

Performance Evaluation of Computer Systems and Networks Lecture notes for CS 681

Varsha Apte
Department of Computer Science and Engineering
IIT Bombay. Autumn - 2005.

Probability Refresher

1 Simple Probability Models:

Example 1: Let bit error rate/probability (b.e.r.) on a link be e , then what is the probability that a packet of length L bits is correctly transmitted? Assume:

(a) All bits need to be correct.

Solution:

$P[\text{packet is correct}] = P[\text{all bits correct}] = (1 - e)^L$
This assumes that bit errors are **independent**.

(b) Error correcting code can correct 2 bit errors.

Solution:

$P[\text{packet is correct}] = P[\leq 2 \text{ bits in error}]$
 $\gamma = P[0 \text{ bit error}] + P[1 \text{ bit error}] + P[2 \text{ bit error}]$
 $\gamma = (1 - e)^2 + Le(1 - e)^{L-1} + {}^L C_2 e^2 (1 - e)^{L-2}$

(c) Suppose packet can go on one of two links with b.e.r. e_1, e_2 respectively. What is the probability of successful transmission?

Solution:

$P[\text{packet goes on link1}] = p_1$
 $P[\text{packet goes on link2}] = p_2$
From case(b), we have:
 $\gamma_1 \dots$ success on link1
 $\gamma_2 \dots$ success on link2

A : packet is successful
 B_1 : Link1 is used
 B_2 : Link2 is used

$P[A|B_1] = \gamma_1$ (this is **conditional probability**)
 $P[A|B_2] = \gamma_2$

$P[A] = P[A|B_1] + P[A|B_2] = \gamma_1 P_1 + \gamma_2 P_2$

This is theorem of **total probability**.

Bernoulli trials: These are events which have only 2 possible outcomes. One is termed success, the other, failure.

Definition of conditinal probability:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

Definition of total probability:

Let $S = B_1, B_2, \dots, B_n$ be the event space where, B_1, B_2, \dots, B_n are mutually exclusive and exhaustive events. Then, by law of total probability, we have,

$$P[A] = \sum_{i=1}^n P(A|B_i) \cdot P(B_i)$$

Bayes' Theorem The event space can be viewed as a collection of **Mutually exclusive and Collectively exhaustive** events. Depending on our intuition or assumptions we can have a prior estimate of the events which can be termed as Prior probability. Given that an event occurred in the random experiment the probabilities for the events under consideration will change. This probability can be termed as the Posterior probability. Bayes theorem helps in determining the posterior probability given some prior information of the events.

Let the event space consists of B_j mutually exculsive and collectively exhaustive events. Knowing the prior probabilities and a event A ocured the posterior probabilities can be calculated by:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

Example:

Suppose, in the two-link case, the packet arrived successfully. What is the probability that it came on link1?

Solution:

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A \cap B_j)}{\sum_i P(A|B_i)P(B_i)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$$

P[Link1 | packet is succesful] :

$$P(B_1|A) = \left(\frac{P(A|B_1)P(B_1)}{\gamma_1 P_1 + \gamma_2 P_2} \right)$$

Law of diminishing returns:

$$P[\text{server is up}] = r = 0.9$$

2 Server system :

$$P[\text{system is up}] = P[\text{at least one server is up}]$$

$$r = 1 - P[\text{both servers are down}]$$

$$r = 1 - (1 - r)^2$$

For an n-server parallel system, system is up when at least one server is up.
 $R_{\text{server}} = 1 - (1 - r)^n$

Efficiency of pure ALOHA

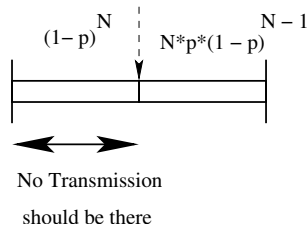


Figure 1: Pure ALOHA

Stations transmit whenever they have a packet. If packet collides, it is retransmitted after random amount of time.

At “steady-state”, let N be the number of stations and p be the probability that a station transmits a packet. We want to know the maximum utilization of the system.

$$P[\text{successful transmission}] = P[\text{no transmission in previous slot}] \times P[\text{no other transmission in this slot}]$$

$$S = (1 - p)^N \cdot Np(1 - p)^{N-1}$$

$$S = Np(1 - p)^{2N-1}$$

The probability of transmission when success is maximum is :

$$\frac{dS}{dp} = N(1 - p)^{2N-1} - Np(2N - 1)(1 - p)^{2N-2} = 0$$

$$p = \frac{1}{2N}$$

$$S_{\text{max}} = \frac{1}{2} \left(1 - \frac{1}{2N}\right)^{2N-1}$$

As N increases, efficiency decreases. The limiting efficiency is:

$$\lim_{N \rightarrow \infty} S = \frac{1}{2e} \sim 0.18 = 18\%$$

(Experiment with the excel sheet.)

Efficiency of slotted ALOHA

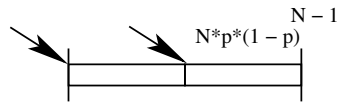


Figure 2: Slotted ALOHA

Stations transmit only at the beginning of a slot.

$P[\text{successful transmission}] = P[\text{max. 1 transmission in the slot}]$

$$S = Np(1-p)^{N-1}$$

For maximum success rate,

$$\frac{dS}{dp} = N(1-p)^{N-1} + Np(N-1)(1-p)^{N-2} = 0,$$

Thus, highest success rate is at $p = \frac{1}{N}$
At this value of p ,

$$S = \left(1 - \frac{1}{N}\right)^{N-1}$$

$$\lim_{N \rightarrow \infty} S = \frac{1}{e} \sim 0.37 = 37\%$$

2 Random Variable:

These map sample space to set of real numbers.

Example: If a packet may go over either link1, or link2, or ... each link can be associated with a real number which would be called 'probability' that the packet takes that link. The question as to which link the packet takes is a random variable.

There are two types of random variables viz. discrete, and continuous. For discrete r.v.s, the associated set of real numbers has discrete valued elements whereas for continuous r.v.s, the associated set of real numbers has continuous valued elements.

For a discrete r.v. X , p.m.f. is defined as :

$$0 \leq P_X(x) \leq 1, \text{ and}$$
$$\sum_{x \in R} P_X(x) = 1$$

The c.d.f. is

$$CDF = F_X(x) = P[X \leq x]$$
$$P[a \leq X \leq b] = F(b) - F(a)$$

Examples: Some common random variable probability distributions are Bernoulli, Binomial, Geometric, and Poisson distributions.

Go-back-N ARQ System

Retransmission Timer = $(n-1)$ frame transmission times. Assume re-transmission can only occur if packet is lost.

If k re-transmission, average delay for packet to be sent (effective service time) is $(1 + kn)$ with the probability $p^k(1 - p)$

$$P(X = 1 + kn) = p^k(1 - p) \text{ for } k = 0, 1, 2, \dots$$

Bernoulli random variable: Packet transmissions are Bernoulli trials.

$$X = \begin{cases} 1, & \text{if packet transmission successful} \\ 0, & \text{otherwise} \end{cases}$$

Binomial r.v.: Number of packets in 1 window that reached correctly.

$$P[X = i] = \binom{n}{i} p^i (1 - p)^{n-i}$$

Geometric r.v.: Number of transmissions required to send a packet correctly.

Modified Geometric r.v.: Number of re-transmissions required to send a packet correctly. In other words, number of failures before the packet was transmitted correctly.

Example: Packet arrive into source in an interval Δt with probability $\lambda \Delta t$.

$P[2 \text{ or more arrivals in } \Delta t] \rightarrow 0$.

What is $P[k \text{ jobs in time interval } t]$?

Solution: Divide time t into n pieces such that $\frac{t}{n} = \Delta t$.

$$\begin{aligned}
P[k \text{ jobs}] &= k \text{ out of } n \text{ slots have job arrivals} \\
&= \binom{n}{k} (\lambda \Delta t)^k (1 - \lambda \Delta t)^{n-k} \\
&= \binom{n}{k} \left(\lambda \frac{t}{n}\right)^k \left(1 - \lambda \frac{t}{n}\right)^{n-k} \\
&= \left(\frac{n!}{(n-k)! n^k}\right) \frac{(\lambda t)^k}{k!} \left[\frac{\left(1 - \frac{\lambda t}{n}\right)^{-n/\lambda t}}{\left(1 - \frac{\lambda t}{n}\right)^k}\right]^{-\lambda t} \\
\lim_{n \rightarrow \infty} &= \frac{\overbrace{n \cdot n - 1 \cdots (n - k + 1)}^{n-k \text{ times}}}{n \cdot n \cdots n} [\dots] \\
\lim_{n \rightarrow \infty} &= \frac{(\lambda t)^k}{k!} e^{-\lambda t} \\
&\sim \text{Poisson}(\lambda t)
\end{aligned}$$

Let $\lambda t = \alpha$.

$$\begin{aligned}
X &\sim \text{Poisson}(\alpha) \\
\Rightarrow P[X = k] &= \frac{\alpha^k e^{-\alpha}}{k!}
\end{aligned}$$

2.1 Bernoulli Trials

These are experiments with two possible outcomes, having probability “p” and “1-p”. Examples of Bernoulli Trials are: tossing a coin, transmitting a single bit across a link, and other success-failure type of trials.

2.2 Random Variables

Mathematically a random variable is a function which maps a sample point to a real number.

$$X \longrightarrow \text{Values}$$

2.3 Probability Mass Function (PMF)

The Probability Mass Function $P_X(x)$ of a random variable X gives the probability of occurrence of each value of the random variable.

$$P_X(X = x) \longrightarrow [0, 1]$$

where

$$x \in \text{Sample Space}(S)$$

Some properties of a PMF are:

- $0 \leq P_X(x) \leq 1$
- $\sum_x P_X(x) = 1$

2.4 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function $F_X(t)$ of a random variable X represents the cumulative probability of X till point ($x = t$).

$$F_X(t) = P(x \leq t) = \sum_{x \leq t} P_X(x)$$

where $t \in \mathbf{R}$ and t can be continuous.

Example: $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

$$P_X(x) = \frac{1}{10} \quad F_X(8.6) = \frac{8}{10}$$

Some properties of a CDF, also referred-to as simply the *Distribution Function*, are the following:

- $0 \leq F_X(t) \leq 1$
- Monotonically nondecreasing.
- $\lim_{t \rightarrow \infty} F_X(t) = 1$
- $\lim_{t \rightarrow -\infty} F_X(t) = 0$

3 Discrete Distributions

3.1 Bernoulli PMF

Parameter: p

$$X \in \{0, 1\}$$

$$P_X(1) = p \quad P_X(0) = 1 - p$$

$$F_X(t) = \begin{cases} 0, & \text{for } t < 0 \\ 1 - p, & \text{for } 0 \leq t < 1 \\ 1, & \text{for } t \geq 1 \end{cases}$$

3.2 Binomial PMF

Parameters: n, p

$X \longrightarrow$ Number of successes i in n Bernoulli Trials

$$X = 0, 1, 2, \dots, n$$

$$P(\text{success}) = p$$

$$P_X(i) = \binom{n}{i} p^i (1 - p)^{n-i}$$

The Poisson PMF can be used as a convenient approximation to the binomial PMF when n is large and p is small:

$$\binom{n}{i} p^i (1-p)^{n-i} \approx \frac{e^{-\alpha} \alpha^i}{i!}$$

where $\alpha = np$

Example: Bit errors in a packet of length L

3.3 Geometric Distribution

Parameter: p

$X \rightarrow$ Number of trials upto and including a success

$$X = 1, 2, 3, \dots, \infty$$

$$P(\text{success}) = p$$

$$P_X(i) = (1-p)^{i-1} p$$

3.3.1 Markov Property (memorylessness)

$X \rightarrow \text{Geom}(p)$ and $P(X = i) = (1-p)^{i-1} p$

$$P[X = k+i | X > i] = (1-p)^{k-1} p \quad k = 1, 2, \dots, \infty \quad (1)$$

Additional number of trials after i unsuccessful trials is denoted by random variable Z .

$$\begin{aligned} P[z = k] &= P[X = k+i | X > i] \\ &= \frac{P[X = k+i]}{P[X > i]} \\ &= \frac{(1-p)^{k+i-1} p}{(1-p)^i} \\ &= (1-p)^{k-1} p \end{aligned}$$

$\Rightarrow Z \rightarrow \text{Geom}(p)$

Forget that i trials were done \Rightarrow Memoryless property.

Example: B.E.R. on a channel is b . B-errors are independent. In a packet, the first 10 bits have come error-free. What is the probability that there would be k error-free bits before an error is seen?

Solution:

$Z \sim \#$ of bits after tenth bit transmitted error-free until first-error is seen.

$Z \sim \text{Geom}(b)$

$$P[z = k] = (1-b)^{k-1} b$$

3.4 Modified Geometric Distribution

Parameter: p

$X \rightarrow$ Number of failures before a success

$$X = 0, 1, 2, \dots, \infty$$

$$P_X(i) = (1-p)^i p$$

Example: while not B do S ;
 X : # of times S is executed

$$\begin{aligned} P(B = \text{true}) &= p \\ P(x = k) &= (1-p)^k p \end{aligned}$$

$X \sim \text{ModGeom}(p)$

3.5 Poission

Parameter: α

$$P_X(i) = \frac{e^{-\alpha} \alpha^i}{i!}$$

3.6 Uniform

Parameter: N

$$X = \{x_1, x_2, x_3, \dots, x_N\}$$

$$P_X(x_i) = \frac{1}{N}$$

$$F_X(t) = \frac{t}{N}$$

3.7 Constant

Parameter: c

$$X = c$$

$$P_X(x) = \begin{cases} 1, & \text{for } x = c \\ 0, & \text{for } x \neq c \end{cases}$$

$$F_X(t) = \begin{cases} 0, & \text{for } t < c \\ 1, & \text{for } t \geq c \end{cases}$$

3.8 Indicator Random Variable

$X = 0$ if event A and $X = 1$ if event \bar{A} .

4 Expectation

The expectation (mean), denoted by $E[X]$, of a random variable X is defined as:

$$E[X] = \sum_i x_i p(x_i) \quad X \rightarrow \text{discrete random variable} \quad (2)$$

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \quad X \rightarrow \text{continuous random variable} \quad (3)$$

4.1 Expectation of Some Common Distributions

4.1.1 Uniform

$$X = \{1, 2, \dots, n\} \quad p_X(x) = \frac{1}{n}$$

$$E[X] = \sum_{i=1}^n \frac{i}{n} = \frac{n+1}{2}$$

4.1.2 Bernoulli pmf

Parameter: p

$$E[X] = p \quad (4)$$

4.1.3 Binomial

Parameters: n, p

$$E[X] = np \quad (5)$$

$$E\left[\sum_i X_i\right] = \sum_i E[X_i] \quad (X_i\text{s need not be independent}) \quad (6)$$

4.1.4 Geometric

Parameter: p

$$p_X(i) = (1-p)^{i-1}p$$
$$E[X] = \frac{1}{p} \quad (\text{prove as homework}) \quad (7)$$

4.1.5 Poisson

Parameter: α

$$p_X(k) = \frac{\alpha^k e^{-\alpha}}{k!}$$

$$\begin{aligned}
E[X] &= \sum_{k=0}^{\infty} \frac{\alpha^k e^{-\alpha}}{k!} \\
&= \alpha e^{-\alpha} \sum_{k=1}^{\infty} \frac{\alpha^{k-1}}{(k-1)!} \\
&= \alpha e^{-\alpha} e^{\alpha} \\
E[X] &= \alpha
\end{aligned} \tag{8}$$

4.1.6 Exponential

Parameter: λ

$$\begin{aligned}
f(x) &= \lambda e^{-\lambda x} \\
E[X] &= \int_0^{\infty} x \lambda e^{-\lambda x} dx \\
&= \frac{1}{\lambda} \int_0^{\infty} u e^{-u} du \\
&= \frac{\Gamma(2)}{\lambda} \\
E[X] &= \frac{1}{\lambda}
\end{aligned} \tag{9}$$

4.2 Examples

Example 1: Consider Array[1..n]

$\lambda = \{1, 2, \dots, n\} \rightarrow$ Number of comparisons for a successful search on the array.

$$\begin{aligned}
P_X(i) &= \frac{1}{n} \\
E[X] &= \frac{n+1}{2}
\end{aligned}$$

Example 2:

$$P_X(i) = \frac{c}{i}, \quad i = 1, 2, \dots, n$$

$$\begin{aligned}
\sum_{i=1}^n \frac{c}{i} &= 1 \\
\Rightarrow c &= \frac{1}{\sum_{i=1}^n \frac{1}{i}} \\
&= \frac{1}{\ln(n) + \alpha}
\end{aligned}$$

$$\begin{aligned}
E[X] &= \sum_{i=1}^n i \cdot \frac{c}{i} \\
&= \frac{n}{\ln n + \alpha}
\end{aligned}$$

This is used to model web traffic (hits).

Example 3: A cache of m out of a total n webpages.

$$P_X(i) = \frac{c}{i}, \quad i = 1, 2, \dots, n$$

$$c = \frac{1}{H_n}$$

There will be m pages in the cache. The the probability P_{hit} that a request hits the cache is:

$$\begin{aligned} P_{hit} &= \sum_{i=1}^m m \frac{c}{i} \\ &= \frac{H_m}{H_n} \\ &= \frac{\ln m + \alpha}{\ln n + \alpha} \end{aligned}$$

5 Probability Generating Function

X is a discrete random variable, $X \geq 0$

The Probability Generating Function is defined as:

$$G_X(z) = \sum_{i=0}^{\infty} p_i Z^i = p_0 + p_1 z + p_2 z^2 + \dots$$

Some properties:

- $G_X(z)$ will converge if $z \leq 1$
- $z = 1 \Rightarrow G_X(z) = 1$
- $G_X(z) = G_Y(z) \forall z \Rightarrow P_X(z) = P_Y(z)$

5.1 Bernoulli PGF

$$P_X(0) = 1 - p \quad P_X(1) = p$$

$$G_X(z) = 1 - p + pz$$

5.2 Binomial PGF

$$G_X(z) = \sum_i^n \binom{n}{i} p^i (1-p)^{n-i} z^i = (1-p + pz)^n$$

5.3 Modified Geometric PGF

$$P_X(i) = (1-p)^i p \quad i = 0, 1, 2, \dots$$

$$G_X(z) = \sum_{i=0}^{\infty} (1-p)^i p z^i = \frac{p}{1 - (1-p)z}$$

5.4 Poisson PGF

$$\begin{aligned}
 P_X(i) &= \frac{e^{-\alpha} \alpha^i}{i!} \\
 G_X(z) &= \sum_{i=0}^{\infty} \frac{e^{-\alpha} \alpha^i}{i!} z^i \\
 &= e^{-\alpha} \sum_{i=0}^{\infty} \frac{(\alpha z)^i}{i!} \\
 &= e^{-\alpha} e^{\alpha z} \\
 &= e^{-\alpha(1-z)}
 \end{aligned}$$

5.5 Geometric PGF

$$\begin{aligned}
 P_X(i) &= (1-p)^{i-1} p \\
 G_X(z) &= \sum_{i=1}^{\infty} (1-p)^{i-1} p z^i \\
 &= p z \sum_{i=1}^{\infty} (z(1-p))^{i-1} \\
 &= \frac{p z}{1-(1-p)z}
 \end{aligned}$$

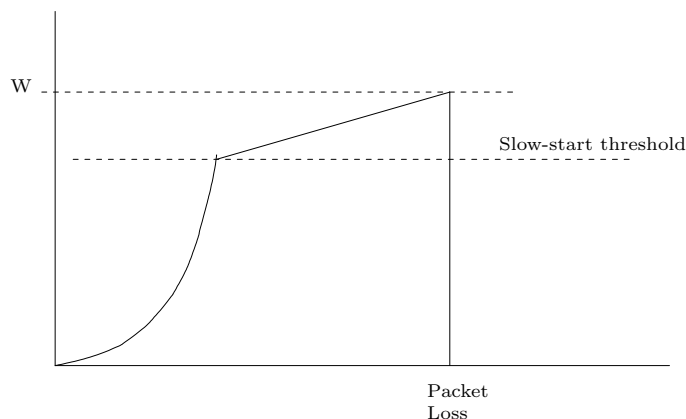
Important: $G_X(z) = G_Y(z) \Rightarrow X, Y \sim \text{same distribution}$

6 TCP model for light/moderate loss link

TCP uses a window-based flow control, If W is the size of window, R the link data rate,

$$Throughput = \frac{W \cdot L}{L/R + RTT}$$

A TCP flow has two phases of operation: slow start and congestion avoidance which works as follows:

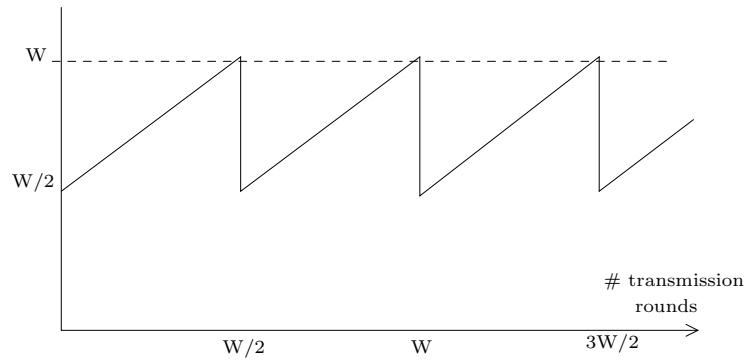


Assumptions:

- the flow spends most of the time in linear(congestion avoidance) mode
- the TCP flow lasts long enough, so that the steady-state analysis is relevant

Let the probability of packet loss be p . Then, average number of packets after which packet loss happens = $1/p$

Also, number of rounds after which packet loss happens is $W/2$.



Number of packets transmitted in $W/2$ packet transmission rounds

= Area under the curve

$$= \left(\frac{3}{8}\right)W^2 = \frac{1}{p}$$

$$\Rightarrow W = \sqrt{\frac{8}{3p}}$$

$$\begin{aligned} \text{TCP throughput} &= \frac{\text{number of bytes sent}}{\text{time taken}} \\ &= \frac{\text{packet size} \times 1/p}{RTT \times W/2} \\ &= \frac{MSS \times 1/p}{RTT \times W/2} \\ &= \frac{MSS}{RTT} \sqrt{\frac{3}{2p}} \end{aligned}$$

where,

MSS = TCP maximum segment size

RTT = Round Trip Time

Thus, TCP throughput $\propto \frac{1}{\sqrt{p}}$

7 Discrete Convolution

X and Y are discrete random variables. $T = X + Y$.

$X \rightarrow P_X$ and $Y \rightarrow P_Y$

$$P_T(t) = \sum_{x=0}^t P_X(x)P_Y(t-x)$$

$$\begin{aligned}
G_T(z) &= \sum_{t=0}^{\infty} P_T(t)z^t \\
&= \sum_{t=0}^{\infty} \sum_{x=0}^t P_X(x)P_Y(t-x)z^t \\
&= \sum_{x=0}^{\infty} \sum_{t=x}^{\infty} P_X(x)z^x P_Y(t-x)z^{t-x} \\
&= \sum_{x=0}^{\infty} P_X(x)Z^x \sum_{t=x}^{\infty} P_Y(t-x)z^{t-x} \\
&= G_X(z)G_Y(z)
\end{aligned} \tag{5}$$

Example: $X_i \rightarrow \text{Poisson}(\alpha_i)$

$$G_{X_i}(z) = e^{-\alpha_i(1-z)}$$

$$T = \sum X_i$$

$$\Rightarrow G_T(z) \rightarrow \text{Poisson}(\sum \alpha_i)$$

For Binomial PGF, Poisson PGF, and the others, please refer to K.Trivedi book.

7.1 Stack Optimization Problem

$P(\text{overflow}) < 0.05$

7.1.1 Case 1: Two stacks, each of size n

$$\begin{aligned}
P(\text{overflow}) &= P(S_1 > n \text{ or } S_2 > n) \\
&= P(S_1 > n) + P(S_2 > n) - P(S_1 > n)P(S_2 > n) \\
&= 2(1-p)^n - (1-p)^{2n}
\end{aligned}$$

7.1.2 Case 2: A single stack of size N , working as two stacks

$$T = S_1 + S_2$$

$$\begin{aligned}
P(\text{overflow}) &= P(S_1 + S_2 > N) \\
&= P(T > N)
\end{aligned}$$

$$\begin{aligned}
G_{S_i}(z) &= \frac{pz}{1-qz} \\
\Rightarrow G_T(z) &= \left(\frac{pz}{1-qz} \right)^2 \\
&= \frac{p^2 z^2}{q} \frac{d}{dz} \left(\frac{1}{1-qz} \right) \\
&= \frac{p^2 z^2}{q} \frac{d}{dz} (1 + qz + q^2 z^2 + q^3 z^3 + \dots) \\
&= \frac{p^2 z^2}{q} (q + 2q^2 z + 3q^3 z^2 + \dots) \\
&= p^2 (z^2 + 2qz^3 + 3q^2 z^4 + \dots)
\end{aligned}$$

$$P_S(i) = (i-1)q^{i-2}p^2 \quad i \geq 2$$

$$\begin{aligned}
P(\text{overflow}) &= P(T > N) \\
&= 1 - P(T \leq N) \\
&= 1 - \sum_{i=2}^N (i-1)q^{i-2}p^2 \\
&= 1 - p^2 \sum_{j=1}^{N-1} jq^{j-1} \\
&= 1 - p^2 (1 + 2q + 3q^2 + \dots + (N-1)q^{N-2}) \\
&= 1 - p^2 \frac{d}{dq} \left(\frac{1-q^N}{1-q} - 1 \right) \\
&= 1 - p^2 \left(\frac{-Nq^{N-1}}{1-q} + \frac{1-q^N}{(1-q)^2} \right) \\
&= q^{N-1} (1 - p + Np)
\end{aligned}$$

8 Continuous Random Variables

Random variables such as processing time, interarrival time may not always be described by discrete values - they are better described by continuous random variables.

For continuous random variables, we define a *distribution function*, i.e. the cumulative distribution function:

$$F_X(x) = P[X \leq x] \quad -\infty < x < \infty$$

The CDF has the same properties as the discrete CDF:

$$\begin{aligned}
0 \leq F_X(x) &\leq 1, -\infty < x < \infty \\
\lim_{x \rightarrow -\infty} &= 0 \\
\lim_{x \rightarrow \infty} &= 1
\end{aligned}$$

and $F_X(x)$ is a *monotone non-decreasing function*.

The *probability density function* of a continuous random variable is given by

$$f_X(x) = \frac{dF_X(x)}{dx}$$

and, for X has a PDF $f_X(x)$,

$$\begin{aligned} P[a < X \leq b] &= \int_a^b f_X(t) dt \\ F_X(x) &= \int_{-\infty}^x f_X(t) dt \end{aligned}$$

The value of a PDF is not a probability. A PDF is a function that must satisfy the following properties:

$$\begin{aligned} f_X(x) &\geq 0, \quad -\infty < x < \infty \\ \int_{-\infty}^{\infty} f_X(x) dx &= 1 \end{aligned}$$

It follows that

$$\begin{aligned} P[X = c] &= \int_c^c f_X(t) dt = 0 \\ P[a < X < b] &= P[a \leq X \leq b] = P[a < X \leq b] = P[a \leq X < b] = F_X(b) - F_X(a) \end{aligned}$$

In other words, for continuous r.v.s, the *area under the curve* of a pdf function is the probability - “pieces” of this area will be probabilities, not the values. So, although $P[X = c] = 0$, $P[c < X < c + \delta c]$ is defined and can be non-zero. The total area under the pdf curve should be 1.

(Figure here)

9 Exponential Distribution

This is a distribution which has some unique properties, which make it a very useful distribution in modeling.

If a random variable X has a distribution given by

$$F(x) = 1 - e^{-\lambda x}, \quad X \geq 0, \lambda > 0$$

then X is said to have the *exponential* distribution with parameter λ . This distribution is also denoted by $EXP(\lambda)$.

The pdf of such a random variable is given by:

$$f(x) = \lambda e^{-\lambda x}$$

The exponential distribution has the *memoryless* property. An example of this is as follows: suppose you are at a phonebooth while somebody is on the phone. Suppose this person has been already talking for 10 minutes. What is the probability that he will hold the phone for 10 more minutes? If the call holding time distribution is exponential (which it often is), the probability

of talking for 10 hold minutes is the same as if the person had just started talking - i.e. it is “memoryless”. Let us prove this mathematically:

Let X be the random variable denoting the holding time. Suppose we know that $X > t$ (“person has already been talking for t time units”). Let Y be the remaining time. Then $X = Y + t$.

$$\begin{aligned}
 P[Y \leq y] &= P[X \leq y + t | X > t] \\
 &= \frac{P[X \leq y + t | X > t]}{P[X > t]} \\
 &= \frac{P[t < X \leq y + t]}{P[X > t]} \\
 &= \frac{F(y + t) - F(t)}{[1 - F(t)]} \\
 &= \frac{(1 - e^{-\lambda(y+t)}) - (1 - e^{-\lambda t})}{e^{-\lambda t}} \\
 &= \frac{e^{-\lambda t}[1 - e^{-\lambda y}]}{e^{-\lambda t}} \\
 &= 1 - e^{-\lambda y}
 \end{aligned}$$

10 Relationship of Poisson distribution and Exponential Distributions

Poisson distribution can be derived in the context of arrivals of requests to a system. Suppose that number of arrivals N_m to a system in an interval $(0, t]$ follows Poisson distribution.

$$\Rightarrow P[N = k] = \frac{(\lambda t)^k e^{-(\lambda t)}}{k!}$$

What is the distribution of the interarrival time (X)? We can pose this question as what is $P[X > t]$. In words, if one arrival happened at time 0, then the probability that there has been no arrival till at least time t is that there were 0 arrivals in the interval $(0, t]$. Thus $P[X > t] = P[N = 0] = e^{-(\lambda t)}$. Thus $P[X > t] = 1 - e^{-(\lambda t)}$. Thus X has $EXP(\lambda)$ distribution.

Poisson and Exponential distributions compliment each other. When the arrival rate is Poisson, the inter-arrival time is Exponential. A Poisson distribution with parameter λ has mean value as λ , whereas an Exponential distribution with parameter λ has mean value as $1/\lambda$.

11 Properties of Poisson and Exponential Distribution

- Sum of Poisson random variables is also Poisson (prove using PGFs)

Thus, if N_i are independent and $\text{poisson}(\lambda_i)$
then $N = \sum_{i=1}^n N_i$ is Poisson $(\sum_{i=1}^n \lambda_i)$.

- Minimum of exponential random variables is also exponential

Thus, if $X_i \sim EXP(\lambda_i)$ and are independent
then $X = \min(X_i) \sim EXP(\sum \lambda_i)$.

12 Distributions

12.1 Hypoexponential Distribution

Sum of exponential stages.

Consider $X = Y + Z$, where $Y \rightarrow \text{EXP}(\lambda_1)$ and $Z \rightarrow \text{EXP}(\lambda_2)$.

Then $X \rightarrow \text{HYPO}(\lambda_1, \lambda_2)$ and it has pdf:

$$f(t) = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} (e^{-\lambda_1 t} - e^{-\lambda_2 t}), \quad t > 0$$

Expectation:

Parameters: $\lambda_1, \lambda_2, \dots, \lambda_n$

$$E[X] = \sum_{i=1}^n \frac{1}{\lambda_i}$$

12.2 Erlang Distribution

Special case of HYPO where all stages have identical distribution.

$$f(t) = \frac{\lambda^r t^{r-1} e^{-\lambda t}}{(r-1)!}, \quad t > 0, \quad \lambda > 0, \quad r = 1, 2, \dots$$

Represented as $\text{ERLANG}(n, \lambda)$ which is equivalent to sum of n $\text{EXP}(\lambda)$.

Expectation:

Parameters: n, λ

$$E[X] = \frac{n}{\lambda}$$

12.3 Hyperexponential Distribution

Mixture of exponential phases where one and only one of the alternative phases is chosen. The pdf of a 2-way hyperexponential random variable is:

$$f(t) = \alpha_1 \lambda_1 e^{-\lambda_1 t} + \alpha_2 \lambda_2 e^{-\lambda_2 t}$$

Expectation:

Parameters: α_i, λ_i where $i = 1, 2, \dots, n$

$$E[X] = \sum_{i=1}^n \frac{\alpha_i}{\lambda_i}$$

Examples

Example 1: Consider `Array[1..n]`

$\lambda = \{1, 2, \dots, n\} \rightarrow$ Number of comparisons for a successful search on the array.

$$\begin{aligned} P_X(i) &= \frac{1}{n} \\ E[X] &= \frac{n+1}{2} \end{aligned}$$

Example 2:

$$P_X(i) = \frac{c}{i}, \quad i = 1, 2, \dots, n$$

$$\begin{aligned} \sum_{i=1}^n \frac{c}{i} &= 1 \\ \Rightarrow c &= \frac{1}{\sum_{i=1}^n \frac{1}{i}} \\ &= \frac{1}{\ln(n) + \alpha} \end{aligned}$$

$$\begin{aligned} E[X] &= \sum_{i=1}^n i - \frac{c}{i} \\ &= \frac{n}{\ln n + \alpha} \end{aligned}$$

This is used to model web traffic (hits).

Example 3: A cache of m out of a total n webpages.

$$P_X(i) = \frac{c}{i}, \quad i = 1, 2, \dots, n$$

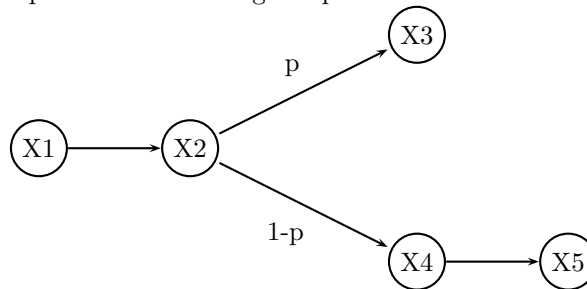
$$c = \frac{1}{H_n}$$

There will be m pages in the cache. The the probability P_{hit} that a request hits the cache is:

$$\begin{aligned} P_{hit} &= \sum_{i=1}^m m \frac{c}{i} \\ &= \frac{H_m}{H_n} \\ &= \frac{\ln m + \alpha}{\ln n + \alpha} \end{aligned}$$

13 Method of partial fractions of LST for mixtures and sums of exponentialdistributions

We are interested in finding the pdf for the following setup.



$X = \text{Total time}$

$$X = X_1 + X_2 + Y_3$$

where

$$Y_3 = X_3 \text{ with probability } p \quad (-60)$$

$$= X_4 + X_5 \text{ with probability } (1 - p) \quad (-59)$$

$$X_i \sim \text{Exp}(\lambda_i)$$

LST:

X is a r.v. with pdf $f(x), x \geq 0$

$$L_X(s) = \int_0^{\infty} e^{-sx} f(x) dx$$

$$Z = X + Y$$

$$L_Z(s) = L_X(s) \cdot L_Y(s)$$

(-62)

LST of $f(x) \Rightarrow \bar{f}(s)$

$$h(x) = \alpha f(x) + (1 - \alpha)g(x)$$

$$\text{i.e. } X = \alpha Y + (1 - \alpha)Z$$

$$\bar{h}(s) = \alpha \bar{f}(s) + (1 - \alpha)\bar{g}(s)$$

$$L_X(s) = \alpha L_Y(s) + (1 - \alpha)L_Z(s)$$

$$E[X^k] = \int_0^{\infty} x^k f(x) dx$$

$$= (-1)^k L_x^k(0)$$

$$= (-1)^k \frac{d^k L_X(s)}{ds} \text{ at } S = 0$$

$$E[X^2] - (E[X])^2 = (\text{Var})(X)$$

So, we have:

$$E[X] = E[X_1] + E[X_2] + p \cdot E[X_3] + (1 - p) \cdot (E[X_4] + E[X_5])$$

$$L_X(s) = L_{X_1}(s) \cdot L_{X_2}(s) \cdot L_{Y_3}(s)$$

$$L_{Y_3}(s) = p \cdot L_{X_3}(s) + (1 - p) \cdot L_{X_4}(s) \cdot L_{X_5}(s)$$

(-74)

Therefore,

$$L_X(s) = p \cdot L_{X_1}(s) \cdot L_{X_2}(s) \cdot L_{X_3}(s) + (1-p) \cdot L_{X_1}(s) \cdot L_{X_2}(s) \cdot L_{X_4}(s) \cdot L_{X_5}(s) \quad (-74)$$

$$X \sim Exp(\lambda_i)$$

$$\begin{aligned} L_{X_i}(s) &= \frac{\lambda_i}{\lambda_i + s} \\ &= L_X(s) \\ &= \frac{p\lambda_1\lambda_2\lambda_3}{(\lambda_1 + s)(\lambda_2 + s)(\lambda_3 + s)} + \frac{(1-p)\lambda_1\lambda_2\lambda_4\lambda_5}{(\lambda_1 + s)(\lambda_2 + s)(\lambda_4 + s)(\lambda_5 + s)} \end{aligned} \quad (-76)$$

This can be seen as

$$L_X(s) = \frac{N(s)}{D(s)} = \frac{\sum_{i=1}^d C_i}{(s + a_i)} \quad (-76)$$

The first term of the RHS can be solved by partial fractions method to give:

$$\begin{aligned} \frac{p\lambda_1\lambda_2\lambda_3}{(\lambda_1 + s)(\lambda_2 + s)(\lambda_3 + s)} &= \frac{C_1}{(\lambda_1 + s)} + \frac{C_2}{(\lambda_2 + s)} + \frac{C_3}{(\lambda_3 + s)} \\ C_1 &= \frac{p\lambda_1\lambda_2\lambda_3}{(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)} \\ C_2 &= \frac{p\lambda_1\lambda_2\lambda_3}{(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_2)} \\ C_3 &= \frac{p\lambda_1\lambda_2\lambda_3}{(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)} \end{aligned} \quad (-79)$$

So, the solution is :

$$f_X(x) = C_1 e^{-\lambda_1 x} + C_2 e^{-\lambda_2 x} + C_3 e^{-\lambda_3 x} + \dots + \sum_{i=1,2,3,4} B_i e^{-\lambda_i x} \quad (-79)$$

14 Observational/Operational Laws

Some measurable quantities of a queueing system are:

- T Observation/measurement period
- A Number of arrivals in time T
- C Number of completions/departures in time T
- B Total time that system was busy in time T

$$\begin{aligned}
\text{Arrival Rate } \lambda &= \frac{A}{T} \\
\text{Throughput } X &= \frac{C}{T} \\
\text{Mean Service Time } \tau &= \frac{B}{C} \\
\text{Utilization } \rho &= \frac{B}{T} \\
&= \frac{B}{C} \cdot \frac{C}{T} \\
&= \tau X
\end{aligned}$$

For a stable system, $X = \lambda$ and $\rho = \lambda\tau$.

Waiting Time: Time spent in queue

Residence/System/Response Time: Waiting Time + Service Time

Queue Length: Number of customers in the system (including in service)

Example: At a router in a network, let average packet length be L , overall bit rate of incoming links be B and the arrival rate of packets be λ . Then for stability:

$$\frac{\lambda L}{B} < 1$$

14.1 Little's Law

Consider a lossless, stable system where:

N Average queue length (number of jobs)

λ Arrival rate

R Average response time

The *Little's Law* states that:

$$N = \lambda R$$

This is an elegant and extremely useful formula which directly relates the average queue length in steady state to arrival rate (or throughput) and average response time in the system.

Intuitively, Little's Law makes sense—for example, if a petrol station, on the average, gets 40 customers in an hour and takes an average of 0.2 hours (12 minutes) to serve a customer, then at any given time there are 8 customers in the petrol station.

14.1.1 Derivation

τ = Packet transmission time

For a Stable system

$$\begin{aligned}
\frac{1}{\lambda} &> \tau \\
\lambda &< \frac{1}{\tau}
\end{aligned}$$

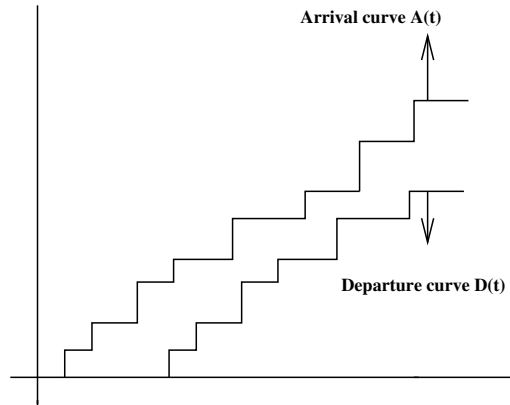


Figure 3: Little's law

Area by adding horizontal rectangles =

$$\left(\frac{\sum_{i=1}^A R_i}{A} \right) A = RA$$

Area by adding vertical rectangles =

$$\left[\sum_{i=1}^k N_i \left(\frac{T_i}{T} \right) \right] T = N\tau$$

where $\sum_{i=1}^k N_i \left(\frac{T_i}{T} \right)$ is average queue length
Equating two areas

$$\begin{aligned} RA &= N\tau \\ N &= \frac{A}{\tau} R \\ N &= \lambda R \end{aligned}$$

15 M/M/1 queue

(Refer to Section 8.2.1 Trivedi 2/e for details)

Consider the system with one server and infinite queue capacity. The arrivals are poisson with rate λ and the service time is exponentially distributed with rate μ .

Let $N(t)$ = number of jobs in the system at time t

Suppose $\Pr[2 \text{ or more jobs arriving / departing in } (t, t + \Delta t)] \rightarrow 0$ for $\Delta t \rightarrow 0$

$\Pr[\text{Arrival in } \Delta t]$ is $\lambda\Delta t$.

$\Pr[\text{Departure in } \Delta t]$ is $\mu\Delta t$.

Then $N(t + \Delta t) = k$ if

A: $N(t) = k$ and no jobs arrive/depart in $(t, t + \Delta t]$ w.p. $1 - \lambda\Delta t - \mu\Delta t$

B: $N(t) = k - 1$ and 1 job arrives in $(t, t + \Delta t]$ w.p. $\lambda\Delta t$

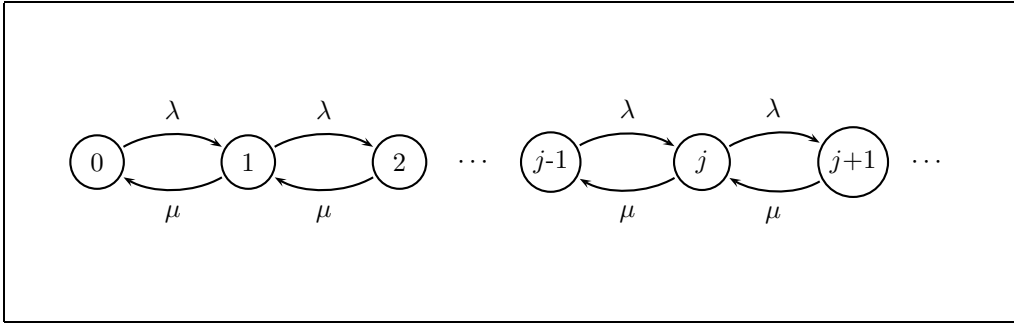


Figure 4: State diagram of $M/M/1$ queue

C: $N(t) = k + 1$ and 1 job departs in $(t, t + \Delta t]$ w.p. $\mu\Delta t$

Unconditional Probability:

Let $P_k(t) = P[N(t) = k]$

$PA = P_k(t)(1 - \lambda\Delta t - \mu\Delta t)$ $PB = P_{k-1}(t)\lambda\Delta t$

$PC = P_{k+1}(t)\mu\Delta t$

$$\begin{aligned} P_k(t + \Delta t) &= PA + PB + PC \\ &= P_k(t) + [\lambda P_{k-1}(t) - (\lambda + \mu)P_k(t) + \mu P_{k+1}(t)]\Delta t \end{aligned}$$

$$\begin{aligned} \text{The rate of change of probability} &= \lim_{\Delta t \rightarrow 0} \frac{P_k(t + \Delta t) - P_k(t)}{\Delta t} = \frac{dP_k(t)}{dt} \\ &= \lambda P_{k-1}(t) - (\lambda + \mu)P_k(t) + \mu P_{k+1}(t) \end{aligned}$$

We are often interested in the state of the system as $t \rightarrow \infty$.

If steady-state exists, $\lim_{t \rightarrow \infty} \frac{dP_k(t)}{dt} \rightarrow 0$ and $\lim_{t \rightarrow \infty} P_k(t) \rightarrow P_k$

where P_k is “equilibrium” or “steady-state” distribution.

$$\begin{aligned}
 (\lambda + \mu)P_k &= \lambda P_{k-1} + \mu P_{k+1} \quad , \quad k = 1, 2, \dots \\
 \lambda P_0 &= \mu P_1 \\
 (\lambda + \mu)P_1 &= \lambda P_0 + \mu P_2 \\
 (\lambda + \mu)P_2 &= \lambda P_1 + \mu P_3 \\
 &\vdots \\
 P_k &= \frac{\lambda}{\mu} P_{k-1}
 \end{aligned}$$

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

Then the above equations can be written as

$$\begin{aligned}
 P_1 &= \frac{\lambda}{\mu} P_0 = \rho P_0 \\
 P_2 &= \frac{\lambda}{\mu} P_1 = \rho P_1 \\
 P_3 &= \frac{\lambda}{\mu} P_2 = \rho P_2 \\
 &\vdots \\
 P_k &= \rho P_{k-1} \\
 &= \rho \cdot \rho P_{k-2} \\
 &\vdots \\
 &= \rho^k P_0 \quad , \quad k = 1, 2, \dots, \infty
 \end{aligned}$$

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

$$\sum_{k=0}^{\infty} \rho^k P_0 = 1$$

$$\Rightarrow P_0 = \frac{1}{\sum_{k=0}^{\infty} \rho^k}$$

$$P_0 = 1 - \rho$$

$$\rho = 1 - P_0 \dots \text{Probability that the server is busy}$$

$$P_k = \rho^k \cdot (1 - \rho)$$

Thus P_k is often interpreted as “Modified Geometric”

If N is the number of jobs in the system at steady state, then

$$E[N] = \frac{1 - (1 - \rho)}{1 - \rho} = \frac{\rho}{1 - \rho} = \text{Avg. Queue length}$$

$$E[R] = \frac{E[N]}{\lambda} = \frac{1}{\mu(1 - \rho)} = \text{Avg. Response time}$$

16 Stochastic Processes

A stochastic process is a family of random variables $X(t)|t \in T$ defined on a given probability space, indexed by the parameter t , where $t \in T$. Values of $X(t)$ define the state space and the parameter t is time in many cases. Based on the combinations of state and parameter types, we have the following classification of stochastic processes:

State \ Parameter	Discrete	Continuous
Discrete	Discrete-parameter chain	Discrete-parameter continuous state process
Continuous	Continuous-parameter chain	Continuous-parameter continuous state process

Examples:

DPDS: N_k : number of jobs in the system seen by the k^{th} arrival

DPCS: waiting time of the k^{th} customer

CPDS: number of jobs in the system at time t

CPCS: remaining work in the system at time t

16.1 Markov Process

$X(t)|t \in T$ is called a markov process if for any $t_0 < t_1 < t_2 < \dots < t_n$

$$P[X(t) \leq x | X(t_n) = x_n, \dots, X(t_1) = x_1, X(t_0) = x_0] = P[X(t) \leq x | X(t_n) = x_n]$$

Further if $P[X(t) \leq x | X(t_n) = x_n] = P[X(t - t_n) \leq x | X(0) = x_n]$,

the markov proces is said to be time-homogenous.

The markov property of a process implies that the distribution of time spent in a state is exponential.

17 Continuous-parameter markov chain

Let the state space be $\{0, 1, 2, \dots\}$ parameter $t \geq 0$

$$P[X(t) = x | X(t_n) = x_n, \dots, X(t_0) = x_0] = P[X(t) = x | X(t_n) = x_n]$$

This process is characterized by

1. The distribution of the initial state of the system $[P_0(0), P_1(0), \dots] = P(0)$
2. Transition probabilities $P_{ij}(v, t) = P[X(t) = j | X(v) = i]$, $0 \leq v \leq t$, $i, j = 0, 1, 2, \dots$

$$P_{ij}(t, t) = 1, \quad \text{if } i = j$$

$$= 0, \quad \text{otherwise}$$

If the markov chain is time-homogeneous, $P_{ij}(v, t)$ depends only on $t - v$.

$$P_{ij}(t) = P[X(t + u) = j | X(u) = i], \quad u \geq 0$$

Define $P_j(t) = P[X(t) = j]$ (unconditional probability of being in state j at time t)
 The following is satisfied:

$$\begin{aligned} \sum_j P_{ij}(v, t) &= 1 \quad \forall i, 0 \leq v \leq t \\ \sum_j P_j(t) &= 1, \quad t \geq 0 \\ P_j(t) &= \sum_i P[X(t) = j | X(v) = i] P[X(v) = i], \quad 0 \leq v < t \\ &= \sum_i P_{ij}(v, t) P_i(v) \\ \text{for } v = 0 \\ &= \sum_i P_{ij}(0, t) P_i(0), \end{aligned}$$

which is completely determined by P_{ij} 's and $P_j(0)$.

If Q is the transition rate matrix (generator matrix) of the CTMC then:

$$\begin{aligned} Q_{ij} &= [q_{ij}] & i \neq j & p_{ii}(0) = 1 \\ Q_{ii} &= -q_i & & p_{ij}(0) = 1 \\ P(t) &= [P_j(t)] & \text{where } q_i &= \sum_{j \in S} q_{ij} \end{aligned}$$

Thus the equation in matrix form:

$$\frac{dP(t)}{dt} = P(t)Q \quad (-96)$$

Given the initial condition $P(0)$, a solution to the above equation is:

$$P(t) = P(0)e^{Qt}$$

We are still interested in steady state:

Theorem: For an irreducible CTMC the following limits always exist and are independent of the initial state i .

$$j = \lim_{t \rightarrow \infty} p_{ij}(t) = \lim_{t \rightarrow \infty} p_j(t), \quad i, j \in I$$

Note: A CTMC is said to be irreducible if every state is reachable from every other state. A state j is reachable from state i if $p_{ij}(t) > 0$ for some $t > 0$.

$$\begin{aligned} 0 &= \sum_{k \neq j} p_k q_{kj} - p_j q_j \\ \Rightarrow p_j q_j &= \sum_{k \neq j} p_k q_{kj} \\ \Rightarrow \text{rate out of state } j &= \text{rate in to state } j \end{aligned}$$

$$\sum p_j = 1$$

M/M/s/∞ queue

State: No of customers in the system

Interarrival $\sim \text{Exp}(\lambda)$

Service $\sim \text{Exp}(\mu)$

The arrival and service times are exponential with rates λ and μ respectively.

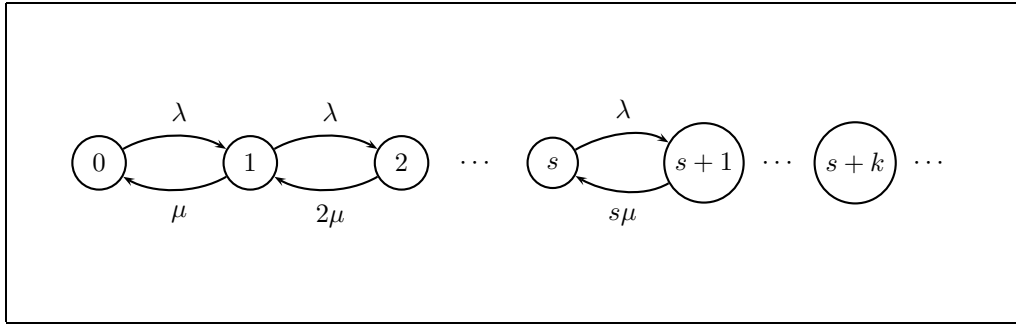


Figure 5: Markov chain for M/M/s/∞ queue

Stable system:- [Flow in = Flow out]

$$\begin{aligned} (\lambda + 2\mu)P_2 &= \lambda P_1 + 3\mu P_3 \\ \lambda P_2 &= 3\mu P_3 \\ \lambda P_i &= (i+1)\mu P_{i+1} \quad 0 \leq i \leq s-1 \\ \lambda P_{s+i} &= s\mu P_{s+i+1} \quad i \geq s \end{aligned}$$

By repeated substitutions:

$$\begin{aligned} P_i &= \frac{(s\rho)^i}{i!} P_0 \quad 0 \leq i \leq s \\ P_{s+i+1} &= \left(\frac{\lambda}{s\mu}\right)^{i+1} \left(\frac{s\rho}{s!}\right)^s P_0 \end{aligned}$$

P_0 can be found by using " Sum of probabilities of being in any state = 1"

$$\sum_j P_j = 1$$

$$P_0 = \left[\sum_{i=0}^s \frac{(s\rho)^i}{i!} + \sum_{i=0}^{\infty} \frac{\rho^{i+1}(s\rho)^s}{s!} \right]^{-1}$$

Comparison between 2 M/M/1 Queues and 1 M/M/2 Queues

where λ is arrival rate and μ is the service rate for the queues.

1. Avg. Utilization of each server:

- 2 M/M/1 Queues: $\frac{\lambda}{2\mu}$
- M/M/2 Queue: $\frac{\lambda}{2\mu}$
- 2. Throughput (if $\lambda < 2\mu$)
 - 2 M/M/1 Queues: λ
 - M/M/2 Queue: λ
- 3. Max Throughput
 - 2 M/M/1 Queues: 2μ
 - M/M/2 Queue: 2μ
- 4. Response time
 - 2 M/M/1 Queues: $\frac{2(2\mu+\lambda)}{4\mu^2-\lambda^2}$
 - M/M/2 Queue: $\frac{4\mu}{4\mu^2-\lambda^2}$

17.1 M/M/s/0 queue: [No waiting Queue nodes]

State: No of busy servers in the system

Interarrival $\sim \text{Exp}(\lambda)$

Service $\sim \text{Exp}(\mu)$

The arrival and service times are exponential with rates λ and μ respectively.

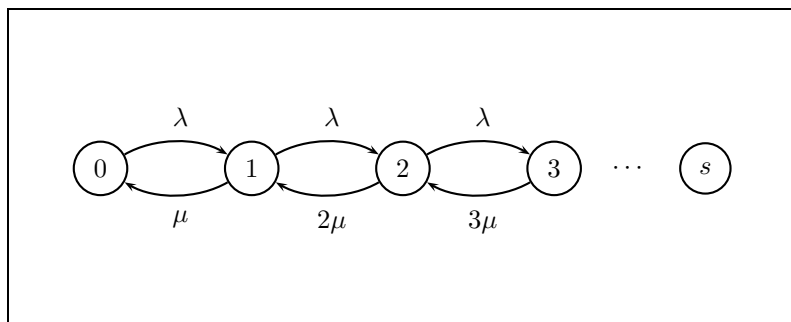


Figure 6: Markov chain for M/M/s/0 queue

Balanced equations:

$$\begin{aligned}
 P_0\lambda &= P_1\mu \\
 P_2\lambda &= P_33\mu \\
 &\vdots \\
 &\vdots \\
 &\vdots \\
 P_{s-1}\lambda &= P_s s\mu
 \end{aligned}$$

The general formula is:

$$P_i = \left(\frac{\lambda}{\mu}\right)^i \frac{P_0}{i!}$$

Blocking Probabilitiy

GSM: Circuit switched network

Measure of interest:

Blocking Probabilitiy $[P_s]$ = Probabilitiy of all lines busy

$$\begin{aligned} \sum_{i=0}^s P_i &= 1 \\ \Rightarrow \sum_{i=0}^s \frac{\rho^i}{i!} P_0 &= 1 \\ \Rightarrow P_0 &= \frac{1}{\sum_{i=0}^s \frac{\rho^i}{i!}} \\ \therefore P_i &= \frac{\frac{\rho^i}{i!}}{\sum_{i=0}^s \frac{\rho^i}{i!}} \end{aligned}$$

The above formula is also called **Erlang-B** formula and gives the Blocking probability.

17.2 M/M/1/n Queueing system

λ - Call arrival rate

$Exp(\mu)$ - Avg. talking time

Measures:

- Waiting / Response time
- Blcking Probability
- Utilization

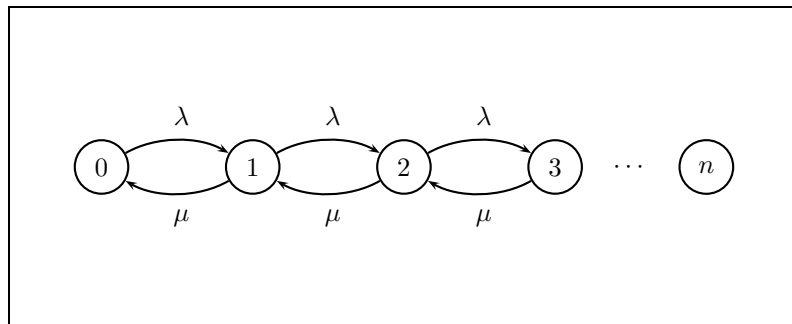


Figure 7: Markov chain for M/M/1/n queue

State: No.of Customers in the queue.

Balanced equations:

$$\lambda P_{i-1} = \mu P_i P_i = \left(\frac{\lambda}{\mu}\right)^i P_0$$

By repeated Substituting:

$$P_i = (\rho)^i P_0$$

Using the theorem "Sum of Probabilities at all states is 1" and *PASTA theorem* we get:

$$P_0 = \frac{1 - \rho}{1 - \rho^{n+1}}$$

Waiting time:

$$E(R) = \frac{E(N)}{\text{Throughput}} = \frac{n}{\mu}$$

Throughput:

$$T = \lambda * (1 - \text{Blockingprobability}) = \lambda * (1 - P_n) = \mu * (1 - P_0)$$

Utilization:

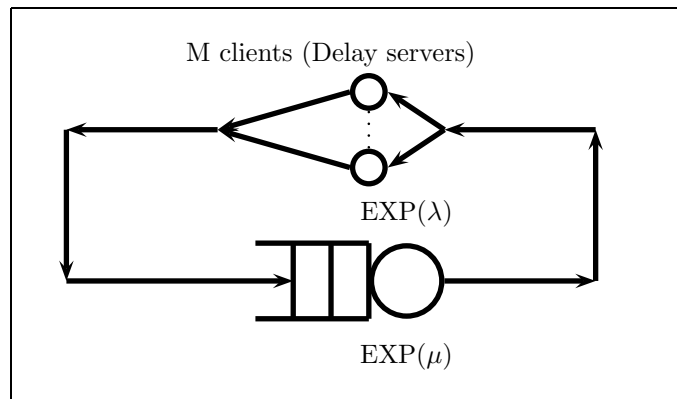
$$\rho = \lambda / \mu E[R] = \frac{E[N]}{1 - P_0}$$

18 Client Server model

Refer Ex. 8.9 in Edition 2.

Think time: Time between different requests from the same user.

i : No. of requests at the server.



$$\begin{aligned} \mu P_i &= (M - i + 1) \lambda P_{i-1} \\ P_i &= \frac{(M - i + 1) \lambda}{\mu} P_{i-1} \\ &= \frac{(M - i + 1) \lambda}{\mu} \cdot \frac{(M - i + 2) \lambda}{\mu} \dots \frac{M \lambda}{\mu} P_0 \\ &= \left(\frac{\lambda}{\mu} \right)^i \cdot \frac{M!}{(M - i)!} \cdot P_0 \\ &= \rho^i \cdot \frac{M!}{(M - i)!} \cdot P_0 \end{aligned}$$

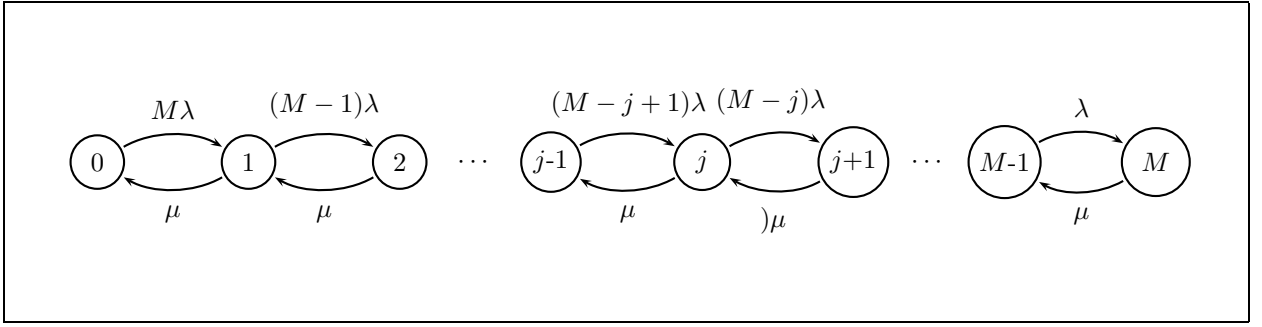


Figure 8: Markov chain for client server system

$$U = 1 - P_0$$

$$P_0 = \frac{1}{\sum_{i=0}^M \frac{\rho^i M^i}{(M-i)!}}$$

$$E[T] = U\mu$$

$$\text{Average request issue rate by 1 client} = \frac{1}{E[R] + \frac{1}{\lambda}}$$

$$\text{Total request arrival rate} = \frac{M}{E[R] + \frac{1}{\lambda}}$$

$$E[T] = \frac{M}{E[R] + \frac{1}{\lambda}}$$

$$E[R] = \frac{M}{E[T]} - \frac{1}{\lambda}$$

$$= \frac{M}{U\mu} - \frac{1}{\lambda}$$

$$= \frac{M\tau}{U} - \frac{1}{\lambda} \text{ At low load, } \lim_{U \rightarrow 0} E[R] = \tau$$

$$\text{At high load, } \lim_{U \rightarrow 1} E[R] = M\tau - \frac{1}{\lambda}$$

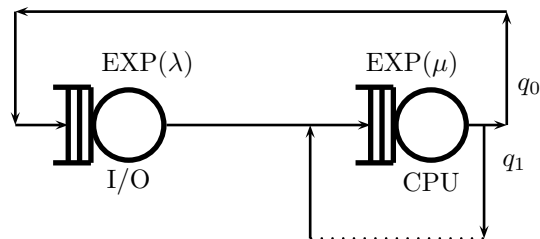
Low load asymptote: τ

High load asymptote: $M\tau - \frac{1}{\lambda}$

$$\begin{aligned}
M\tau - \frac{1}{\lambda} &= \tau \\
M\tau &= \tau + \frac{1}{\lambda} \\
M^* &= 1 + \frac{1}{\lambda\tau} \\
&= 1 + \frac{\mu}{\lambda}
\end{aligned}$$

M^* is called the “saturation number”.

19 Cyclic Queuing Model of a Multiprogramming System



Refer Ex. 8.6 of 2nd edition, Ex. 8.6 of 1st edition.

n : limit of multiprogramming.

n is fixed. It is the limit met at peak load.

$$\begin{aligned}
T_{\text{system}} &= q_0 T_{\text{cpu}} \\
T_{\text{cpu}} &= P_{\text{busy}} \cdot \mu + P_{\text{idle}} \cdot 0 \\
&= U_{\text{cpu}} \cdot \mu \\
U_{\text{cpu}} &= P(\text{number of jobs at cpu} > 0) \\
&= 1 - P_0
\end{aligned}$$

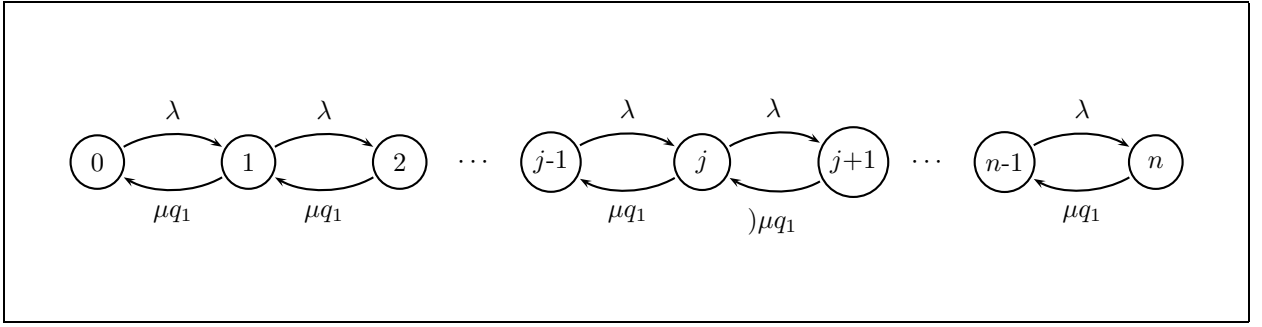


Figure 9: Markov chain for a multiprogramming system

$$\mu q_1 P_i = \lambda P_{i-1}, \quad i=1,2,\dots,n$$

$$P_i = \frac{\lambda}{\mu q_1} P_{i-1} \\ = \rho P_{i-1},$$

where $\rho = \frac{\lambda}{\mu q_1}$

$$P_i = \rho^i P_0, \quad i = 0,1,\dots,n$$

$$\sum_{i=0}^n P_i = 1$$

$$P_0 = \frac{1}{\sum_{i=0}^n \rho^i} \\ = \begin{cases} \frac{1-\rho}{1-\rho^{n+1}}, & \text{for } \rho \neq 1 \\ \frac{1}{n+1}, & \text{for } \rho = 1 \end{cases}$$

$$\sum_{i=0}^n = \begin{cases} \frac{1-\rho^{n+1}}{1-\rho}, & \text{for } \rho \neq 1 \\ n+1, & \text{for } \rho = 1 \end{cases}$$

$$U_0 = \begin{cases} 1 - \left(\frac{1-\rho}{1-\rho^{n+1}} \right) = \frac{\rho-\rho^{n+1}}{1-\rho^{n+1}}, & \text{for } \rho \neq 1 \\ \frac{n}{n+1}, & \text{for } \rho = 1 \end{cases}$$

$$T_{\text{system}} = U_{\text{cpu}} \mu q_0 \\ = \frac{U_{\text{cpu}}}{\frac{1}{\mu q_0}} \\ = \frac{U_{\text{cpu}}}{E[\text{total CPU time before exit per job}]}$$

Also,

$$\begin{aligned}
T_{\text{system}} &= \frac{U_{\text{I/O}}}{\text{E}[\text{total I/O time before exit per job}]} \\
U_{\text{I/O}}\lambda + T_{\text{system}} &= U_{\text{cpu}}\mu \\
U_{\text{I/O}}\lambda + U_{\text{cpu}}\mu q_0 &= U_{\text{cpu}}\mu \\
U_{\text{I/O}}\lambda &= -T_{\text{system}} + \frac{T_{\text{system}}}{q_0} \\
U_{\text{I/O}}\lambda &= T_{\text{system}} \left(\frac{1 - q_0}{q_0} \right) \\
&= \frac{q_1}{q_0} T_{\text{system}}
\end{aligned}$$

$$\begin{aligned}
T_{\text{system}} &= \frac{q_0 \lambda U_{\text{I/O}}}{q_1} \\
&= \frac{U_{\text{I/O}}}{\left(\frac{q_1}{q_0}\right) \left(\frac{1}{\lambda}\right)} \\
&= \frac{U_{\text{I/O}}}{\text{E}[\text{total I/O time before exit per job}]}
\end{aligned}$$

Number of visits to I/O has a modified geometric distribution with parameter p_0 .

$$\begin{aligned}
\rho &= \frac{\lambda}{\mu q_1} \\
&= \frac{1}{\mu q_0} \cdot \frac{\lambda q_0}{q_1} \\
&= \frac{\frac{1}{\mu q_0}}{\frac{q_1}{\lambda q_0}} \\
&= \frac{\text{E}[B_{\text{cpu}}]}{\text{E}[B_{\text{I/O}}]}
\end{aligned}$$

20 Response time distribution in queuing models

M/M/1 response time distribution

If an arriving job finds k in system,

$$R = S'_1 + S_2 + S_3 + \dots + S_k + S$$

where S'_1 represents the remaining service of the job in service.
 S_k represents service time for the k th job in the queue.

Each job has a service distribution of $Exp(\mu)$. This reduces to $Erlang(k + 1, \mu)$.

$$L_R(s|N = k) = \left(\frac{\mu}{\mu + s} \right)^{k+1}$$

$$\therefore L_R(s) = \sum_{k=0}^{\infty} \left(\frac{\mu}{\mu + s} \right)^{k+1} \rho^k (1 - \rho)$$

$$= \frac{\mu(1 - \rho)}{\mu(1 - \rho) + s}$$

$$\therefore L_R(s) \sim Exp(\mu(1 - \rho))$$

Therefore Response time distribution is also Exponential with parameter $(\mu(1 - \rho))$.

Defective distribution

For M/M/1/n, some packets wait for $t = \infty$ (if queue is full) as

$$\int_0^{\infty} f(t) dt < 1$$

there must be no weight at $f(\infty)$, but we have with some probability. Hence it is called a Defective distribution.

M/M/s/ ∞ :

Let R denote waiting time. $P(R < t) = P(\text{No queueing})P(R < t \text{— No queueing}) + (1 - P(\text{No queueing}))P(R < t \text{— Queueing})$

If there is not queueing, response time = service time, which is = $EXP(\mu)$

If there is queueing, response time = waiting time + service time

Let k be the number of customers seen in queue (in addition to s in service), by arriving customer.

If $k = 0$, that is all servers busy, but none in queue, you have to wait until the first among the s in service, departs. That is, you have to wait for Y time, where Y is minimum of s , $EXP(\mu)$ random variables. Thus Y is $EXP(s\mu)$.

Similarly, if you see 1 in queue, you have two wait for $Y + Y$ time.

Thus, if you see k in queue, you have to wait for sum of $k + 1$, $EXP(s\mu)$ random variables. That is, $ERLANG(k + 1, s\mu)$ distribution.

$$F_{w|N=k+s}(t) = P(\text{waitingtime} < t | k \text{ in queue, given queueing happens}) \sim Erlang(k + 1, s\mu)$$

Conditional probability that there are k in queue, given that there is queueing, is unconditional probability of $s + k$ in system, divided by probability that number of customers in system is $\geq s$. This is =

$$\frac{p_{s+k}}{\sum_{k=0}^{\infty} p_{s+k}}$$

This can be shown to be = $(1 - \rho)\rho^k$.

Then, conditional lablaca transform of Waiting time distribution, given queueing =

$$L_{w|queueing}(x) = \sum_{k=0}^{\infty} (1-\rho)\rho^k \left(\frac{s\mu}{s\mu+x} \right)^{k=1}$$

$$= \frac{s\mu - \lambda}{(s\mu - \lambda) + x}$$

$$\therefore L_{w|queueing}(x) \sim Exp(s\mu - \lambda)$$

Finally, remember that response time given queueing = waiting time as above + service time.
That is, Response time = Hypo Exponential distribution with parameters μ and $s\mu - \lambda$.
Finally, apply law of total probability with probability of queueing vs not queueing.

Lecture 23 (13-09-2005)

Discrete-Time Markov Chains

(Refer to Chapter 7, Trivedi for details)

A Markov process with discrete state space and discrete parameter space is known as a Discrete-Time Markov Chain (DTMC).

- Random variables $\{X_n \mid n = 0, 1, 2, \dots\}$ represent the state of the system at time step n . X_n takes discrete values.
- From the memoryless property, we have $P(X_n = i_n \mid X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}) = P(X_n = i_n \mid X_{n-1} = i_{n-1})$
- Given the present state of the system, the future is independent of its past.
- Let $p_j(n) = P(X_n = j)$ denote the probability that the system is in state j at time step n .
- Let $p_{jk}(m, n) = P(X_n = k \mid X_m = j)$, $0 \leq m \leq n$ denote the probability that the process makes a transition from state j at step m to state k at step n .
- For homogeneous Markov chains, $p_{jk}(n) = P(X_{m+n} = k \mid X_m = j)$
- One-step transition probabilities: $p_{jk} = p_{jk}(1) = P(X_n = k \mid X_{n-1} = j)$, $n \geq 1$
- Zero-step transition probabilities:

$$p_{jk}(0) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$$

Example 1: State 0 represents webserver is down and state 1 represents webserver up. The system is observed at the end of each day.

Sum of outgoing arcs is unity.

Let d be the number of steps in which, starting from any given state, the same state can be reached. A DTMC is said to be *periodic* if $d > 1$ and *aperiodic* otherwise. The DTMC of Example 1 is aperiodic.

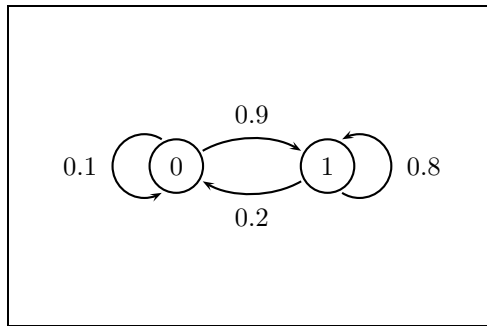


Figure 10: State diagram of Example 1

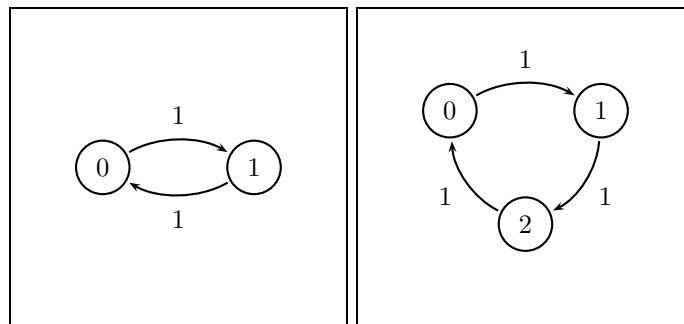


Figure 11: Examples of periodic DTMCs

One-step Transition Probability Matrix

$$P = [p_{ij}] = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

$$0 \leq p_{ij} \leq 1$$

If $\sum_j p_{ij} = 1$ then the matrix is a stochastic matrix.

Chapman Kolmogorov Equation

$$\begin{aligned}
 p_{ij}(m+n) &= \sum_k p_{ik}(m) \cdot p_{kj}(n) \\
 p_{ij}(n) &= \sum_k p_{ik}(1) \cdot p_{kj}(n-1) \\
 \Rightarrow P(n) &= \sum_k p_{ik} \cdot p_{kj}(n-1) \\
 &= P \cdot P(n-1) \\
 &= P \cdot P \cdot P(n-2) \\
 &\vdots \\
 &= P^n
 \end{aligned}$$

$$P(n) = [p_{ij}(n)]$$

Unconditional probability at n_{th} step

$$\begin{aligned} p_j(n) &= \sum_i p_i(0)p_{ij}(n) \\ \mathbf{p}(n) &= [p_0(n), p_1(n), \dots, p_j(n), \dots] \\ \Rightarrow \mathbf{p}(n) &= \mathbf{p}(0)P(n) \\ \text{and } \mathbf{p}(n) &= \mathbf{p}(0)P^n \end{aligned}$$

Limiting State Probabilities In the condition that the Markov chain is irreducible, aperiodic, recurrent and non-null, the following limit exists and is independent of initial probability:

$$v_j = \lim_{n \rightarrow \infty} p_j(n), \quad j = 0, 1, \dots$$

Now we have:

$$\begin{aligned} \lim_{n \rightarrow \infty} p_j(n) &= \lim_{n \rightarrow \infty} \sum_i p_i(n-1)p_{ij} \\ \Rightarrow v_j &= \sum_i v_i p_{ij} \\ \mathbf{v} &= \mathbf{v}P \\ \text{where } \sum_i v_i &= 1 \quad \text{and} \quad v_j \geq 0 \end{aligned}$$

Ex: Page 378. Q . 2.

Lecture 24 (15-09-2005)

Some Definitions and explanations

(Refer Kishore Trivedi 2nd Ed. pg 318 onwards)

Transient State: A state i is said to be ‘transient’ (or ‘nonrecurrent’) if and only if there is a positive probability that the process will not return to this state.

For example, if we model a program as a Markov chain, then all but the final state will be transient states. Otherwise, the program has an infinite loop. In general, for a finite Markov chain, we expect that after a sufficient number of steps the probability that the chain is in any transient state approaches zero independent of the initial state.

Let X_{ji} be the number of visits to the state i , starting at j . Then it can be shown that:

$$E[X_{ji}] = \sum_{n=0}^{\infty} p_{ji}(n)$$

It follows that if the state i is a transient state, then $\sum_{n=0}^{\infty} p_{ji}(n)$ is finite for all j , hence $p_{ji}(n)$ approaches 0 as n approaches infinity.

Recurrent State: A state i is said to be ‘recurrent’ if and only if, starting from state i , the process eventually returns to state i with probability **one**.

$E[X_{ij}] = \sum_{n=0}^{\infty} p_{ji}(n)$ is infinite.

For recurrent states, the time to reentry is important. Let f_{ij} be the conditional probability that the first visit to state j from state i occurs in exactly n steps. If $i = j$, then we refer to f_{ij} as the probability that the first return to state i occurs in exactly n steps. These probabilities are related to the transition probabilities by:

$$p_{ij}(n) = \sum_{k=1}^n f_{ij} p_{jj}(n-k), \quad n \geq 1.$$

Let f_{ij} denote the probability of ever visiting state j , starting from state i . Then:

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}(n)$$

It follows that i is recurrent if $f_{ii} = 1$ and transient if $f_{ii} < 1$.

If $f_{ii} = 1$, define the mean recurrence time of state i by:

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}(n)$$

A recurrent state i is said to be **recurrent non-null** if its mean recurrence time μ_i is finite and is said to be **recurrent null** if its mean recurrence time is infinite.

Period: For a ‘recurrent’ state i , $p_{ii}(n) > 0$ for some $n \geq 1$. The period of state i , denoted by d_i is defined as the greatest common divisor of the set of positive integers n such that $p_{ii}(n) > 0$.

A recurrent state i is said to be **aperiodic** if its period $d_i = 1$, and **periodic** if $d_i > 1$.

A state i is said to be an **absorbing** state if and only if $p_{ii} = 1$

Irreducible Markov Chain: A Markov chain is said to be ‘irreducible’ if every state can be reached from every other state in a finite number of steps. In other words, for all $i, j \in I$, there is an integer $n \geq 1$ such that $p_{ij}(n) > 0$. Feller has shown that all states of an irreducible Markov chain are of the same type. Thus, if one state of an irreducible chain is aperiodic then so are all the states and such a Markov chain is called ‘aperiodic’.

The n -step transition probabilities $p_{ij}(n)$ of finite, irreducible, aperiodic Markov chains become independent of i and n as $n \rightarrow \infty$

$$v_j = \lim_{n \rightarrow \infty} p_{ij}(n)$$

Assume that for a given Markov chain the limiting probabilities v_j exist for all state $j \in I$ (where v_j does not depend on the initial state i). Then, $\sum_{j \in I} v_j \leq 1$. Furthermore, either all $v_i = 0$ (this can happen only for a chain with an infinite number of states) or $v_i = 1$. The latter case represents a **steady-state**.

$$v_j = \sum_i v_i p_{ij}, \quad j = 0, 1, 2, \dots$$

In matrix notations:

$$v = vP$$

For aperiodic DT Markov chain,

$$v_j = \lim_{n \rightarrow \infty} p_j(n) \text{ exists.}$$

For irreducible, aperiodic DT Markov chain,

$$v_j = \lim_{n \rightarrow \infty} p_j(n) \text{ exists and is independent of initial probability.}$$

For irreducible, recurrent non-null, aperiodic DT Markov chain, limiting probability exists, is independent of initial probability and is given by:

$$v = vP, \quad \sum v_i = 1$$

Also see the solved example of section 7.5.1 on page 326 of K.T. ed.-2 book.

DONE TILL HERE

21 M/G/1 queuing system

(Refer to Section 7.7 Trivedi 2/e)

This is a single-server queuing system whose arrival process is Poisson with the average arrival rate λ . The job service times are independent and identically distributed with the distribution function F_B and pdf f_B . Jobs are scheduled for service in their order of arrival; that is, scheduling is FCFS.

Let $N(t)$ denote the number of jobs in the system at time t . If $N(t) \geq 1$, then a job is in service, and since the general service time distribution need not be memoryless, besides $N(t)$, we also require knowledge of time spent by the job in service in order to predict the future behaviour of the system. It follows that the stochastic process $N(t)|t \geq 0$ is **NOT** a Markov chain.

To simplify the state description, we take a snapshot of the system at times of departure of jobs.

$$\begin{aligned} Y_{n+1} &= \text{No. of arrivals during } (n+1)^{\text{th}} \text{ service.} \\ P[Y_{n+1} = a_j] &= a_j \\ X_{n+1} &= Y_{n+1}, \text{ if } X_n = 0 \\ X_{n+1} &= X_n - 1 + Y_{n+1}, \text{ if } X_n > 0 \\ P_{ij} &= P[X_{n+1} = j | X_n = i] \\ P_{ij} &= P[X_n - 1 + Y_{n+1} = j | X_n = i] \text{ if } i > 0 \\ P_{ij} &= P[Y_{n+1} = j - i + 1 | X_n = i] = a_{j-i+1} \text{ if } j \geq i - 1 \\ P_{0j} &= P[Y_{n+1} = j] = a_j, \text{ if } i = 0 \end{aligned}$$

One-step Transition Probability Matrix

$$P = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ 0 & 0 & 0 & a_0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix} \quad (-118)$$

Limiting State Probabilities $\tau = \text{mean}$, $\lambda = \text{rate}$

$\lambda\tau < 1 \Rightarrow$ recurrent non-null

$\lim_{n \rightarrow \infty} P[X_n = j] = v_j$

$\bar{v} = \bar{v}P, \sum v_i = 1$

$$\bar{v} = [v_0, v_1, v_2, \dots]P$$

$$\begin{aligned} v_j &= v_0 a_j + \sum_{i=1}^{j+1} v_i a_{j-i+1}, j = 0, 1, 2, \dots \\ &= v_0 a_j + [v_1 a_j + v_2 a_{j-1} + \dots + v_{j+1} a_0] \end{aligned}$$

$$v_j z^j = \sum_{j=0}^{\infty} v_0 a_j z^j + \sum_{j=0}^{\infty} \sum_{i=1}^{j+1} v_i a_{j-i+1} z^j$$

$$G(z) = v_0 G_A(z) + \sum_{i=1}^{\infty} \sum_{j=i-1}^{\infty} v_i a_{j-i+1} z^{j-i+1} z^{i-1}$$

$$= v_0 G_A(z) + \sum_{i=1}^{\infty} v_i z^{i-1} \sum_{k=0}^{\infty} a_k z^k$$

$$= v_0 G_A(z) + \frac{G_A(z)}{z} \sum_{i=1}^{\infty} v_i z^i$$

$$= v_0 G_A(z) + \frac{G_A(z)}{z} [G(z) - v_0]$$

$$G(z) = \frac{v_0(z-1)G_A(z)}{z - G_A(z)}$$

$$G(1) = 1, \sum v = 1$$

$$G'(1) = ?$$

$$G(1) = v_0 + v_1 + \dots$$

$$G(1) = \lim_{z \rightarrow 1} \frac{v_0(z-1)G_A(z)}{z - G_A(z)}$$

$$= \lim_{z \rightarrow 1} v_0 \left[\frac{(z-1)G'_A(z) + G_A(z)}{1 - G'_A(z)} \right]$$

$$= 1$$

Now, with $\rho = G'_A(z)$, it follows that

$$\frac{v_0}{1 - G'_A(z)} = 1$$

$$\therefore v_0 = 1 - \rho$$

Lecture

Memory referencing behaviour

Independent Reference Model

A program's address space typically consists of continuous pages represented by the indices $1, 2, \dots, n$. For the purpose of studying a program's reference behaviour, it can be represented by the reference string $w = x_1, x_2, \dots, x_t, \dots$. Successive references are assumed to form a sequence

of independent, identically distributed random variables with the pmf:

$$P(X_t = i) = \beta_i, \quad 1 \leq i \leq n; \quad \sum_{i=1}^n \beta_i = 1.$$

The number of page frames is in most cases smaller than the number of pages referenced. Let n denote the number of pages and m denote the number of page frames available. Clearly, $1 \leq m \leq n$. The internal state of the paging algorithm at time t denoted by $q(t)$, is an ordered list of the m pages currently in main memory. We assume that on a page fault, the rightmost page in the ordered list $q(t)$ will be replaced. On the page hit, there is no replacement but the list $q(t)$ is updated to $q(t+1)$, reflecting the new replacement priorities. It is clear that the sequence of states $q(0), \dots, q(n), \dots, q(t), \dots$ forms a discrete-time homogeneous Markov chain with the state space consisting of $n!/(n-m)!$ permutations over $1, 2, \dots, n$.

LRU paging algorithm

As an example, consider the LRU paging algorithm with $n = 3$ and $m = 2$. $q(t) = (i, j)$ implies that the page indexed i was more recently used than page indexed j , and, therefore, page j will be the candidate for replacement. The state space I is given by:

$$I = (1, 2), (2, 1), (1, 3), (3, 1), (2, 3), (3, 2)$$

Let the current state $q(t) = (i, j)$. Then the next state $q(t+1)$ takes one of the following values:

$$q(t+1) = \begin{cases} (i, j), & \text{if } x_{t+1} = i, \text{ with associated probability } \beta_i, \\ (j, i), & \text{if } x_{t+1} = j, \text{ with associated probability } \beta_j, \\ (k, i), & \text{if } x_{t+1} = k, k \neq i, k \neq j, \text{ with associated probability } \beta_k. \end{cases}$$

Then the transition probability matrix P is given by:

$$P = \begin{bmatrix} \beta_1 & \beta_2 & 0 & \beta_3 & 0 & 0 \\ \beta_1 & \beta_2 & 0 & 0 & 0 & \beta_3 \\ 0 & \beta_2 & \beta_1 & \beta_3 & 0 & 0 \\ 0 & 0 & \beta_1 & \beta_3 & \beta_2 & 0 \\ \beta_1 & 0 & 0 & 0 & \beta_2 & \beta_3 \\ 0 & 0 & \beta_1 & 0 & \beta_2 & \beta_3 \end{bmatrix}$$

It can be verified that the above Markov chain is irreducible and aperiodic; hence a unique steady-state probability vector v exists. This vector is obtained by solving the system of equations:

$$\begin{aligned} v &= vP, \quad \text{and} \\ \sum_{(i,j)} v_{(i,j)} &= 1 \end{aligned} \tag{-139}$$

Solving this system of equations, we get:

$$v_{(i,j)} = \frac{\beta_i \beta_j}{1 - \beta_i}$$

A page fault occurs in state (i, j) , provided that a page other than i or j is referenced. The associated conditional probability of this event is $1 - \beta_i - \beta_j$; we hence regard $r_{(i,j)} = 1 - \beta_i - \beta_j$ to state (i, j) . The steady-state page fault probability is then given by:

$$E[Z] = F(LRU) = \sum_{(i,j) \in I} (1 - \beta_i - \beta_j) \frac{\beta_i \beta_j}{1 - \beta_i}$$

Lecture

Independent Reference Model

LRU Stack Model

(Refer Sec. 7.5.2.2 of K.T. book ed.1)

Number of page frames allocated to the program: m

Maximum no. of pages available for referencing: n

Position of page at t^{th} reference: D_t

D_t s are i.i.d. random variables. Their pmf is given as:

$$P(D_t = i) = a_i, \quad i = 1, 2, \dots, n, t \geq 1, \text{ and } \sum_{j=1}^n a_j = 1$$

The distribution function is:

$$P(D_t \leq i) = A_i = \sum_{j=1}^i a_j, \quad i = 1, 2, \dots, n, t \geq 1$$

Page fault will occur at a time t provided $D_t > m$. Thus, the page fault probability is given as:

$$F(LRU) = P(D_t > m) = 1 - P(D_t \leq m) = 1 - A_m$$

We study the movement of a tagged page (say y) through the LRU stack as the time progresses. Define the random sequence $E_0 E_1 E_2 \dots E_t \dots$ such that $E_t = i$ if page y occupies the i th position in stack s_t . Clearly $1 \leq E_t \leq n$ for all $t \geq 1$. Thus the sequence above is a discrete parameter, discrete state stochastic process. The position of the page y in stack s_{i+1} is determined by the next reference r_{t+1} and the position of page y in stack s_t , but not its position in previous stacks. Thus, the sequence is a discrete-parameter Markov chain. Furthermore, the chain is homogeneous.

It would be seen that:

$$\begin{aligned}
p_{i1} &= P(E_{t+1} = 1 | E_t = i) \\
&= P(r_{t+1} = y) = P(D_{t+1} = i) = a_i, \quad 1 \leq i \leq n, \\
p_{ii} &= P(E_{t+1} = i | E_t = i) \\
&= P(D_{t+1} < i) = A_{i-1}, \quad 2 \leq i \leq n, \\
p_{i,i+1} &= P(E_{t+1} = i + 1 | E_t = i) \\
&= P(D_{t+1} > i) = 1 - A_i, \quad 1 \leq i \leq n - 1,
\end{aligned}$$

and

$$p_{i,j} = 0, \quad \text{otherwise}$$

One-step Transition Probability Matrix

$$P = \begin{bmatrix} a_1 & 1 - A_i & 0 & 0 & \cdots & \cdots & 0 \\ a_2 & A_1 & 1 - A_2 & \cdot & \cdots & \cdots & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdots & \cdots & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdots & \cdots & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdots & \cdots & \cdot \\ a_i & \cdot & \cdot & A_{i-1} & 1 - A_i & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ a_n & 0 & \cdot & 0 & \cdot & \cdots & A_{n-1} \end{bmatrix} \quad (-141)$$

The chain is aperiodic and irreducible if we assume that $a_i > 0$ for all i . Then the steady-state probability vector $v = (v_1, v_2, \dots, v_n)$ is obtained from the system of equations:

$$\begin{aligned}
v_1 &= \sum_{i=1}^n v_i a_i, \\
v_i &= v_{i-1}(1 - A_{i-1}) + v_i A_{i-1}, \quad 2 \leq i \leq n, \\
\sum_{i=1}^n v_i &= 1
\end{aligned}$$

Solving the above equations, we get,

$$\begin{aligned}
v_1 &= v_{i-1} = v_2, \quad 2 \leq i \leq n, \\
v_1 &= v_i a_1 + v_2 \sum_{i=2}^n a_i = v_1 a_1 + v_2(1 - a_1), \quad \text{and} \\
v_1 &= v_2
\end{aligned}$$

Since summation on v_1 should be 1, we get,

$$v_i = \frac{1}{n}, \quad i = 1, 2, \dots, n.$$

Thus, the position of the tagged page, in the steady-state, is independent of its initial position and is equally likely to be in any stack position. This implies that each page is equally likely to be referenced in the long run. Therefore, the LRU stack model is not able to cater to the nonuniform page-referencing behavior of real programs although it does reflect the clustering effect.

Lecture

More precise model of Slotted ALOHA

(Ref: Kishore Trivedi book 2nd Edition: pg. 373)

(Ref: Bertsekas and Gallager)

There are m nodes, out of which

- n are 'backlogged', i.e nodes which tried to send a packet, unsuccessfully
- $m - n$ 'not backlogged', i.e nodes which do not have packets to send

Assumptions:

- First transmission: Arrivals to a station in a slot are Poisson(λ/m)
- Retransmission: A backlogged node retransmits with probability q_r
- Stations have no buffer, so arrivals to backlogged nodes are dropped

Probability that an unbacklogged station has a packet to transmit in a slot
 $= P[\text{arrivals in previous slot}] = 1 - e^{-\lambda/m} = q_a$

$$P[i \text{ backlogged nodes transmit in a slot}] = Q_r(i, n) = \binom{n}{i} q_r^i (1 - q_r)^{n-i}$$

$$P[i \text{ unbacklogged node first transmit}] = Q_a(i, n) = \binom{m-n}{i} q_a^i (1 - q_a)^{m-n-i}$$

We define a DTMC: number of backlogged users (at slot intervals)

$$P_{n,n-1} = (\text{no unbacklogged transmission}), (\text{exactly 1 backlogged transmission}) \\ = Q_a(0, n) \cdot Q_r(1, n)$$

$$P_{n,n} = (\text{no unbacklogged transmissions}), (\text{more than 1 backlogged transmission}), \\ \text{OR } (1 \text{ unbacklogged, } 0 \text{ backlogged}) \\ = Q_a(0, n)[1 - Q_r(1, n)] + Q_a(1, n)Q_r(0, n)$$

$$P_{n,n+1} = Q_a(1, n)[1 - Q_r(0, n)] \quad (\text{exactly 1 unbacklogged and atleast 1 backlogged})$$

$$P_{n,n+i} = Q_a(i, n) \quad [2 \leq i \leq m - n \text{ unbacklogged}]$$

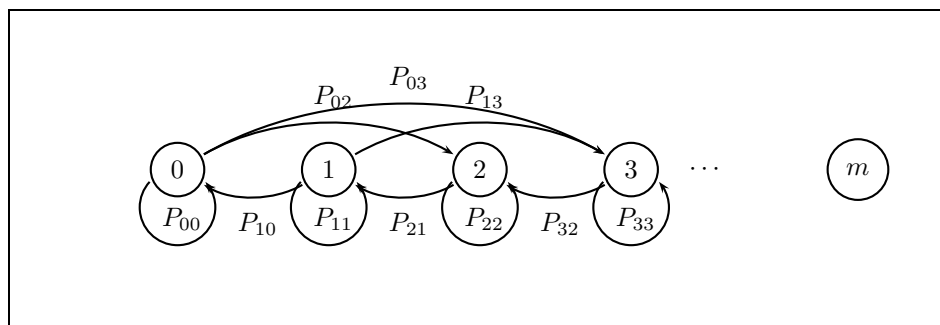


Figure 12: Markov chain for Slotted ALOHA

$$\mathbf{v} = \mathbf{vP} \text{ where } \mathbf{P} = [P_{ij}]$$

$$\sum v_i = 1$$

$$\mathbf{E}[\text{backlogged nodes}] = \sum_{i=0}^n i v_i$$

Delay =

$$\left[\frac{\sum_{i=0}^n i v_i}{\text{Throughput}} \right]$$

Throughput = ?

Throughput, Probability of Success, etc. calculations in next lecture.

Lecture

More precise model of Slotted Aloha.. continued....

We shall now calculate the expected number of successful transmission per unit time.

$$\begin{aligned} P_{succ}(n) &= Q_a(1, n)Q_r(0, n) + Q_a(0, n)Q_r(1, n) \\ &= G(n)e^{-G(n)} \end{aligned}$$

$$(1 - x)^y = e^{-xy} \quad \sim \text{Poisson } (G(n)) \text{ for } x \text{ small}$$

$$\begin{aligned} G(n) &= \text{Expected no. of transmission in a slot} \\ &= nq_n + (m - n)q_a \end{aligned}$$

$$P_{succ} = \sum_n V_n P_{succ}(n)$$

$$\text{Packet Delay} = \frac{E[N]}{P_{succ}}$$

$\mathbf{E}[\text{succ. trans. per unit time}] = 1 \cdot P_{succ} + 0 \cdot (1 - P_{succ}) = P_{succ}$
--

P-K Formula

Alternative Derivation:

λ : arrival rate

τ : mean service time

B : service time

For $M/G/1$ system, we have,

$$E[W] = \frac{\lambda\tau^2(1 + C_B^2)}{2(1 - \rho)} \quad \text{where } C_B \text{ is the covariance}$$

$$C_B^2 = \frac{\sigma_B^2}{\tau^2}$$

$$\therefore E[W] = \frac{\lambda(\tau^2 + \sigma_B^2)}{2(1 - \rho)}$$

$$\sigma_B^2 = E[B^2] - \tau^2$$

$$\therefore E[W] = \frac{\lambda E[B^2]}{2(1 - \rho)} = \frac{\lambda \overline{X^2}}{2(1 - \rho)}$$

W_i : waiting time of i^{th} customer

R_i : Residual service time seen by the i^{th} customer

X_i : Service time of i^{th} customer

N_i : Number of customers found in queue by the i^{th} customer

$$W_i = R_i + X_{i-1} + X_{i-2} + \dots + X_{i-N_i}$$

$$E[W_i] = W = R + E[N] \cdot E[X]$$

$$= R + \tau N_q$$

Now, by Little's Law

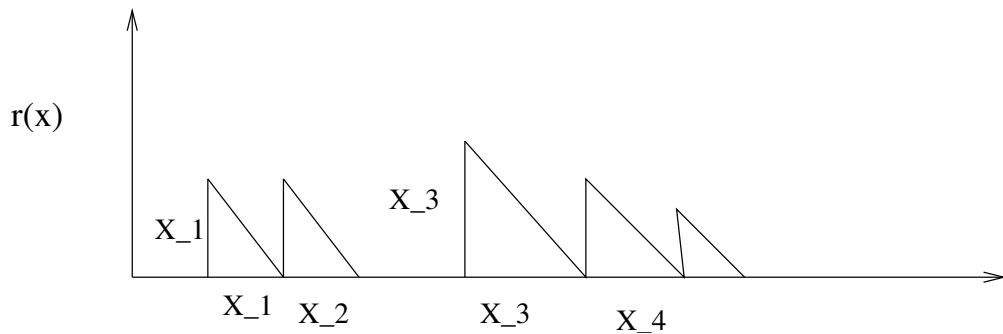
$$N_q = \lambda W$$

$$W = R + \lambda \tau W$$

$$W(1 - \rho) = R$$

$$W = \frac{R}{1 - \rho}$$

$r(x)$: residual service at time x



Assumption:

Time Average = Instantaneous Average

Interval[0,t] ... $\frac{\int_0^t r(x)dx}{t}$ is the average.

$$\begin{aligned}
 R &= \lim_{t \rightarrow \infty} \frac{\int_0^t r(x)dx}{t} \\
 &= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{M(t)} \frac{1}{2} X_i^2 \\
 &= \lim_{t \rightarrow \infty} \frac{M(t)}{t} \frac{1}{2} \frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)}
 \end{aligned}$$

$M(t)$: No. of service completions in time t

$$\begin{aligned}
 \therefore R &= \frac{\lambda \overline{X^2}}{2} \\
 \therefore W &= \frac{\lambda \overline{X^2}}{2(1-\rho)}
 \end{aligned}$$

M/G/1 with vacations

Here, we study M/G/1 systems that go on 'vacation' i.e. temporary periods of unavailability

X : service time

V : vacation periods

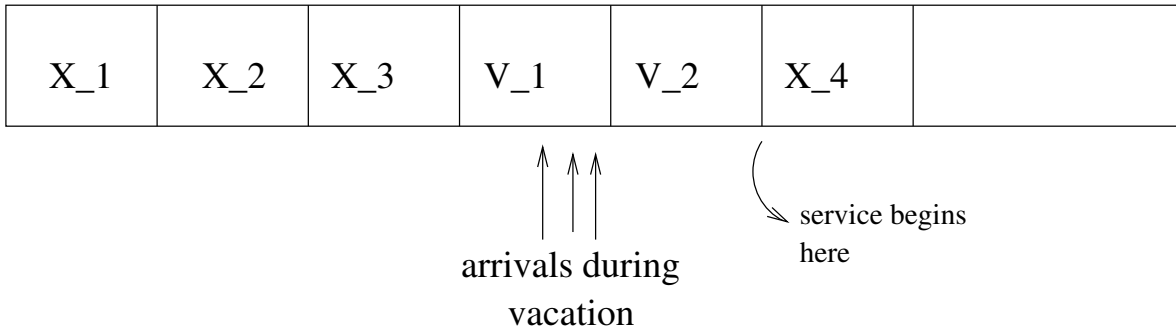


Figure 13: M/G/1 with vacations

Again:

$$W = \frac{R}{1-\rho}$$

R : mean residual service or vacation time

The Waiting time can be evaluated as:

$$\lim_{t \rightarrow \infty} \frac{\int_0^{\tau} r(x) dx}{t} = \lim_{t \rightarrow \infty} \frac{1}{t} \frac{\sum_{i=1} M(t) X_i^2}{2} + \frac{1}{t} \sum_{i=1}^{L(t)} \frac{V_i^2}{2}$$

where, $L(t)$ is the no. of vacations

$$W = \frac{\lambda \bar{X}^2}{2(1-\rho)} + \lim_{t \rightarrow \infty} \frac{L(t)}{t} \frac{\sum_{i=0}^{L(t)} V_i^2}{2(1-\rho)L(t)}$$

Now, fraction of time spent in vacations = $t(1-\rho)$

$$\therefore \lim_{t \rightarrow \infty} L(t) \bar{V} = \lim_{t \rightarrow \infty} t(1-\rho)$$

Lecture

FDM, SFDM and TDM

Consider there are m users. Data packets are sent by stations. Overall packet arrival rate per station is λ . Effectively, packet arrival rate per station is $\frac{\lambda}{m}$.

Also consider that the system operates with unit transmission time if entire bandwidth is available.

Slot duration is 1 unit.

1. Frequency Division Multiplexing (FDM):

$\frac{1}{m}$ of the total bandwidth is available to each user.

\Rightarrow Transmission time = m

$$\rho = \frac{\lambda}{m} \cdot m = \lambda$$

For an $M/D/1$, we calculate the waiting time and response time as under:

$$W_{FDM} = \frac{(\frac{\lambda}{m})m^2}{2(1-\lambda)} = \frac{m\lambda}{2(1-\lambda)}$$

$$R_{FDM} = \frac{m\lambda}{2(1-\lambda)} + m$$

2. Synchronous Frequency Division Multiplexing (SFDM)

m is the duration of vacation.

We calculate the waiting time and response time as under:

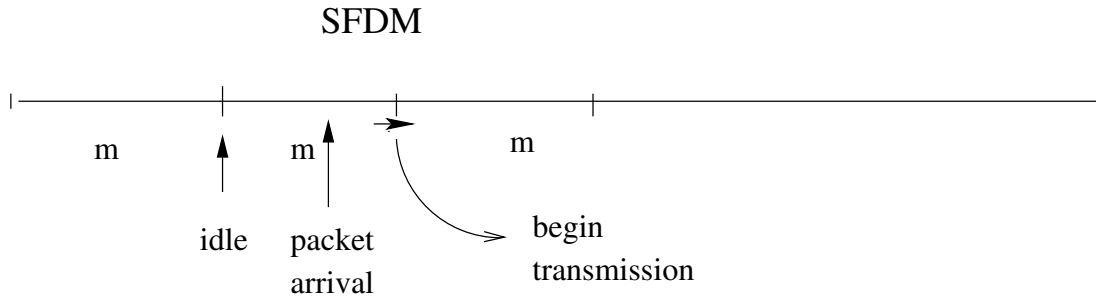


Figure 14: SFDM

$$\begin{aligned}
 W_{SFDM} &= \frac{m\lambda}{2(1-\lambda)} + \frac{m^2}{2m} \\
 &= \frac{m\lambda}{2(1-\lambda)} + \frac{m}{2}
 \end{aligned}$$

$$R_{FDM} = \frac{m\lambda}{2(1-\lambda)} + \frac{3m}{2}$$

3. Time Division Multiplexing:

Each user has a unit duration slot. There are m users. So, the frame is of duration m . From waiting time point of view, the system behaves as if service time $\sim m$, and vacation time $\sim m$.

Figure here.

Again, we calculate the waiting time and response time as under:

$$\begin{aligned}
 W_{SFDM} &= \frac{m\lambda}{2(1-\lambda)} + \frac{m}{2} \\
 R_{FDM} &= \frac{m\lambda}{2(1-\lambda)} + \frac{m}{2} + 1
 \end{aligned}$$

Delay analysis of ARQ systems

Go-Back-N protocol:

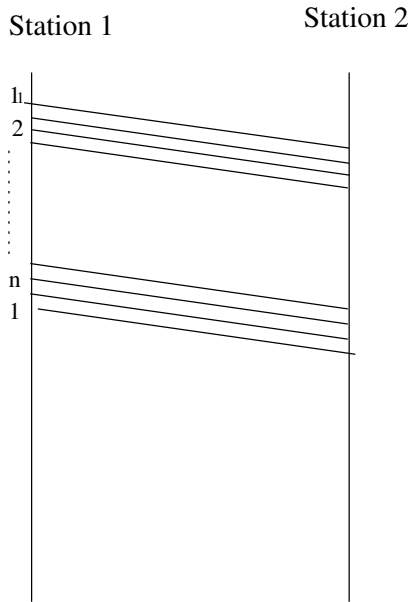


Figure 15: Go-back-N ARQ Model

The packet transmission time is of unit duration. The sender times out after $n - 1$ packet transmissions (after current one).

Let the probability of time-out (or failure in transmission) be p .

Also, ‘service’ of packet i begins only after ‘service’ of $(i - 1)^{th}$ packet ends.

Number of transmissions	Time taken	Probability
1	1	$(1-p)$
2	$1 + n$	$(1-p)p$
3	$1 + 2n$	$(1-p)p^2$
·	·	·
·	·	·
·	·	·
k	$1 + (k-1)n$	$(1-p)p^{k-1}$
(k+1)	$1 + kn$	$(1-p)p^k$

We now have the distribution of time taken for successful transmission.

Probability function of X i.e. time taken is

$$P[X = 1 + kn] = p^k(1 - p) \quad k = 0, 1, 2, \dots, \infty$$

The mean is

$$\bar{X} = \sum_{k=0}^{\infty} (1 + kn) \cdot (1 - p) \cdot p^k$$

And the second moment is

$$\bar{X}^2 = \sum_{k=0}^{\infty} (1 + kn)^2 \cdot (1 - p) \cdot p^k$$

Lecture

Refer sections 9.1, 9.2, 9.3 of K.Trivedi book, ed. 1 and 2.

Tandem Queues

Consider a two-stage tandem network as shown in the figure. The system consists of two nodes with respective service rates μ_0 and μ_1 . The external arrival rate is λ and has Poisson distribution. The service-time distribution at both nodes is exponential.

The system can be modeled as a stochastic process whose states are specified by pairs (k_0, k_1) , $k_0 \geq 0, k_1 \geq 0$, where k_0 and k_1 are number of jobs at server 0 and server 1 respectively.

Since all interevent times are exponentially distributed (by our example), it follows that the stochastic process is a Markov chain with the state diagram as shown.

Figure here.

Let $P(k_0, k_1)$ be the joint probability of k_0 jobs at the first server and k_1 jobs at the next server, in the steady state. Equating the rates of flow into and out of the state, we obtain the following balance equations:

$$(\mu_0 + \mu_1 + \lambda)P(k_0, k_1) = \mu_0 P(k_0 + 1, k_1 - 1) + \mu_1 P(k_0, k_1 + 1) + \lambda P(k_0 - 1, k_1), \quad k_0 > 0, k_1 > 0.$$

For the boundary states, we have:

$$\begin{aligned} (\mu_0 + \lambda)P(k_0, 0) &= \mu_1 P(k_0, 1) + \lambda P(k_0 - 1, 0), & k_0 > 0 \\ (\mu_0 + \lambda)P(0, k_1) &= \mu_0 P(1, k_1 - 1) + \mu_1 P(0, k_1 + 1), & k_1 > 0 \\ \lambda P(0, 0) &= \mu_1 P(0, 1) \end{aligned}$$

The normalization is provided by:

$$\sum_{k_0 \geq 0} \sum_{k_1 \geq 0} P(k_0, k_1) = 1$$

It is easily shown by direct substitution that the following equation is the solution to the above balance equations:

$$P(k_0, k_1) = (1 - \rho_0)\rho_0^{k_0}(1 - \rho_1)\rho_1^{k_1}$$

where, $\rho_0 = \lambda/\mu_0$ and $\rho_1 = \lambda/\mu_1$. The condition for stability of the system is that both ρ_0 and ρ_1 are less than unity.

Burke's theorem: Burke has shown that the output of an $M/M/1$ queue is also Poisson with rate λ .

Thus, the tandem queues in the figure behave like 2 independent $M/M/1$ queues. The joint distribution is the product of individual distributions.

Expected number of jobs in the system is:

$$E[N] = \frac{\rho_0}{1 - \rho_0} + \frac{\rho_1}{1 - \rho_1}$$

Average Response time is:

$$E[R] = \frac{E[N]}{\lambda}$$

Even with queues with branching and feedback, the joint distribution is like above. The tandem queues would behave like independent queues. However, at the feed back point, the arrival no longer remains Poisson distributed. The effective arrival rate changes and so does the other metrics like no. of jobs in the system, response time, etc.

CPU and Disk example

The transition diagram is as below:

$$\rho_0 = \frac{\lambda_0}{\mu_0}, \quad \rho_1 = \frac{\lambda_1}{\mu_1}$$

$$\begin{aligned} P(k_0, k_1) &= P(k_0)P(k_1) \\ &= (1 - \rho_0)\rho_0^{k_0}(1 - \rho_1)\rho_1^{k_1} \\ \lambda_0 &= \lambda + \lambda_1 \\ \lambda_1 &= \lambda_0 P_1 \\ \lambda_0 &= \lambda_0 P_1 + \lambda \\ \lambda_0 &= \frac{\lambda}{1 - P_1} = \frac{\lambda}{P_0} \\ \lambda_1 &= \lambda_0 P_1 = \frac{\lambda P_1}{P_0} \end{aligned}$$

Average number of visits before departure $\sim v_0, v_1$.

P[No. of CPU bursts = k] = $(1 - P_0)^{k-1} P_0$

Before departure =

$$\rho_0 = \frac{\lambda_0}{\mu_0} = \left(\frac{\lambda}{P_0}\right) \left(\frac{1}{\mu_0}\right) = \lambda_0 \left(\frac{1}{P_0 \mu_0}\right) \rho_1 \quad = \frac{\lambda_1}{\mu_1} = \frac{\lambda \rho_1}{P_0 \mu_1} = \lambda \left(\frac{P_1}{P_0 \mu_1}\right)$$

These queues are behaving like:
 Figure here.
 Expected number of jobs in the system:

$$E[N] = \frac{\rho_0}{1 - \rho_0} + \frac{\rho_1}{1 - \rho_1}$$

Expected response time:

$$E[R] = \frac{1}{\lambda} \left[\frac{\rho_0}{1 - \rho_0} + \frac{\rho_1}{1 - \rho_1} \right] = \left[\frac{1/P_0\mu_0}{1 - \frac{\lambda}{\mu_0 P_0}} + \frac{\frac{P_1}{P_0\mu_1}}{1 - \frac{\lambda P_1}{P_0\mu_1}} \right] = \left[\frac{1}{P_0\mu_0 - \lambda} + \frac{1}{\frac{P_0\mu_1}{P_1} - \lambda} \right]$$

Lecture

Open central server network example

Refer K.Trivedi books sections 9.1, 9.2 and 9.3

Consider the open central server queuing model of a computer system shown in figure 18.

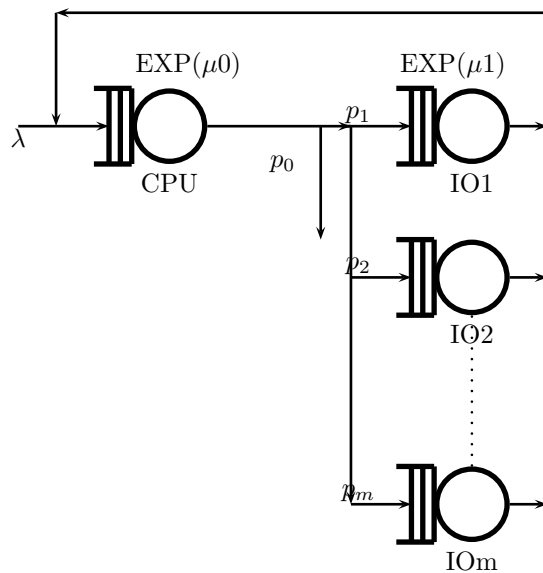


Figure 16: Open central server network

The k-step transition probability matrix is as under:

$$P = \begin{bmatrix} 0 & p_1 & p_2 & \cdots & p_m & p_0 \\ 1 & 0 & 0 & & 0 & 0 \\ 1 & \cdot & \cdot & & 0 & 0 \\ 1 & \cdot & \cdot & & 0 & \cdot \\ 1 & \cdot & \cdot & & 0 & \cdot \\ 1 & \cdot & \cdot & & 0 & 0 \\ 1 & 0 & 0 & & 0 & 0 \\ 0 & 0 & 0 & & 0 & 1 \end{bmatrix}$$

The one-step transition probability is of more interest and is given as under:

$$X = \begin{bmatrix} 0 & p_1 & p_2 & \cdots & p_m \\ 1 & 0 & 0 & & 0 \\ 1 & \cdot & \cdot & & 0 \\ 1 & \cdot & \cdot & & 0 \\ 1 & \cdot & \cdot & & 0 \\ 1 & \cdot & \cdot & & 0 \\ 1 & 0 & 0 & & 0 \end{bmatrix}$$

Assuming $\rho_0 \neq 0$, the average number of visits (or visit counts) V_j made to node j is as under:

$$V_j = \begin{cases} 1/p_0, & j = 0 \\ p_j/p_0, & j = 1, 2, \dots, m. \end{cases}$$

Since λ is the total arrival rate to the entire system, the overall arrival rate λ_j to node j is given by:

$$\lambda_j = \begin{cases} \lambda/p_0, & j = 0 \\ \lambda p_j/p_0, & j = 1, 2, \dots, m. \end{cases}$$

The utilization, ρ , of node j is given by

$$\rho_j = \lambda_j m u_j = \lambda V_j / \mu_j$$

Expected number of jobs in the system would then be:

$$E[N] = \sum_{j=0}^m \frac{\rho_j}{1 - \rho_j}$$

And the expected Response time would be:

$$E[R] = \frac{E[N]}{\lambda} = \sum_{j=0}^m \frac{E[B_j]}{1 - \lambda E[B_j]}$$

where,

$E[B_j]$: average total service demand at queue j , and

$$E[B_j] = \frac{V_j}{\mu_j}$$

Jackson's result: Jackson has shown that the joint probability of k_j customers at node j ($j = 0, 1, \dots, m$) is given by:

$$p(k_0, k_1, k_2, \dots, k_m) = \prod_{j=0}^m p_j(k_j)$$

This formula implies that the queue lengths are mutually independent and in the steady state, and the steady-state probability of k_j customers at node j is given by the $M/M/1$ formula:

$$p_j(k_j) = (1 - \rho_j)\rho_j^{k_j}$$

Example: Consider an open queuing network with $(n + 1)$ stations, where the i^{th} station consists of c_i exponential servers, $i = 0, 1, \dots, m$ each with mean service time of $1/\mu_i$ seconds. External Poisson sources contribute γ_i jobs/second to the average rate of arrival to the i^{th} station.

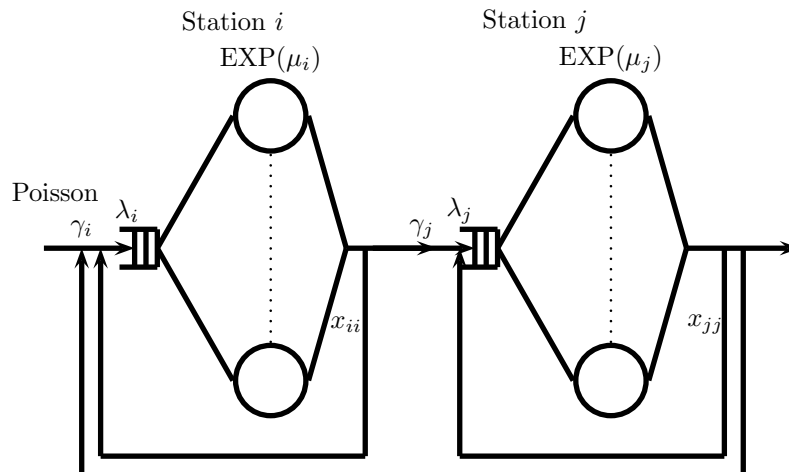


Figure 17:

Then the total arrival rate is given as:

$$\lambda = \sum_{i=0}^m \gamma_i$$

The one-step transition probability matrix would be:

$$X = [x_{ij}]$$

such that

$$1 - \sum_{j=0}^m x_{ij} > 0, \text{ for atleast one } i.$$

According to Jackson's result, each queue behaves like an $M/M/1$ with λ_i as derived.

$$\lambda_j = \gamma_j + \sum_{i=0}^m x_{ij} \lambda_i, i = 0, 1, \dots, m$$

Also, for each queue $i, \lambda_i < c_i \mu_i$

The number of jobs, $P(k_i)$, at station i , is

$$p(k_0, k_1, \dots, k_m) = p(k_0)p(k_1) \dots p(k_m)$$

Lecture

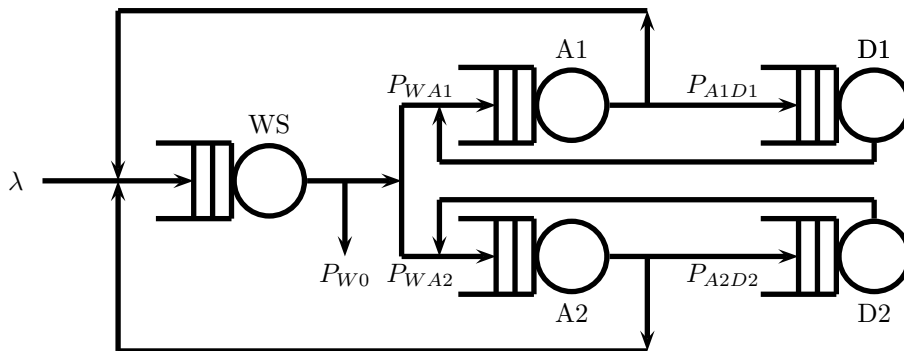


Figure 18:

Consider the system as shown in figure 18. Let the service requirement per visit of the individual servers be $\tau_{WS}, \tau_{A1}, \tau_{A2}, \tau_{D1}, \tau_{D2}$.

Assumption: Arrivals are Poisson and Service times are Exponential.

We find the average number of visits per server, i.e. $V_W, V_{A1}, V_{A2}, V_{D1}, V_{D2}$ as:

$$\begin{aligned}
V_{A1} &= P_{WA1}V_W + P_{A1D1}V_{A1} \\
V_{A2} &= P_{WA2}V_W + P_{A2D2}V_{A2} \\
V_{D1} &= P_{A1D1}V_{A1} \\
V_{D2} &= P_{A2D2}V_{A2} \\
V_W &= \frac{1}{P_{W0}}
\end{aligned}
\tag{-153}$$

By back-substitution, we get:

$$\begin{aligned}
V_{A1} &= P_{WA1}V_W + P_{A1D1}V_{A1} = P_{WA1}V_W/(1 - P_{A1D1}) = P_{WA1}/(P_{W0}(1 - P_{A1D1})) \\
V_{A2} &= P_{WA2}V_W + P_{A2D2}V_{A2} = P_{WA2}V_W/(1 - P_{A2D2}) = P_{WA2}/(P_{W0}(1 - P_{A2D2})) \\
V_{D1} &= P_{A1D1}V_{A1} = \frac{P_{A1D1}P_{WA1}}{P_{Q0}(1 - P_{A1D1})} \\
V_{D2} &= P_{A2D2}V_{A2} = \frac{P_{A2D2}P_{WA2}}{P_{Q0}(1 - P_{A2D2})} \\
V_W &= \frac{1}{P_{W0}}
\end{aligned}
\tag{-157}$$

The effective arrival rates at individual servers in terms of (request/sec)*(visits/request) would be:

$$\begin{aligned}
\lambda_W &= V_W\lambda \\
\lambda_{A1} &= V_{A1}\lambda \\
\lambda_{A2} &= V_{A2}\lambda \\
\lambda_{D1} &= V_{D1}\lambda \\
\lambda_{D2} &= V_{D2}\lambda
\end{aligned}
\tag{-161}$$

The average service demand at each of the servers would be:

$$\begin{aligned}
E[B_W] &= V_W\tau_W \\
E[B_{A1}] &= V_{A1}\tau_{A1} \\
E[B_{A2}] &= V_{A2}\tau_{A2} \\
E[B_{D1}] &= V_{D1}\tau_{D1} \\
E[B_{D2}] &= V_{D2}\tau_{D2}
\end{aligned}
\tag{-165}$$

Response time on individual servers would be:

$$\begin{aligned}
 E[R_W] &= \left(\frac{\tau_W}{1 - \lambda_W \tau_W} \right) \\
 E[R_{A1}] &= \left(\frac{\tau_{A1}}{1 - \lambda_{A1} \tau_{A1}} \right) \\
 E[R_{A2}] &= \left(\frac{\tau_{A2}}{1 - \lambda_{A2} \tau_{A2}} \right) \\
 E[R_{D1}] &= \left(\frac{\tau_{D1}}{1 - \lambda_{D1} \tau_{D1}} \right) \\
 E[R_{D2}] &= \left(\frac{\tau_{D2}}{1 - \lambda_{D2} \tau_{D2}} \right)
 \end{aligned}
 \tag{-169}$$

Total Response time for a request entering the system is the summation of total time taken at each of these servers. i.e.

$$\begin{aligned}
 E[R] &= \sum_i \text{No. of visits on server 'i'} * \text{Response time on server 'i' per visit} \\
 E[R] &= \sum_i V_i * E[R_i]
 \end{aligned}
 \tag{-170}$$