Statistical Simulation of Queues

Sunidhi Taneja

ID-20804148

Under the guidance of Prof. Ravi Mazumdar

Queues can be modeled mathematically in several ways. This project aims at looking at some of the important queues from two such ways: from congestion point of view and workload point of view. The basic approach behind the simulation is discussed and some of the important parameters like expected number of customers in the queue, expected delay etc. are analysed. The known theoretical results are validated by comparing them with the simulation results. For all the models, we assume unlimited buffer space, infinite population of users and FIFO(First In First Out) service discipline, unless specified. Also, we assume that the system is work conserving.

Random Number Generators

Uniform Distribution

The function numpy.random.random() is used to generate a random number between 0 and 1. To test that the random number generator simulates the values generated by a $Uniform(0,1)$ random variable, we generate a very large number of random numbers and then compare the simulated cumulative distribution function(CDF) with the theoretical CDF for a uniform random variable, they should be same. For a $Uniform(0,1)$ random variable, $Pr(U \leq x) = x \ \forall \ x \ \epsilon \ (0,1)$. Figure 1 shows that the simulated CDF and theoretical CDF are in agreement with each other (red and blue lines are overlapping).

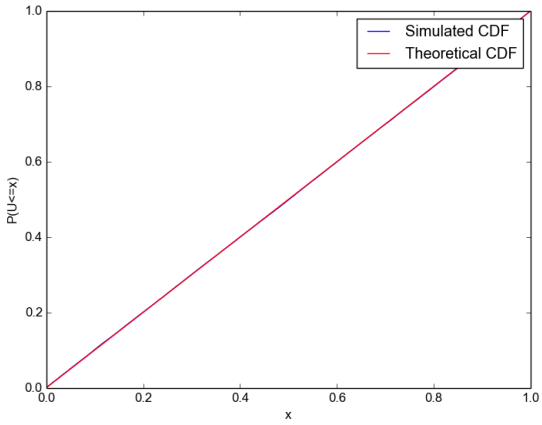

Figure 1: Simulated and Theoretical CDF for $Uniform(0,1)$ random variable

Exponential Distribution

To generate exponential random variables with parameter λ, we use the relation

$$X = -\frac{\log(1-U)}{\lambda},$$

where U is a $Uniform(0,1)$ random variable. To prove that X is an exponential random variable, let $F_X(a)$ be the distribution function of X. Now,

$$F_X(a) = P(X \leq a) = P(-\frac{\log(1-U)}{\lambda} \leq a).$$

Now, $\lambda > 0$ and log is an increasing function, it follows that $-\frac{\log(1-U)}{\lambda} \leq a$ if and only if $U \leq 1 - e^{-a\lambda}$. Hence, $F_X(a) = P(U \leq 1 - e^{-a\lambda}) = 1 - e^{-a\lambda}$, which is the CDF of an exponential random variable with rate λ.
Now, if U is a uniform random variable, then so is $1 - U$. We have used U in place of $1 - U$ in the simulations. To simulate an exponential random variable, we generate a large number of uniform random numbers and find the value of X as given in the expression for X above.

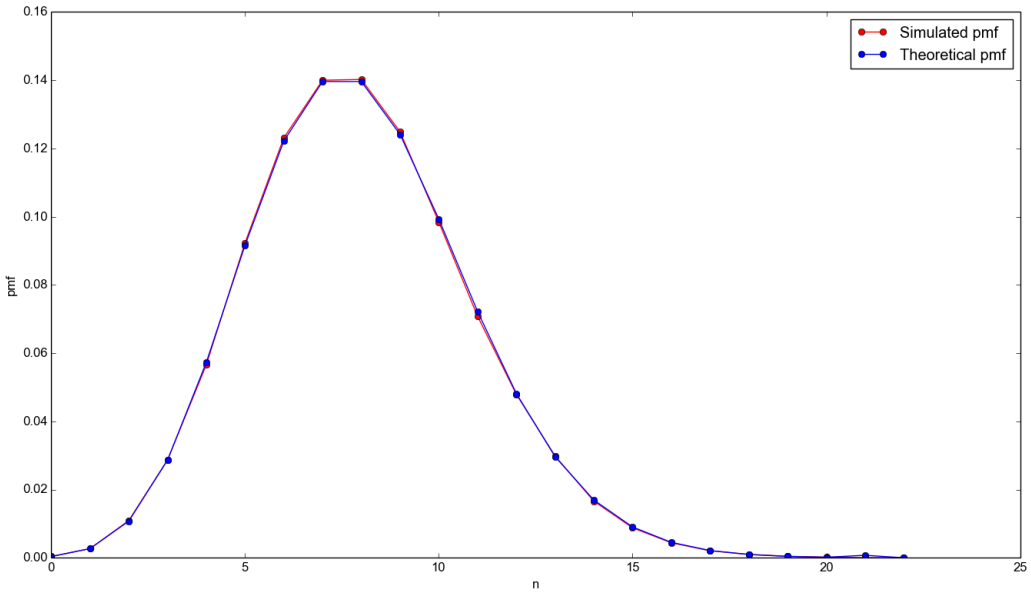

Figure 2: Simulated and Theoretical pmf for Poisson Random Variable

Poisson Distribution

We generate a Poisson random variable Y with parameter λ using the relation

$$Y = min_n \prod_{i=1}^{n} U_i \leq \exp^{-\lambda},$$

where U_i are i.i.d. samples from $Uniform(0,1)$ distribution. We generate a large number of values of Y by using the above expression and then find the probability mass function for the values generated. To show that Y is $Poisson(\lambda)$, we compare the pmf for simulated values with the theoretical pmf values, which is shown in Figure 2. In the figure, we see that the two plots overlap.

Erlang Distribution

An Erlang-k random variable $X(k,\mu)$ can be generated as follows:

$$X = -\frac{\log \prod_{i=1}^{k} U_i}{\mu},$$

where $U_1, U_1, ..., U_k$ are independent $Uniform(0,1)$ random variables.

Congestion Perspective

M/M/1 Queue

An M/M/1 queue has a single server, traffic arrival times determined by a Poisson process and job service times determined by an exponential distribution. We generate an arrival time array and a service time array for packets keeping the total simulation time fixed. The single server serves customers one at a time starting from the first customer in the queue. On completion of service, the customer leaves the queue, which determines the departure times. The waiting time (or delay) is the time the packet has to wait from its arrival till it finishes its service and departs.
The number of customers in the queue at time t (Q_t) is given by:

$$Q_t = Q_0 + A(0,t] - D(0,t], \forall t \geq 0,$$

where $A(0,t]$ and $D(0,t]$ denote the number of arrivals and departures in $(0,t]$ respectively. We have assumed Q_0 to be 0. Figure 3 shows the queue length vs time in an $M/M/1$ queue for $\lambda = 4$ and $\mu = 5$, where λ and μ are arrival and service rates respectively.

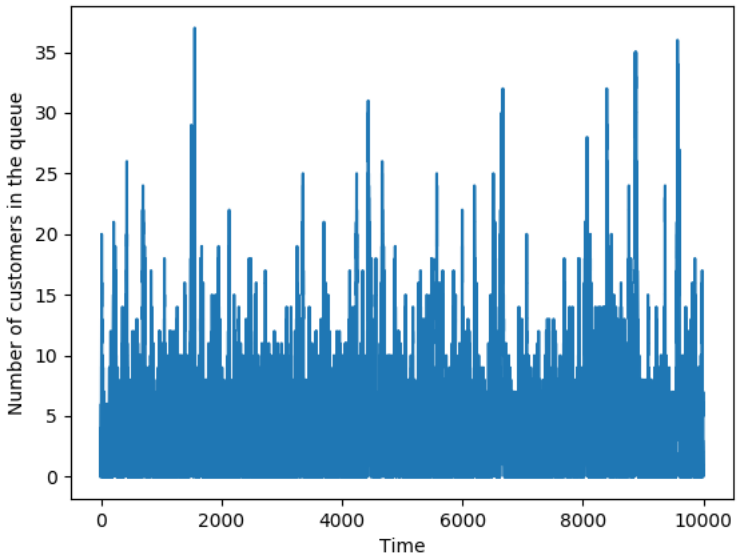

Figure 3: Queue length vs time in M/M/1 queue

π_n is the probability of having n customers in the queue. Figure 4 gives π_n vs n for $\lambda = 4$ and $\mu = 5$ (given by blue plot) and red plot gives the corresponding plot as given by the known formulae $\pi_n = \rho^n(1-\rho)$, where $\rho = \frac{\lambda}{\mu}$. From this formulae, the expected number of customers in the queue is given by $\frac{\rho}{1-\rho} = 4$ and the value obtained from simulation is 4.3385.

The following table gives the expected number of customers in the queue ($E([Q])$ and expected delay ($E[W]$) for three different pairs of λ and μ when $\rho = \frac{1}{2}$. From theory, we know that for an $M/M/1$ queue, $E(Q) = \frac{\rho}{1-\rho} = 1$ and $E[W] = \frac{1}{\mu-\lambda}$ and the same can be validated by the results given in the table. Thus, keeping ρ constant, the expected number of customers in the queue remains constant but the expected delay decreases with increase in arrival and service rates. We also observe that the results very closely conform to the Little's formula given by

$$E(Q) = \lambda E(W).$$

λ	μ	$E[Q]$	$E[W]$
1	2	0.9989	0.9890
3	6	1.0144	0.3354
5	10	1.0039	0.2017

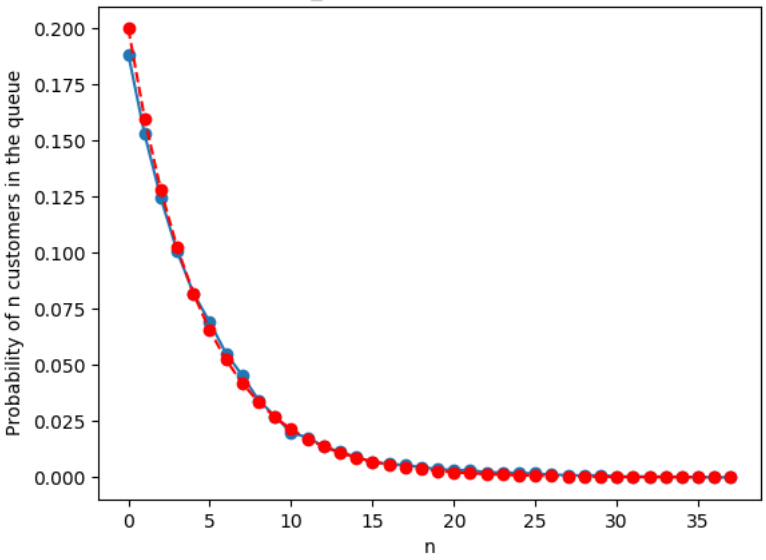

Figure 4: π_n vs n for M/M/1 queue ($\lambda = 4$, $\mu = 5$)

$\mathbf{M/E_k/1}$

For simulating $M/E_k/1$ queue, we use the same simulation approach as in $M/M/1$ except that the service times are now given by Erlang-k random variable.

Figure 5 gives π_n vs n for $\lambda = 4, \mu = 5$ and $k = 1, 2, 3$. The expected number in the system for different k is given by the following table:

k	$E[Q]$
1	3.8402
2	3.5322
3	2.8663

With increase in k, $E[Q]$ decreases and for $k = 1$, the queue is an $M/M/1$ queue and the simulation shows that the value of $E[Q]$ is close to the value obtained from theoretical results ($E[Q] = 4$). Also, the π_n vs n graph indicates that the probability of having low number of customers is more for higher k and the behaviour is reverse for high number of customers, which is indicated by behaviour of $E[Q]$ above as well.

Figure 6 gives the expected number in the system for different values of utilization (ρ), when $k = 1, 5, 10$ and for $M/D/1$ queue as well for which the service times are deterministic. According to Pollaczek-Khinchine formulae given by:

$$E[Q] = \rho + \frac{\rho^2(C^2 + 1)}{2(1 - \rho)},$$

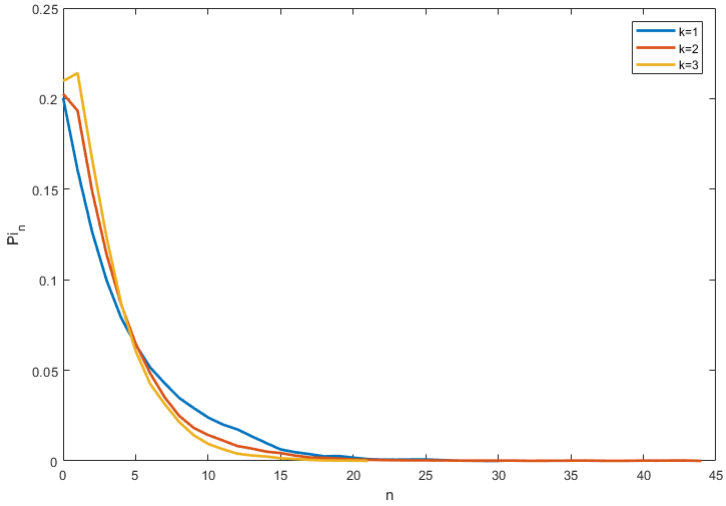

Figure 5: π_n vs n for $\lambda = 4, \mu = 5$ and $k = 1, 2, 3$ for $M/E_k/1$ queue.

where C^2 is the coefficient of variation of service time distribution, the expected number in the system increases with increase in ρ and with increase in k, it tends towards the expected number for $M/D/1$ queue (since mean and variance of an Erlang-k random variable are given by $\frac{k}{\lambda}$ and $\frac{k}{\lambda^2}$ respectively). The same can be observed in the figure.

GI/M/1

The same simulation approach as in $M/M/1$ is used except that the inter-arrival times are now drawn from Erlang-k distribution. The following table gives the average number of customers in the system seen by arriving customers and time average of the number of customers in the system for different values of k, $\lambda = 0.2$ and $\mu = 0.25$.

k	Average number of customers seen at arrival times	Time average of the number of customers
1	3.6939	3.7204
2	2.8942	3.0155
3	2.3344	2.7281
5	2.3228	2.6785
20	1.9462	2.3871

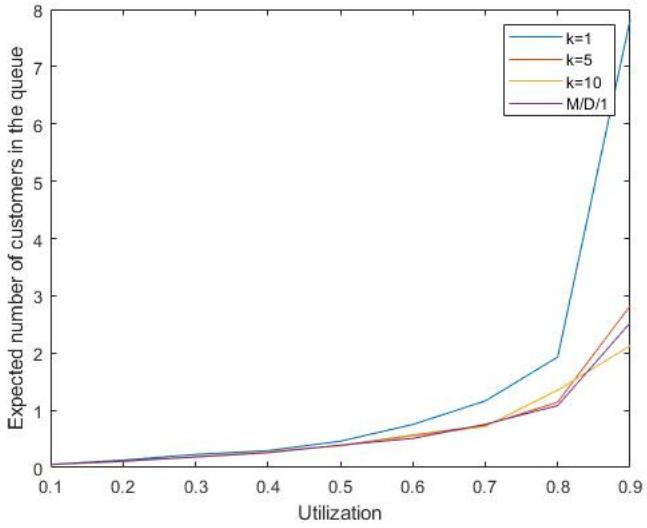

Figure 6: Expected number in the system vs utilization ρ, when $k = 1, 5, 10$ and for $M/D/1$ queue with $\lambda = 4$ and $\mu = 5$

For $k = 1$, the queue is an $M/M/1$ queue and using PASTA(Poisson Arrivals See Time Averages) property, both the averages should be same and this is validated by our simulation. Also, with increase in k, both the averages decrease and the difference between them increases and PASTA property isn't valid anymore.

Workload Perspective

$M/M/1$ queue is now analyzed from workload point of view. Now, the service time array is called the incoming workload. When the service discipline is FIFO, the jobs are executed as explained above. But for LCFS (Last Come First Served) under a preemptive priority serving schedule, as the name suggests, the server serves the last arrived packet first and if a packet is already in service when the last packet arrives, it's service is stopped and continued after the server is finished serving the new packet. Thus, for determining the departure times for LCFS under a preemptive priority serving schedule, we start with the end of the array. Now, the remaining workload (the unfinished) work (W_t) at time t is given by:

$$W_t = W_0 + X(0, T] - c \int_0^t 1_{[W_s > 0]} ds \ \forall t \geq 0.$$

Here, $X(0,T]$ denotes the amount of work arriving in time t and c is the server speed which we have taken as 1. W_0 is assumed to be 0. The remaining workload doesn't depend on the service discipline but only on the amount of work coming in the system and the rate at which the server is serving. For $\lambda = 4$ and $\mu = 5$, average workload in the system for FIFO as well as LCFS under a preemptive priority serving schedule is 0.8183. Expected number of customers in the queue for FIFO is 3.9852. Expected number of customers in the queue for LCFS under a preemptive priority serving schedule is 1.9245. Expected number of customers obtained from Little formula using average workload obtained from simulation is 3.2731.

References

[1] https://www.win.tue.nl/ marko/2WB05/lecture8.pdf

[2] S. Ross. A First Course in Probability. 8th Edition. Pearson Prentice Hall, 2010.

[3] R.R. Mazumdar. Performance Modeling, Loss Networks, and Statistical Multiplexing. Morgan and Claypool, 2010.