CS 444 Operating Systems

Queueing Theory

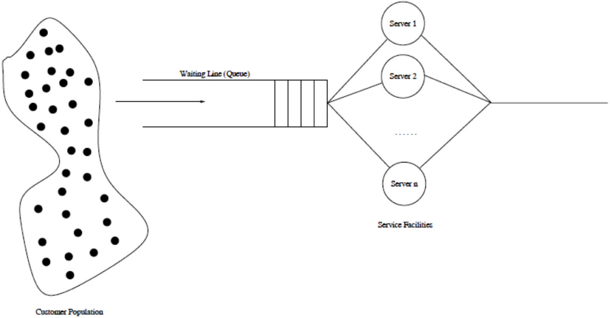
J. Holly DeBlois

November 6, 2024

Applications of Queueing Theory

- Queueing analysis is applicable whenever there are
  - A population of customers
  - A service facility
  - A waiting line
- Examples
  - Processes, CPU, READY queue
  - Customers, call center, waiting queue
- Things we want to know
  - Waiting time in the queue
  - Service time
  - Response time (turnaround time)
    - Waiting time plus service time
  - Utilization of the service facility
  - Queue lengths

Model Description



- Customers arrive in a random fashion
- The service facility has one or more servers
  - One customer per server at a time
- Service time is also random

Assumptions

- Customer population is infinite
- The inter-arrival time of customers is an independent and identically distributed (iid) random variable
- The service time for each customer is also iid
- The length of the queue can be infinite or finite

Review Probability and Statistics

- A continuous random variable X can be described by
- Either its distribution function $F(x)$ — cumulative distribution function, cdf

$$F(x) = \Pr[X \leq x] \qquad F(-\infty) = 0 \qquad F(\infty) = 1$$

- Or its density function $f(x)$ — probability density function, pdf

$$f(x) = \frac{d}{dx}F(x) \qquad F(x) = \int_{-\infty}^{x} f(y)dy \qquad \int_{-\infty}^{\infty} f(y)dy = 1$$

- For a discrete random variable, replace integration by summation

Properties of Distributions

- Mean of a continuous distribution

$$\mathrm{E}[X] = \mu_x = \int_{-\infty}^{\infty} x f(x) dx$$

- Mean of a discrete distribution

$$\mathrm{E}[X] = \mu_x = \sum_{\text{all } k} k \Pr[X = k]$$

- Second moment

$$\mathrm{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx \qquad \mathrm{E}[X^2] = \sum_{\text{all } k} k^2 \Pr[X = k]$$

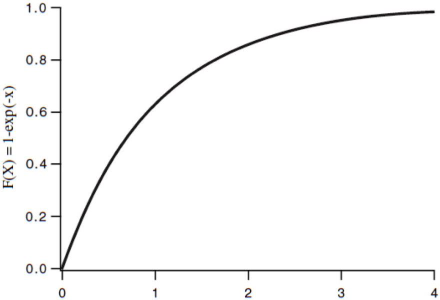
- Variance

$$\mathrm{V}[X] = \mathrm{E}\left[(X - \mu_x)^2\right] = \mathrm{E}[X^2] - \mu_x^2$$

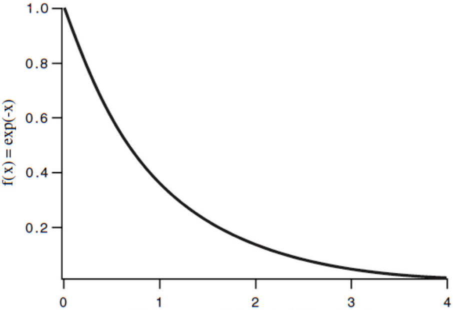
The Exponential Distribution

- $\lambda > 0$, the arrival rate
- $1/\lambda$, the inter-arrival time
- CDF
$$F(x) = 1 - e^{-\lambda x},\ x \geq 0$$
- PDF
$$f(x) = \lambda e^{-\lambda x}$$
- Mean
$$\mathrm{E}[X] = \frac{1}{\lambda}$$

The Exponential Distribution



(a) Exponential probability distribution



(b) Exponential probability density

- CDF
- PDF

Geometric Distribution

- A discrete distribution
- Bernoulli trials with success rate p
- The probability to succeed at the k-th trial

$$\Pr[X = k] = (1-p)^{k-1}p,\ k = 1, 2, 3, \ldots$$

- Memoryless: the probability of success at the k-th trial is the same, regardless of the value of k
- Exponential distribution is the continuous version of geometric distribution

Exponential Distribution is Memoryless

- Waiting for an event to happen
- After the clock has started ticking
  - At any given moment, the probability that we need to wait for T additional amount of time is the same, regardless of how long we have been waiting
  - Random arrival
- Memoryless
  - Assume $b > a > 0$, waiting for a, b, or $(b - a)$ amounts of time
  - $\Pr[T > b | T > a] = \Pr[T > b - a]$
- Geometric and exponential distributions are the only memoryless distributions

Kendall Notation for Queueing Systems

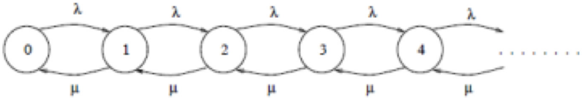
- A/B/m/N S
- A: the distribution of inter-arrival time
- B: the distribution of service time
- m: the number of servers
- N: the length of the queue
  - Omitted if N is infinite
- S: service discipline
  - Omitted if S is FIFO

- The exponential distribution is commonly used for the inter-arrival and service time
  - Designated as M, Markov process
- The most simple queueing system is M/M/1
  - Two exponential distributions, with parameters λ and μ

Steady State

- After running for a long time, the system tends to reach a stable state
- In general, it is possible to analytically calculate the properties of the M/M/m systems at the steady state
- A Markov chain has
  - A set of n states
  - An $n \times n$ matrix of probabilities of transitions from states to states
- Example: Count the number of heads when tossing a coin repeatedly, starting with 0, adding 1 for head, and subtracting 1 for tail; the states are $\ldots, -2, -1, 0, 1, 2, \ldots$
- Memoryless: The probability to move from state X to state Y depends on X only, independent of the state before X

M/M/1 as a Markov Chain



- The states are the numbers of customers in the system
- The states $k = 0, 1, 2, \ldots$
- $P_k(t)$: the probability of the system in state k at time t
- At steady state, $P_k = \lim_{t \to \infty} P_k(t)$

$$\begin{aligned}
\frac{dP_k(t)}{dt} &= (\lambda P_{k-1}(t) + \mu P_{k+1}(t)) - (\lambda P_k(t) + \mu P_k(t)) \\
0 &= \mu P_1 - \lambda P_0 \\
0 &= \lambda P_0 + \mu P_2 - \lambda P_1 - \mu P_1 \\
0 &= \lambda P_{k-1} + \mu P_{k+1} - \lambda P_k - \mu P_k
\end{aligned}$$

Solution of M/M/1 Steady State

$$P_k = \left(\frac{\lambda}{\mu}\right)^k P_0$$

$$\sum_{k=0}^{\infty} P_k = 1$$

$$P_0 = 1 - \frac{\lambda}{\mu}$$

Utilization is

$$1 - P_0 = \frac{\lambda}{\mu} = \rho$$

Queue length is

$$N = \sum_{k=0}^{\infty} kP_k = \frac{\rho}{1-\rho}$$