ECE560: Computer Systems Performance Evaluation

Lecture #11-
Queueing Systems (I)

Instructor: Dr. Liudong Xing

Administration Issues

- Homework #4 assigned
  - Due: **March 4, Monday**
- Annotated Bibliography
  - Due: **March 22, Friday**
  - Refer to Section 2.2 in the Project Description for the guidelines
- Midterm Exam on **March 6, Wednesday**
  - Review session on March 4, Monday

2.2 Annotated Bibliography

- The annotated bibliography is a list of papers that are relevant to your project. For each paper, you must give the complete citation, which includes …. In addition, you must write a 30-70 word summary for each paper describing its contents and how it is relevant to your project. This summary must not be a simple repetition of the paper's abstract. The goal of this annotated bibliography is to show that you have adequately researched the previous peer-reviewed work that has been done in the area of your proposed project.

Review of Lecture#10

- Discrete-time Markov chains
  - One-step, n-step transition probabilities (matrix); homogeneous
  - $\Pi(n) = \Pi(0)\ P^n$
- **Ergodic**
  - irreducible: you can get from every state to every other
  - aperiodic: every state has period 1. For each state there are paths back to that state of various lengths
  - for which all states are positive recurrent: for each state, upon leaving the state you will return with probability 1 and within a finite mean time.
  - Stationary probability distribution = Long-run (limiting) probability distribution
  - $\Pi = \Pi^*P$ and $\sum_i \Pi_i = 1$
  - Balance equations: Rate entering = Rate leaving

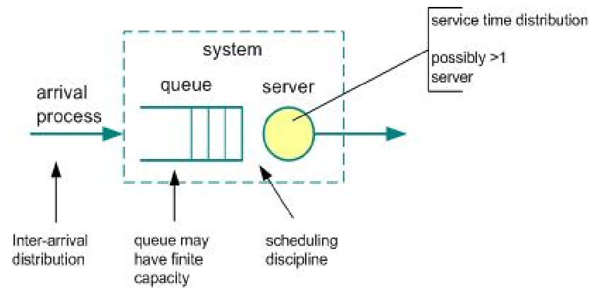
Topics

- Overview of queueing systems
- Performance measures
- D/D/1 queueing systems
- M/M/1 queueing systems

Related reading:
Allen's Ch. 5.0 ~ 5.2

Introduction to Queueing Systems

- What is a queueing system?



- Arrivals to an empty queue get immediate service
- Arrivals to a busy system are held in the queue until server is free
- Arrival rate is typically drawn from a r.v. distribution, e.g. Poisson with a rate λ
- Service rate is computed using the rate of processing for the device and is in units per time and is often denoted by μ

Applications of Queueing Systems

- Supermarket checkout line
- Bank teller line
- Batch jobs waiting on a CPU
- Traffic lights
- Operating systems task scheduling
- Planes to take off or land
- Interactive inquiry system
- Airline reservation system

Queueing Systems

- Performance evaluation with queueing systems involve two steps:
  1. Modeling process
  2. Mathematical solution of the model

Kendall Notation

Standard notation for queueing systems:

A/B/c/K/m/Z

- **A**: arrival process or inter-arrival time distribution
  - 'M' = Poisson arrival process
  - 'D' = Deterministic (constant) arrival rate
  - 'G' = General arrival process
- **B**: service process or service time dist.
  - 'M' = Exponential service time dist.
  - 'D' = Deterministic (constant) service time
  - 'G' = General service time
- **c**: number of servers
- **K**: the capacity of the system (queue+server(s)) (default: ∞)
- **m**: total job/customer population (default: ∞)
- **Z**: scheduling discipline (default: FIFO)

Examples

- D/D/1 queue:
  - Single server FIFO queue
  - No capacity/population restriction
  - Constant inter-arrival time
  - Constant service time
- M/M/1 queue:
  - Single server FIFO queue
  - No capacity/population restriction
  - Poisson arrivals
  - Exponential service time
- M/G/ ∞ queue:
  - Infinite server queue
  - Poisson arrivals
  - General service time

  Why "M"?

  "M" means that the process has the "Markov property", i.e., the process is "memory-less".

Notation

- **λ**: average arrival rate of new jobs
- **E[τ]**: average inter-arrival time (=1/λ)
- **μ** : average service rate
- **W_s**: average service time (=1/μ)
- **W_q**: average time a job spends in the queue (= average waiting time)
- **W**: average time a job spends in the system (= average system time/response time/sojourn time)
- **L_q**: average number of jobs in the queue (= average queue length)
- **L**: average number of jobs in the system
- **c**: number of identical servers

- **More**: see Table 5.1.1 on P*252*

Performance Measures of Queueing Systems

Performance Measures of Queueing Systems (I)

- Average number of jobs/customers in the system (L)
- Average time spent in the system (W: average response time)
- Average number of jobs in the queue (L_q)
- Average time spent in the queue (W_q: average waiting time)

Little's Law/Formula/Theorem

$$L = \lambda W$$
$$L_q = \lambda W_q$$

holding for all queueing systems

Little's Law

$$L = \lambda W$$
$$L_q = \lambda W_q$$

- Rigorous proof: Ref. [42] by Little, Ref. [56] by Stidham
- Intuition:
  - Pick a "typical customer"
  - When the customer arrives to the queueing system, the customer should find L customers waiting
  - When the customer leaves the system, the customer has been in the system for W units of time
  - Implying λW customers should have arrived while the customer was in the system
  - In the steady state, the number of customer left behind on departure should equal the number found on arrival, i.e., $\lambda W = L$.

Performance Measures of Queueing Systems (II)

- $p_n(t)$: probability that there are n customers in the system at time t
- π_n: steady-state probability that there are n customers in the queueing system
- Throughput (γ): rate at which jobs successfully depart from the system
- Blocking probability (P_B, for the finite buffer/queue size): probability an arriving job is turned away due to a full buffer

Performance Measures of Queueing Systems (III)

- Traffic intensity/offered load (α):

$$\alpha = W_s / E[\tau]$$

  - W_s: average service time per server
  - $E[\tau]$: average inter-arrival time for all customers/jobs entering the system and not just for the customers serviced by a particular server, unless there is only one server
  - A measure of the required number of servers
- Server utilization (ρ): $\rho = \alpha/c$
  - Represents average fraction of the time that each server is busy assuming traffic is evenly distributed to each server
  - Probability that a given server is busy as observed by an outsider observer
  - A measure of congestion

An Example

- Consider a D/D/1 queueing system with
  - A constant inter-arrival time of 20 seconds
  - A constant service time of 10 second

Then: the server is busy half of the time

$$\rho = \alpha = 10/20 = 0.5$$

If the server is replaced by one with a constant service time of 15 seconds, then it is busy three-fourth of the time

$$\rho = \alpha = 15/20 = 0.75$$

If the server is replaced by one with a constant service time of 30 seconds, then the server must provide 30 seconds of service every 20 seconds, impossible! Two servers must be provided to keep up!

$$\rho = \alpha = 30/20 = 1.5$$

Important Queueing Systems

- D/D/1 queues

- M/M/1 queues
- M/M/1/N queues
- M/M/c queues
- M/M/∞ queues
- M/M/1/k/k queues

Birth-death queueing systems

- M/G/1 queues
- M/D/1 queues
- GI/M/1 queues
- GI/M/c queues

Embedded Markov chain queueing systems

D/D/1 Queues

- A deterministic (non-random) queue has
  - Deterministic arrival rate λ
    - Constant inter-arrival time $1/\lambda$
  - Deterministic service rate μ
    - Constant service time $1/\mu$
  - 1 serve
  - Infinite length buffer
  - $\alpha=\rho=\lambda/\mu$

  If arrival rate is less than service rate, then there is no waiting in the queue

  ($\rho<1$: probability of server being busy)

  If arrival rate is greater than service rate, the queue will move towards having an infinite waiting time

  ($\rho>1$: infinite queue length $\rightarrow \infty$)

  (finite queue will be overflowed)

Agenda

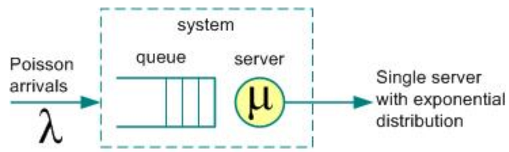
- Overview of queueing systems
- Performance measures
- D/D/1 queueing systems
- **M/M/1 queueing systems**
  - The most basic and important queueing model!

Related reading:

Allen's Ch. 5.0 ~ 5.2

M/M/1 Queues

- An M/M/1 queue has
  - Poisson arrivals with a rate λ
  - Exponential service times with a mean of $1/\mu$, so μ is the average service rate
  - 1 server
  - An infinite length buffer/queue



- Fits the birth-and-death process
  - A birth is a customer arrival
  - A death occurs when a customer leaves the system after completing service

M/M/1: Poisson Arrival Process

- Let N(t) denote the number of arrivals in interval (0,t). Then,

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

- Let τ denote the time between two Poisson arrivals. Then,

$$\Pr[\tau \le t]=1-e^{-\lambda t}$$

- The rate λ is the average number of arrivals per unit of time, and $1/\lambda$ is the average inter-arrival time
- For 2 disjoint intervals (t1, t2) and (t3, t4). The number of arrivals in (t1, t2) is independent of the number of arrivals in (t3, t4) – independent increments!
- Examples:
  - Customers arriving to a bank
  - Packets arriving to a buffer
  - Transactions arriving at a server
  - Read/write requests to a disk controller

M/M/1: Exponential Service Time Distribution

- Let X denote the service time of a job. If X is exponentially distributed with average service time 1/μ. Then,
  - In a small time interval Δt, the probability that a service completion will occur is proportional to the size of the interval:

    $$\Pr[1 \text{ completion in } \Delta\mathrm{t}] = \mu\Delta t + o(\Delta t)$$

  - In Δt, the probability of more than 1 service completion is negligible:

    $$\Pr[>1 \text{ completion in } \Delta\mathrm{t}] = o(\Delta t)$$

  - Service completions are independent of other service completions and also independent of the service completion time since the last service completion (independent/stationary increments)

$$\boxed{\Pr[X \le t] = 1 - e^{-\mu t}}$$

Performance Evaluation of Queueing Systems

Performance Measures of Interest

- Traffic intensity (α)
- Server utilization (ρ)
- π_n: steady-state probability that there are *n* customers in the queueing system
- Throughput (γ): rate at which jobs successfully depart from the system
- Average number of jobs in the system (L)
- Average time in the system (W)
- Average number of jobs in the queue (L_q)
- Average time in the queue (W_q)

Performance Measures of M/M/1

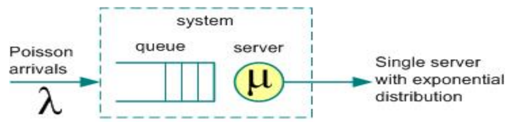
- Traffic intensity/offered load (α):

$$\alpha = W_s / E[\tau]$$
$$= (1/\mu)/(1/\lambda) = \lambda/\mu$$

- Server utilization (ρ):

$$\rho = \alpha/c = \alpha = \lambda/\mu$$

Performance Measures of M/M/1 (Cont'd)



- $p_n(t)$:= probability that the system has n customers at time t
- π_n := steady state probability that there are n customers in the system
- By similar reasoning for birth-and-death process, the differential-difference equation which describe the state of the queue as a function of time:

$$\frac{d}{dt} p_n(t) = -(\lambda + \mu) p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t)$$
$$\frac{d}{dt} p_0(t) = -\lambda p_0(t) + \mu p_1(t)$$

- If we are interested in the steady state behavior, we set

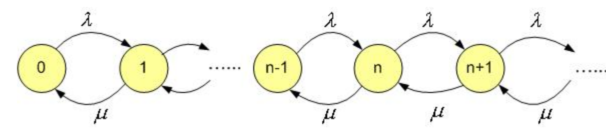
$$\frac{d}{dt} p_n(t) = 0 \quad \text{and} \quad \lim_{t\to\infty} p_n(t) = \pi_n \quad \forall n$$

Then, the steady-state equations:

$$0 = -(\lambda + \mu)\pi_n + \lambda\pi_{n-1} + \mu\pi_{n+1}$$
$$0 = -\lambda\pi_0 + \mu\pi_1$$

Derivation of M/M/1 Queue (II)

- A different way to obtain steady-state probabilities is to look at the state-transition diagram



- In a steady state, the average rate at which the system enters a state must be equal to the average rate at which it leaves the state
- Then, we obtain Balance Equations:

State	Rate out=Rate in
0	$\lambda\pi_0 = \mu\pi_1$
1	$(\lambda+\mu)\pi_1 = \lambda\pi_0 + \mu\pi_2$
2	$(\lambda+\mu)\pi_2 = \lambda\pi_1 + \mu\pi_3$
…	…………
n	$(\lambda+\mu)\pi_n = \lambda\pi_{n-1} + \mu\pi_{n+1}$

M/M/1: Solution to Steady-State Probabilities

- By adding each two consecutive equations:

$$\lambda\pi_0 = \mu\pi_1 \qquad \pi_1 = \frac{\lambda}{\mu}\pi_0$$

$$\lambda\pi_1 = \mu\pi_2 \qquad \text{i.e.,} \qquad \pi_2 = \frac{\lambda}{\mu}\pi_1$$

....

$$\lambda\pi_n = \mu\pi_{n+1} \qquad \pi_{n+1} = \frac{\lambda}{\mu}\pi_n$$

- Thus:

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0$$

- All probabilities have to sum up to one:

$$\sum_{n=0}^{\infty} \pi_n = 1$$

- Therefore:

$$\pi_0 \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = 1 \Rightarrow \pi_0 = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n} = 1 - \frac{\lambda}{\mu}$$

Performance Measures of the M/M/1 Queues (Cont'd)

- Server utilization (ρ): $\rho = \alpha/c = \lambda/\mu$
- Steady-state probability that the system has n customers (π_n):

$$\pi_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho,$$

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0 = \rho^n \pi_0 = \rho^n (1-\rho)$$

It is a Geometric distribution!

Review (L#6)

- Geometric r.v.: is a r.v. that counts the number of independent Bernoulli trials until the first success is encountered.
  - P.m.f:

$$P\{X=0\}=p$$
$$P\{X=1\}=qp$$
$$P\{X=2\}=q^2 p$$
......

In general, for $k=0,1,2,...$
$$P\{X=k\}=q^k p$$

Performance Measures of Interest

- ✓ Traffic intensity (α)
- ✓ Server utilization (ρ)

$$\alpha=\rho=\lambda/\mu$$

- ✓ π_n: steady-state probability that there are *n* customers in the queueing system

$$\pi_0=1-\frac{\lambda}{\mu}=1-\rho,$$
$$\pi_n=\left(\frac{\lambda}{\mu}\right)^n\pi_0=\rho^n(1-\rho)$$

- Throughput (γ): rate at which jobs successfully depart from the system
- Average number of jobs in the system (L)
- Average time in the system (W)
- Average number of jobs in the queue (L_q)
- Average time in the queue (W_q)

Performance Measures of the M/M/1 Queues (Cont'd)

- Throughput γ = rate at which jobs depart from the system

$$\begin{aligned}\gamma &= \mu P[>0 \text{ jobs in the system}] \\ &= \mu(1 - P[0 \text{ jobs in the system}]) \\ &= \mu(1-\pi_0) = \mu(1-(1-\rho)) \\ &= \mu\rho = \lambda\end{aligned}$$

Performance Measures of the M/M/1 Queues (Cont'd)

- Average number of jobs in the system (L):

$$L = \sum_{n=0}^{\infty} n\pi_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n$$

$$= (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1} = \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}$$

- Average time in the system (W):

With Little's Law:

$$W = L/\lambda = \frac{\rho}{1-\rho}/\lambda = \frac{1}{\mu-\lambda}$$

Performance Measures of the M/M/1 Queues (Cont'd)

- Average number of jobs in the queue (L_q):

$$L_q = L - (1 * P[\text{Server is not empty}]$$
$$= L - (1 - P[0 \text{ jobs in the system}])$$
$$= L - (1 - \pi_0) = L - (1 - (1 - \rho))$$
$$= L - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}$$

- Average time in the queue (W_q):

With Little's Law:

$$W_q = L_q / \lambda = \frac{\rho^2}{1-\rho}\frac{1}{\lambda}$$

or

$$W_q = W - W_s = \frac{\rho}{(1-\rho)\lambda} - \frac{1}{\mu} = \frac{\rho^2}{1-\rho}\frac{1}{\lambda}$$

Summary (M/M/1)

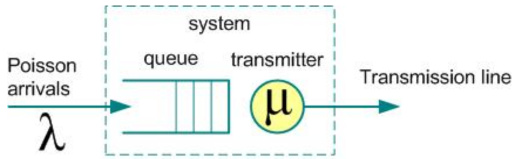
- Performance measures:

$$\alpha = \rho = \lambda / \mu$$
$$\pi_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho, \ \pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0 = \rho^n (1-\rho)$$
$$\gamma = \mu P[> 0 \text{ jobs in the system}]$$
$$= \mu(1 - P[0 \text{ jobs in the system}])$$
$$= \mu(1 - \pi_0) = \mu(1 - (1 - \rho)) = \mu\rho = \lambda$$

$$L = \sum_{n=0}^{\infty} n\pi_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n$$
$$= (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1} = \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}$$

$$W = L/\lambda = \frac{\rho}{1-\rho}/\lambda = \frac{1}{\mu - \lambda}$$
$$L_q = L - (1 * P[\text{Server is not empty}]$$
$$= L - (1 - P[0 \text{ jobs in the system}])$$
$$= L - (1 - \pi_0) = L - (1 - (1 - \rho))$$
$$= L - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}$$
$$W_q = L_q / \lambda = \frac{\rho^2}{1-\rho}\frac{1}{\lambda} \text{ or } W_q = W - W_s = \frac{\rho}{(1-\rho)\lambda} - \frac{1}{\mu} = \frac{\rho^2}{1-\rho}\frac{1}{\lambda}$$

Example: Applying the M/M/1 Results to a Single Network Link



- Poisson packet arrivals with rate $\lambda = 2000$ packets/sec
- Link capacity C=1.545 MB/sec
- Approximate the packet length distribution by an exponential with mean L=515 B

- What is the mean service time W_s? The transmitter utilization ρ? Average number of packets in the system L? Average time spent in the system W?

A Characteristic of M/M/1 System

- Calculate W/W_s:

$$\frac{W}{W_s} = \frac{\text{average time to complete service}}{\text{average service time}}$$
$$= \frac{\rho/(1-\rho)/\lambda}{1/\mu} = \frac{1}{1-\rho}$$

- Graph of W/W_s *versus* ρ:
  - Figure 5.2.2 in Textbook

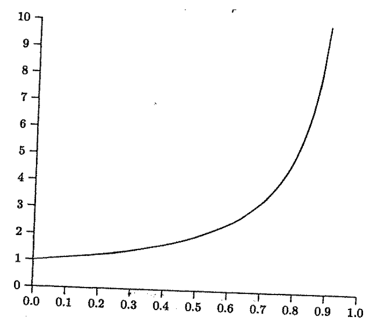


Figure 5.2.2. $\frac{W}{W_s}$ versus ρ for M/M/1 queueing system.

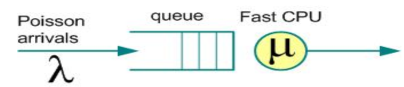
A Characteristic of M/M/1 System (Cont'd)

- W/W_s is a measure of response time
  - the smaller W/W_s , the better response time
- The response time is very sensitive to minor changes as server utilization $\rho \rightarrow 1$
- High utilization and good response time are incompatible goals
- A stretch factor of 5 is often considered the limit of acceptable performance

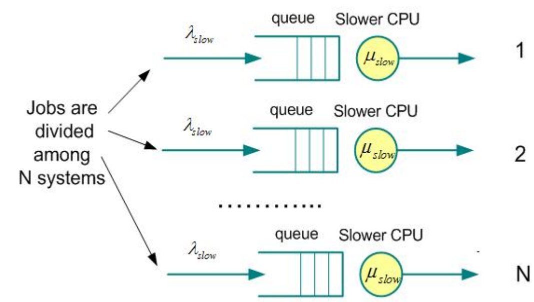
$$\boxed{\frac{W}{W_s} < 5 \Rightarrow \frac{1}{1-\rho} < 5 \Rightarrow \rho < 0.8}$$

Example

- Have one fast computer



- Proposal: divide workload among N slower machines:



$$\lambda_{slow} = \frac{\lambda}{N}, \quad \mu_{slow} = \frac{\mu}{N}$$

Q1: Is the proposed system an improvement? Why or why not?

Solution to Q1

- For N-Slower Machine System:

$$W_s(slow) = \frac{1}{\mu_{slow}} = \frac{1}{\mu/N} = \frac{N}{\mu} = NW_s(fast)$$

$$\frac{W}{W_s} = \frac{1}{1-\rho} \Rightarrow W = \frac{W_s}{1-\rho}$$

$$\therefore W(slow) = \frac{W_s(slow)}{1-\lambda_{slow}/\mu_{slow}} = \frac{N/\mu}{1-\lambda/\mu} = NW(fast)$$

Average response time will INCREASE *N* fold, even though the *N*-Slower CPUs together process the same number of jobs per unit of time as before.

Q2: How fast would the slower machine need to be in order to give customers the SAME average response time W?

Solution to Q2

$$\frac{W}{W_s} = \frac{1}{1-\rho} \Rightarrow W = \frac{W_s}{1-\rho}$$

$$\therefore W(slow) = \frac{W_s(slow)}{1-\lambda_{slow}/\mu_{slow}} = \frac{1/\mu_{slow}}{1-\lambda/N\mu_{slow}}$$

$$W(fast) = \frac{W_s(fast)}{1-\lambda/\mu} = \frac{1/\mu}{1-\lambda/\mu}$$

For equal response time:

$$W(slow) = W(fast)$$

$$\frac{1/\mu_{slow}}{1-\lambda/N\mu_{slow}} = \frac{1/\mu}{1-\lambda/\mu}$$

$$\frac{\mu_{slow}}{\mu} = \frac{1-\lambda/\mu}{1-\lambda/N\mu_{slow}}$$

$$\mu_{slow}(1-\lambda/N\mu_{slow}) = \mu(1-\lambda/\mu)$$

$$\mu_{slow} - \lambda/N = \mu - \lambda$$

$$\mu_{slow} = \mu - \lambda(1-1/N)$$

$$\boxed{\mu_{slow}/\mu = 1 - \frac{\lambda}{\mu}(1-1/N)}$$

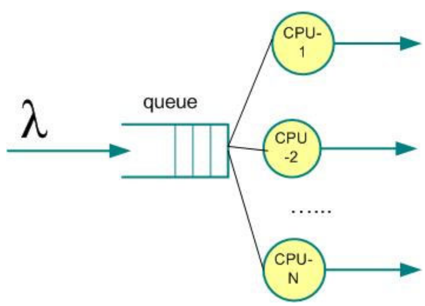
Hands-On Problem

- Assume current system has a utilization of $\rho=0.8$ and it is to be replaced with $N=10$ slower processors. How fast would the slower processors need to be in order to give the SAME average response time W as the original system? How about when ρ=0.5?

Q3: Is there any other multiprocessor architecture that is superior?

Solution to Q3

- YES, as we will see later (M/M/c)



Next Topics

- Birth-and-death queueing systems (Cont'd)
  - M/M/1/N, M/M/c, M/M/∞, M/M/1/k/k Queues

Things to Do

- Read Allen's Ch. 5.0 ~ 5.2
- Prepare for midterm exam
- Annotated Bibliography (refer to Section 2.3 in project description)
  - due March 22 (Friday)