COSC 223: Probability and Computing
Project 2: Queueing

Juhwan Jeong and Ian McClaugherty

April 13, 2019

1 Introduction

In this project, we are studying how variability in job size and interarrival time affects mean response time on a single-server queueing system with a First Come First Serve Policy. Essentially, a queue is an ordered line. Queues are common in everyday life. For example, one often encountered queue is the checkout line in a grocery store. The checkout line happens to be an example of the First Come First Serve Policy that we are using in this project: the first person in the line is the first to get checked out. In other words, customers check out in the order in which they show up in line.

Queueing is also common in computer science and has many important applications. Tasks like CPU scheduling, reading from a disk, and network routing all rely on queues. The queue we studied is a single-server queue. Having only one server means that only one job can be processed at a time. If other jobs arrive while a job is on the server, they have to wait in the queue. First Come First Serve in our system means that the first job to arrive is the first to get serviced – the rest line up in the queue in the order of their arrival times.

Two queue parameters that are crucial to consider are job size and interarrival time. Job size is the average amount of time that a job spends on a server when it is being processed. Interarrival time is the average amount of time that passes between consecutive job arrivals to the queueing system. In the experiments described below, we simulate a single-server queueing system by drawing on our knowledge of continuous random variables and probability distributions. Job sizes and interarrival times are generated using the exponential distribution and a two-phase hyperexponential distribution, which are commonly used distributions in queueing theory. By drawing from these distributions, we are able to run simulations that are reflective of the variability in job size and interarrival time that one may encounter when working with real queueing applications. In conducting our research, our goal is to understand how this variability affects the performance of a queueing system. The practical benefit of this knowledge is that we can apply it to real-world queueing applications to make predictions about performance when we know some basic assumptions about the kinds of jobs we intend to run.

2 Experimental Setup

In the first experiment, we simulated the queueing system for 1,000,000 jobs. Then, we discarded the first 10,000 response times to avoid any transient effects that come from starting with an empty server and queue. The interarrival times were drawn from an exponential distribution with parameter λ, where λ ranged from 0.01 to 1 with 0.01 increments. The job sizes were drawn from a hyperexponential distribution, where the variance was either 1, 10, 20, or 50. Given each variance, the following series of calculations was performed to acquire values of μ_1, μ_2, and p.

First, we compute $E[X]$, where X is a random variable that represents job size. In class, we used integration by parts to show that the expected value of an exponential with rate λ is $\frac{1}{\lambda}$. To compute the expected value of the hyperexponential, we first multiply the expected value of each exponential by its corresponding probability, then sum those products:

$$\begin{aligned} E[X] &= p\int_0^\infty x\mu_1 e^{-\mu_1 x}\ dx + (1-p)\int_0^\infty x\mu_2 e^{-\mu_2 x}\ dx \\ &= p(\frac{1}{\mu_1}) + (1-p)(\frac{1}{\mu_2}) \\ &= \frac{p}{\mu_1} + \frac{1-p}{\mu_2} \end{aligned}$$

Because we are assuming that $E[X] = 1$ and that $\frac{p}{\mu_1} = \frac{1-p}{\mu_2}$, we can compute the values for μ_1 and μ_2:

$$\begin{aligned} E[X] = \frac{p}{\mu_1} + \frac{1-p}{\mu_2} &= 1 \\ \Rightarrow \frac{p}{\mu_1} = \frac{1-p}{\mu_2} &= \frac{1}{2} \\ \Rightarrow \mu_1 = 2p, \mu_2 &= 2(1-p) \end{aligned}$$

To compute $E[X^2]$, we evaluate the following integral:

$$E[X^2] = p\int_0^\infty x^2\mu_1 e^{-\mu_1 x}\ dx + (1-p)\int_0^\infty x^2\mu_2 e^{-\mu_2 x}\ dx$$

Solving this integral requires using integration by parts, and the technique is very similar to the one we used in class to solve the integral for $E[X]$. Consider the following choices for u and dv for the first term in the sum (which uses rate μ_1):

$$\begin{aligned} u &= x^2 \\ du &= 2x\ dx \\ dv &= \mu_1 e^{-\mu_1 x} \\ v &= -e^{-\mu_1 x} \end{aligned}$$

Now, we can write the first term as follows:

$$= p[-x^2 e^{-\mu_1 x}\Big|_{x=0}^{\infty} - \int_0^\infty -e^{-\mu_1 x} 2x\ dx]$$

We can turn this integral into the integral for $E[X]$ by rewriting the 2 as $\frac{2\mu_1}{\mu_1}$ and then pulling $\frac{2}{\mu_1}$ out in front:

$$= p[-x^2 e^{-\mu_1 x}\Big|_{x=0}^{\infty} + \frac{2}{\mu_1}\int_0^\infty x\mu_1 e^{-\mu_1 x}\ dx]$$

As $x \to \infty$, the exponential grows much faster than $-x^2$, so $-x^2 e^{-\mu_1 x}\Big|_{x=0}^{\infty}$ evaluates to zero. We know the integral evaluates to $\frac{1}{\mu_1}$. Now, we have

$$= p[\frac{2}{\mu_1}(\frac{1}{\mu_1})] = \frac{2p}{\mu_1^2}$$

We can perform the same steps for the second term in the expression for $E[X^2]$. The final result is below:

$$E[X^2] = \frac{2p}{\mu_1^2} + \frac{2(1-p)}{\mu_2^2}$$

The values of our parameters μ_1, μ_2, and p depend upon the variance. Using the variance equation, we found an expression for p in terms of the variance.

$$\begin{aligned} Var[X] &= E[X^2] - (E[X])^2 \\ &= \frac{2p}{\mu_1^2} + \frac{2(1-p)}{\mu_2^2} - (1)^2 \end{aligned}$$

Substituting $\mu_1 = 2p$ and $\mu_2 = 2(1-p)$ yields:

$$\begin{aligned} &= \frac{1}{2p} + \frac{1}{2(1-p)} - 1 \\ &= \frac{1}{2p(1-p)} - 1 \end{aligned}$$

Let $Var[X] = v$. We will solve for p in terms of v.

$$\begin{aligned} & v = \frac{1}{2p(1-p)} - 1 \\ & v[2p(1-p)] = 1 - 2p(1-p) \\ & 2p(1-p) + 2vp(1-p) - 1 = 0 \\ & 2p - 2p^2 + 2vp - 2vp^2 - 1 = 0 \\ & p^2(-2-2v) + p(2v+2) - 1 = 0 \\ & p^2(-2-2v) - p(-2-2v) - 1 = 0 \\ & p^2 - p + \frac{1}{2+2v} = 0 \end{aligned}$$

Using the quadratic formula, we get:

$$\begin{aligned} & p = \frac{-(-1) \pm \sqrt{(-1)^2 + \frac{4}{2+2v}}}{2} \\ & p = \frac{1 \pm \sqrt{1 + \frac{2}{1+v}}}{2} \end{aligned}$$

We simulated the queueing system for each possible pair of λ and variance values and recorded the mean response time. The results were graphed on a mean response time vs. λ graph.

In the second experiment, we simulated the queueing system for 1,000,000 jobs again; this time, the interarrival times were drawn from a hyperexponential distribution with balanced means and the job sizes were drawn from an exponential distribution with $\lambda = 1$. As before, we discarded the first 10,000 response times. For the hyperexponential distribution, the expected value ranged from 1.01 to 4.00, with 0.01 increments (Note the expected values of interarrival times are at least 1, which is the expected value of job sizes, as we study the case where processing rate is strictly greater than the arrival rate). For each expected value μ, the variance was set to be either μ^2, $10\mu^2$, $20\mu^2$, $50\mu^2$. These variance values were used to ensure that the ratio between μ and σ remained at 1, $\sqrt{10}$, $\sqrt{20}$, and $\sqrt{50}$ to accommodate for the fact that the mean is changing. Given each mean and variance, the following series of calculations were performed to acquire values of μ_1, μ_2, and p.

The calculations are similar to those above for experiment 1, the key difference being that $E[X]$, the expected value of interarrival times, is not fixed at 1. We let

$$E[X] = \frac{p}{\mu_1} + \frac{1-p}{\mu_2} = l$$

where l represents the expected value. Because our hyperexponential still has balanced means,

$$\begin{aligned} \frac{p}{\mu_1} &= \frac{1-p}{\mu_2} = \frac{l}{2} \\ \Rightarrow \mu_1 &= \frac{2p}{l}, \mu_2 = \frac{2(1-p)}{l} \end{aligned}$$

These values help us solve for $E[X^2]$ in terms of l and p:

$$\begin{aligned} E[X^2] &= \frac{2p}{\mu_1^2} + \frac{2(1-p)}{\mu_2^2} \\ &= \frac{2pl^2}{4p^2} + \frac{2(1-p)l^2}{4(1-p)^2} \\ &= l^2(\frac{1}{2p} + \frac{1}{2(1-p)}) \\ &= \frac{l^2}{2p(1-p)} \end{aligned}$$

Now, we can solve for p in terms of the variance v and expected value l:

$$\begin{aligned}
&Var[X] = v = E[X^2] - (E[X])^2 \\
&v = \frac{l^2}{2p(1-p)} - l^2 \\
&2pv(1-p) = l^2 - 2pl^2(1-p) \\
&2pv - 2p^2v + 2pl^2 - 2p^2l^2 - l^2 = 0 \\
&p^2(-2v - 2l^2) - p(-2v - 2l^2) - l^2 = 0 \\
&p^2 - p + \frac{l^2}{2v + 2l^2} = 0 \\
&p = \frac{1 \pm \sqrt{1 + \frac{4l^2}{2v+2l^2}}}{2}
\end{aligned}$$

Again, we simulated the queueing system for each pair of expected value and variance values and recorded the mean response time. The results were graphed on a mean response time vs. expected value graph.

3 Results and Discussion

In experiment 1, we compute λ and μ as the equivalent of arrival rate and processing rate, respectively. Using these quantities, we were able to make predictions on what the result might look like. The following two equations were used to draw a relationship between $E[T]$ and λ. First,

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

since $\mu = 1$. Then, from Little's Law, we have

$$\begin{aligned}
E[N] &= \frac{\lambda}{1-\lambda} = \lambda E[T] \\
\Rightarrow E[T] &= \frac{1}{1-\lambda}
\end{aligned}$$

As depicted in the relationship above, we expect the mean response time to increase as λ increases from 0 to 1.

In experiment 2, we compute mean response time for each pair of expected value and variance of interarrival times. As expected value increases, the rate at which jobs arrive decreases. In turn, jobs will not get queued up as much, and mean response time will decrease. Therefore, we predict that as expected value of interarrival times increases, the mean response time will decrease.

As for variance, high variance will lead to higher chance of generating a value significantly lower or greater than the expected value.

In experiment 1, we generate service times from the same distribution with varying variances. When we have a distribution with a high variance, generating significantly shorter jobs will not have a great effect on the mean response time - the jobs will be processed quickly and it will not lead to longer queues. Generating longer jobs than usual, however, will significantly impact the mean response time since longer jobs can lead to longer queues, increasing the response time not only of the job itself, but also of the jobs that are to follow. Therefore, generating service times with higher variance will lead to greater mean response time.

In experiment 2, we generate interarrival times from the same distribution with varying variances. Opposite to experiment 1, when we have a distribution with a high variance, generating significantly longer interarrival time will not have a great effect on the mean response time - since jobs are less likely to get queued up with long interarrival times. Generating shorter jobs than usual, however, will significantly impact the mean response time since shorter interarrival times will lead to jobs getting queued up - elongating response times of multiple jobs. Therefore, generating interarrival times with higher variance will also lead to greater mean response time. With these predictions in mind, we performed our experiments.

The results of experiment 1 are as follows:

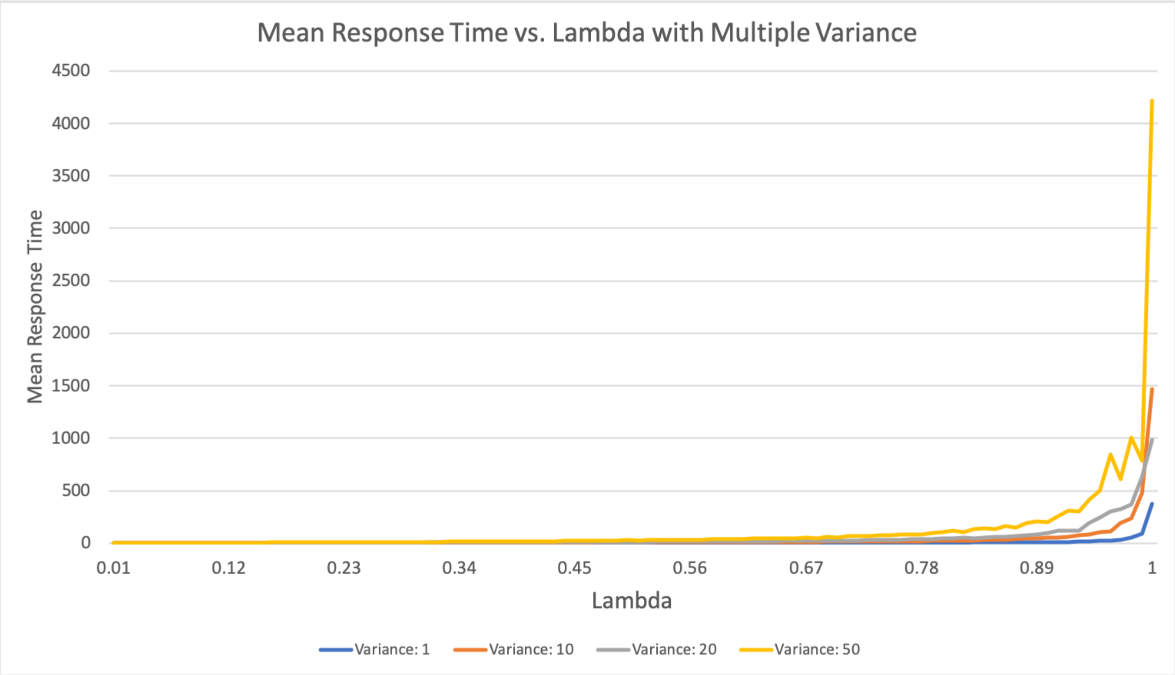

Figure 1: Graph of mean response time vs. λ from experiment 1 from $\lambda = 0$ to 1, where λ is the rate of the interarrival times distribution.

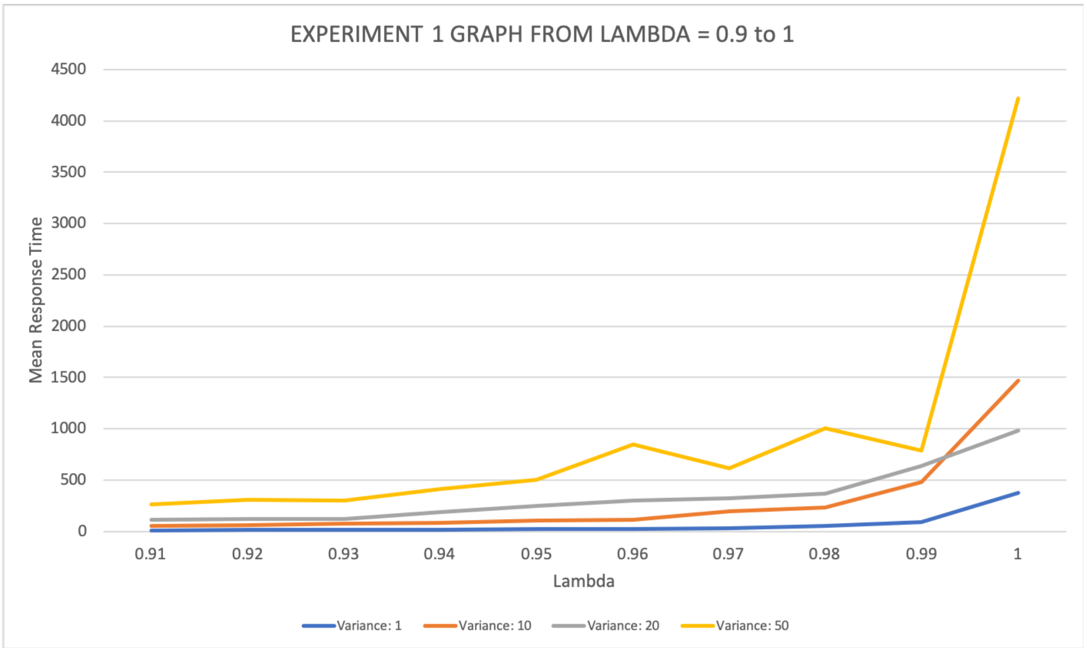

Figure 2: Graph of mean response time vs. λ from experiment 1 from $\lambda = 0.9$ to 1, where λ is the rate of the interarrival times distribution. (Zoomed in version of Figure 1.)

In experiment 1, the graphs showed behaviors we expected. The curves had similar shape to $E[T] = \frac{1}{1-\lambda}$ and the mean response time increased as λ approached 1 (the processing rate). This is because as interarrival times approach job sizes, jobs start to get queued up, and the mean response time increases. In addition, graphs with higher variances had higher mean response times as expected. This is due to the fact that with higher variance, the system is more likely to encounter a long service time, which leads to jobs getting queued up, lengthening response times of multiple jobs.

The results of experiment 2 are as follows:

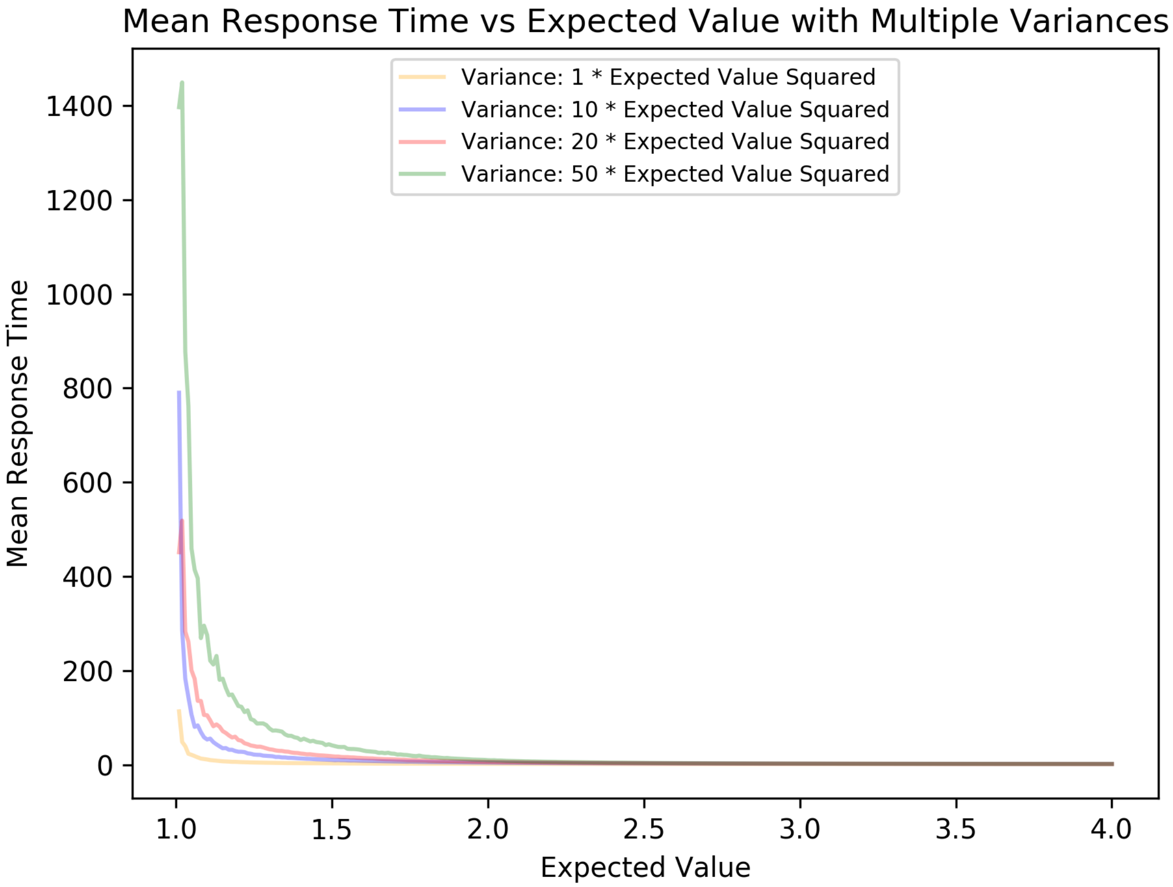

Figure 3: Graph of mean response time vs. expected value for expected values between 1.0 and 4.0 from experiment 2.

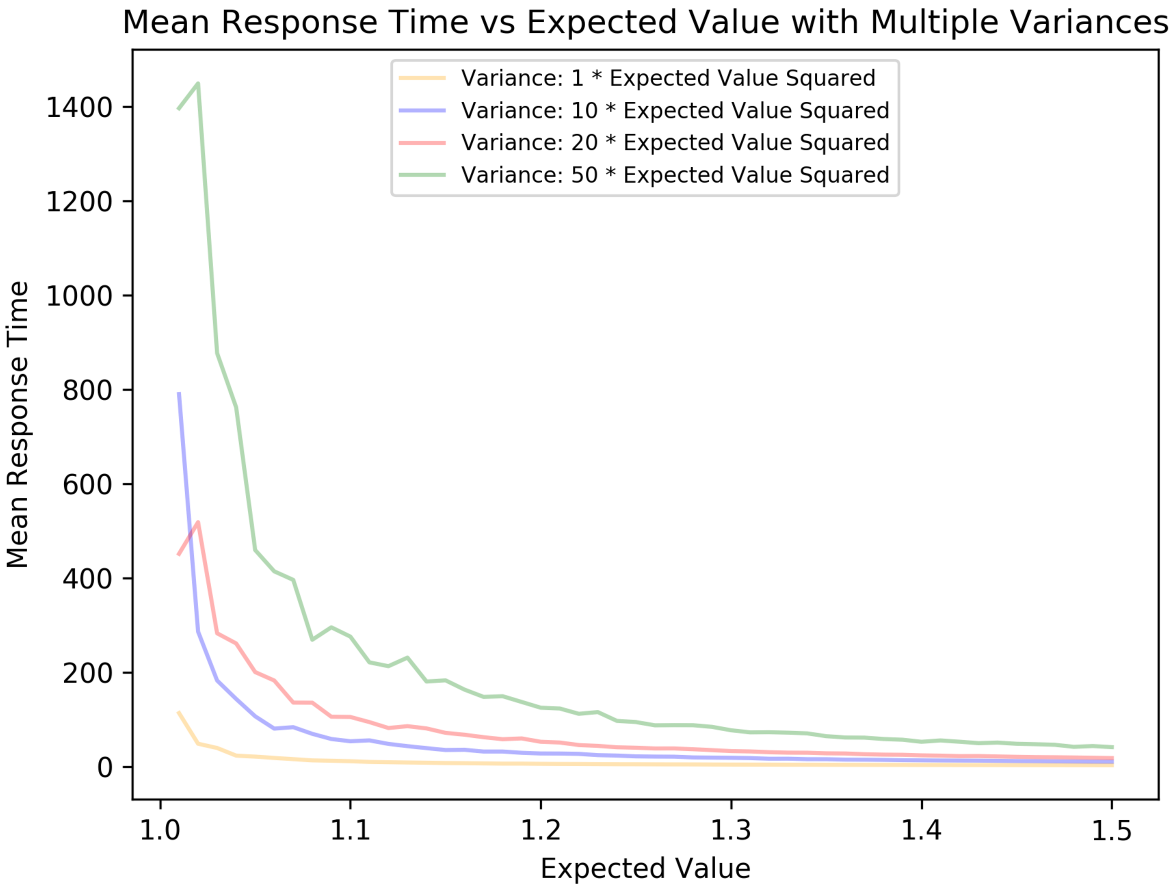

Figure 4: Graph of mean response time vs. expected value for expected values between 1.0 and 1.5 from experiment 2. (Zoomed in version of Figure 3.)

In experiment 2, the graph exhibits behavior as predicted; as expected value of interarrival times increases, the mean response time decreases. When the interarrival times are comparable in length to service times, jobs are more likely to get queued up, which increases the mean response time. As interarrival times get longer, however, jobs get queued up less, and mean response time decreases drastically. In addition, graphs with higher variances had higher mean response times as expected. This is due to the fact that with higher variance, system is more likely to encounter a short interarrival time, which leads to jobs getting queued up, lengthening response times of multiple jobs.

One significant point to note is that the graphs take on exponential shape; in experiment 1, the mean response times don't spike until lambda is near 1; in experiment 2, the mean response times decreases significantly when expected value is near 1 but levels off quickly.

4 Conclusion

Our experiment results show that as arrival rate approaches processing rate, the mean response time increases exponentially. Higher variance in distribution of service times and interarrival times both lead to greater mean response time as well.

Several assumptions were made in our experiments. In both experiments, we assumed that a service time of one job is independent of the service times of its previous job and its subsequent job. Similarly, we conveniently made the assumption that there is no concept of locality in the distribution of interarrival times; however, in both cases, these assumptions may be unrealistic. Had we not made these assumptions, designing the distributions of service times and interarrival times would have been significantly more complex.

Consequently, one follow-up question from this project is - can we design distributions that incorporate locality to model real life queueing systems more closely (possibly by incorporating requestGenerator code from Project 1)? For example, we can imagine designing distributions such that consecutive interarrival times (or service times) are not independent of another and test our queueing system on them. Another interesting question is - how does this queueing system's performance with hyperexponential distributions compare to that with Pareto distributions? In class, we discussed that we use the hyperexponential distribution for its simplicity in modeling over the Pareto distribution, the distribution that resembles service times in real life more closely. The question then becomes, with a few assumptions to simplify modeling the Pareto distribution, can we run our experiments with the Pareto distributions and how do the mean response times compare?