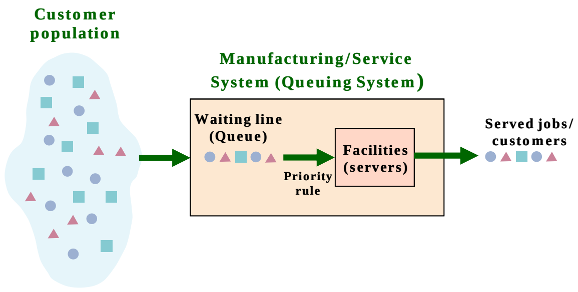
ENE 3031
Computer Simulation
Week 3: Queueing Theory

Chuljin Park
Assistant Professor
Industrial Engineering
Hanyang University

1

Basic Queuing Model



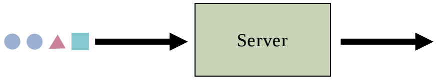
HANYANG UNIVERSITY

Examples of Queues

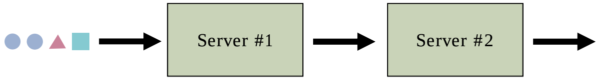
Arrivals	Servers	Queue
Shoppers	Clerks	Checkout line
Patients	Doctor	Waiting room
Patients	Operating teams	Waiting list
Customers	Stock	Backorders
Machine breakdowns	Repair persons	Broken machines
Automobiles	Intersection	Traffic jams

HANYANG UNIVERSITY

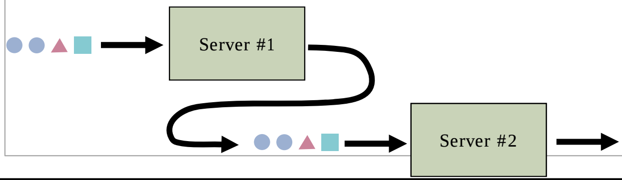
- Single-Server Single-Stage



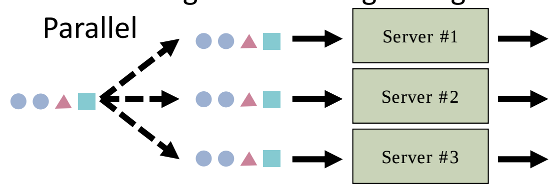
- Single-Server Multiple-Stage



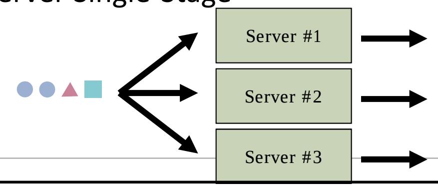
- Single-Server Queues in Series



- Several Single-Server Single-Stage in Parallel



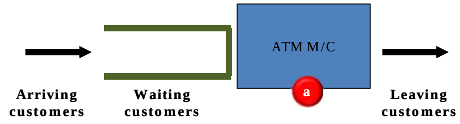
- Multiple-Server Single-Stage



Simple Queuing Model

- Single-server

TIME 2



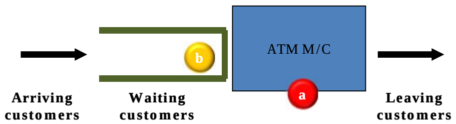
Arr.	Start	End
2	2	5

HANYANG UNIVERSITY

Simple Queuing Model

- Single-server Single-stage

TIME 3



Arr.	Start	End
2	2	5
3	5	6

HANYANG UNIVERSITY

Simple Queuing Model

- Single-server Single-stage

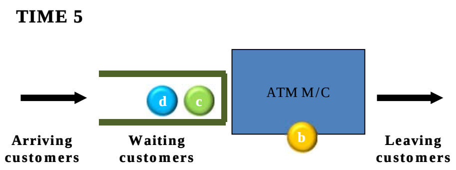
TIME 4

Arr.	Start	End
2	2	5
3	5	6
4	6	8

HANYANG UNIVERSITY

Simple Queuing Model

- Single-server Single-stage

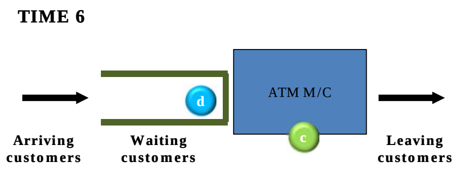


Arr.	Start	End
2	2	5
3	5	6
4	6	8
5	8	

HANYANG UNIVERSITY

Simple Queuing Model

Single-stage



Arr.	Start	End
2	2	5
3	5	6
4	6	8
5	8	…
…	…	…

HANYANG UNIVERSITY

Performance

- How does the system perform?
  - Utilization of servers
  - Average waiting time in queue
  - Average staying time in the system
  - Average number of customers in queue
  - Average number of customers in the system

HANYANG UNIVERSITY

Basic Performance Measures

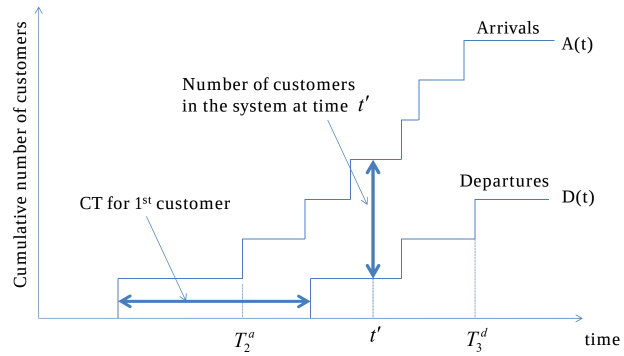
- A(t): Number of arrivals until time t
- D(t): Number of departures until time t
- L(t): Number of customers in system at t
- L_Q(t): Number of customers in queue at t

HANYANG UNIVERSITY 12

Basic Performance Measures

- a: time when the first customer enters to the empty system
- b: time when the last customer leaves from the system (the system just becomes empty).
- WIP(a,b): Number of customers in system per unit time from time a to time b.
- CT(a,b): Time spent in system per customer from time a to time b.

A(t) and D(t)



WIP(a,b)

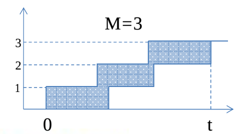
- Consider a time interval (*a,b*) such that the system **starts empty** and **returns to empty**
- Let *L*(*t*)=*A*(*t*)-*D*(*t*), number of customers in the system at time *t*

$$WIP(a,b) = \frac{1}{b-a}\int_a^b L(t)dt = \frac{1}{b-a}\int_a^b \left(A(t)-D(t)\right)dt$$

CT(a,b)

- *M*: number of customers that arrive to (or depart from) the system during the interval (*a,b*)
- T_i^d : departure time for the ith customer
- T_i^a : arrival time for the ith customer

$$CT(a,b) = \frac{1}{M}\sum_{i=1}^{M}\left(T_i^d - T_i^a\right)$$



$$= \frac{1}{M}\int_a^b \left(A(t)-D(t)\right)dt$$

CT(a,b) and WIP(a,b)

$$WIP(a,b)=\frac{1}{b-a}\int_a^b \left(A(t)-D(t)\right)dt$$

$$CT(a,b)=\frac{1}{M}\int_a^b \left(A(t)-D(t)\right)dt$$

$$\Rightarrow WIP(a,b)=\frac{M}{b-a}CT(a,b)=\hat{\lambda}\cdot CT(a,b)$$

Throughput: Average number of customers arriving to (departing from) the system per unit time

HANYANG UNIVERSITY

Example: ATM case - Hand Simulation-

Let's consider an ATM process. We observed 1) interarrival times and 2) service times for 6 customers as follows:

i	Interarrival time Between i-1 and i	Service time for i
1	2	3
2	1	1
3	1	2
4	1	1
5	2	2
6	3	1

Find estimates of 1) Cycle time, 2) WIP, 3) Throughput, and 4) waiting time of the process.

HANYANG UNIVERSITY

Performance Measure (infinite T)

- Long-run time-average number of customers in system (L)

$$L=\lim_{T\to\infty} WIP_{(0,T)}=\lim_{T\to\infty}\frac{1}{T}\int_0^T L(t)dt=\lim_{T\to\infty}\frac{1}{T}\int_0^T \left(A(t)-D(t)\right)dt$$

- Long-run average time spent in system per customer (W)

$$W=\lim_{T\to\infty} CT_{(0,T)}=\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\left(T_i^d-T_i^a\right)$$

HANYANG UNIVERSITY

Performance Measure (infinite T)

- Long-run time-average number of customers in queue (L_Q)

$$L_Q=\lim_{T\to\infty}\frac{1}{T}\int_0^T L_Q(t)dt$$

- Long-run average time spent in queue per customer (W_Q)

$$W_Q=W-E[S] \text{ where } S \text{ is a service time.}$$

HANYANG UNIVERSITY

Little's Law

- Throughput Rate (*TH*)
  - The **number of *completed* jobs** leaving the system **per unit of time**
- For a system satisfying steady-state conditions,

$$L = \lambda W$$

$$(\text{cf. } WIP(0,T) = \hat{\lambda} \times CT(0,T))$$

Steady-State Behavior

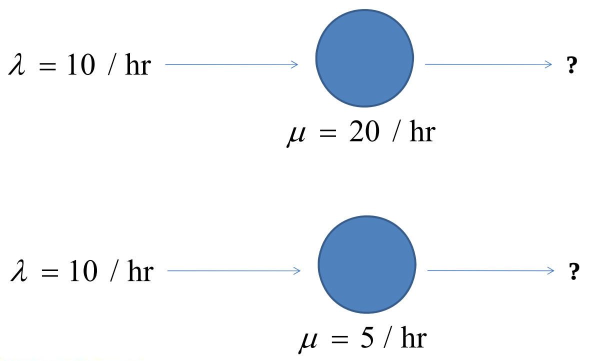
Steady-State: the probability that the system in a given state is independent of t.

	Arrival	Service
Rate	Arrival Rate λ The mean number of arrivals during a time period	Service Rate μ The mean number of customers serviced during a time period
Time	Mean Inter-arrival Time The expected amount of time between two sequential arrivals	Mean Service Time The expected amount of time needed to service a customer

Examples

- If six customers visit a store during a hour on average, the arrival rate would be expressed as 6 customers/hour, and mean inter-arrival time would be equal to 10 minutes/customer (=1/6 hour/customer)
- If a cashier can attend , on an average 5 customers in an hour, the service rate would be expressed as 5 customers/hour, and mean service time would be equal to 12 minutes/customer (=1/5 hour/customer)

Steady-State Behavior (Stability)



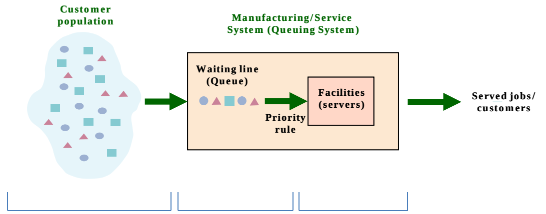
Steady-State Behavior (Stability)

- Stability (server utilization or traffic intensity)
  - Proportion of time that a server is busy.

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

  - For the stable system (normally operating), the system need $\boxed{\rho < 1}$.

Characteristics of Basic Queuing Model



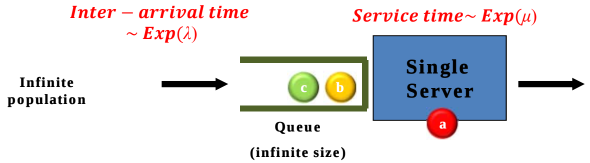
Characteristics of Queuing Systems

- Arrivals
  - Population size
  - Arrival distribution (inter-arrival time)
- Services
  - Number of servers
  - Service distribution (service time)
- Queue
  - Queue size (finite? Infinite?)
  - Service priority among customers
- Customer behavior in queue
  - Balking – customers do not join if a line is long
  - Reneging – customer quit the line if waiting too long

Queue Notation

- A / B / c / (E / F / Queue discipline)
  - Symbols for A (distribution type of inter-arrival times) and B (distribution type of service times) :
    - D – deterministic
    - M – exponential
    - G – general
  - c: number of identical and parallel servers
  - E: system capacity
  - F: size of population
  - Example 1: M/M/1(/∞/∞)
  - Example 2: There are10 PC's and 5 pagers for students waiting for a PC. (a) students can still wait for a PC even if a lab runs out of pagers. M/M/10 (b) If not, M/M/10/15

M/M/1(∞/ ∞)



- Arrival process is the Poisson process with rate λ.
  - Inter-arrival times follow the exponential distribution with parameter λ.
- Service times follow the exponential distribution with parameter μ.
- First-In First-Out

HANYANG UNIVERSITY Chuljin Park 29

M/M/1(∞/ ∞)

- Sever Utilization $(stability)$ $\rho = \frac{\lambda}{\mu}$
- $L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$
- $W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda}$
- $L_Q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$
- $W_Q = \frac{L_Q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}$
- $P_n = P(L(t) = n) = (1-\rho)\,\rho^n$

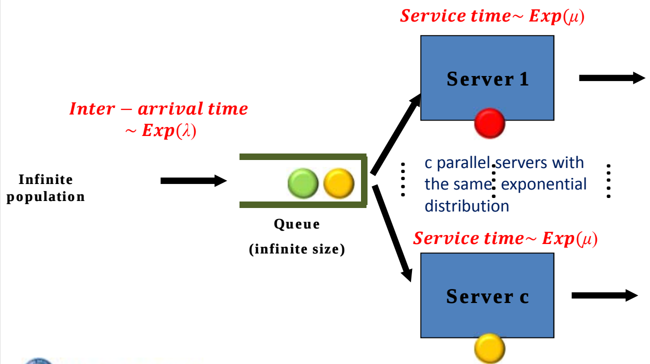
HANYANG UNIVERSITY Chuljin Park 30

Example: Hair-Styling Shop

- Consider a single-chair unisex hair-styling shop.
- Interarrival time ~ Exp(2): Exponential with mean 1/2 hour
- Service time ~ Exp(3): Exponential with mean 1/3 hour
  - Find server utilization, L, L_Q, W, W_Q.
  - Find the probability that the hair stylist is busy

HANYANG UNIVERSITY Chuljin Park 31

M/M/c(∞/ ∞)



HANYANG UNIVERSITY Chuljin Park 32

M/M/c(∞/ ∞)

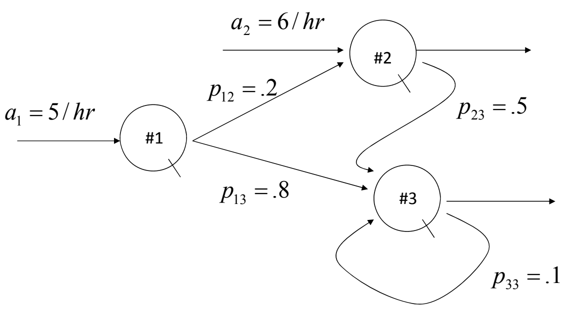
- $\rho = \frac{\lambda}{c\mu}$
- $L_Q = \frac{(\frac{\lambda}{\mu})^c \rho}{c!(1-\rho)^2}\left(\sum_{n=0}^{c-1}\frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^c}{c!}\cdot\frac{1}{1-\frac{\lambda}{c\mu}}\right)^{-1}$
- $W_Q = \frac{L_Q}{\lambda}$
- $W = W_Q + \frac{1}{\mu}$
- $L = \lambda W = \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu}$

Example: Hospital

- Patients arrive according to a Poisson process with intensity of 2 patients per hour.
- The service time (the doctor's examination and treatment time of a patient) follows an exponential distribution with its mean of 20 minutes.

➢ ***What is the number of doctors to make the average wait time before the service for a patient no bigger than 30 minutes?***

Queueing Networks (Rough-cut Modeling)



Overall Arrival Rate

- We use the fact that
  - Arrival rate into a queue = departure rate out of a queue
  - The overall arrival rate into a queue is the *sum* of all the arrival rates
  - # servers does not matter

Overall arrival rate for station j

Probability that an entity moves from station i to station j

$$\lambda_j = a_j + \sum_{\text{all queues } i} \lambda_i p_{ij}$$

Outside arrival rate for station j

Sum of all internal arrival rates

Example

$$\lambda_j = a_j + \sum_{\text{all queues } i} \lambda_i p_{ij}$$

$$\lambda_1 = a_1 = 5$$

$$\lambda_2 = a_2 + \sum_{i=1}^{3} \lambda_i p_{i2} = 6 + 0.2\lambda_1$$

$$\lambda_3 = a_3 + \sum_{i=1}^{3} \lambda_i p_{i3} = 0.8\lambda_1 + 0.5\lambda_2 + 0.1\lambda_3$$

$$\lambda_1 = 5, \lambda_2 = 7, \lambda_3 = 8\tfrac{1}{3}$$

Comments

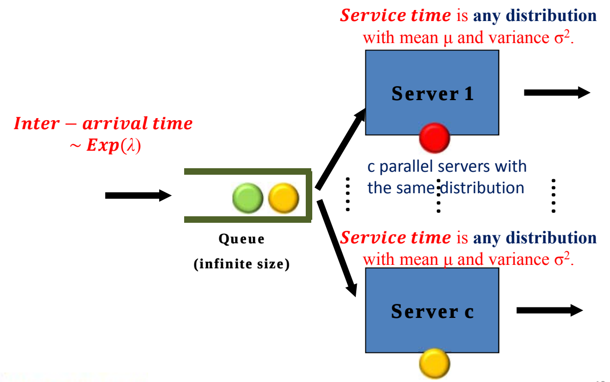
- Result assumes no capacity restriction
- Result does not depend on service rate at each queue [but must be fast enough to keep up]
- The number of servers does not matter.
- If
  - external arrival processes are Poisson,
  - Service times are exponential,
  - Infinite queue and probabilistic routing

  Then each queue behaves like an independent M/M/c Queue!

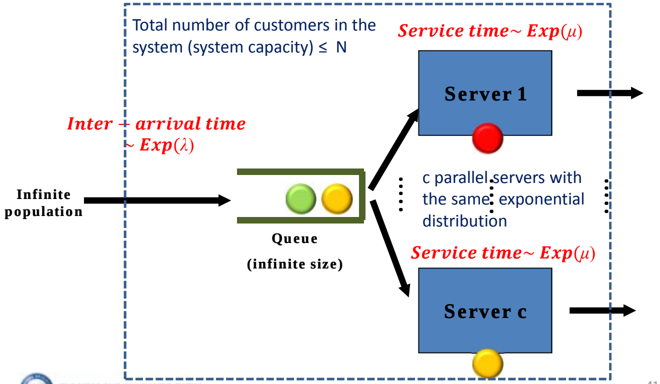
R-C example

- A production line consists of two stations (station 1 and 2) and one rework station (station 3). An engineer recorded some time study data between 8am and 6pm over one week (5 days).
  - # of arrivals to station 1: 1000
  - Average service time of station 1: 1/15 hr (2 servers)
  - Average service time of station 2: 1/24 hr (1 server)
  - Average service time of station 3: 1/8 hr
  - 20% jobs are found to have defects at station 1 and sent to rework station. Only 50% are salvaged at rework station and sent to station 2.
  - Note that we don't know the distribution of number of arrivals and service times.
- The question is the following:
  - One extra server is available now. I'd like to know which station to put him.

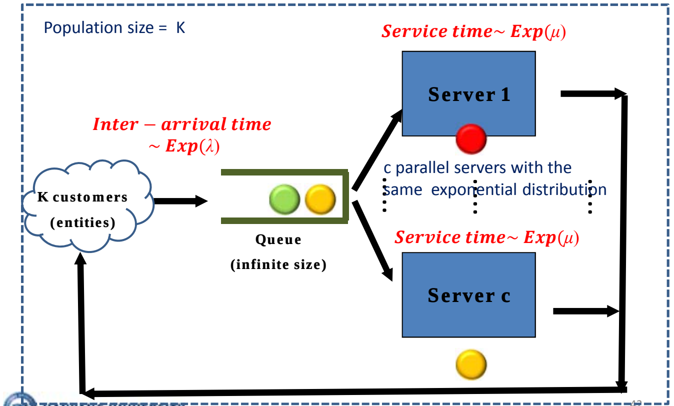
Advanced Part 1: M/G/c(∞/ ∞)



Advanced Part 2: M/M/c(N/ ∞)



Advanced Part 3: M/M/c(K/K)



Solving a simple Queueing problem

(0. Decide whether the system is steady-state or not.)

1. If it is steady-state, formulate the problem and find the type of the queueing systems (i.e., M/M/1/ (∞/ ∞)).
2. Set performance measures you are interested in (i.e., server utilization)
3. If any formula (or computer program) exist, then find the solution using them.

Example: Milling machines

- There are two workers who are responsible for 10 milling machines.
- The machines run on the average for 20 minutes, then requires 5-minute service period, both times exponentially distributed.

➢ ***Compute the various measures of performance for this system.***

Summary

- Queueing theory is very useful to examine some systems in a wide range of applications (manufacturing/service/SCM/IT).
- However, real systems are much more complicated with complex structures and Queueing is not enough to cover them.

Next Class

- Hand Simulation