Queueing Theory

CS 450: Operating Systems
Michael Lee <lee@iit.edu>
Agenda

- What is it?
- Probability refresher
  - Probability distributions and stochastic processes
- Queueing theory
  - Basic model
  - Little’s Law
  - M/M/1 queueing system
Queueing Theory?
Thinking about scheduling

- The design of a scheduler can be considered from different angles:
  1. As a practical set of policies driven by heuristics and experimentation
     - e.g., tuning the rules and “magic numbers” used by a MLFQ scheduler based on perceived system responsiveness and empirical data
  2. As a theoretical exercise in mathematical modeling and analysis
     - Helps to ensure rigor in our calculations, and to provide a more solid foundation for reasoning about policies and desired outcomes
Queueing theory

- The mathematical study of wait queues
  - e.g., using probability distributions to describe job behavior and stochastic processes to model queueing systems
- Important: rigor does not guarantee correctness!
  - Models are only as good as the assumptions they’re based on
    - e.g., if we assume constant-length (deterministic) jobs, but jobs are exponentially distributed, our results won’t reflect reality
Applications of queueing theory

- Emergency services
- Project management
- Telecommunications and Networking
- Logistics and Transportation
- OS Scheduling
- Etc.
Tip of the iceberg

- Queueing theory was “invented” by Agner Erlang in 1909 in a paper featuring a proof concerning telephone traffic
- 100+ years of development, with extant open problems
- In depth coverage in CS 555: *Analytic Models and Simulation of Computer Systems*
Probability refresher
Probability theory

- Mathematical analysis of experiments with random outcomes

- Given the set of all possible outcomes $\Omega$ (the sample space), assign to each outcome $\omega \in \Omega$ a probability $P(\omega) \in [0, 1]$ reflecting its likelihood

- The probabilities of all outcomes sum to 1: $\sum_{\omega \in \Omega} P(\omega) = 1$

- An event $E$ is a subset of $\Omega$, with probability $P(E) = \sum_{\omega \in E} P(\omega)$
Random variable

- A random variable is a function that maps the sample space onto numeric values; e.g., $X: \Omega \to \mathbb{N}$

  - The event $E$ where $X = n$ is the set $\{\omega \in \Omega \mid X(\omega) = n\}$

  - The probability of this event is $P(X = n) = p(n) = \sum_{\omega \in E} P(\omega)$

- Discrete r.v.s map events onto a countable set (e.g., $\mathbb{N}$, $\mathbb{Z}$)

- Continuous r.v.s map events onto an uncountable set (e.g., $\mathbb{R}$)
Discrete vs. Continuous

- The function $P$ for a discrete r.v. $X$, called its *probability mass function*, can be evaluated for distinct values $n \in \text{range}(X)$; e.g., $P(X = n)$.

- The function $P$ for a continuous r.v. $X$ can not be evaluated for distinct values, and so we define $f$, its *probability density function* (PDF), where:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

- For both discrete and continuous R.V.s, we can define a *cumulative distribution function* (CDF) $F$, where:

$$F(n) = P(X \leq n) = \sum_{x \leq n} P(X = x) \quad \text{or} \quad \int_{-\infty}^{n} f(x)dx$$
E.g., triple coin toss

$X_1(\omega) = \# \text{ of tails in } \omega$

$X_2(\omega) = \begin{cases} 
1, & \text{if } (\#\text{ tails in } \omega) \geq 2, \\
0, & \text{otherwise}
\end{cases}$
E.g., triple coin toss

\[ X_1(\omega) = \# \text{ of tails in } \omega \]

Sample space \((\Omega)\)

\[
P(X_1 = 0) = \frac{1}{8}
\]

\[
P(X_1 = 1) = \frac{3}{8}
\]

\[
P(X_1 = 2) = \frac{3}{8}
\]

\[
P(X_1 = 3) = \frac{1}{8}
\]

\[
F(2) = P(X_1 \leq 2) = \sum_{x \leq 2} p(x) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}
\]
Statistics of discrete R.V.s

- Expected value (mean): \[ E(X) = \sum_{x \in \mathcal{R}(X)} x \cdot p(x) \quad \text{(discrete } X) \]
  \[ = \int_{-\infty}^{\infty} x \cdot f(x) \, dx \quad \text{(continuous } X) \]

- Variance: \[ \sigma^2 = E((X - E(X))^2) = E(X^2) - E(X)^2 \]

- Standard deviation: \[ \sigma = \sqrt{\sigma^2} = \sqrt{E((X - E(X))^2)} \]
Multiplication & Addition rules

- For any two independent events

  - Multiplication rule: \( P(A \text{ and } B) = P(A) \cdot P(B) \)
    
    - e.g., probability of rolling “snake-eyes” with two 6-sided dice:
      \[
      P(X = 1) \cdot P(X = 1) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}
      \]
    
    - Addition rule: \( P(A \text{ or } B) = P(A) + P(B) \)
      
      - e.g., probability of rolling two or four with a 6-sided dice:
        \[
        P(X = 2) + P(X = 4) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}
        \]
Two discrete distributions
Geometric distribution

- Models the number of Bernoulli trials (independent experiments that can either fail or succeed) needed to get one success
  - Each trial has success rate $p$

- PMF: $P(X = n) = (1 - p)^n p, \quad n = 0, 1, 2, \ldots$

- $E(X) = \frac{1 - p}{p}, \quad \sigma^2 = \frac{1 - p}{p^2}$

- E.g., average number of six-sided dice rolls until we get a specific face:
  - $E(X; p = \frac{1}{6}) = \frac{1 - \frac{1}{6}}{\frac{1}{6}} = 5$
Geometric distribution

\[ P(X = x) \]

\[ P(X \leq x) \]

\[ p = 0.5, \ p = 0.3, \ p = 0.1 \]
Poisson distribution

- Models the number of events occurring in a fixed time interval given the average arrival rate $\lambda$ is known, and if each event occurs independently

- **PMF:** $P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad n = 0, 1, 2, \ldots$

- $E(X) = \lambda, \quad \sigma^2 = \lambda$

- E.g., if we know that an average of 10 buses per hour arrive at a stop, what is the likelihood that only 5 buses arrives in an hour?

- $P(X = 5; \lambda = 10) = \frac{10^5}{5!} e^{-10} \approx 0.0378$
Poisson distribution

$\lambda = 5, \lambda = 10, \lambda = 15$
Two continuous distributions
Gaussian (Normal) distribution

- Models a “bell curve” with specified mean (μ) and variance (σ²)

- PDF: \[ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Exponential distribution

- Models the amount of time elapsing between success independent events, given the average arrival rate $\lambda$

  - PDF: $f(t) = \lambda e^{-\lambda t}, \quad t \geq 0$
  - CDF: $F(t) = 1 - e^{-\lambda t}$

  - $E(X) = \frac{1}{\lambda}$, $\sigma^2 = \frac{1}{\lambda^2}$

- E.g., if we know that an average of 10 buses per hour arrive at a stop, what is the likelihood that we will wait $\leq 5$ minutes for the next bus?

  - $F\left(\frac{1}{12}; \lambda = 10\right) = 1 - e^{-\frac{10}{12}} \approx 0.5654$
Key property: memoryless

- I.e., the amount of time we have to wait until the next event does not depend on how much time has already elapsed!

- i.e., $P(X > t + \Delta t \mid X > t) = P(X > \Delta t)$

- E.g., Given exponential bus inter-arrival times, with $P(X > 20 \text{ min}) = 0.3$

- If you’ve already waited 15 minutes for a bus, how likely is it that the bus won’t arrive for another 20 minutes?

- $P(X > 35 \mid X > 15) = P(X > 20) = 0.3$
E.g., Gaussian vs. Exponential

- Exponential ($\mu=0.1$)
- Gaussian ($\mu=10$, $\sigma^2=2.5$)
Stochastic processes
Stochastic process

- A stochastic process is a collection of random variables \( \{F_t, \ t \in T\} \) defined over the same sample space

- \( t \) is typically a time parameter

  - so \( F_t \) may describe how some system behaves over time period \( t \)
Poisson process

- Series of r.v.s \( \{N_t, \ t \geq 0\} \) where:
  - \( N_t \) models the number of arrivals in time interval \([0, \ t]\)
  - \( N_t \) is described by a Poisson distribution with param \( \lambda t \)
  - Time between arrivals is exponentially distributed with rate \( \lambda \)
  - Connects the Poisson & Exponential distributions
Markov Chain

- Sequence of r.v.s, $X_1, X_2, X_3$, such that:

$$P(X_{t+1} = x \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_2 = x_2, X_1 = x_1)$$

$$= P(X_{t+1} = x \mid X_t = x_t)$$

- I.e., next state depends only on the current state
  - Future is independent of the past
E.g., predicting the weather

“transition matrix”

\[
P = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}
\]

\[ p_{ij} = P(X_{t+1} = j \mid X_t = i) \]

\[
P(X_{t+1}=\text{sunny} \mid X_t=\text{rainy}) = p_{20} = 0.3
\]

P(X_{t+2}=\text{sunny} \mid X_t=\text{rainy})?

\[
= p_{20}p_{00} + p_{21}p_{10} + p_{22}p_{20} = 0.35
\]
E.g., predicting the weather

\[ p_{20}^{(2)} = p_{20}p_{00} + p_{21}p_{10} + p_{22}p_{20} \]

\[ p_{ij}^{(2)} = \sum_{k \in S} p_{ik}p_{kj} = (P \times P)[i][j] \]

\[ P \times P = P^2 = \begin{pmatrix} 0.45 & 0.37 & 0.18 \\ 0.31 & 0.43 & 0.26 \\ 0.35 & 0.41 & 0.24 \end{pmatrix} \]

\[ p_{ij}^{(n)} = P^n[i][j] \]

\[ P^2 = \begin{pmatrix} 0.45 & 0.37 & 0.18 \\ 0.31 & 0.43 & 0.26 \\ 0.35 & 0.41 & 0.24 \end{pmatrix} \]

\[ P^3 = \begin{pmatrix} 0.398 & 0.392 & 0.210 \\ 0.350 & 0.412 & 0.238 \\ 0.364 & 0.406 & 0.230 \end{pmatrix} \]

\[ P^4 = \begin{pmatrix} 0.380 & 0.399 & 0.220 \\ 0.364 & 0.406 & 0.230 \\ 0.369 & 0.404 & 0.227 \end{pmatrix} \]

\[ P^5 = \begin{pmatrix} 0.374 & 0.402 & 0.224 \\ 0.369 & 0.404 & 0.227 \\ 0.370 & 0.404 & 0.226 \end{pmatrix} \]

\[ P^6 = \begin{pmatrix} 0.372 & 0.403 & 0.225 \\ 0.370 & 0.404 & 0.226 \\ 0.371 & 0.403 & 0.226 \end{pmatrix} \]

\[ P^7 = \begin{pmatrix} 0.371 & 0.403 & 0.226 \\ 0.371 & 0.403 & 0.226 \\ 0.371 & 0.403 & 0.226 \end{pmatrix} \]
E.g., predicting the weather

\[ P^7 = \begin{pmatrix} 0.371 & 0.403 & 0.226 \\ 0.371 & 0.403 & 0.226 \\ 0.371 & 0.403 & 0.226 \end{pmatrix} \]

\[ \lim_{k \to \infty} P^k \] converges to a **steady-state distribution**

all rows are equal to the same vector \( \pi \), where

\[ \pi = \pi \times P \text{ and } \sum_{i \in S} \pi_i = 1 \]
E.g., predicting the weather

\[ \pi = \begin{bmatrix} 0.371 & 0.403 & 0.226 \end{bmatrix} \]

Independent of starting state:

\[ P(X_t=\text{sunny}) = 0.371 \]

i.e., fraction of sunny days \( \approx 37\% \)
E.g., predicting the weather

\[\pi = [0.371 \quad 0.403 \quad 0.226]\]

For every state, rate of flow out \(=\) rate of flow in

e.g., for \(S_0\):

rate out \(=\) \((0.371)(0.1 + 0.3)\)
\(=\) 0.148

rate in \(=\) \((0.403)(0.2) + (0.226)(0.3)\)
\(=\) 0.148

i.e., the system is in equilibrium
Queueing theory
Basic model

- Arriving customers
- Wait queue
- Server
- Leaving customers
- Queueing system
Queueing parameters

\[ \lambda \] (arrival rate)

\[ \mu \] (service rate)

\[ T_s = 1/\mu \] (service time)

\[ T_q = T \] (turnaround time) = \[ T_q + T_s \]

\[ L \] (total customers)

\[ L_q \] (waiting customers)

\[ T_q \] (wait time)
Not (typically) constants!

- Queues we are interested in typically have parameters that vary over time.
- Mathematically, we would describe them using probability distributions.
- We use $\lambda$, $\mu$ to refer to the expected values (aka averages) of their respective distributions.
- A typical queueing theory application: given expected values and/or distributions of $\lambda$ and $\mu$, derive other parameters.
Stable system

In a stable system, queue cannot grow unboundedly!

Define ratio $\rho = \frac{\lambda}{\mu}$ as server utilization, and require $\rho < 1$
Steady state / Equilibrium

- Given a stable system, queueing theory models are often only interested in describing long-term, “steady state” behavior.

- i.e., after running the queueing system for some time, over a period we should find # of customers arriving = # of customers departing.

in a steady state, \( \lambda = \text{system throughput} \)
Little’s Law

- In a stable queueing system, $L = \lambda T$

  - I.e., the average number of customers in the system is equal to the product of the average arrival rate and the average turnaround time

  - A useful result that is true regardless of the distributions of parameters!

- Can be applied to just the waiting queue: $L_q = \lambda T_q$

- Or just the server: $\rho = \lambda T_s$
Intuition for Little’s Law \((L = \lambda T)\)

- Suppose the price for a customer to use the system is \$1 per time unit

  - Option 1 (LHS): Each customer can pay an ongoing cost per time unit while in the system.

    - Total income per time unit = \$L

  - Option 2 (RHS): Each customer can pay a lump sum when leaving for the total time spent in the system (\(T\)).

    - \(\lambda\) = throughput in steady state, so total income per time unit = \$\lambda T
e.g., 35th St. Jimmy John’s:

12 customers arrive per hour,
Average time spent in store = 15 minutes.

Average # customers in store?

\[ L = \lambda T = \frac{12}{\text{hour}} \times \frac{1}{60 \text{ min}} \times 15 \text{ min} = 3 \]
e.g., Customer appreciation day!

100 customers arrive per hour,
Average line length = 15

Average wait time?

\[ T = \frac{L}{\lambda} = 15 \times \frac{1 \text{ hour}}{100} = 0.4 \text{ hour} = 9 \text{ min} \]
e.g., Packet switching system with 2 inputs:

\[ \lambda_1 = 200 \text{ packets/s}, \; \lambda_2 = 150 \text{ packets/s}, \]

On average 2,500 packets in system.

Mean packet delay?

\[ T = \frac{L}{\lambda_1 + \lambda_2} = \frac{2,500}{200 + 150} \approx 7.1 \text{s} \]
Kendall’s notation: $A/S/c/k/n/d$

- Shorthand for describing important aspects of a queuing model:
  - $A$: inter-arrival time distribution
  - $S$: service time distribution
  - $c$: number of servers available
  - $k$: waiting line capacity (default = $\infty$)
  - $n$: customer population size (default = $\infty$)
  - $d$: scheduling discipline (default = FCFS)
Kendall’s notation: distributions

- Options for inter-arrival and service distributions:
  - **D**: Deterministic (fixed)
  - **M**: Markovian/Memoryless (exponential distribution)
  - **G**: General/arbitrary distribution (possibly known mean & variance)

- E.g., **M/M/1** = exponential inter-arrival & service distributions, 1 server, infinite capacity and population, FCFS scheduling discipline
M/M/1 queueing system
M/M/1 system

- We can use $L$ (# of customers) to describe the state of the M/M/1 queueing system.
- We can model transitions between these states using a “birth-death” process (a special type of Markov chain), where $\lambda$ and $\mu$ are the infinitesimal rates of flow between states.
Birth-Death process

- $P(L_t = n)$ is the probability of $L = n$ at time $t$

- We are interested in the steady-state distribution:

$$P(L = n) = p_n = \lim_{t \to \infty} P(L_t = n \mid L_0 = i), \quad i = 0, 1, 2, \ldots$$

- I.e., $p_n$ is the probability of $L = n$ after a long period of time
  (and irrespective of starting state)
Deriving $p_n$

- At equilibrium, the rate of flow out of = the rate of flow in to each state
- Giving us the balance equations:
  \[
  \lambda p_0 = \mu p_1 \\
  (\lambda + \mu)p_n = \lambda p_{n-1} + \mu p_{n+1}, \quad n = 1, 2, \ldots
  \]
- Latter is a second order recurrence relation with solution of form:
  \[
  p_n = c_1 x_1^n + c_2 x_2^n, \quad n = 0, 1, 2, \ldots
  \]
- Where $x_1$ and $x_2$ are roots of the equation $\mu x^2 - (\lambda + \mu) x + \lambda = 0$
Deriving $p_n$

- $\mu x^2 - (\lambda + \mu)x + \lambda = 0$ has two roots: $x = 1$ and $x = \lambda/\mu = \rho$

- Solutions to recurrence relation are of form $p_n = c_1 + c_2\rho^n$, $n = 0, 1, 2, \ldots$

- We know that: \[\sum_{n=0}^{\infty} p_n = 1\], i.e., \[\sum_{n=0}^{\infty} (c_1 + c_2\rho^n) = 1\]

- $c_1$ must be 0, and we have \[\sum_{n=0}^{\infty} c_2\rho^n = 1\]

only converges if $\rho < 1$; i.e., $\lambda < \mu$
Deriving $p_n$

- Assuming $\rho < 1$, \( \sum_{n=0}^{\infty} c_2 \rho^n = \frac{c_2}{1 - \rho} = 1 \)

- I.e., $c_2 = 1 - \rho$

- Giving us \( P(L = n) = p_n = (1 - \rho) \rho^n \)

- Probability of system being in any state is dependent on $\rho$ alone!
e.g., M/M/1 queue over at JJ’s

Average of 15 customers arriving per hour
Average service time of 2.5 minutes per customer

How likely is it for there to be 5 customers in the store?

\[ \rho = \frac{\lambda}{\mu} = \frac{15}{24} = 0.625 \]

\[ P(L = 5) = p_5 = (1 - 0.625)0.625^5 \approx 0.0358 \]
e.g., M/M/1 queue over at JJ’s

Average of 15 customers arriving per hour
Average service time of 2.5 minutes per customer

How likely is it for there to be \( \leq 5 \) customers in the store?

\[
P(L \leq 5) = \sum_{n=0}^{5} (1 - 0.625)0.625^n \approx 0.9404
\]
Expected value of $L$?

- Can derive directly from distribution of $L$
  - $(1 - \rho)\rho^n$ is just the geometric distribution with parameter $1 - \rho$
  - Expectation is $E(L) = \frac{\rho}{1 - \rho}$
- Or can derive it directly using a useful property of M/M/* queues: PASTA
PASTA

- **PASTA** property: Poisson Arrivals See Time Averages

  - i.e., customers arriving will on average encounter the same number of customers in the system as predicted by the steady state average

  - also: customers arriving will be faced with the same average service times as predicted by the steady state average

- Seems intuitive but not always true of other distributions!
Assume $E(L) = 5$ people in store

- i.e., to the outside observer, there are an average of 5 people in the store
- given Poisson arrivals, new customers on average also see 5 people in the store
Not true in general!

- consider deterministic system:
  - arrival times = 1, 3, 5, 7, …
  - service time = 1 (constant)
  - $E(L) = 1/2$

- but arriving customers always see 0 in store!
Mean value approach

- We can compute $E(L)$ directly (without deriving the distribution), using Little’s law and PASTA

- Start by considering $E(T)$ (average time spent in system)

  - $E(T) = \text{avg } \# \text{ customers} \times \text{avg service time} + \text{avg remaining service time}$

  - by PASTA:
    
    $E(L) + \frac{1}{\mu} + \frac{1}{\mu}$

  - i.e., $E(T) = E(L) \left( \frac{1}{\mu} + \frac{1}{\mu} \right)$
Mean value formulae

\[ E(T) = E(L) \frac{1}{\mu} + \frac{1}{\mu} \]

- By Little’s law, \( E(L) = \lambda E(T) \)

\[ E(T) = \frac{1}{\mu(1 - \frac{\lambda}{\mu})} = \frac{1}{\mu(1 - \rho)} \]

\[ E(L) = \frac{\lambda}{\mu(1 - \rho)} = \frac{\rho}{(1 - \rho)} \]

\[ E(T_s) = \frac{1}{\mu} \]

\[ E(T_q) = E(T) - E(T_s) = \frac{\rho}{\mu(1 - \rho)} \]

Agrees with distribution-based analysis
Powerful (and surprising?) results

- Expected values of all M/M/1 system parameters are entirely dependent on the relationship of arrival and service times
- Applicable to a vast number of different domains!
  - But: important to understand M/M/1 assumptions
    - And remember: Little’s law applies to all queues, regardless of arrival/service distributions
e.g., M/M/1 queue over at JJ’s

Average of 15 customers arriving per hour
Average service time of 2.5 minutes per customer

What is the **average number** of customers in the store?

\[
\rho = \frac{\lambda}{\mu} = \frac{15}{24} = 0.625 \quad E(L) = \frac{\rho}{1 - \rho} = \frac{0.625}{1 - 0.625} \approx 1.667
\]