Notes on Queueing Theory

Dr. Deep Medhi, University of Missouri-Kansas City

Chapter 2:
Stochastic Processes, B-D Model and Queues

In this section, we provide brief overview of stochastic processes, and then go into birth-and-death model and queueing analysis. You may want to consult the book by Allen [1] (used often in CS 394) for more material on stochastic processes etc.

1. Stochastic Processes

Let t be a parameter, assuming values in a set T. Let $A(t)$ be a random or stochastic variable for every $t \in T$. The family $\{A(t)|t \in T\}$ is called a stochastic process.

We think about stochastic events that occur over time, i.e, t is in time-space. In terms of measurement, we can quantify time as either a continuous variable or a discrete variable. (for example the ALOHA protocol uses continuous time model while slotted ALOHA uses discrete time model.) Similarly, the state space (values for the stochastic process) can also be continuous (C) or discrete (D). The possible combinations are:

```
Time    state-space
C       D            to  discuss
C       C            e.g. brownian motion - not covered
D       C            Waiting time of n-th arrival in a queue
D       D            to discuss
```

In discrete state space, the stochastic process is called a chain with values denoted, e.g., $S = \{0, 1, ..., m\}$.

2. Discrete-time Markov chain

A stochastic process $\{A_n, n \geq 0\}$ is called a Markov chain if for every $x_i \in S$, we have

$$Pr\{A_n = x_n | A_{n-1} = x_{n-1}, ...A_0 = x_0\} = Pr\{A_n = x_n | A_{n-1} = x_{n-1}\}.$$

(In this definition, we use time to be discrete.) What this means is that for a Markov chain, the probability at time n depends only on the previous state and nothing before that. This is known as the memoryless property of a Markov chain.

Now, what is the probability of the process being in state j given that it was in state i in the preceding time? This is the transition probability from state i to j. It is written as:

$$p_{ij} = Pr\{A_n = j | A_{n-1} = i\}$$

Given, you are at state i at time $n-1$, the probabilities of moving to all states (in the next time slot) must add up to 1, i.e.

$$\sum_{j=0}^{\infty} p_{ij} = 1, \quad \text{for each } i.$$

A Markov chain is called temporally homogeneous if :

$$Pr\{A_n = j | A_{n-1} = i\} = Pr\{A_{n+m} = j | A_{n+m-1} = i\}.$$

The transition probability is then denoted by p_{ij}. For all possible values of i, j, one can denote he the transition probability as a matrix with elements (p_{ij}).

The transition in n-step is given by

$$p_{ij}(n) = Pr\{A_n = j | A_0 = i\}.$$

Since a Markov chain has stationary transition probabilities, we have

$$p_{ij}(n) = Pr\{A_{m+n} = j | A_m = i\} \quad \text{for all } m \geq 0 \text{ and } n > 0.$$

3. Continuous-time, Markov Chain

Let $\{X(t), 0 \leq t < \infty\}$ be a Markov process with countable state space $S = \{0, 1, 2, ...\}$ over continuous time-space t. For, example, $X(t)$ can be the number of customers in the system at time t. For continuous time, discrete space (Markov chains) the transition probability is denoted by,

$$p_{ij}(t) = Pr\{x(t+u) = j | x(u) = i\}, \quad t > 0, \quad i, j \in S$$

Note,

$$\sum_{j} p_{ij}(t) = 1, \quad \text{for each } i.$$

Now, we can consider all the possible cases for i, j at time t giving us a matrix of information. Thus, in matrix notation,

$$P(t) := [p_{ij}(t)] := \text{ transition probability matrix.}$$

4. Chapman-Kolmogorov equation

So far we have mentioned one-step transition probabilities, i.e., probability of A_n given A_{n-1}. C-K equation provides a relation for multiple steps as follows:

$$p_{ij}(n+m) = \sum_k p_{ik}(n) p_{kj}(m). \tag{1}$$

where $n, m = 0, 1, 2, ...$

For continuous time, this can be written as:

$$p_{ij}(s+t) = \sum_k p_{ik}(s) p_{kj}(t). \tag{2}$$

where $s, t \geq 0$.

In matrix notation,

$$P(s+t) = P(s)P(t).$$

Set the initial transition probability matrix as

$$P(0) = I \quad \text{(identity matrix)}$$

i.e.,

$$p_{ij}(0) = 1 \text{ if } i = j; \text{ else } 0 \text{ if } i \neq j.$$

Define,

$$q_{ij} := \lim_{t \to 0} \frac{p_{ij}(t) - p_{ij}(0)}{t} = \lim_{t \to 0} \frac{p_{ij}(t)}{t}, \quad \text{for } j \neq i.$$

$$q_{ii} := \lim_{t \to 0} \frac{p_{ii}(t) - p_{ii}(0)}{t} = \lim_{t \to 0} \frac{p_{ii}(t) - 1}{t}.$$

Note that q's are instantaneous rate. In matrix notation,

$$Q = \lim_{t \to 0} \frac{P(t) - I}{t}.$$

where, $Q = [q_{ij}]$. Now, recall the relation

$$\sum_j p_{ij}(t) = 1.$$

If we move 1 from RHS to LHS of this equation and then divide by t and let $t \to 0$, we get

$$\lim_{t \to 0} \{p_{i1}(t)/t + ... + (p_{ii}(t) - 1)/t + ... + p_{ik}(t)/t + ...\} = 0$$

which implies

$$q_{i1} + ... + q_{ii} + ... + q_{ik} + ... = 0.$$

In short, we have

$$\sum_{j \neq i} q_{ij} + q_{ii} = 0. \tag{3}$$

Typically, we denote $q_{ii} = -q_i$, thus, we have

$$\sum_{j \neq i} q_{ij} = q_i.$$

The matrix, Q, is known as the transition density matrix, or, infinitesimal generator, or a rate matrix.

If the state space S is finite, then

$$Q = \begin{pmatrix} -q_0 & q_{01} & ... & q_{0m} \\ q_{10} & -q_1 & ... & q_{1m} \\ & & ... & \\ q_{m0} & q_{m1} & ... & -q_m \end{pmatrix}.$$

This is another way to write the C-K equation in terms of rate:

$$\begin{aligned} P'(t) = \frac{d(P(t))}{dt} = QP(t). \\ \text{or,} \quad \frac{d(P(t))}{dt} = P(t)Q. \end{aligned} \tag{4}$$

Now, consider the vector $\pi(t) := \{\pi_0(t), \pi_1(t), ...\}$. This is the probability vector of the state of the system at time t, i.e., probability of being at state 0 in time t is $\pi_0(t)$ etc.

Now, we move to the steady-state situation, i.e, what is the system going to be like in steady-state as $t \to \infty$.

It is easy to see that

$$\pi(t) = \pi(0)P(t).$$

Also,

$$\frac{d(\pi(t))}{dt} = \pi(0)P'(t) = \pi(0)P(t)Q = \pi(t)Q. \tag{5}$$

If we denote the steady-state probability vector by π, i.e.

$$\pi = \lim_{t \to \infty} \pi(t)$$

then, from the relation (5) and letting $t \to \infty$, we get (since, derivative of a constant is zero),

$$0 = \pi Q$$

with the requirement that probabilities sum up to 1

$$\pi_1 + \pi_2 + ... = 1.$$

if e is a column vector of the same dimension as π with each component being 1, we can write this in compact form as $\pi e = 1$.

Note that

$$0 = \pi Q, \quad \pi e = 1 \tag{6}$$

is a linear system of equation which has a unique solution (under certain conditions).

To illustrate, suppose that we have a system that takes three values (0, 1, 2). The probabilities are:

$$\pi = (\pi_0, \pi_1, \pi_2)$$

and $\pi Q = 0$ is

$$[\pi_0, \pi_1, \pi_2] \begin{bmatrix} -q_0 & q_{01} & q_{02} \\ q_{10} & -q_1 & q_{12} \\ q_{20} & q_{21} & -q_2 \end{bmatrix} = [\,0 \quad 0 \quad 0\,]$$

and $\pi e = 1$ is

$$\pi_0 + \pi_1 + \pi_2 = 1.$$

If we know q's, then we can solve for π.

5. Birth-and-Death Process

This is a special case of continuous-time Markov chain. You can only transition to a neighbor (one step away) or stay where you are in an infinitesimal time period, and the rate is state-dependent, one defined for arrival and the other for departure.

$$\begin{aligned} q_{i,i+1} &= \lambda_i, \quad i = 0, 1, 2, ... \\ q_{i,i-1} &= \mu_i, \quad i = 1, 2, ... \\ q_{i,j} &= 0, \quad \text{for } j \neq i+1, i-1, i. \end{aligned} \tag{7}$$

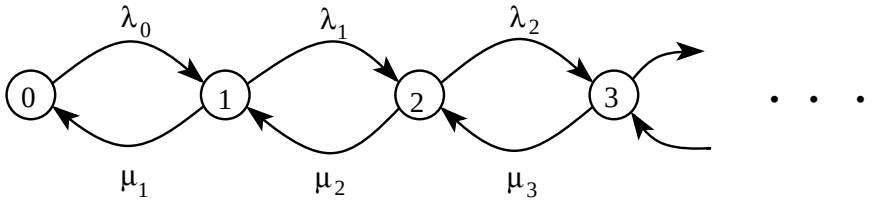
Due to this situation and the requirement that the row sum to be zero (from (3)), we have

$$q_i = -q_{ii} = \lambda_i + \mu_i, \quad \text{for } i \text{ not in boundary state .}$$

Now, what does the matrix Q look like? We essentially get a band around the diagonal with one element on each side (tridiagonal matrix); all of the other elements are zero.

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

Pictorially, the state-transition diagram is as follows:



Now, let's go back to C-K equation. For the birth-and-death process, we can rewrite the C-K equation (from (4)) as

$$\begin{aligned} p'_{ij}(t) &= -(\lambda_j + \mu_j)\, p_{ij}(t) \;+\; \lambda_{j-1} p_{i,j-1}(t) \;+\; \mu_{j+1} p_{i,j+1}(t), \quad \text{for } j \geq 1 \\ p'_{i0}(t) &= -\lambda_0 p_{i0}(t) + \mu_1 p_{i,1}(t). \end{aligned} \tag{8}$$

Suppose that λ_j's and μ_i's are non-zero. Then, the Markov chain is irreducible (means every state can be reached from every other state by chaining), it can be shown that in steady-state

$$\lim_{t\to\infty} p_{ij}(t) = \pi_j$$

exists and are independent of the initial state i.

Revisiting C-K equation above and using the fact that the derivative of a constant is zero, we have

$$0 = -(\lambda_j + \mu_j)\, \pi_j \;+\; \lambda_{j-1}\pi_{j-1} \;+\; \mu_{j+1}\pi_{j+1}, \quad \text{for } j \geq 1 \tag{9}$$

$$0 = -\lambda_0 \pi_0 + \mu_1 \pi_1. \tag{10}$$

These are also known as balance equations. In fact, the above is nothing but $\pi Q = 0$, specialized for the B-D model. The nice thing is that this linear system of equations is easily solvable analytically.

From (10), we have

$$\pi_1 = \frac{\lambda_0}{\mu_1}\pi_0.$$

For $j = 1$ from (9), we have

$$0 = -(\lambda_1 + \mu_1)\, \pi_1 \;+\; \lambda_0 \pi_0 \;+\; \mu_2 \pi_2$$

which implies [using (10)],

$$\mu_2 \pi_2 = \lambda_1 \pi_1 + \mu_1 \pi_1 - \lambda_0 \pi_0 = \lambda_1 \pi_1.$$

Thus,

$$\pi_2 = \frac{\lambda_1}{\mu_2}\pi_1 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2}\pi_0.$$

Generalizing, we get

$$\pi_j = \frac{\lambda_0 \lambda_1 ... \lambda_{j-1}}{\mu_1 \mu_2 ... \mu_j}\pi_0, \qquad j \geq 1. \tag{11}$$

Recall that all the probabilities must sum to 1, i.e.,

$$\pi_0 + \pi_1 + \pi_2 + ... = 1.$$

which, using the previous relation, gives

$$\pi_0(1 + \lambda_0/\mu_1 + \lambda_0\lambda_1/(\mu_1\mu_2) + ...) = 1$$

Thus, as long as the sum is convergent, we can calculate π_0, and thus, all the other probabilities, π_j, due to (11). This is important since this probabilities help us determine something about the system behavior as we will see soon.

Aside: for the system to converge, we need the condition

$$\sum \frac{\lambda_0\lambda_1 \cdots \lambda_{j-1}}{\mu_1\mu_2 \cdots \mu_j} < \infty.$$

[A finite note is that the balance equation can be easily written by looking at the steady-state equations, since anything out of a state should some up to anything coming in, i.e. the “flow conservation" idea.]

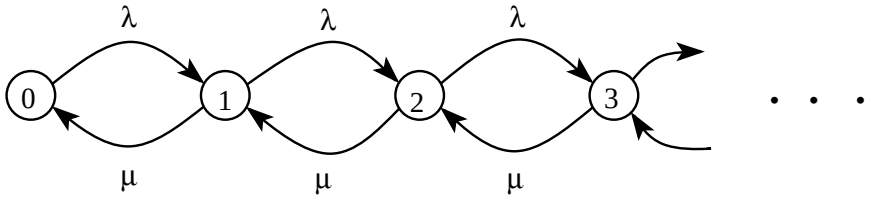
6. Special case: M/M/1 Queue

An important special case of B-D process is the case where transition rates are state independent and are fixed, one for birth another for death, i.e.,

$$\lambda_j = \lambda$$

$$\mu_j = \mu.$$

For this example, due to Poisson property (we will visit shortly), the interarrival time is exponentially distributed with mean $1/\lambda$ and the service time is exponentially distributed with mean $1/\mu$. Also, λ known as the average arrival rate. Remember, we are still talking about Markov chains. This is a special case and is identified as the M/M/1 queue (known as Kendall’s notation) where first M is for the arrival, the second M is for the service time and the third entry is to denote we have one server. The state-transition diagram can be shown as below:



For this special case, we get from (11),

$$\pi_j = \left(\frac{\lambda}{\mu}\right)^j \pi_0, \quad j = 1, 2, ...$$

Now, $\pi e = 1$ is

$$\pi_0 \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + ...\right] = 1.$$

For convergence in this case, we need the condition $\sum(\lambda/\mu)^j < \infty$. Since this is a geometric series, convergence condition is satisfied by the requirement that $\lambda/\mu < 1$. We, thus, have

$$\pi_0 = 1 - \frac{\lambda}{\mu}.$$

and hence,

$$\pi_j = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^j, \quad j = 1, 2, ...$$

If we write $\rho = \lambda/\mu$, which is known as utilization or traffic intensity, then

$$\pi_j = (1 - \rho)\rho^j.$$

The convergence condition also has a simple physical interpretation. Since $\lambda/\mu = \rho < 1$, this means the average arrival rate should be less than the average service rate; if it is other way around, the system will overflow.

What has the M/M/1 queue anything to do with network design and analysis? We have mentioned earlier that to get a handle on network modeling and performance, we have to have some idea about traffic. The simplest network we can think of is just a *network link*. So, here λ would refer to the arrival rate to a network link while μ will refer to the average service rate of the link with the link being one server. Specifically, for data networks, we can consider the arrival rate to be in packets/sec. Let

$$N(t) = \text{ number of packets in the link at time } t$$

Q – What is the average number of packets in the link?

$$N = \lim_{t\to\infty} E\{N(t)\} = \sum_{j=0}^{\infty} j\pi_j = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}.$$

Thus, using the steady-state probabilities, we can obtain the result for N. What about other measures such as average packet delay?

To obtain this, we need the use the following result known as *Little's Law*:

```
avg. number of packets in system in steady state (N)
  = avg. arrival rate * avg. packet delay in system
                                    in steady state (T)
```

That is,

$$N = \lambda T.$$

Rearranging, we get

$$T = N/\lambda = \frac{1}{\mu - \lambda} = \frac{\rho}{\lambda(1-\rho)}.$$

Average waiting time in the queue, W, is

$$W = \text{ Avg. waiting time in System – Avg. waiting time in Service } = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}.$$

Again, by Little's Law which is also applicable in queues, the average number of customers in queue is,

$$N_q = \lambda W = \frac{\rho^2}{1-\rho}.$$

Thus, we have shown above four performance measures that may be of interest for a network link which can be obtained once you know the average arrival rate and the average service rate.

Remark: An important assumption about the M/M/1 is that the input population is considered to be infinite. In generic terms, the term 'customer' is used rather than packets since such models are applicable in other areas besides communication networks.

We have discussed so far delay results for the average value only. How about 90-th percentile value? For M/M/1, we have the following:

90-th percentile system delay = $T\,ln(10)$.

90-th percentile queueing delay = $max\,\{T\,ln(10\rho), 0\}$.

Example # 1:

Suppose average arrival rate (λ) is 100 pps, and the service rate (μ) is 200 pps, then the average number of packets in the system is $N = \lambda/(\mu - \lambda) = 100/(200 - 100) = 1$.

The following is a table for average number in the system and the average delay (in the system) for different values of λ and μ (Note the impact on delay when λ and μ are scaled):

λ	μ	N	T
100	200	1	10 ms
150	200	3	20 ms
175	200	7	40 ms
1000	2000	1	1 ms
1500	2000	3	2 ms
1750	2000	7	4 ms

7. On exponential distribution

This is a good time to quickly go over a couple of things about exponential distribution. The pdf of the exponential distribution with parameter λ is given by

$$f(x) = \lambda \mathrm{e}^{-\lambda x}, \quad x > 0.$$

The mean and the variance are given by

$$E(X) = \frac{1}{\lambda}, \qquad V(X) = \frac{1}{\lambda^2}.$$

8. Poisson Process

A pure-birth homogeneous process, i.e., $\mu_j = 0, \lambda_j = \lambda$, is a Poisson process. Another way:

A stochastic process $\{A(t)|t \geq 0\}$ taking non-negative integer values is said to be a Poisson process with rate λ if

1. $\{A(t)\}$ is a counting process.
2. The number of arrival that occur in disjoint time interval are independent.
3. The number of arrivals in interval of length τ is Poisson distributed with parameter $\lambda\tau$

$$Pr\{A(t+\tau) - A(t) = n\} = exp(-\lambda\tau)(\lambda\tau)^n/n!, \quad n = 0, 1, ...$$

Note that exp is the exponential function, $exp(x) = \mathrm{e}^x$, where e is the Euler number (=2.718281828...).

Important properties of a Poisson process:

1) Interarrival times are independent and exponentially distributed with parameter λ.

$$\tau_n = t_{n+1} - t_n$$

$$Pr\{\tau_n \leq s\} = 1 - exp(-\lambda s) = 1 - \mathrm{e}^{-\lambda s}.$$

2) Sum of two independent Poisson processes is a Poisson process with rate being the sum of the two. This holds for more than two also.

It is important to note that the Poisson arrival can also be described by a pure-birth homogeneous process.

9. Revisiting a network link

In the section on M/M/1, we mentioned why the results are applicable for a data network link. We need to reiterate that M/M/1 results are applicable for Poisson arrival (discussed above) and exponentially distributed service time.

Packet arrivals characterized by Poisson arrival is a reasonable assumption, although this is NOT always a good one. However, this helps us get started on getting a handle on doing some analysis.

Typically, when we think of a network link we often have some idea about the rate such as 56 Kbps etc. This is a deterministic rate. Now, how do we get exponentially distributed service time? In a link model, the other factor that comes into the picture is the size of the packet that has arrived. Here is where we make another *assumption*: the packet length is exponentially distributed with mean length $1/\hat{\mu}$ bits. Now if we denote the speed of a link by C bps; then, the average service rate of packets is

$$\mu = \hat{\mu}C$$

which is exponentially distributed. Thus, we can use the M/M/1 model for exponentially distributed packet size for studying a network link.

Example # 2:

If link speed is 1.5 Mbps (i.e., a T1-link), and the average packet size is 1 Kbits, then the average service rate μ is 1.5 Mbps/1Kbits = 1500 packets per sec (pps).

10. A view on Statistical Multiplexing

The next thing to realize is that all the packets that arrive to a data network link is statistically multiplexed. This is a key behavior of packet-switched networks. This also results in queueing of packets if all the capacity of a link is used up at a certain time – this points to the average packet delay we discussed earlier.

Consider m statistically identical and independent Poisson packet streams each with an arrival rate of λ/m packets per sec, transmitted over a communication link with exponentially distributed service time with mean service rate μ. In case of single “pipe", we have total Poisson arrival as $m\lambda/m = \lambda$ due to one of properties of Poisson process mentioned earlier. Using M/M/1 model, we get the avg. delay to be

$$T_a = \frac{1}{\mu - \lambda}.$$

If, in, the link is divided into m separate pipes (partitioned), then the service rate of each pipe is μ/m, i.e., $1/m$-th of the rate of the original pipe. In this case, the avg. delay (again using M/M/1 on one partition):

$$T_b = \frac{1}{\mu/m - \lambda/m} = \frac{m}{\mu - \lambda}.$$

Thus, splitting the line into m pipes increases the avg delay by m times! This is an interesting phenomenon about statistical multiplexing; as a rule of thumb, it is better to have a fatter link than partition the link into lower speeds since it results in more delay on average.

What about the effect on average number of packets due to the scenarios discussed here?

11. Inverse problem: Find service rate for a given average delay

We now pay attention to the simplest design problem when just a single network link is considered. We know that the average delay for M/M/1 model is given by:

$$T = \frac{1}{\mu - \lambda}.$$

Suppose, you are given λ, the offered traffic in arrival rate (based on forecasted data) and the requirement that the average link delay should not be more than T_{max}. To meet this delay, T_{max}, we need

$$\frac{1}{\mu - \lambda} \leq T_{max}$$

which implies:

$$\mu \geq \frac{1}{T_{max}} + \lambda.$$

Thus, we get a bound on the minimum service rate required to meet the requirement. Since service rate corresponds to the link speed, this mean we can determine minimum link speed required if the average packet size is known. Although this is a simple result, it ties offered traffic and QoS requirement to the service rate to be provided.

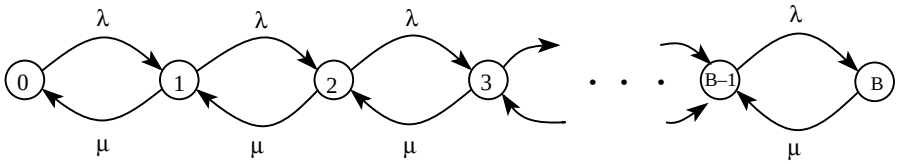
Example # 3:

Suppose that the average arrival rate is 500 pps, and we are given that the average packet size is 4Kbits. If the system goal is to provide an average delay of 30 ms or less, what should be the link speed?

Here $30ms \geq 1/(\mu - \lambda) = 1/(\mu - 500)$. This implies that the service rate μ is 533.33 pps or higher. Since, the average packet size is 4 Kbits, the link speed needed is at least 2.13 Mbps or higher.

12. Finite Buffer System: M/M/1/B

So far, we have talked about M/M/1. Often, in reality, we do not have an infinite queue to hold all the packets for transfer due to finite buffer size. Thus, another model of interest is the finite buffer system. How is the system behavior under finite buffer scenario? Assume that the size of the finite buffer is B, including server (but customer population is still infinite). That means, you can queue only up to $B - 1$. If there is an arrival when the buffer is full, then that arrival is dropped (lost); this is packet dropping due to buffer overflow. Again, B-D model can be used here by catering specifically to this case. Here, the state-transition diagram is as follows:



and we have

$$\lambda_j = \begin{cases} \lambda, & \text{if } 0 \le j \le B-1 \\ 0, & \text{otherwise.} \end{cases}$$

$$\mu_j = \begin{cases} \mu, & \text{if } 1 \le j \le B \\ 0, & \text{otherwise.} \end{cases}$$

It can be derived (from B-D model or from the special case M/M/1) that

$$\pi_j = \pi_o \prod_{i=0}^{j-1} \frac{\lambda}{\mu} = \pi_0 \left(\frac{\lambda}{\mu}\right)^j, \quad j \le B \quad (\lambda/\mu \ne 1).$$

and,

$$\pi_0 = \left[1 + \sum_{j=1}^{B} \left(\frac{\lambda}{\mu}\right)^j\right]^{-1}$$
$$= \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{B+1}} = \frac{1-\rho}{1-\rho^{B+1}}, \qquad (\rho = \lambda/\mu \ne 1)$$

For $\rho = 1$, the following is true:

$$\pi_j = \frac{1}{B+1}, \quad j \le B.$$

Note that the packet loss probability is given by π_B.

The mean number of packets in the system:

$$N = \sum_{j=0}^{B} j\pi_j = \frac{\rho}{1-\rho} - \frac{(B+1)\rho^{B+1}}{1-\rho^{B+1}} \quad (\rho \ne 1)$$

and

$$N = B/2 \qquad \text{if } \rho = 1.$$

The mean number of packets at the server, $E\{N_s\}$, is:

$$E\{N_s\} = Pr\{N=0\}E\{N_s|N=0\} + Pr\{N>0\}E\{N_s|N_s>0\}$$
$$= \pi_0 \times 0 + (1-\pi_0) \times 1.$$

Thus, the mean number of packets in the queue (for $\rho \ne 1$):

$$N_q = N - E\{N_s\} = N - (1-\pi_0) = \frac{\rho}{1-\rho} - \rho\frac{1+B\rho^B}{1-\rho^{B+1}}.$$

while for $\rho = 1$, we have

$$N_q = B/2 - B/(B+1).$$

Effective arrival rate is given by

$$\lambda' = \lambda \sum_{j=0}^{B-1} \pi_j = \lambda(1-\pi_B).$$

The mean response time (average delay) is

$$T = N/\lambda' = N/\left[\lambda(1-\pi_B)\right].$$

The mean waiting time is

$$W = N_q/\lambda' = N_q/\left[\lambda(1-\pi_B)\right].$$

13. M/M/m model: m parallel servers

Another model of interest in communication networks is m parallel servers. This model comes up in the case of a communication link which is divided into m circuits or channels. An arriving customer can take any of the circuits on arrival, OR, has to wait if all the servers (circuits) are busy. Infinite population assumption is made here. [this model is close to a telephone network link, but not quite; we'll see later where the difference is].

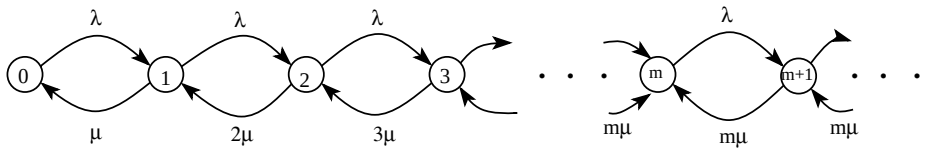
This is again a B-D model, where

$$\lambda_j = \lambda,$$

and

$$\mu_j = \begin{cases} j\mu, & \text{if } 1 \le j \le m \\ m\mu, & \text{for } j > m. \end{cases}$$

The state-transition diagram is



Note that for M/M/m model, the utilization, ρ, is

$$\rho = \frac{\lambda}{m\mu} < 1.$$

Here, we can show (from B-D model) that

$$\pi_j = \begin{cases} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}\pi_0, & \text{for } 1 \le j \le m \\ \left(\frac{\lambda}{\mu}\right)^j \frac{1}{m!m^{j-m}}\pi_0, & \text{for } j > m. \end{cases}$$

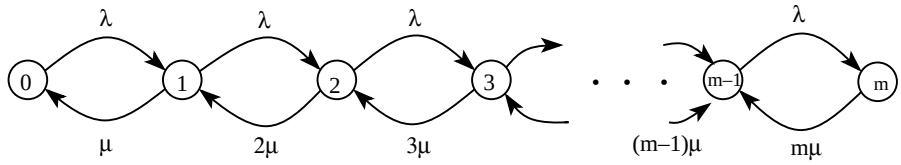
Further,

$$\pi_0 = \left[\frac{(\lambda/\mu)^m}{m!(1-\lambda/(m\mu))} + \sum_{j=0}^{m-1} \frac{(\lambda/\mu)^j}{j!} \right]^{-1}.$$

Given π's, different performance measures can be derived.

14. M/M/m/m model: m parallel servers, loss system

In this case, the buffer size is m (including for servers) with m parallel servers. I.e., the buffer size is the same as the number of servers; thus, if all the servers are occupied, then an arriving customer is rejected. This is the case in the telephone network link model - if all the circuits are busy, an arriving call is rejected (instead of being queued up if one of them become free later). This is a very important system and is sometimes known as the *loss* system. The state-transition diagram is



Here,

$$\lambda_j = \begin{cases} \lambda, & \text{if } 0 \le j \le m-1 \\ 0, & \text{otherwise.} \end{cases}$$

$$\mu_j = \begin{cases} j\mu, & \text{if } 1 \le j \le m \\ 0, & \text{otherwise.} \end{cases}$$

In this case,

$$\pi_j = (\lambda/\mu)^j \frac{1}{j!}\pi_0, \quad j = 1, ..., m.$$

Since sum of probabilities is 1, we can show that

$$\pi_0 = \left[\sum_{j=0}^{m} \frac{(\lambda/\mu)^j}{j!}\right]^{-1}.$$

Specifically, the probability of an arriving call being blocked is the probability of being in state m:

$$\pi_m = \frac{\frac{(\lambda/\mu)^m}{m!}}{\sum_{j=0}^{m} \frac{(\lambda/\mu)^j}{j!}}.$$

This is the well-known *Erlang-B Blocking/loss* formula and is usually denoted by $E(\lambda/\mu, m)$. Further, we wrote

$$a = \lambda/\mu$$

to denote the offered load in Erlang (this is NOT utilization) giving us:

$$E(a, m) = \frac{\frac{a^m}{m!}}{\sum_{j=0}^{m} \frac{a^j}{j!}}.$$

The average number in the system, N, is given by

$$N = a(1 - E(a, m)).$$

It is interesting to note that the Erlang-B loss formual has the following recurrence relation:

$$E(a, m) = \frac{aE(a, m-1)}{m + aE(a, m-1)}$$

with the starting point $B(a, 0) = 1$.

This can also be computed iteratively: Write

$$E(a, m) = 1/d(a, m).$$

Then,

$$d(a, m) = m\, d(a, m-1)/a + 1.$$

Thus, the iteration to compute blocking given offered load of a erlangs and the number of channel, m, is:

procedure erlangb (a, m)

$d = 1$

for $j = 1, ..., m$

  $d = j * d/a + 1$

endfor

$b = 1/d$

return (b)

Note that here the offered load (traffic) is given in a erlangs. Recall our discussion at the end of chapter 1. We stated there that typically for telephone networks, we provide offered traffic as

Offered Load = Number of call attempts/ hour * average call holding time

On close scrutiny, this is λ/μ, where $1/\mu =$ average call holding time. Again note that the assumption of arrival is Poisson. Further, we assume call holding time to be exponentially distributed.

Example # 4:

Suppose that the call arrival rate is on average 10 per hour, and the average duration of calls is 6 minutes, then the offered load in erlangs is $a = 10/60 \times 6 = 1$erl.

If now the link capacity is 1 ($= m$), then the call blocking probability is

$$E(1, 1) = 1/(1 + 1) = 0.5$$

If now the capacity is increased to 2 ($= m$), then the call blocking probability reduces to $E(1, 2) = 0.2$.

Note that Erlang-B formula has non-linear property. For example if offered load is 2 erl, and the capacity is 2 channels, then the call blocking probability is $E(2, 2) = 0.4$ which is smaller than when 1 erl load was offered to link capacity of 1 channels (see above).

Example # 5:

This example shows how increase in call holding time impacts blocking.

Suppose that the average call arrival rate is 100 calls/hour, and the call holding time is 3 min, then the offered load is 5 erl. If the link capacity is 10, then call blocking probability is $E(5, 10) = 0.018$.

Now suppose that the average call duration time increases (somehow) to 30 minutes; then the offered load is 50 erl; Thus, the call blocking probability increases to $E(50, 10) = 0.8$. Note that in this example, the average call arrival rate did NOT increase at all. This show that the change in the average call holding time can also impact on network call blocking performance (a phenomenon observed with more users using the telephone for Internet dial-up). Now, if we wanted to keep the call blocking at the initial value of 0.018, then for 50 erl of load, we will need at least 61 channels! (This is obtained using the Inverse Erlang-B formula discussed later.)

15. M/M/∞ model: infinite servers

This is the limiting case of M/M/m model with $m = \infty$. The balance equation reduces to

$$\lambda \pi_{j-1} = j\mu\pi_j, \quad j = 1, 2, ...$$

Thus,

$$\pi_j = (\lambda/\mu)^j \frac{1}{j!}\pi_0, \quad j = 1, 2, ...$$

From the condition, $\sum_j \pi_j = 1$, we obtain that

$$\pi_0 = exp(-\lambda/\mu).$$

Thus,

$$\pi_j = (\lambda/\mu)^j \frac{exp(-\lambda/\mu)}{j!}, \quad j = 0, 1, 2, ...$$

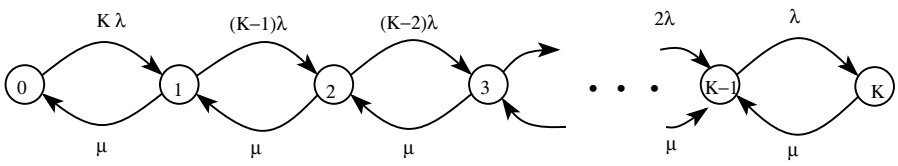
Thus, in steady-state, the number in the system is Poisson distributed with parameter λ/μ. The average number in the system (i.e. number of busy channels) is

$$N = \frac{\lambda}{\mu}.$$

However, we know that this nothing but a for erlang offered load. Thus, this gives us another interpretation for the definition of offered load in erlangs: *the average number of busy servers (channels) if there were infinite number of servers available*.

16. M/M/1/ /K model: single server, finite population

In this case, we have a finite population, K. The state-transition diagram is:



In this case,

$$\lambda_j = \begin{cases} \lambda(K-j), & \text{if } 0 \le j \le K \\ 0, & \text{otherwise.} \end{cases}$$

$$\mu_j = \begin{cases} \mu, & \text{if } 1 \le j \le K \\ 0, & \text{otherwise.} \end{cases}$$

Steady-state probabilities are given by

$$\pi_j = \frac{K!}{(K-j)!}\left(\frac{\lambda}{\mu}\right)^j \pi_0, \qquad 0 \le j \le K,$$

where

$$\pi_0 = \left[\sum_{j=0}^{K}\left(\frac{\lambda}{\mu}\right)^j \frac{K!}{(K-j)!}\right]^{-1}.$$

This model is often applicable for performance evaluation of a computer system where finite population is a good assumption.

17. Inverse Erlang-B: given a and blocking, find the number of channels

This is another inverse design problem for a network link where the offered load is given in erlangs and the acceptable blocking level is also given, and we are to find the number of channels required in the link to meet the acceptable level of blocking (known as grade-of-service).

The problem is to find an integral minimum m for given a and acceptable blocking b:

$$min\ \{m \mid E(a,m) \le b\}.$$

In this case, the inverse is not easy to calculate as in the M/M/1 case with delay. Thus, an algorithmic approach is needed. We provide below a rough sketch:

```
Given offered_load and b_goal, estimate number of trunks
=======================================================
```

```
assign tolerance

estimated_trunk = (int) offered_load

b_test = erlangb(offered_load, estimated_trunk)

if { b_test > b_goal } then
   b_low = b_test
   c_low = estimated_trunk
   while { b_test > b_goal } do
      estimated_trunk += 20
      b_test = erlangb(offered_load, estimated_trunk)
      b_high = b_test
      c_high = estimated_trunk
   endwhile
else
   b_high = b_test
   c_high = estimated_trunk
   while { b_test < b_goal } do
      estimated_trunk -= 20
      b_test = erlangb(offered_load, estimated_trunk)
      b_low = b_test
      c_low = estimated_trunk
   endwhile
endif

diff_b = b_low - b_goal

while {  abs(diff_b) >  tolerance && diff_c <> 1 }  do
   c_mid = (int) (( c_low + c_high )/2.0 )
   b_test = erlangb(offered_load, c_mid)
   diff_b = b_test - b_goal
   diff_c = c_high - c_low
   if { diff_b > 0.0 } then
      c_low = c_mid
      b_low = b_test
   else
      c_high = c_mid
      b_high = b_test
   endif
endwhile

if {diff_b < 0.0 } then
  no_trunks = c_mid
else
  no_trunks = c_mid+1
endif
```

Jagerman [5] presents a much elegant method where the number of channels is assumed to take non-integral values.

18. M/G/1: single non-exponential server, infinite population

So far, we have considered exponential server case. However, the service distribution could be non-exponential. Thus, results for non-exponential service time are also desirable. There are several results for this one; however, the mathematical derivation is quite complex and beyond the scope of this course. As such, we list a couple of results for M/G/1 system below:

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)}$$

where $\overline{X^2}$ is the second moment of the service time distribution.

$$N = \rho + \frac{\lambda^2 \overline{X^2}}{2(1-\rho)}.$$

$$N_q = \frac{\lambda^2 \overline{X^2}}{2(1-\rho)}.$$

$$T = \frac{1}{\mu} + W.$$

Another important special case of M/G/1 (besides M/M/1) is M/D/1, i.e., the service time is deterministic. The deterministic service time is applicable, for example, in ATM (Asynchronous Transfer Mode) networks where cell size are fixed and thus service rate is fixed (we will address at some point later whether Poisson arrival is a good assumption for ATM traffic, or for other networks such as Internet). Results for M/D/1 are listed below:

$$W = \frac{\rho}{2\mu(1-\rho)}, \qquad N = \rho + \frac{\rho^2}{2(1-\rho)},$$

$$T = \frac{2-\rho}{2\mu(1-\rho)}, \qquad N_q = \frac{\rho^2}{2(1-\rho)}.$$

19. G/M/1 and G/G/1

G/M/1 refers to the case where the inter-arrival time is non-exponential while the service time is still exponential. Finally, G/G/1 means both inter-arrival time and service time are non-exponential (not necessarily the same distribution).

20. An example of a non-exponential distribution: hyper-exponential

The pdf for two-stage hyperexponential distribution (H_2) is given by

$$f(x) = \alpha_1 \mu_1 \mathrm{e}^{-\mu_1 x} + \alpha_2 \mu_2 \mathrm{e}^{-\mu_2 x}, \qquad \alpha_1, \alpha_2 \geq 0 \text{ and } \alpha_1 + \alpha_2 = 1$$

The mean is

$$E(X) = \frac{\alpha_1}{\mu_1} + \frac{\alpha_2}{\mu_2}.$$

The variance is

$$V(X) = \left(2 \sum_{i=1}^{2} \frac{\alpha_i}{\mu_i^2} \right) - \left(\frac{\alpha_1}{\mu_1} + \frac{\alpha_2}{\mu_2} \right)^2.$$

The distribution is considered to be balanced if:

$$\frac{\alpha_1}{\mu_1} = \frac{\alpha_2}{\mu_2}.$$

The squared co-efficient of variation (c^2) is given by:

$$c^2 = \frac{V(X)}{[E(X)]^2} = \frac{E(X^2)}{[E(X)]^2} - 1.$$

(note: $E(X^2)$ is the second moment).

How about generating hyperexponential distribution? Given c^2 and μ, a hyperexponential with *balanced* mean can be generated as follows:

Calculate α_1 and α_2 as follows:

$$\alpha_1 = \frac{1}{2} \left(1 - \sqrt{\frac{c^2 - 1}{c^2 + 1}} \right)$$

$$\alpha_2 = 1 - \alpha_1$$

and μ_1 and μ_2 as follows:

$$\mu_1 = 2\alpha_1 \mu$$

$$\mu_2 = 2\alpha_2 \mu.$$

21. Considering M/M/1 and M/D/1 together

In the figures below, we plot N and T for different values of ρ for the systems M/M/1 and M/D/1. Notice the difference as $\rho \to 1$.