VJ Davey

CSci 426 Notes Queueing Theory

1. Markov Chain

2. **Kendall's Notation**
   Kendall Notation is *the standard system used to describe and classify a queueing node*. It was proposed by D.G. Kendall in 1953. It originally follows a pattern A—S—c, where A denoted the time between arrivals to the queue, S denotes service time, and c denoted the number of servers at the node. Since its original proposal, it has been extended to A—S—c—K—N—D, where K is the capacity of the node, N is the size of the population of jobs to serve, and D is the queueing discipline.
   When the final parameters K, N, and D are not specified, it is assumed that $K = \infty$,$N = \infty$, and $D = FIFO$.

   (a) *The Arrival Process, A - Generally assumed to be Markovian/Memoryless (represented by the symbol M) at most basic level*. There are other kinds of arrivals including batch Markov(M^X), Markovian arrival process(MAP), Degenerate distributions(D), etc.

   (b) *The Service Time Distribution, S - Generally assumed to be Markovian/Memoryless (represented by symbol M) at most basic level*. There are other kinds of service time distributions including bulk Markov(M^Y), Degenerate distribution(D), General distribution(G), etc.

   (c) *The Number of Servers, c*

   (d) *The Number of Places in the System, K - This is the capacity of the system.* It is the maximum number of customers allowed inside the node including those in service. Generally assumed to be unlimited unless otherwise mentioned.

   (e) *The Calling Population, N - The size population where the customers come from.* This is generally assumed to be unlimited unless otherwise mentioned.

   (f) The Queue's Discipline, D - The service priority for the jobs in the system's queue. Generally assumed to be FIFO unless otherwise mentioned.

3. **M/M/1 Queue**
   The M/M/1 queue is a single server queue where job arrivals are random and job service times have an exponential distribution. This is the most elementary queueing model, and it is attractive object of study since closed form expressions can be obtained for many metrics of interest in this model.
   The state space of the model corresponds to the number of customers in the system, including any in service.
   Arrivals occur at a rate λ and move the process from state i to $i+1$.
   Service times have an exponential distribution with rate parameter μ and $1/\mu$ is the mean service time.
   A single server serves customers one at a time from the front of the queue, according to a FIFO discipline. When the service is complete, the customer leaves the queue and the number of customers in the system reduces by one.
   The buffer is of infinite size, and there is no limit to the number of customers it can contain.

The model is considered stable only if $\lambda < \mu$. If, on average, arrivals happen faster than service completions, the queue will grow to an infinite size and the system wont be stable. Various performance measures can be computed explicitly for the M/M/1 queue. **We write $\rho = \lambda/\mu$ for the utilization of the buffer and require $\rho < 1$ for the queue to be stable**. ρ represents the average proportion fo time which the server is occupied.
The number of customers in the system is equal to deducing which state of the Markov chain that the process is in. *The probability that the process is in state i is*

$$\pi_i = (1-\rho)\rho^i$$

The number of customers in the system is geometrically distributed with parameter $1-\rho$. Thus the average number of customers in the system is $\rho/(1-\rho)$, and the variance of the number of customers in the system is $\rho/(1-\rho)^2$. This result holds for any work conserving service regime, such as processor sharing.

4. Little's Law
5.