIGURATION

- **Multi-channel, Single phase**



QUEUE CONFIGURATION

- **Single channel, Multi phase**
- **Multi channel, Multi phase**

SERVICE TIMES

- **Non-random: constant**
  - Example: Automated car wash
- **Random**
  - Mean service rate = μ (6 customers / hour)
  - Mean service time = $1/\mu$ (10 minutes for one customer)

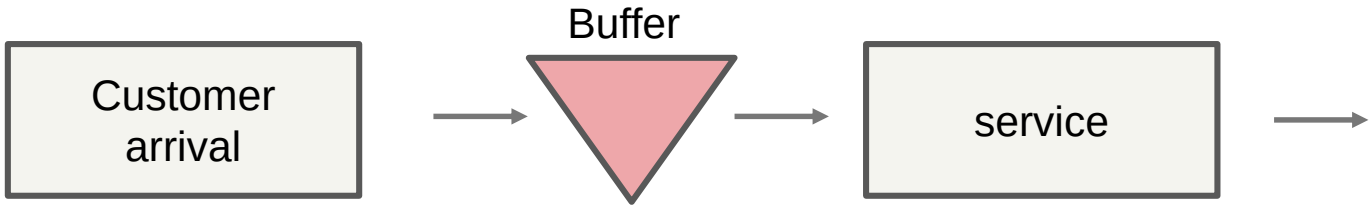
COMMON QUESTIONS

- On average, how many customers are waiting in the line?
- What is the average waiting time for a customer / average total time spend in the system
- How many servers (channels) are needed to keep the average wait within certain limit?
- How many servers (channels) are needed to minimize the total cost?

THE ROLE OF VARIABILITY

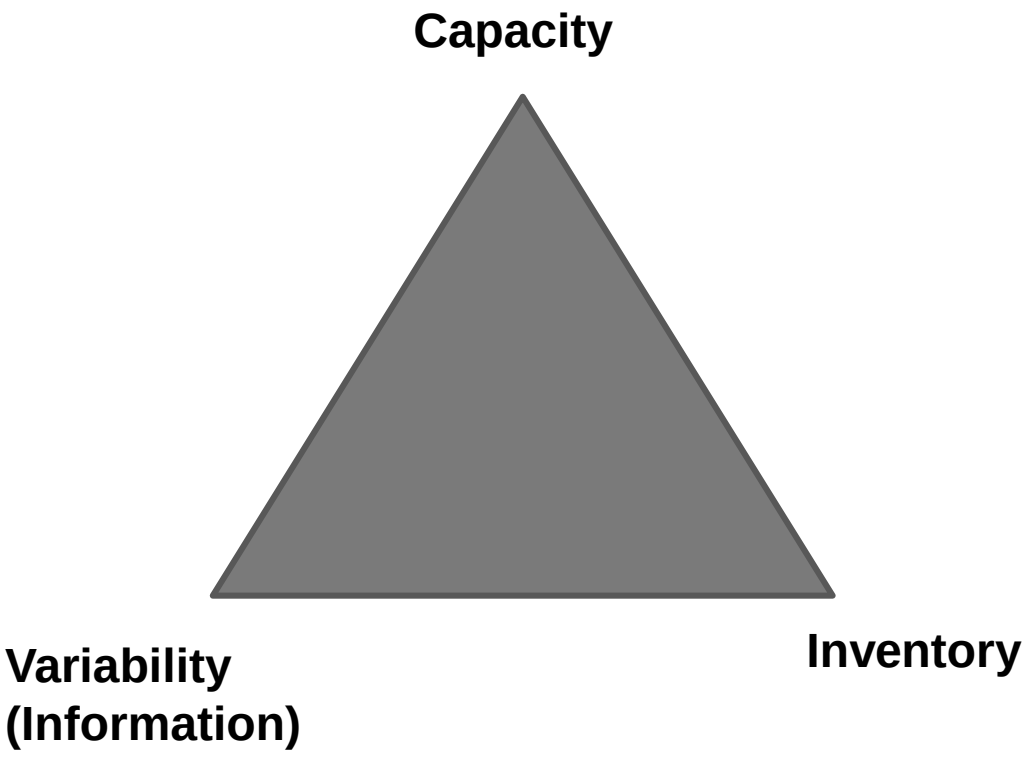
- A local clinic with just one doctor from 7 am to 8 am
- The clinic estimates that a treatment takes 5 minutes to complete
- On average, one patient arrives every 5.5 minutes.
- Question: What will be the average number of customers waiting in line to get treated?

ROLE OF VARIABILITY



- **Number in the queue increases in**
- **The utilization rate**
- **The customer arrival variability**
- **The service variability**

OM TRIANGLE



THE BASIC QUEUEING MODEL

- **We will focus on the scenario when**
  - The arrival is a Poisson process
  - The service time is exponential distributed or deterministic
  - Customer population is homogeneous and infinite
  - Queue capacity is infinite
  - Customers are well behaved (no balking or reneging)
  - Arrivals are served FCFS (FIFO)
  - The total service rate is greater than the arrival rate

POISSON PROCESS

- Number of arrivals during a time interval of length T is characterize by a Poisson distribution
- If the mean number of arrivals is λ per unit of time, then the probability of having x arrivals during T units of time is

$$P(x) = \frac{e^{-\lambda T}(\lambda T)^x}{x!}$$

- Poisson distribution is discrete
- Time between arrivals (inter-arrival time) is exponential distributed:

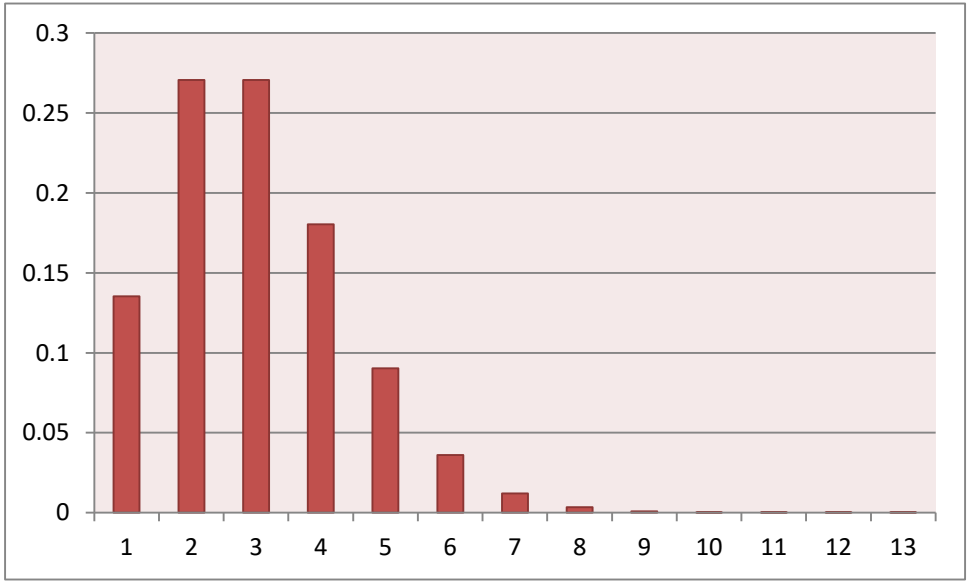
$$f(\mathrm{t}) = \lambda e^{-\lambda t}$$

- Mean of interarrival time is $1/\lambda$

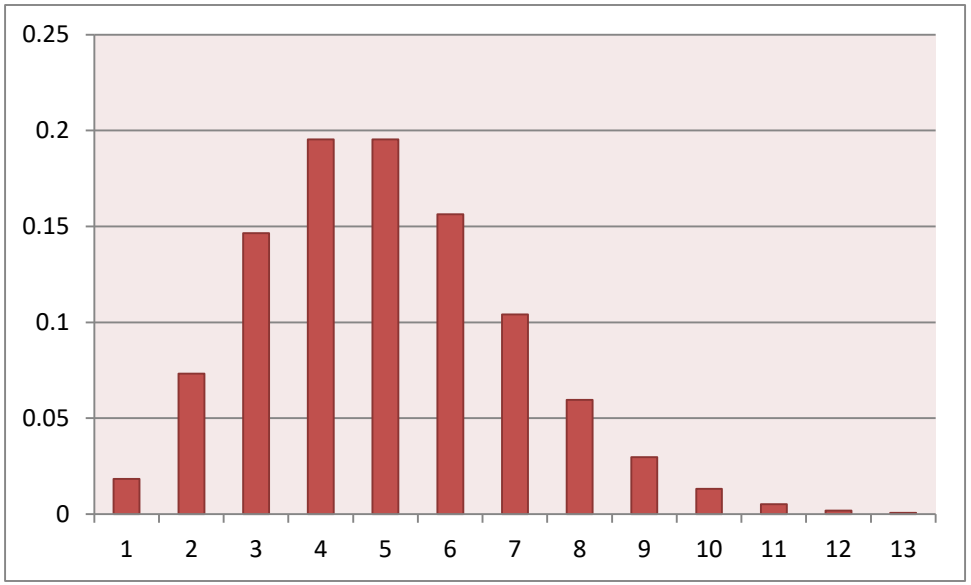
POISSON DISTRIBUTION

- **T = 1**

$\lambda = 2$



$\lambda = 4$



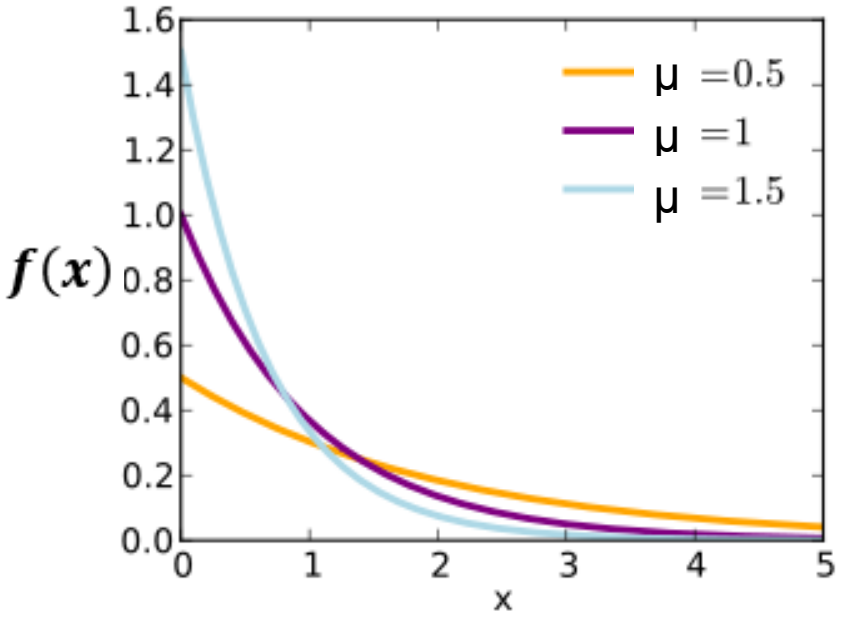
EXPONENTIAL DISTRIBUTION

- **Continuous distribution:**

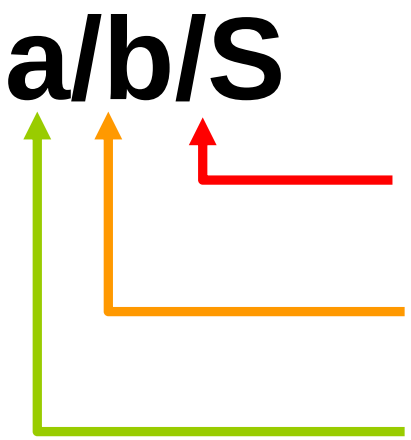
$$P(t \leq x) = 1 - e^{-\mu x}$$

$$f(x) = \mu e^{-\mu x}$$

- **Usually used to characterize interarrival time, service time, life time of a wearable product (light bulb), etc.**



KENDALL'S NOTION FOR QUEUES



Number of servers or channels.

Service time distribution.

Arrival time distribution.

M = Exponential (Poisson Arrival)

G = General distribution

D = Deterministic

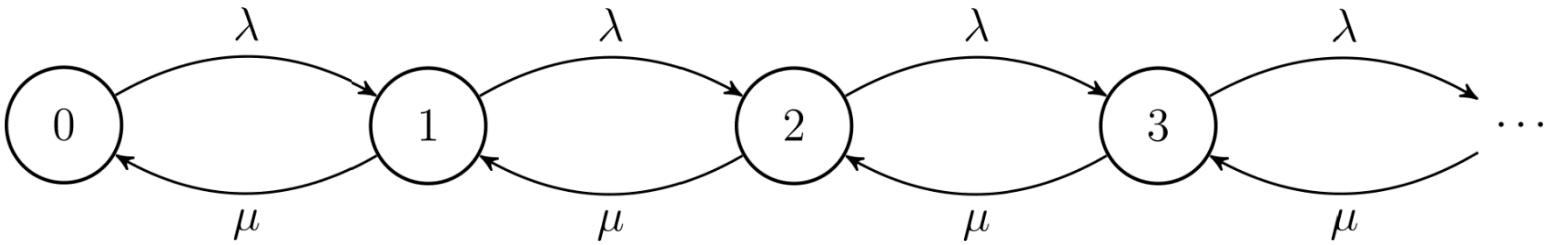
TYPES OF QUEUEING MODELS

- **Simple (M/M/1)**
  - Example: Information counter at the mall
- **Multi-channel (M/M/S)**
  - Example: Airline ticket counter, bank counter
- **Constant Service (M/D/1)**
  - Example: Automated car wash

LONG-RUN AVERAGE PERFORMANCE MEASURES

- Average waiting time = W_q
- Average queue length = L_q
- Average time in the system = W_s
- Average number of customers in the system = L_s
- Probability that there are n customers in the system = P_n
- System utilization rate = ρ

BIRTH-DEATH PROCESS



Birth-death process is a specific type of continuous-time Markov chain.

GENERAL QUEUEING EQUATIONS

- $\rho = \frac{\lambda}{S\mu} = 1 - P_0$
- $L_q = \lambda W_q$
- $L_s = \lambda W_s$
- $W_s = W_q + \frac{1}{\mu}$

Little's Law: Inventory = Throughput x Flow time

- $L_s = L_q + \frac{\lambda}{\mu}$

Given one of W_s, W_q, L_s, L_q, we can solve for the rest

M/M/1 MODEL EQUATIONS

- Utilization $\rho = \frac{\lambda}{\mu}$
- Average # of customers in the system $L_s = \frac{\lambda}{\mu - \lambda}$
- Average time in the system $W_s = \frac{1}{\mu - \lambda}$
- Average # of customers in the queue $L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$
- Average time in the queue $W_q = \frac{\lambda}{\mu(\mu - \lambda)}$
- Probability of having *n* customers in the system $P_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$

M/D/1 MODEL EQUATIONS

- Utilization $\rho = \frac{\lambda}{\mu}$
- Average # of customers in the system $L_s = \frac{\lambda(2\mu-\lambda)}{2\mu(\mu-\lambda)}$
- Average time in the system $W_s = \frac{2\mu-\lambda}{2\mu(\mu-\lambda)}$
- Average # of customers in the queue $L_q = \frac{\lambda^2}{2\mu(\mu-\lambda)}$
- Average time in the queue $W_q = \frac{\lambda}{2\mu(\mu-\lambda)}$

M/M/S MODEL EQUATIONS

- Utilization $\rho = \frac{\lambda}{S\mu}$
- The formulation for of W_s, W_q, L_s, L_q are complicated, we usually lookup into tables for L_q (See L_q table.pdf) and use the general equations to solve for the rest

EXAMPLE: CUSTOMERS IN LINE

Western National Bank is considering opening a drive-through window for customer service. Management estimates that customers will arrive at the rate of 15 per hour. The teller can service customers at the rate of one every three minutes or 20 per hour

Questions: Find

1. **Utilization of the teller**
2. **Average number in the waiting line**
3. **Average number in the system**
4. **Average waiting time in line**
5. **Average waiting time in the system**

EXAMPLE: CUSTOMERS IN LINE

Solution:

1. Utilization rate = $\rho = \frac{\lambda}{\mu} = \frac{15}{20} = 0.75 = 75\%$
2. Average number in the line $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{15^2}{15(20-15)} = 2.25$
3. Average number in the system $L_s = L_q + \frac{\lambda}{\mu} = 3$
4. Average waiting time in line $W_q = \frac{L_q}{\lambda} = \frac{2.25}{15} = 0.15$ hour
5. Average waiting time in the system $W_s = \frac{L_s}{\lambda} = \frac{3}{15} = 0.2$ hour

EXAMPLE: CUSTOMERS IN LINE

Because of the limited space availability and a desire to provide an acceptable level of service, the bank manager would like to ensure, with 95 percent confidence, that no more than 3 cars will be in the system. What is the present level of service for the three-car limit? What is the minimum service rate to ensure 95% service level?

What is $P_0 + P_1 + P_2 + P_3$

What is the minimum μ to make sure $P_0 + P_1 + P_2 + P_3 \geq 0.95$

EXAMPLE: CUSTOMERS IN LINE

Solution:

Use $P_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$**, with** $\lambda = 15, \mu = 20$

- $P_0 + P_1 + P_2 + P_3 = \left(1 - \frac{\lambda}{\mu}\right)\left(1 + \left(\frac{\lambda}{\mu}\right)^1 + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3\right)$
- $= 0.25\ (1 + 0.75 + 0.75^2 + 0.75^3) = 0.684$
- With $\left(1 - \frac{\lambda}{\mu}\right)\left(1 + \left(\frac{\lambda}{\mu}\right)^1 + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3\right) = 0.95$, solve μ
  - Use trial and error
  - Use computer software or online tools, such as
  - http://www.wolframalpha.com/
  - $\rho = \frac{\lambda}{\mu} = 0.47,\ \mu = 32$

EXAMPLE: DETERMINE THE NUMBER OF SERVERS

In the service department of Glenn-Mark Auto Agency, mechanics requiring parts for auto repair or service present their request forms at the parts department counter. The parts clerk fills a request while the mechanic waits. Mechanics arrive in a random (Poisson) fashion at the rate of 40 per hour, ad a clerk can fill requests at the rate of 20 per hour (exponential). If the cost a clerk is $6 per hour and cost for a mechanic waiting in line is $12 per hour, determine the optimum umber of clerks to staff the counter.

EXAMPLE: DETERMINE THE NUMBER OF SERVERS

Solution: We will need at least 3 clerks. Why?

$\frac{\lambda}{\mu}$ = 2, If we hire 3 clerks, $L_q = 0.888$

- Cost of clerks per hour = 3 * $6 = $18
- Cost of mechanics waiting = 0.889 * $12 = $10.67
- Total cost = $18 + $10.656 = $28.67

If we hire 4 clerks, $L_q = 0.173$

- Cost of clerks per hour = 4 * $6 = $24
- Cost of mechanics waiting = 0.173 * $12 = $2.08
- Total cost = $24 + $2.08 = $26.08

SINGLE-SERVER QUEUEING APPROXIMATION INPUT PROCESS

λ = 15 customers per hour

- Expected time between arrivals in minutes
- $E\{a\} = 1/\lambda = 1/15$ hr = 4 min

Times between customer arrivals (i.e., it is called inter-arrival times) may vary

- with standard deviation $\sigma\{a\}$

Ca = coefficient of variation of inter-arrival time

- = standard deviation divided by mean
- $= \sigma\{a\}/E\{a\} = \sigma\{a\}\lambda$

SINGLE-SERVER QUEUEING APPROXIMATION SERVICE PROCESS

μ = 0.5 customers per minute

- Expected time for service to take place

$$E\{s\} = 1/\mu = 2 \text{ min}$$

Service times are not always the same

(i.e., they fluctuate even for the exact same task)

- standard deviation $\sigma\{s\}$

Cs = coefficient of variation in service times

= standard deviation divided by mean

= $\sigma\{s\}/E\{s\}$

= $\sigma\{s\}\mu$ [dimensionless]

SINGLE-SERVER QUEUEING APPROXIMATION POLLACZEK-KHINCHIN (PK) FORMULA

$$I_q \cong \frac{\rho^2}{1-\rho} * \frac{C_a^2 + C_s^2}{2}$$

"=" for special cases "≈" in general

l_q = average queue length (excl. the one in service)

r = average utilization

= average throughput / average capacity = λ / μ

C_a = coefficient of variation of inter-arrival time = $\sigma\{a\}/E\{a\}$

C_s = coefficient of variation of service times = $\sigma\{s\}/E\{s\}$

IMPACT OF UTILIZATION ($\rho = \lambda/\mu$)

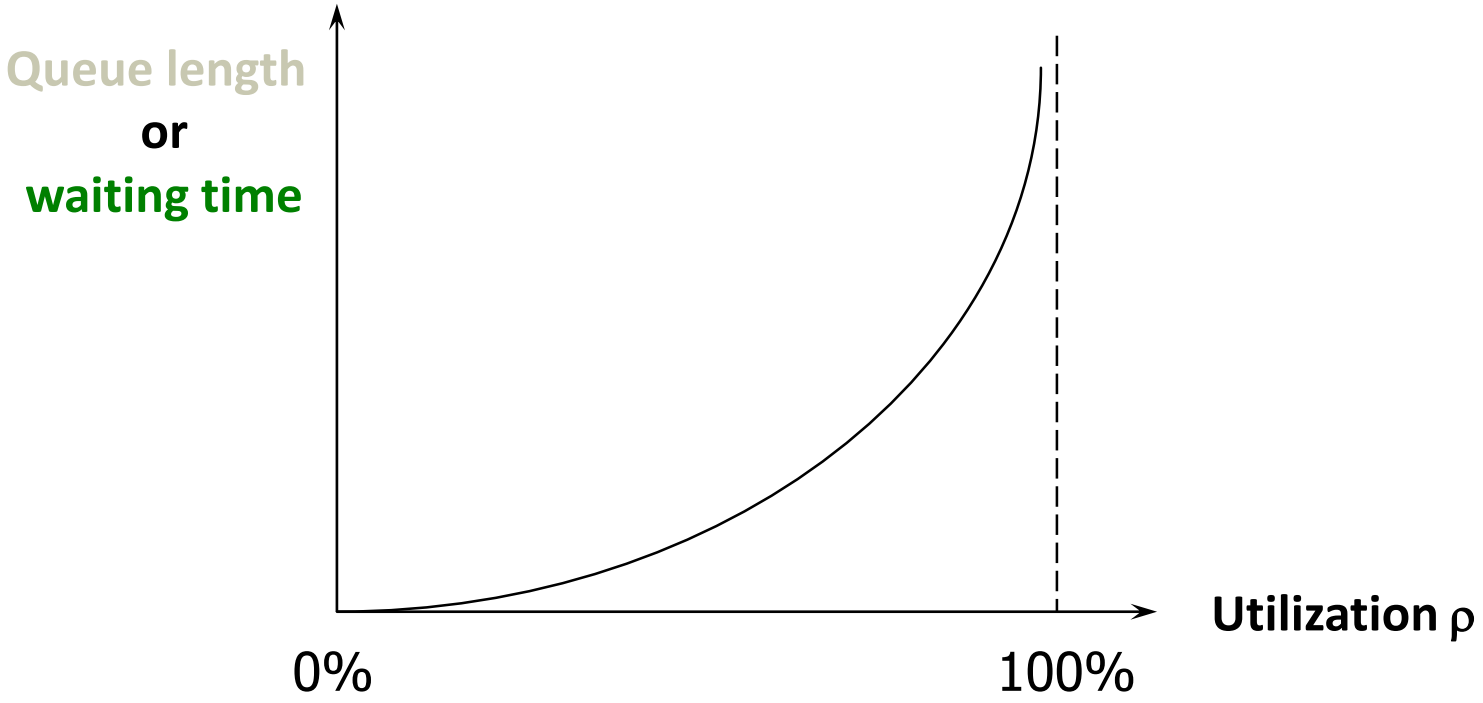
Impact on queue length (inventory)

$$I_q \cong \frac{\rho^2}{1-\rho} * \frac{C_a^2 + C_s^2}{2}$$

Impact on waiting time (flow time)

By Little's Law:

$$T_q = I_q/\lambda$$



SUMMARY

- Understand what is a queueing model
- M/M/1, M/M/s, M/D/1 and their performance measures