Lecture 5. Queuing Theory

Sim, Min Kyu, Ph.D.
mksim@seoultech.ac.kr

서울과학기술대학교 SEOULTECH SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

BK21 FOUR 데이터사이언스와 비즈니스포텐셜

Warming up - Exponential distribution

Warming up - Example 1

- Suppose that the amount of time one spends in a bank is exponentially distributed with mean of ten minutes, that is $\lambda = 1/10$.
  - What is the probability that a customer will spend more than fifteen minutes in the bank?
  - What is probability that a customer will spend more than fifteen minutes in the bank given that she is in the bank after ten minutes?

Warming up - Example 2

- Consider a post office that is run by two clerks.
  - Suppose that when Mr. Smith enters the system he discovers that Mr. Jones is being served by one of the clerks and Mr. Brown by the other.
  - Suppose also that Mr. Smith is told that his service will begin as soon as either Jones or Brown leaves.
  - If the amount of time that a clerk spends with a customer is exponentially distributed with mean $1/\lambda$, what is probability that, of the three customers, Mr. Smith is the last one to leave the post office?

Queuing Theory

- Motivation: You want to study the service flow of the post office. What information you need to have?

Standard variables in queuing theory

- Inter-arrival times: How frequently customers arrive?
- Inter-departure times (Service times): How fast the servers do the jobs?
- Number of servers
- Queue size: How large waiting space?

Standard Notation (Kendall's)

- **G**: General Distribution
- **M**: Exponential Distribution
- **D**: Deterministic (constant)

- Examples
  - $G/G/1/\infty$
  - $M/M/1/\infty$
  - $G/D/2/2$

Illustration - G/D/2/2 (barber shop)

Stability and Traffic intensity - Motivation (w/ one server)

1. Suppose there are 30 customers coming to post office in an hour on average, and the clerk can handle 20 customers in an hour on average. Is this system ( )?
2. Suppose mean of inter-arrival times is 2 minutes and the mean of service times is 3 minutes, is this system Stable?
3. What is condition for *stability*?
4. Inverse relationship between 1 and 2.
5. Inverse relationship between rate and inter-(blah)times?

Definition - stable

- What is being **"stable"**?

Stability and Traffic intensity

Definition - stability and traffic intensity

- A queuing system is said to be **unstable** if the number of waiting customers for the system diverges to infinity.
- A queuing system is said to be **stable** unless unstable.
- The **traffic intensity** ρ is the ratio of demand to capacity in the queuing system.

- Property : Stability of queuing system with infinite waiting spaces.
  - If $\rho < 1$,
  - If $\rho = 1$,
  - If $\rho > 1$,

- Particularly for $M/M/1/\infty$, the traffic intensity ρ is defined to be $\rho = \frac{\lambda}{\mu} = \frac{1/\mathbb{E}(U)}{1/\mathbb{E}(V)}$, where λ is arrival rate, μ is departure (service) rate, U is inter-arrival times, and V is inter-departure times.

Definition - Utilization of server (u)

- The fraction of time that a server is busy

blank

blank

Kingman's high-traffic approximation formula

Assume the traffic intensity $\rho < 1$ but ρ is close to 1. The long-run average waiting time W_q in a queue is following:

$$EW_q = EV(\frac{\rho}{1-\rho})(\frac{c_a^2 + c_s^2}{2}),$$

where V is service time, c_a and c_s are coefficient of variation of inter-arrival times and service times, respectively.

Example

1. Suppose inter-arrival time follows an exponential distribution with mean of 3 minutes and service time follows an exponential distribution with mean of 2 minutes. What is the expected waiting time per customer?

2. Suppose inter-arrival time is constant 3 minutes and service time is also constant 2 minutes. What is the expected waiting time per customer?

Little's Law

$$L = \lambda W$$

blank

Example

- Atlanta is the place where two interstate highways, I-75 and I-85, merge and cross each other.
  - As a traffic manager of Atlanta, you would like to estimate the average time it takes to drive from the north confluence point to the south confluence point.
  - On average, 100 cars per minute enter the merged area from I-75 and 200 cars per minute enter the same area from I-85.
  - You also dispatched a chopper to take a aerial snapshot of the merged area and counted how many cars are in the area.
  - It turned out that on average 3000 cars are within the merged area.
  - What is the average time between entering and exiting the area per vehicle?

blank

Summary

- Arrival rate λ
- Service rate μ
- Traffic intensity ρ
- Utilization level $u = \min(\rho, 1)$

- Throughput (TH): Rate of outflow.
  (e.g. Production level of manufacturing line)

Illustration - 1 server

- ρ=λ/μ
- Throughput $= \lambda$ if $\rho \leq 1$
- Throughput $= \mu$ if $\rho > 1$
- Or, Throughput $= min(\lambda, \mu)$

Illustration - 2 servers in parallel

- ρ=
- Throughput $= \lambda$ if $\rho \leq 1$
- Throughput $= \mu_1 + \mu_2$ if $\rho > 1$
- Or, Throughput $= min(\lambda, \mu_1 + \mu_2)$

Illustration - 2 servers in sequence

- ρ=
- Throughput $= \lambda$ if $\rho \leq 1$
- Throughput $= min(\mu_1, \mu_2)$ if $\rho > 1$
- Or, Throughput $= min(\lambda, min(\mu_1, \mu_2)) = min(\lambda, \mu_1, \mu_2)$

Illustration - n servers in sequence

- ρ=
- Throughput $= min(\lambda, \mu_1, \mu_2, ..., \mu_n)$
- Most slowest server screws up all system!
- Most slowest server is most busy server.
- Revisit Kingman's equation.
- "Bottleneck" station.

blank

Exercise 1

- A small call center has three phone lines that are answered by two operators, John and Paul. That is, the call center can hold up to three calls (two in service and one in hold) and when the call center already has three calls, new calls will hear busy signals and have to call again to be connected. The call processing times by John are iid (identical independent distribution), having exponential distribution with mean 2 minutes. The call processing times by Paul are iid, having exponential distribution with mean 4 minutes. An arriving call to an empty system is always processed by John. Suppose that three calls (A, B, and C) arrive at an empty center at 8am with John taking call A, Paul taking call B, and call C waiting.
  - *(a)* What is the probability that call A is still in service at 8:10am?
  - *(b)* What is the probability that all three calls still in the system at 8:04am?
  - *(c)* What is the probability that call C leaves the system before call B?
  - *(d)* How likely is call A to finish before call B?
  - *(e)* What is the probability that call B will be the last one among them to leave the call center?
  - *(f)* What is the probability that call C will be the last one among them to leave the call center?

(Solution)

Before we begin, let's define some notation let X_A, X_B, X_C be the times from 8AM to completion for each of the calls. Note that $X_A \sim exp(\lambda_J)$, also $X_B \sim exp(\lambda_P)$, where $\lambda_J = 1/2$ and $\lambda_P = 1/4$. And let Y_C be the service time of call C, note at this point we don't know its distribution.

(a) This is the following probability:

$$\mathbb{P}\{X_A > 10\} = 1 - F_{X_A}(10) = 1 - (1 - e^{-\frac{1}{2}10}) = e^{-5}$$

(b) This is the probability that neither A or B have hung up at this time, recall the two calls are independent, so, this is

$$\mathbb{P}\{X_A > 4, X_B > 4\} = \mathbb{P}\{X_A > 4\}\,\mathbb{P}\{X_B > 4\} = (1-F_{X_A}(4))(1-F_{X_B}(4))$$

(d) (Note, we are solving this before (c)) Here we are looking at the probability that the minimum of two independent exponentials (X_A and X_B) is X_A, from class we have that

$$\mathbb{P}\{X_A < X_B\} = \frac{\lambda_J}{\lambda_J + \lambda_P} = \frac{1/2}{1/2 + 1/4} = \frac{2}{3}$$

(c) This is the probability that A completes service at John's phone, then C completes service at John's phone, before Paul is done with B, in this case we know that $Y_C \sim exp(\lambda_J)$ and the probability we want is

$$\begin{aligned}
\mathbb{P}\left(X_A < X_B, X_C < X_B\right) &= \mathbb{P}\left(Y_C < X_B | X_A < X_B\right) \mathbb{P}\left(X_A < X_B\right) \\
&= \left(\frac{\lambda_J}{\lambda_J + \lambda_P}\right)\left(\frac{\lambda_J}{\lambda_J + \lambda_P}\right) \\
&= \left(\frac{1/2}{1/2 + 1/4}\right)\left(\frac{1/2}{1/2 + 1/4}\right) \\
&= \frac{4}{9}
\end{aligned}$$

(e) Note that this the exact same event as part (c).

(f) Here we have two possible orderings ABC and BAC, we use total probabilities.

$$\begin{aligned}
\mathbb{P}\{X_C > X_B, X_C > X_A\} &= \mathbb{P}\{X_C > X_B > X_A\} + \mathbb{P}\{X_C > X_A > X_B\} \\
&= \mathbb{P}(X_A < X_B, X_B < X_C) + \mathbb{P}(X_B < X_A, X_A < X \\
&= \mathbb{P}\{X_A < X_B\}\,\mathbb{P}\{X_B < X_C | X_A < X_B\} \\
&+ \mathbb{P}\{X_B < X_A\}\,\mathbb{P}\{X_A < X_C | X_B < X_A\} \\
&= \mathbb{P}(X_A < X_B)\mathbb{P}(X_B < Y_C | Y_C \sim exp(\lambda_J)) \\
&+ \mathbb{P}(X_B < X_A)\mathbb{P}(X_A < Y_C | Y_C \sim exp(\lambda_P)) \\
&= (2/3)(1/3) + (1/3)(2/3) \\
&= \frac{4}{9}
\end{aligned}$$

Exercise 2

- A small bank is staffed by a single server. It has been observed that, during a normal business day, the inter-arrival times of customers to the bank are iid having exponential distribution with mean 3 minutes. Also, the the processing times of customers are iid having the following distribution (in minutes): $\mathbb{P}(X = 1) = 1/4; \mathbb{P}(X = 2) = 1/2; \mathbb{P}(X = 3) = 1/4$; An arrival finding the server busy joins the queue. The waiting space is infinite.
  - *(a)* What is long-run fraction of time that the server is busy?
  - *(b)* What is long-run average waiting time of each customer in the queue?
  - *(c)* What is long-run average service time of each customer?
  - *(d)* What is long-run average time that each customer spend in the system?
  - *(e)* What is long-run average number of customers in the bank?
  - *(f)* What is long-run average number of customers in the queue?
  - *(g)* What is long-run average number of customers in the middle of service?
  - *(h)* Are your answer for (a) and (g) same? Should it be same? Discuss shortly why or why not.
  - *(i)* What is TH(Throughput) of this system?

(Solution)

(a)

$$\mathbb{E}[X] = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} = 2$$

If the system is stable the long run average time the system is busy is ρ, so calculating,

$$\rho = \frac{1/3}{1/2} = \frac{2}{3}$$

so the long run average time the server is busy is 2/3.

(b) Since arrivals are exponential, so $c_a^2 = 1$. For service side,

$$\mathbb{E}[X] = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} + 3 \cdot \frac{1}{4} = 2$$

$$\mathbb{E}[X^2] = 1 \cdot \frac{1}{4} + 2^2 \cdot \frac{1}{2} + 3^2 \cdot \frac{1}{4} = 4.5$$

$$Var[X] = 4.5 - 4 = 0.5$$

$$c_s^2 = \frac{Var[X]}{\mathbb{E}[X]^2} = \frac{1/2}{2^2} = \frac{1}{8}$$

Now we are ready to apply Kingman's formular,

$$\mathbb{E}[W_q] \approx \mathbb{E}[V] \left(\frac{\rho}{1-\rho}\right) \left(\frac{c_a^2 + c_s^2}{2}\right)$$

$$= 2 \left(\frac{2/3}{1/3}\right) \left(\frac{1+1/8}{2}\right) = \frac{9}{4}\ minutes$$

(c) $E[X] = 2$ from (a)

(d) $W_{sys} = W_q + W_{svc} = (ans\ of\ (b)) + (ans\ of\ (b)) = 9/4 + 2 = 17/4$

(e) Using Little's formula, $L_{sys} = \lambda W_{sys} = 1/3 \cdot 17/4 = 17/12$

(f) Using Little's formula, $L_q = \lambda W_q = 1/3 \cdot 9/4 = 9/12$

(g) Using Little's formula, $L_{svc} = \lambda W_{svc} = 1/3 \cdot 2 = 2/3$

(h) The single server is busy if and only if there is a customer in the server. And there can be only one customer at most in the service. Thus,
$\big($the fraction of time server is busy$\big) = \big($expected number of customer in the service$\big)$

(i) Since the system is stable, the throughput is determined by the arrival rate.

$$Thoughput = \frac{1}{3}$$

Exercise 3

- We want to decide whether to employ a human operator or buy a machine to paint steel beams with a rust inhibitor. Steel beams are produced at a constant rate of one every 14 minutes. A skilled human operator takes an average time of 700 seconds to paint a steel beam, with a standard deviation of 300 seconds. An automatic painter takes on average 40 seconds more than the human painter to paint a beam, but with a standard deviation of only 150 seconds.
  - *(a)* What is the arriving rate to either a human operator or the machine painter?
  - *(b)* Estimate the expected waiting time in queue of a steel beam for each of the operators.
  - *(c)* The expected numbers of steel beams waiting in queue in each of the two cases.
  - *(d)* Which system is better? Why?

(Solution)

(a) The arriving rate to either a human operator or the machine painter is 1/14 beams per minutes or 1/840 beams per second.

(b) Note that for human operator, $\rho = \frac{700}{840} = \frac{5}{6}$. Use Kingman's approximation, For human operator, expect waiting time is

$$\begin{aligned} E[W_{human}] &= E[V](\frac{\rho}{1-\rho})(\frac{c_a^2 + c_s^2}{2}) \\ &= 700 * (\frac{5/6}{1-5/6})(\frac{0/840^2 + 300^2/700^2}{2}) \\ &= 321.43 \text{ seconds} \end{aligned}$$

For machine operator, $\rho = \frac{740}{840} = \frac{37}{42}$ Expect waiting time is

$$\begin{aligned} E[W_{machine}] &= E[V](\frac{\rho}{1-\rho})(\frac{c_a^2 + c_s^2}{2}) \\ &= 740 * (\frac{37/42}{\quad})(\frac{0/840^2 + 150^2/740^2}{\quad}) \end{aligned}$$

(c) The expect number of steel beams waiting in queue for human operator is $L_{human} = \lambda W_{human} = \frac{1}{840} \cdot 321.43 = 0.383$. And the expect number of steel beams waiting in queue for machine operator is $L_{machine} = \lambda W_{machine} = \frac{1}{840} \cdot 112.5 = 0.134$

(d) Machine is better, because it has less queue time and less queue size in the long run. Machine is slower on average, but it has less variation, hence better in overall performance.

Exercise 4

- Suppose you are modeling the Seoultech's undergraduate population. Assume that the undergraduate population is in steady-state.
  - *(a)* Every year 2500 students come to Seoultech as first year students. It takes 4.5 years on average for each student to graduate. How many undergraduate students does Seoultech have on average at a certain point of time?
  - *(b)* In addition to the first year students, an unknown number of students come to Seoultech as transferred students every year. It takes 2.5 years on average for a transferred student to graduate. Suppose you know that the average size of Seoultech undergraduate population at a certain point of time is 13250 students. How many students come to Seoultech as transferred students every year?

(Solution)

(a) Using Little's Law, $\mathbb{E}[L_{gt}] = \lambda_{gt}\mathbb{E}[W_{gt}] = 2500 \times 4.5 = 11250$ student.

(b) We know that the total number of students, L, is defined as $E[L] = E[L_{gt}] + E[L_{tf}] = 13250$ so we can once again use Little's Law within the previous formula, $11250 + \lambda_{tf}\mathbb{E}[W_{tf}] = 13250$. Hence, $\lambda_{tf} = 2000/2.5 = 800$ students per year.

Exercise 5

- An auto collision shop has roughly 10 cars arriving per week for repairs. A car waits outside until it is brought inside for bumping. After bumping, the car is painted. On the average, there are 15 cars waiting outside in the yard to be repaired, 10 cars inside in the bump area, and 5 cars inside in the painting area. What is the average length of time a car is in the yard, in the bump area, and in the painting area? What is the average length of time from when a car arrives until it leaves?

(Solution)

Since none of the areas are over-flooded by cars in the long run, we assume system is stable. we use Little's Law to solve this problem

$$\begin{aligned} W_{yard} &= L_{yard}/\lambda = 15/10 = 1.5\ weeks \\ W_{bump} &= L_{bump}/\lambda = 10/10 = 1\ weeks \\ W_{paint} &= L_{paint}/\lambda = 5/10 = 0.5\ weeks \end{aligned}$$

On average, it takes 1.5+1+0.5=3 weeks from when a car arrives until it leaves.

```
"Faber est suae quisque fortunae."
```