Introduction to Queuing Theory

Basics of Probability

- Random Variable
  - A function that reflects the result of a random experiment
    - Result of “toss a single die” can be described by a random variable that can assume the values one through six
    - No. of requests that arrive at an airline reservation system in one hour
    - Time interval between the arrivals of two consecutive jobs in a computer system

- Random Variable definition Clarified
  - Ask 10 persons for a yes/no (1/0) reply to a query
  - Define the variable X to represent the number of 1 replies
  - The original sample space has 2^{10} elements
  - Sample space of X is the set of integers 1…10
  - Thus X defines a mapping from the original sample space to a new sample space, usually a set of real numbers

Discrete Random Variables

- A random variable that can assume only discrete values
- The random variable is described by the possible values it can assume and by the probabilities of each of these values
- Set of these probabilities is called the probability mass function (pmf) of the random variable
- For example, if the possible values of a random variable X are the non-negative integers, then the pmf is given by the probabilities:
  - P_k = P(X=k), for k=0,1,2,… the probability that the random variable X assumes the value k
- The following must hold
  - P(X=k) >= 0 and $\sum_{\text{all } k} P(X=k) = 1$
- For example, the following pmf results from the experiment “toss a single die”
  - P(X=k) = 1/6, for k=1,2,…,6

Discrete Random Variables

- Bernoulli Random Variable
  - A random experiment that has two possible outcomes, like tossing a coin (k=0,1). Pmf of X is:
  - P(X=1) = p, P(X=0) = 1-p with 0<p<1
- Binomial Random Variable
  - Experiment with two possible outcomes is carried out *n* times where successive trials are independent. The random variable X is the number of times the outcome 1 occurred. Pmf of X is:
  - P(X=k) = $\binom{n}{k} p^{k} (1-p)^{n-k}, \; k = 0,1,\ldots,n$

Discrete Random Variables

- Geometric Random Variable
  - Experiment with two possible outcomes is carried out several times, the random variable X represents the number of trials it takes for the outcome 1 to occur (current trial included). Pmf of X is:
  - $P(X=k) = p(1-p)^{k-1}$, $k = 1,2,\ldots$
- Poisson Random Variable
  - X represents occurrence of k events. Pmf is given by:
  - $P(X=k) = \frac{\alpha^k}{k!}.e^{-\alpha}, k = 1,2,\ldots$

Moments

- Mean or expected value $\overline{X} = E[X] = \sum_{\text{all } k} k.P(X = k)$
- A function of a random variable is another random variable with expected value of
  - E[f(X)]= $\sum_{\text{all } k} f(k).P(X = k)$
- n^{th} moments $\overline{X^n} = E[X^n] = \sum_{\text{all } k} k^n.P(X = k)$
  - i.e., the expected value of the n^{th} power of X.
- n^{th} central moment:

$$\overline{(X - \overline{X})^n} = E[(X - E[X])^n] = \sum_{\text{all } k} (k - \overline{X})^n.P(X = k)$$

- n^{th} central moment is the expected value of the n^{th} power of the difference between X and its mean. The first central moment is equal to zero.

Moments

- The second central moment is called the variance of X: $\sigma_X^2 = \mathrm{var}(X) = \overline{(X - \overline{X})^2} = \overline{X^2} - \overline{X}^2$
  - σ_X is called the standard deviation.
- Coefficient of variation is the normalized standard deviation $c_X = \dfrac{\sigma_X}{\overline{X}}$

Properties of several Discrete Random Variables

Random Variable	Parameter	Mean	Variance
Bernoulli	p	p	p(1-p)
Binomial	n,p	np	np(1-p)
Geometric	p	1/p	$(1-p)/p^2$
Poisson	α	α	α

Continuous Random Variables

- X can assume all values in the interval [a,b] where $-\infty \le a < b \le +\infty$
- Described by its distribution function (also called CDF or cumulative distribution function)
  - $F_X(x) = P(X \le x)$, which specifies the probability that the random variable X takes values less than or equal to x.
  - $F_X(x) \le F_X(y)$ for $x < y$; $P(x < X \le y) = F_X(y) - F_X(x)$
  - Probability density function (pdf) $f_X(x)$ can also be used instead of the distribution function provided the latter is differentiable.

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

Continuous Random Variables

- $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx$, $P(X=x) = \int_{x}^{x} f_X(x)dx = 0$
- $P(X>x_3) = \int_{x_3}^{\infty} f_X(x)dx$
- Mean or expected value: $\overline{X} = E[X] = \int_{-\infty}^{\infty} x.f_X(x)dx$
- Expected value of a function of X: $E[g(X)] = \int_{-\infty}^{\infty} g(x).f_X(x)dx$
- n^{th} moment $\overline{X^n} = E[X^n] = \int_{-\infty}^{\infty} x^n.f_X(x)dx$
- n^{th} central moment – Similarly defined
- Variance $\sigma_X^2 = \mathrm{var}(X) = \overline{(X-\overline{X})^2} = \overline{X^2} - \overline{X}^2$

Normal Distribution

- CDF of a normally distributed random variable

$$F_X(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\overline{X})^2}{2\sigma_x^2}\right) du$$

- pdf is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\overline{X})^2}{2\sigma_X^2}\right)$$

- Standard normal distribution is defined by setting $\overline{X} = 0$ and $\sigma_X = 1$
- CDF: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du$
- pdf: $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

Exponential Distribution

- CDF of an exponentially distributed random variable X. λ or μ denotes a parameter.

$$F_X(x) = \begin{cases} 1 - \exp(-\frac{x}{\overline{X}}), & 0 \le x < \infty, \\ 0, & \text{otherwise} \end{cases}$$

$$\text{with } \overline{X} = \begin{cases} \frac{1}{\lambda}, \text{if X represents interarrival times} \\ \frac{1}{\mu}, \text{if X represents service times} \end{cases}$$

- For exponentially distributed random variable with parameter λ,

$$\text{pdf}: f_X(x) = \lambda e^{-\lambda x} \text{ for } x >= 0, \text{mean}: \overline{X} = \frac{1}{\lambda},$$

$$\text{Variance}: \text{var}(X) = \frac{1}{\lambda^2}, \text{coefficient of variation}: c_X = 1$$

- Exponential distribution is completely determined by its mean value

Properties of Exponential Distribution

- Only continuous distribution that is memoryless
- $P(X \leq u+t \mid X > u) = 1-\exp(-\frac{t}{\overline{X}}) = P(X \leq t)$
- Let buses arrive with exponentially distributed interarrival times and identical mean. If you have already been waiting for *u* units of time for the bus to come, the probability of a bus arrival within the next *t* units is the same as if you had just come to the bus stop.
- Relation to discrete Poisson random variable
  - If the interarrival times are exponentially distributed and successive interarrival times are independent with identical mean $\overline{X}$, then the random variable representing the number of arrivals in a fixed interval of time [0,t) has a Poisson distribution with parameter $\alpha=t/\overline{X}$

Merging and Splitting of Poisson Processes and Property of corresponding Distributions

- If n Poisson processes with distributions for the interarrival times $1-e^{-\lambda_i t}, 1 \le i \le n,$ into one single process, the result is a Poisson process with interarrival times having the distribution, $1-e^{-\lambda t}$ where $\lambda = \sum_{i=1}^{n} \lambda_i$
- If a Poisson process with interarrival time distribution $1-e^{-\lambda t}$ is split into *n* processes so that the probability that the arriving job is assigned to the i^{th} process is q_i, $1 \le i \le n$, then the i^{th} subprocess has an interarrival time distribution of $1-e^{-q_i \lambda t}$, i.e., *n* Poisson processes are created.

Multiple Random Variables

- Joint probability mass function of discrete random variables $X_1, X_2, .., X_n$:

  $P(X_1=x_1, X_2=x_2, .., X_n=x_n)$ represents the probability that $X_1=x_1, X_2=x_2, .., X_n=x_n$.
- In the continuous case, joint distribution function:

  $F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \le x_1, X_2 \le x_2, .., X_n \le x_n)$ represents the probability that $X_1 \le x_1, X_2 \le x_2, .., X_n \le x_n$. Here $\mathbf{X}=(X_1, X_2, .., X_n)$ is the n-dimensional random variable and $\mathbf{x}=(x_1, x_2, .., x_n)$.
- Random variables $X_1, X_2, .., X_n$ are independent if
  - $P(X_1=x_1, X_2=x_2, .., X_n=x_n) = P(X_1=x_1).P(X_2=x_2). ..P(X_n=x_n)$ in the discrete case and
  - $P(X_1 \le x_1, X_2 \le x_2, .., X_n \le x_n) = P(X_1 \le x_1).P(X_2 \le x_2). ..P(X_n \le x_n)$ in the continuous case.

Conditional Probability

- Probability for $X_1=x_1$ under the conditions that $X_2=x_2$, $X_3=x_3$,…, $X_n=x_n$ is given by:

$$P\left(X_1 = x_1 | X_2 = x_2, \ldots X_n = x_n\right) = \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P(X_2 = x_2, \ldots, X_n = x_n)}$$

- For continuous random variables:

$$P\left(X_1 \le x_1 | X_2 \le x_2, \ldots X_n \le x_n\right) = \frac{P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)}{P(X_2 \le x_2, \ldots, X_n \le x_n)}$$

Random/Stochastic Processes

- A stochastic process is defined as a family of random variables $\{X_t: t \in T\}$, where each random variable X_t is indexed by parameter $t \in T$, usually called the time parameter if $T \subseteq R+ = [0, \infty)$.
- Set of all possible values of X_t (for each $t \in T$) is called the state space S of the stochastic process.
- If a countable, discrete parameter set T is considered, the S.P. is called *discrete parameter* process. T is represented by a subset of $N_0 = \{0,1,\ldots\}$. Else, it is called a *continuous parameter* process.
- The state space can also be continuous or discrete. If discrete, the S.P.s are called as chains.

Markov Process and Markov Chain

- An S.P. $\{X_t: t \in T\}$ constitutes a Markov Process if for all $0=t_0<t_1<..t_n<t_{n+1}$ and all $s_i \in S$, the conditional CDF of X_{tn+1} depends only on the last previous value X_{tn} and not on the earlier values X_{t0}, X_{t1},…X_{tn-1}.
- When we consider discrete state spaces, we deal with Continuous Time Markov Chains (CTMC), Otherwise: Discrete Time Markov Chains (DTMC)
- For DTMC, following property must hold for all $n \in N_0$ and all $s_i \in S$:
- $P(X_{n+1}=s_{n+1}|X_n=s_n,X_{n-1}=s_{n-1},\ldots,X_0=n_0) = P(X_{n+1}=s_{n+1}|X_n=s_n)$

Discrete Time Markov Chains

- Given an initial state s_0, the DTMC evolves over time according to one-step transition probabilities.
- Let S = {0, 1, …}, we can write the conditional pmf of one-step transition probability from state i to state j as:

$$p_{ij}^{(1)}(n) = P(X_{n+1} = s_{n+1} = j \mid X_n = s_n = i)$$

- If the conditional pmf is independent of epoch n (called the homogeneous case),

$$p_{ij}^{(1)} = p_{ij}^{(1)}(n) = P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i)$$

- We drop the superscript to denote one step transition probability of a homogeneous DTMC as p_{ij}

Discrete Time Markov Chains

- From initial state i, DTMC goes to some state j (including the possibility of j=i) so that

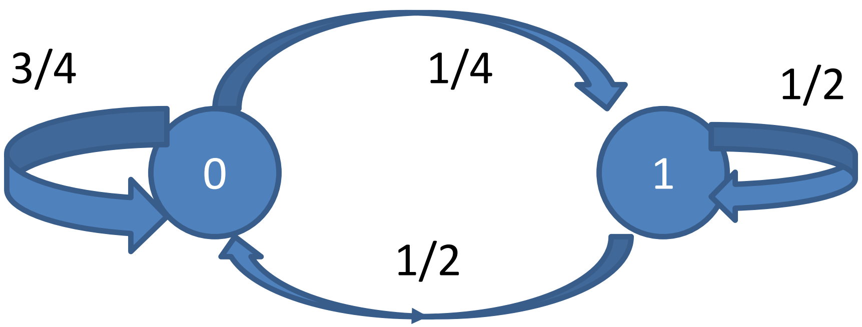
$$\sum_j p_{ij} = 1, \text{ where } 0 \le p_{ij} \le 1$$

- Usually represented in a transition matrix **P**:

$$P = P^{(1)} = [p_{ij}] = \begin{pmatrix} p_{00}\ p_{01}\ p_{02}\cdots\cdots \\ p_{10}\ p_{11}\ p_{12}\cdots\cdots \\ p_{20}\ p_{21}\ p_{22}\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots \end{pmatrix}$$

Finite State Discrete Time Markov Chains

- A Finite State Discrete Time Markov Chains (FSDTMC) can be represented as a state transition diagram.
- Consider S ={0, 1} and $P^{(1)} = \begin{pmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{pmatrix}$
- Corresponding state transition diagram:



State Sojourn Time

- Transition behavior reflects memoryless property
  - Depends on the current state
  - Not on the history that led to the current state
  - Also not on the time already spent in the current state
- Probability of leaving the current state *i* is given by $(1-p_{ii}) = \sum_{i \neq j} p_{ij}$
- Repetitive application of this leads to a description of random experiment in the form of a sequence of Bernoulli trials with success probability $(1-p_{ii})$

State Sojourn Time

- State sojourn time R_i during a single visit to state *i* is a geometrically distributed random variable with pmf: $P(R_i = k) = (1 - p_{ii}) p_{ii}^{k-1}, \forall k \in N^+$
- Expected sojourn time $E[R_i]$, i.e., mean number of time steps the process spends in state *i* per visit: $E[R_i] = \frac{1}{1 - p_{ii}}$

Continuous Time Markov Chains

- State transitions may occur at arbitrary instants of time.
- Parameter T is represented by a set of non-negative real numbers R_0^+
- A stochastic process $\{X_t : t \in T\}$ constitutes a CTMC if for arbitrary $t_i \in R_0^+$, with $0=t_0 < t_1 < .. < t_n < t_{n+1}$, $\forall n \in N$ and $\forall s_i \in S = N_0$ for the conditional pmf, the following holds:

$$P(X_{t_{n+1}} = s_{n+1} | X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \ldots, X_{t_0} = s_0) = P(X_{t_{n+1}} = s_{n+1} | X_{t_n} = s_n)$$

- Since exponential distribution is the only memoryless continuous-time distribution, the state sojourn times of a CTMC are exponentially distributed (under certain assumptions)

Continuous Time Markov Chains

- RHS of the last equations is referred to as transition probability $p_{ij}(u,v)$ of the CTMC to move from state *i* to state *j*:

$$p_{ij}(u,v) = P(X_v = j \mid X_u = i)$$

- Unlike homogeneous DTMC, we cannot have a transition matrix since the time parameter is continuous
- An infinitesimal generator matrix **Q** of the transition probability matrix $\mathbf{P}(t)=[p_{ij}(0,t)]=[p_{ij}(t)]$ is used
- $\mathbf{Q}=[q_{ij}]$, $\forall i,j \in S$, contains the transition rates q_{ij} from any state i to any other state j, where $i \neq j$ of a given CTMC. The elements q_{ii} on the main diagonal of Q are defined by

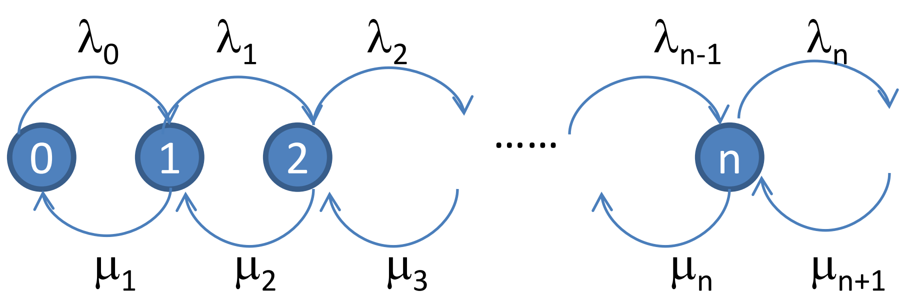
$$q_{ii} = -\sum_{j,\, j \neq i} q_{ij}$$

Continuous Time Markov Chains

- Q satisfies the property: $\mathbf{0} = \boldsymbol{\pi}\mathbf{Q}$
- Here $\boldsymbol{\pi}$ denotes state probabilities
- CTMC sojourn time: Random variables denoting the sojourn times are exponentially distributed with mean equal to $1/(-q_{ii})$

Birth Death Process

- Can be either discrete time or continuous time process
- Set of integers as the discrete state space
- State transitions take place only between neighboring states
- We focus on CTMC
- State in which the population size is k, is denoted by E_k. A transition from E_k to E_{k+1} denotes a "birth" while a transition from E_k to E_{k-1} denotes a "death"
- λ_k is the rate at which birth occurs when the population is of size k. μ_k is defined similarly.



Solution for a Birth Death Process

- Generator matrix for a one-dimensional birth-death process as shown in the last figure:

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \cdots \\ & . & & & \\ . & & & & \\ . & & & & \\ & & & & \end{pmatrix}$$

- The transition rates λ_k, k≥0 are state-dependent birth rates and μ_l,l ≥ 1, are state-dependent death rates.

Solution for a Birth Death Process

- From the equation $\mathbf{0}=\pi\mathbf{Q}$ for CTMC,

  $0 = -\pi_0\lambda_0 + \pi_1\mu_1$

  $0 = -\pi_k(\lambda_k+\mu_k) + \pi_{k-1}\lambda_{k-1} + \pi_{k+1}\mu_{k+1}$, $k\geq 1$
- We get: $\pi_1=(\lambda_0/\mu_1)\pi_0$, $\pi_2=(\lambda_0\lambda_1/\mu_1\mu_2)\pi_0$
- In general, $\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}},\ k \geq 1$
- Since $\sum_i \pi_i = 1$, $\pi_0 = \frac{1}{1+\sum_{k=1}^{\infty}\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}} = \frac{1}{\sum_{k=0}^{\infty}\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}}$
- Condition for convergence of the series,

  $\exists k_0$, such that $\forall k>k_0\ \lambda_k/\mu_k<1$

Queuing Systems

- Kendall's Notations
  - A/B/m – queuing discipline
    - Here A indicates the distribution of inter-arrival times and B denotes the distribution of service times. m is the number of servers
    - A/B = M denotes exponential distribution
    - A/B = G denotes general distribution
  - Queuing discipline could be FCFS, LCFS, etc.
  - Average arrival rate is denoted as λ and mean service time is denoted as μ.

Performance Measures of Queuing Systems

- Probability of the Number of jobs in the system: π_k = P[there are *k* jobs in the system]
- Response time *T* is the <u>total time a job spends in the system</u>
- Waiting time W, is the <u>time a job spends in the queue waiting to be served</u>. $\overline{T} = \overline{W} + \frac{1}{\mu}$

Performance Measures of Queuing Systems

- Queue length Q is the <u>number of jobs in the queue</u>
- <u>Number of jobs in the system</u> K whose mean is given by $\overline{K} = \sum_{k=1}^{\infty} k.\pi_k$
- Little's Theorem states that: $\overline{K} = \lambda\overline{T}$, and $\overline{Q} = \lambda\overline{W}$
- The theorem holds for all queuing disciplines and arbitrary G/m queues.

Markovian Queues: M/M/1 queue

- M/M/1 Queue: Arrival process is Poisson, service times are exponentially distributed and there is one server
- Can be modeled as a birth-death process with birth rate (arrival at) λ and death rate (service rate) μ.
- Unlike general birth-death processes, λ and μ do not depend on the current state.
- $\lambda < \mu$ for the queuing system to be stable
- Steady state probability of the system being empty: $$\pi_0 = \frac{1}{1+\sum_{k=1}^{\infty}\prod_{i=0}^{k-1}\frac{\lambda}{\mu}} = \frac{1}{\sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k} = \frac{1}{1+\frac{\lambda/\mu}{1-\lambda/\mu}} = 1-\lambda/\mu$$

M/M/1 queue

- Probability that there are k jobs in the system:

$$\pi_k = \pi_0 \left(\frac{\lambda}{\mu}\right)^k = \left(1-\frac{\lambda}{\mu}\right).\left(\frac{\lambda}{\mu}\right)^k, \quad k \geq 0$$

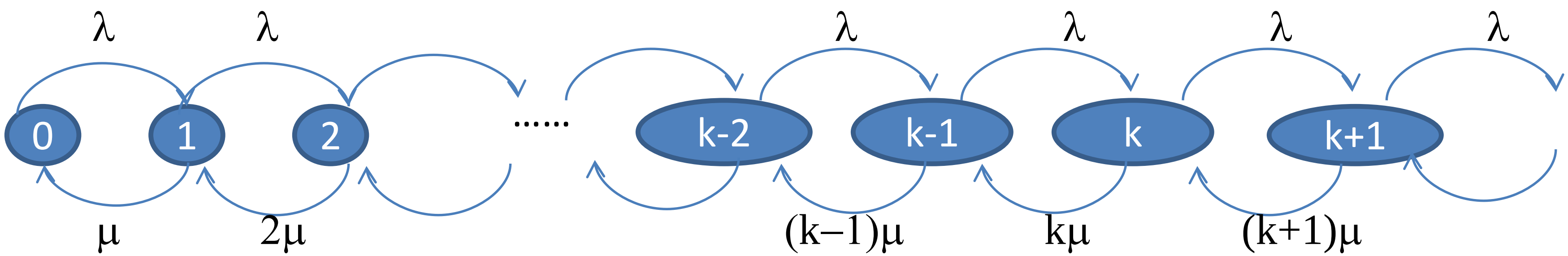
- Defining utilization $\rho=\lambda/\mu$, we get $\pi_0=1-\rho$ and $\pi_k=(1-\rho)\rho^k$
- Mean number of jobs: $\overline{K} = \sum_{k=1}^{\infty} k.(1-\rho)\rho^k = (1-\rho)\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}$
- Using Little's theorem, mean response time:

$$\overline{T} = \frac{\overline{K}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1/\mu}{1-\rho}$$

- Mean waiting time: $\overline{W} = \overline{T} - \frac{1}{\mu} = \frac{1/\mu}{1-\rho} - 1/\mu = \frac{\rho/\mu}{1\text{-}\rho}$
- Using Little's theorem, mean queue length:

$$\overline{Q} = \lambda\overline{W} = \lambda.\frac{\rho/\mu}{1-\rho} = \frac{\rho^2}{1-\rho}$$

M/M/∞ Queue



- $\lambda_k = \lambda$, k=0,1,2,...
- $\mu_k = k\mu$ for k= 1, 2, 3,…

$$\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}$$

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu}} = \frac{1}{\sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu}} = \frac{1}{\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}} = e^{-\lambda/\mu}$$

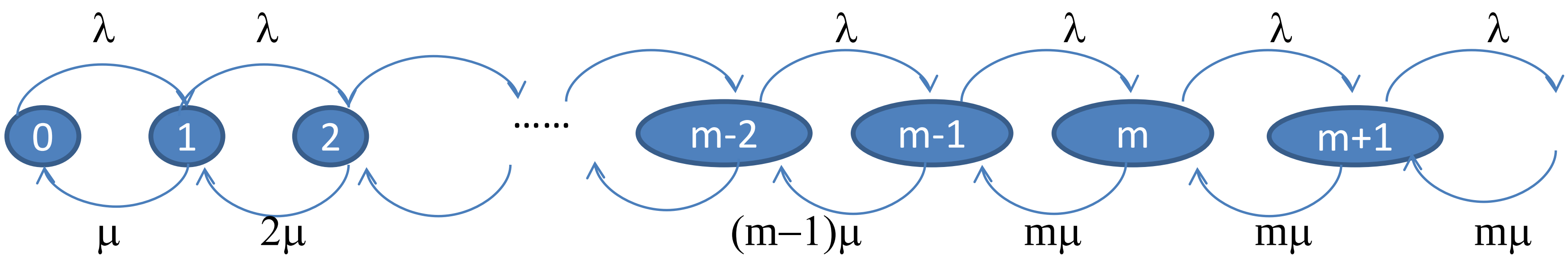
Hence, $\pi_k = e^{-\lambda/\mu} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}$

$$\overline{K} = \frac{\lambda}{\mu} \qquad \overline{T} = \frac{\overline{K}}{\lambda} = 1/\mu$$

M/M/m queue

- There are m servers
- The job at the head of the queue is routed to any server that is available
- Computations are different since as long as less than m servers are busy, a job does not wait in a queue.
- This is different from a case with m different M/M/1 queues. There, if a job comes to a queue with a busy server, it has to experience waiting time even if some of the other servers are free

M/M/m Queue



- $\lambda_k = \lambda$, k=0,1,2,...
- $\mu_k = \min[k\mu, m\mu] = k\mu$ for $0 \le k \le m$; $m\mu$ for $k \ge m$

$$\pi_k = \pi_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}, \quad k \le m$$

$$\pi_k = \pi_0 \cdot \prod_{i=0}^{m-1} \frac{\lambda}{(i+1)\mu} \prod_{j=m}^{k-1} \frac{\lambda}{m\mu} = \pi_0 \cdot \left(\frac{\lambda}{\mu}\right)^k \frac{1}{m! m^{k-m}}, \quad k \ge m$$

$$\pi_k = \pi_0 \cdot \frac{(m\rho)^k}{k!}, \quad k \le m$$
$$= \pi_0 \cdot \frac{(\rho)^k m^m}{m!}, \quad k \ge m$$

- $\rho = (\lambda / m\mu) < 1$

M/M/m Queue

$$\pi_0 = \left[1 + \sum_{k=1}^{m-1} \frac{(m\rho)^k}{k!} + \sum_{k=m}^{\infty} \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}}\right]^{-1} = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}\right]^{-1}$$

- Probability of queuing (P_Q) is

$$P\{\text{queuing}\} = \sum_{k=m}^{\infty} \pi_k = \sum_{k=m}^{\infty} \pi_0 \frac{(m\rho)^k}{m!} \frac{1}{m^{k-m}} = \frac{(m\rho)^m}{m!(1-\rho)}.\pi_0$$

- Values of π_0 is used from the first equation: This is Erlang's C formula
- Mean number of jobs in the system $\overline{K} = m\rho + \frac{\rho}{1-\rho}.P_Q$
- Mean queue length: $\overline{Q} = \frac{\rho}{1-\rho}.P_Q$
- By Little's Theorem, Mean response time

$$\overline{T} = \frac{\overline{K}}{\lambda} = \frac{m\rho}{\lambda} + \frac{\rho}{\lambda(1-\rho)}.P_Q = \frac{1}{\mu} + \frac{\rho}{\lambda(1-\rho)}.P_Q = \frac{1}{\mu} + \frac{P_Q}{m\mu-\lambda}$$

- By Little's Theorem, Mean waiting time: $\overline{W} = \frac{\overline{Q}}{\lambda} = \frac{\rho P_Q}{\lambda(1\text{-}\rho)}$

Problems

a) Find the average queue length, average number of jobs in the system, average queuing delay and average system delay for an M/M/1 queue with average inter-arrival time of 3 minutes and average service time requirement 2 minutes.

b) Determine the above values if the average inter-arrival time is 1.5 minutes and average service time requirement is 1 minute.

c) Find the queuing probability if the number of servers is 2 for the arrival and service rates mentioned above.

d) For a single VCPU of a Virtual Machine, the time taken for executing processes to completion is exponentially distributed with a mean of 5 seconds. The inter-arrival times between processes to the VM are exponentially distributed with mean 8 seconds. Determine the average time for a process to complete under (i) M/M/1 and (ii) M/M/∞ assumptions. (iii) How many VMs should be used for reducing the average process completion time by 50% with respect to M/M/1 queue?