STOCHASTIC SIMULATION

# ASSIGNMENT 2 - MULTIPLE QUEUES AND MULTIPLE SERVERS

November 30, 2018

Student ID: 12297127 & 11037466
Jordan Earle & Nathalie van Sterkenburg

# Contents

# 1 Abstract

In order to determine the dynamics of a queue system, a number of experiments were performed and the results analyzed from a theoretical and practical standpoint. First the theoretical system was defined and the theoretical system dynamics of a FIFO M/M/n system were compared to the theoretical system dynamics of a FIFO M/M/1 system. It was found that the M/M/n system should be equal to or faster than the M/M/1 system. Once this was theoretically proven, the system was simulated and the theoretical results were compared to the experimental results. The system load was varied with a fixed arrival rate, and the systems were analyzed. The experimental results were confirmed by the theoretical results and it was found that for small values of the system load, that the number of iterations in the experiment needed to achieve the same statistical significance was higher than those with higher system loads.

Next the M/M/1 system was modified to observe the effects of a shortest job first scheduling method. The dynamics of the system were observed and it was found that for the shortest jobs first, the number of people in the queue and the average wait time was shorter, but it is possible that those in the queue had been there for a long period of time due to the jobs never being the shortest job in the system.

Finally, different service rate distributions were altered. At first the system was changed to a deterministic service rate (M/D/n) and the effects were observed. It was found that the systems perform slightly worse than those with a Markov service rate. This is likely because without the stochastic distribution, the M/D/n model did not have a chance to receive the shorter service times that the M/M/n system experienced. The final distribution modeled was a longtail service distribution, which was an example of a hyper-exponential distribution. This longtail distribution caused a much lower number in the queue and lower waiting time than for the original FIFO experiment. There was an increase in the number of runs needed to achieve the same statistical significance, which could be in part due to the increased variance in the system since the distribution was being selected from 2 peeks.

# 2 Introduction

Queuing theory is often used in many different fields including communication networks, computer systems, machine plants and in any other processes where there is a set of tasks waiting for a service to be performed. The tasks to be undertook can follow different distributions, some may be deterministic, while others may be stochastic. The queue length, or even the allowance of a queue, needs to be considered, as well as the number of 'service centers', which can allow for additional resources to be spent in addition to different models to be implemented. As such the optimization of resources in the system can be of key importance when determining how to schedule tasks.

There are many different service disciplines for a queuing system and managing the system to optimize the objective. Some disciplines include First in First Out (FIFO), shortest jobs first Priority, and Last in First Out (LIFO) to name a few. The choices made for determining how to model the arrivals and services can also effect how the system, such as queue length and waiting time, responds to a given service discipline.

Understanding queues and predicting queue behaviour is important in many different fields such as the ones mentioned above, for example to determine how many servers have to be available. Therefore in this experiment, the theory of the effectiveness of queuing systems, such as queue length and waiting time, was explored, showing how for M/M/n queues, the average wait time is shorter than for a single M/M/1 queue with the same loading characteristics for a FIFO system. Once these characteristics had been confirmed theoretically, the systems were simulated in python to confirm the theoretical results. Once the theoretical results were confirmed, the number of servers was modified to determine how the system reacted to this with the same loading characteristics and the system loads were also varied to determine how the system load effected the number of runs required to keep the same statistical significance in the output.

Next the effects of alternative service disciplines was examined by modifying the M/M/1 queue to have shortest job first scheduling. The effects of this service discipline were compared to the previous M/M/1 queue with FIFO scheduling. Finally the effects of a different service rate distribution were examined. The service rates were modified to be deterministic rather than Markovian in one case, and to have a long tailed distribution where a distribution of 75% of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25% an exponential distribution with an average service time of 5.0. The M/D/n and M/D/1 systems were then analyzed and compared with the original distributions to determine the effects on the waiting time and the queue length.

# 3 Theory

Before discussing queuing theory, it is good to quickly discuss Markov chains. A Markov chain is a memory-less, probabilistic system [5]. This means probabilities play a role in determining the next state, but more importantly, this means that the next state of the system is only dependent on the current state, no matter what states it may have had in the past. This is important because queues are an example of a Markovian system. A visualization of the Markov chain for an M/M/1 queue (more on queue notation later) is given in figure 1. With the mathematics of Markov chains, steady state probabilities for queue systems can be determined, which will be done later in the theory section.

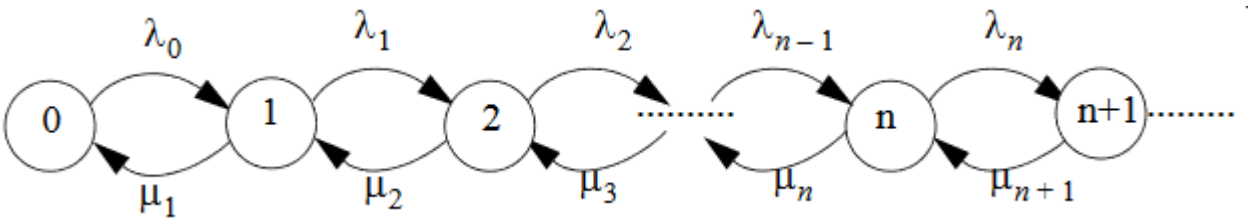



**Figure 1:** A visualization of the Markov chain of an M/M/1 queue.

Queuing theory was originally published in 1909 by Agner Krarup Erlang [1], where he modeled the number of telephone calls arriving at a telephone exchange using a Poisson process. Queuing theory deals with customers who have a service that has to be performed by a server. When no server is available upon arrival, a customer can either leave or wait in line. In queuing systems Kendal notation is often used to denote what properties a system has. The notation looks as follows: (A/B/m/N - S). In this notation the first letter (A) denotes how the arrival times are handled, the second letter (B) denotes how service times are handled. Some common distributions are Markov (M), Deterministic (D), Erlang-k ($E_k$) and Hyper-k ($H_k$). The third letter (m) denotes the number of servers in the system. A fourth letter (N) may be used, which denotes the length of the waiting line if finite (if infinite the forth letter is omitted), and can also be followed by a fifth letter (S) which can also be used, represents the service discipline. If the letter is omitted, the system is assumed to be using First in First Out (FIFO).

In this experiment it is assumed that all customers join the queue and none of them leave before having their service performed. The two most important variables in the system are the arrival rate $\lambda$ of the customers and the service rate $\mu$, also called the capacity of a server, of each customer [2]. This means the (mean) time between two arrivals is $\frac{1}{\lambda}$ and the (mean) service time is $\frac{1}{\mu}$. Here mean is between brackets, because both the arrivals and services can

theoretically be either stochastic or deterministic, though arrivals are generally stochastic. In this experiment the queues that are looked at are M/M/n queues and M/D/n queues, where Kendall's notation is used. The M's used stands for Markov chain and means the arrival time or service time is a stochastic Poisson process. This means $\frac{1}{\lambda}$ and $\frac{1}{\mu}$ are means, and that the actual arrival and service time follow an exponential distribution. The D stands for deterministic, meaning the service and arrival times are fixed at $\frac{1}{\lambda}$ and $\frac{1}{\mu}$. In the two types of queues looked at in the experiment, the arrival time is always stochastic and the service time is looked at for both a stochastic and a deterministic case. The number of servers is varied between 1, 2 and 4. These three variables are combined in the system load $\rho$, which is defined as:

$$\rho = \frac{\lambda}{n\mu}. \tag{1}$$

If $\rho < 1$, than the arrival rate is smaller than the service rate and as the system evolves the average waiting time and average number of customers in the queue diverge to certain values $W$ and $Q$ respectively. But when $\rho > 1$, the arrival rate is larger than the service rate, and as the system evolves the waiting times and number of customers in line keep increasing as more customers arrive than that are being serviced and the waiting time and number of customers diverge indefinitely. For this experiment it was assumed that $\rho$ is always smaller than 1.
In this experiment the system load, the mean service time and the number of servers are set. This means that if we change one of these variables, the arrival rate changes along with it. Double the number of servers means there will be double the number of arrivals.

In this experiment there are two variables that are measured, the average waiting time in the queue and the average number of customers in the queue. It will now be described how these values can be derived theoretically [5, 4]. The first step is to determine the steady state probability that $k$ customers are in the queue, this probability is denoted $p_k$. As described by Andreas Willig [5], this probability is

$$p_k = \begin{cases} p_0 \frac{(n\rho)^k}{k!} \text{ for } k \leq n \\ p_0 \frac{\rho^k n^n}{n!} \text{ for } k \geq n \end{cases} \tag{2}$$

where $p_0$ is defined as

$$p_0 = \left[ \sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \left( \frac{(np)^n}{n!} \right) \left( \frac{1}{1-\rho} \right) \right]^{-1}. \tag{3}$$

This result can then be used to compute the mean number of customers in the queue [4] using the formula

$$\overline{Q} = \sum_{k=n+1}^{\infty} (k-n)\,\rho_k, \tag{4}$$

which results in

$$\overline{Q} = P_Q \frac{\rho}{1-\rho} \tag{5}$$

where

$$P_Q = \frac{(n\rho)^n}{n!} \frac{p0}{1-\rho} \tag{6}$$

is the Erlang-C formula which denotes the probability that all servers are busy when a customer arrives. From $\overline{Q}$, the average waiting time in the queue can be determined [4] using Little's law [5],

$$\overline{W} = \frac{\overline{Q}}{\lambda} = P_Q \frac{\rho}{\lambda(1-\rho)}. \tag{7}$$

For $n = 1$ the equations for average number of customers and average waiting time reduce to

$$\overline{Q} = \frac{\rho}{\mu - \lambda} \tag{8}$$

$$\overline{W} = \frac{\rho^2}{\mu - \lambda} \tag{9}$$

The selection of what customer in a queue will be serviced next, the service discipline, is

another important factor that influences how a queue evolves. While there are many models to choose from, only two service disciplines were explored in this experiment. The first one is first in first out (FIFO) scheduling. The customer who has been in line longest will be next to receive service. The other discipline is shortest job first scheduling. The moment a server becomes available, the queue is searched for the customer with the lowest service time and this customer is serviced next.

Queuing theory shows that for FIFO scheduling, the average waiting times are shorter for M/M/n queues with a system load $\rho$ and a processor capacity $\mu$ than for a single M/M/1 queue with the same characteristics, and thus a higher arrival rate. This can be shown by computing the average waiting times for $n = 1$ and $n = 2$ using equations 8 and 5 and calculating that $\overline{W}_{n=1} \geq \overline{W}_{n=2}$. This can be done in the same way to compare $\overline{W}_{n=1}$ with the mean waiting time for any other $n > 1$.
This can also be thought of intuitively. A single server has half the arrival rate of a system with two servers. This means the arrivals for the two server system will be better distributed, which leads to less idle time. And all idle time could have been used to instantly service customers, so all idle time increases waiting times for the rest of the system. Another thing is that when there's a customer with a very long service time, a single server system is blocked for the whole duration of this service, which can mean a lot of customers have to wait a long time. This will also affect a two server system, but because there is still a server handling services quickly, the consequences are less severe.

In order to attain a high and known statistical significance in the simulation, a confidence interval of 99% was selected to ensure that the means were of good quality and a range of $l < 0.1 \cdot mean$ was desired. As such, the technique outlined in Ross, Ch.8.2[3] was used. The length of the range (l) had to be 10:

$$2\zeta_{\alpha\,2} S/\sqrt{k} < l \tag{10}$$

where S is the sample standard deviation, $\zeta$ is the umber of standard distributions away from the mean of a normal distribution, k is the number of iterations, and l is the length of the tolerance range. Once this was achieved, the measured mean had a confidence interval length of $\overline{X} \pm \zeta_{\alpha\,2} S/\sqrt{k}$ where $\overline{X}$ is the measured mean in question.

# 4 Experimental Methods

In order to determine the dynamics of a queuing system with different system properties, the theoretical results were reproduced in the theory section and a queuing system was implemented to verify the theoretical results. The simulation was then utilized to investigate the waiting times and the queue length as a function of system load $\rho$ and number of servers $n$. The theoretical results discussed above showed that for FIFO scheduling the average wait time was shorter for M/M/n queues with the same loading characteristics as a single M/M/1 system. One of the goals was to reproduce this result with the simulation. Another goal was to compare the two service disciplines described in the theory and the final goal was to compare different service rate distributions. To achieve those goals, four different experiments were performed and compared. Each experiment is described below.

## 4.1 Simulation

The simulation was written in Python notebook using the Simpy package and was set to run for a certain amount of time-steps. During this time, customers were generated in an infinite loop. In this loop, it was calculated when the next customer would arrive, the loop was put in a time out for this amount of time before generating a new customer and calculating the next time between arrivals.
A generated customer would yield a request to the servers to have their service performed and would wait until this request was excepted. The request was excepted right away if a server was available, but if all servers were blocked by other customers, the request would be placed in a waiting list. In the FIFO experiments, when a server came free, the request with the longest waiting time was excepted, but in the shortest first priority experiment, when a server came free, the request with the shortest service time was excepted. In the FIFO code, the service time was calculated the moment a customer entered service, but in the shortest job first code, this had to be changed around so the service time was determined when the customer was generated. Once the request for service was excepted, the customer would start service at a server and block this server for the duration of the service.
Each server started out as idle and whenever a request came this request was excepted and the customer was appointed to the first server available. This server would then be blocked and would be inaccessible to other customers. This was achieved by putting the server in a time out for the duration of the service. After this the server would be available again and the request from the first customer in the queue would be excepted or the server would stay idle

if there was no customer in the queue.

## 4.2 Experiments

Four experiments were performed where each experiment had either an altered service discipline of service rate distribution. $\mu$ was kept constant at 0.75 for the first three experiments. How $\mu$ was handled in the fourth experiment will be discussed later. All experiments were done for $\rho$ ranging from 0.1 up to 0.9 with steps of 0.1. The first experiment and the last two experiments were done for $n = 1$, $n = 2$ and $n = 4$ and for the second experiment $n$ was kept at a constant of 1. For each set of $\rho$ and $n$ in each experiment, the queue simulation was run for a duration of 10000 time-steps, during which there was a startup period of length 100 time-steps, followed by a measurement period of length 400, followed by a rest period of 100. This was done until the end of the simulation.
A startup period was necessary for the system to evolve beyond its starting conditions and the rest periods were necessary to avoid correlations between observations. The two values that were measured were the waiting time and the number of customers in the queue. A measurement of the length of the queue was done at every event. These events were either the arrival of a customer or the completion of a service. The start of a service would always coincide with either one of those events, and therefore the start of service was not an event at which was measured. The waiting time was determined for every customer when their service started. In order to decrease the correlation in the data, batch sampling was utilized, using a pause period between each measurement to decrease the correlation. At the end of the simulation iteration,the batch was analyzed to produce the mean's and the variance for the iteration.
To increase the statistical significance, multiple simulations were run for each $\rho$ and $n$. After every simulation, the mean and variance of the simulations were calculated and simulations were added until the statistical significance range $l$ was lower than $0.1 \cdot mean$. At this point the relative error was 0.1 which was considered to be accurate. However, as the experiments were being run, it turned out that for $\rho = 0.1$ it was nearly impossible to reach $l < 0.1 \cdot mean$ for both waiting time and number of customers in the queue, because the $mean$s of those values were too small. The computational power needed to reach $l < 0.1 \cdot mean$ was too much to handle for this experiment. Because the $mean$s at $\rho = 0.1$ were so small, the conclusion was drawn that the static value $l = 0.01$ was accurate enough for $\rho = 0.01$, even if this meant that the relative error was larger than preferable. The preferred accuracy was too small too be

feasible. So for every $\rho = 0.1$, $l$ was set to 0.01. For bigger $\rho$, it was actually computationally impossible to reach $l = 0.01$ and setting $l$ as $0.1 \cdot mean$ was a lot more feasible. Unfortunately this means that when it comes to statistical significance, two measures were used in this experiment. This was, however, unpreventable because of the big difference in $mean$s between $\rho = 0.1$ and $\rho = 0.9$.

The first experiment was an M/M/n queue with FIFO scheduling. So both the arrival times and service times for each customer were determined using an exponential distribution and they had means of $\frac{1}{\lambda}$ and $\frac{1}{\mu}$ respectively. The data was used to determine the mean number of customers as a function of system load and the mean waiting time as a function of system load $\rho$. This was done for $n = 1$, $n = 2$ and $n = 4$ and the results were compared to each other. The values were also compared to the theoretical results that were determined using equations 5, 7, 8 and. 9.

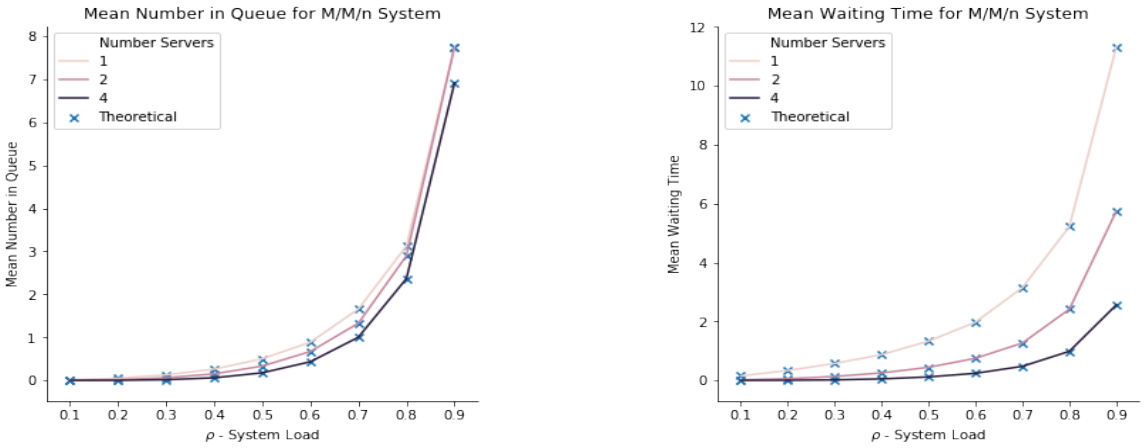
The second experiment was an M/M/n queue with shortest jobs first scheduling. This experiment was only run for $n = 1$ and the outcomes were compared to the outcomes of experiment 2 for $n = 1$, both for average number of customers as a function of system load as well for average waiting time as a function of system load.

The third experiment was an M/D/n queue with FIFO scheduling. So instead of $\frac{1}{\mu}$ being an average service time, it was now the exact service time for each customer. This was again done for $n = 1$, $n = 2$ and $n = 4$ and the mean number of customers as a function of system load and the mean waiting time as a function of system load were determined.

The last experiment was an M/M/n queue with FIFO scheduling but with a long-tail distribution for the service rate. Normally the service time of a customer was chosen randomly from an exponential distribution with an average value of $\frac{1}{\mu}$, but for a long-tail distribution the service time was drawn randomly from one of two exponential distributions, with two different mean values. There was a 75% chance of the service time being drawn from a distribution with mean 1 and a 25% chance of a service time being drawn from a distribution with mean 5. Because a $\mu$ was still needed to compute the arrival rate, the weighted average of 1 and 5 was taken and used for $\mu = \frac{1}{0.75 \cdot 1 + 0.25 \cdot 5} = 0.5$. The mean number of customers as a function of system load and the mean waiting time as a function of system load were determined for $n = 1$, $n = 2$ and $n = 4$.

# 5 Results and Discussion

The theoretical results clearly show that for a system with a FIFO scheduling and the same system loading characteristics, the average wait times should be shorter for an M/M/n queue than an M/M/1 queue. In order to test these results, a queue system was implemented in python, and the results were graphed against the expected theoretic values. These can be seen in Figure 2. The theoretical results closely match the observed values, showing the system dynamics match the expected dynamics. Once this was confirmed, it can be easily seen that the average wait time of the M/M/n system is indeed less than that of the M/M/1 system.

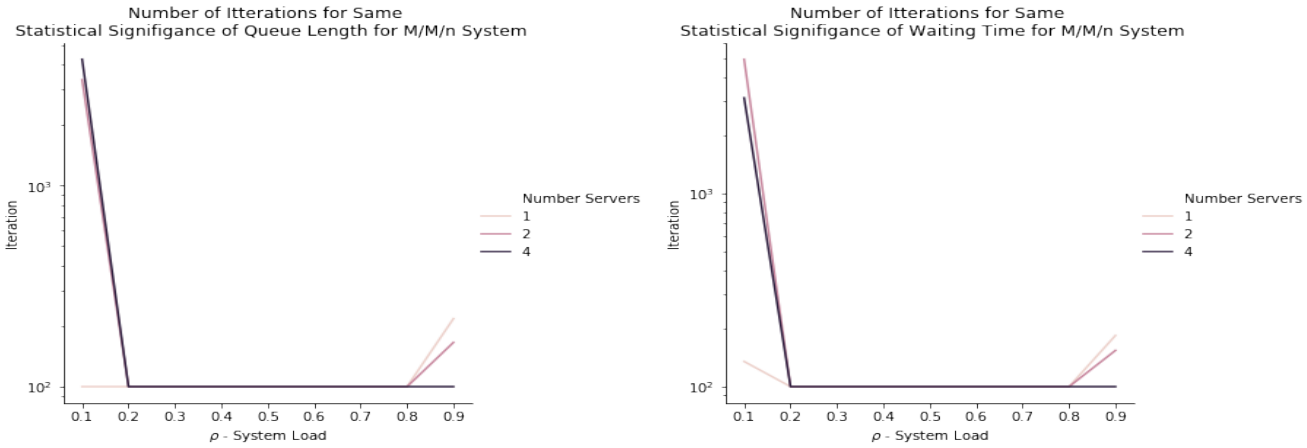



**(a)** **(b)**

**Figure 2:** M/M/n FIFO Queuing System simulated and theoretical average queue length and wait times for various system loads.

In order to determine how the system would respond when seeking a certain statistical significance for the waiting time and length of the queue, the number of iterations as a function of the system load $\rho$ was investigated. For this investigation, the confidence interval used was 99%, and the length of the range (tolerance bands) sought was $0.1 \cdot (mean)$. This was in order to ensure that the tolerance bands were small enough not to overshadow the mean. It should be noted that for $\rho = 0.1$ the computational expense required was significant enough that the simulations did not finish after 520,000 iterations. In order to achieve a reasonable computation expense, the value for the length of the tolerance bands for $\rho = 0.1$ was set to 0.01. Figure 3 shows the results of the experiment.

It can be seen from Figure 3 that for small and large values of $\rho$ the computational expense (number of iterations) to ensure the same statistical significance was much higher than for those in the middle of the range. For the lower values of $\rho < 0.2$ the variance is relatively



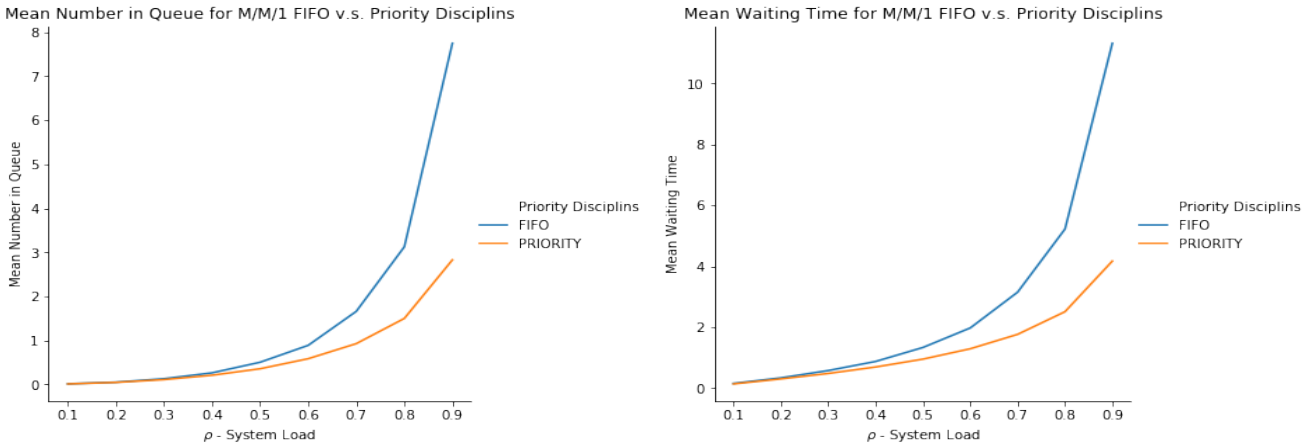

**(a)** **(b)**

**Figure 3:** M/M/n FIFO Queuing System number of iterations required to achieve the same statistical significance (l = 0.1) average queue length and wait times for various system loads.

higher than the mean, and requires a large number of iterations to bring the variance down to allow it to fall within the required significance range. For higher values of $\rho > 0.8$ the system gets close to the point where the average number of customers and average waiting time diverge and it can be seen that at this point the system converges to the average slower, therefore the system has more variance. This caused it to take longer to converge to the required significance range.

Next the effects of a different scheduling method (Shortest Priority) on a M/M/1 queuing system were examined. The FIFO method was substituted for Shortest Priority (shortest job first) and the number in the queue and the waiting time were compared to the standard FIFO method. This can be seen in Figure 4. From the figure it can be seen that for a M/M/1 queuing system that wait time and the number in the queue were both shorter than for the FIFO method. This could be because the shorter jobs were taken care of first, allowing the wait times for those jobs to be less, and letting the longer jobs be seen to as soon as they were the shortest. This would keep the queue shorter, but possibly keep longer jobs in the system indefinitely if they are always longer than another.

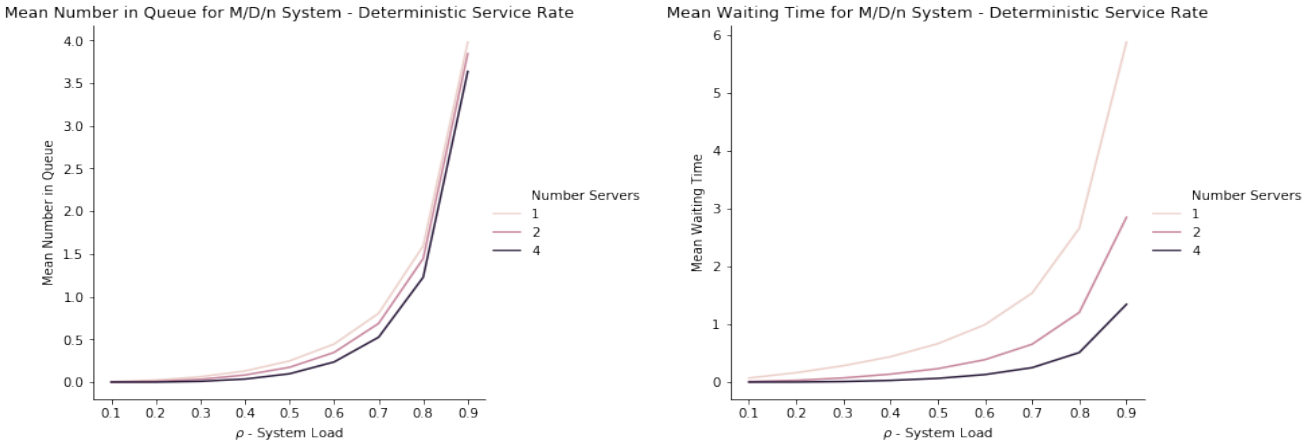
Next the system was altered so that the service time was a deterministic, an M/D/n system. The system was modeled and the effects of a deterministic service time were observed. The observed data can be seen in Figure 5. From the figures it can be seen that the systems perform slightly worse than those in Figure 2. This is likely because without the stochastic distribution, the M/D/n model did not have a chance to receive the shorter service times that the M/M/n system experienced. The systems still performed better when the number of




**(a)** **(b)**

**Figure 4:** M/M/n Shortest Priority Queuing System simulated and theoretical average queue length and wait times for various system loads.

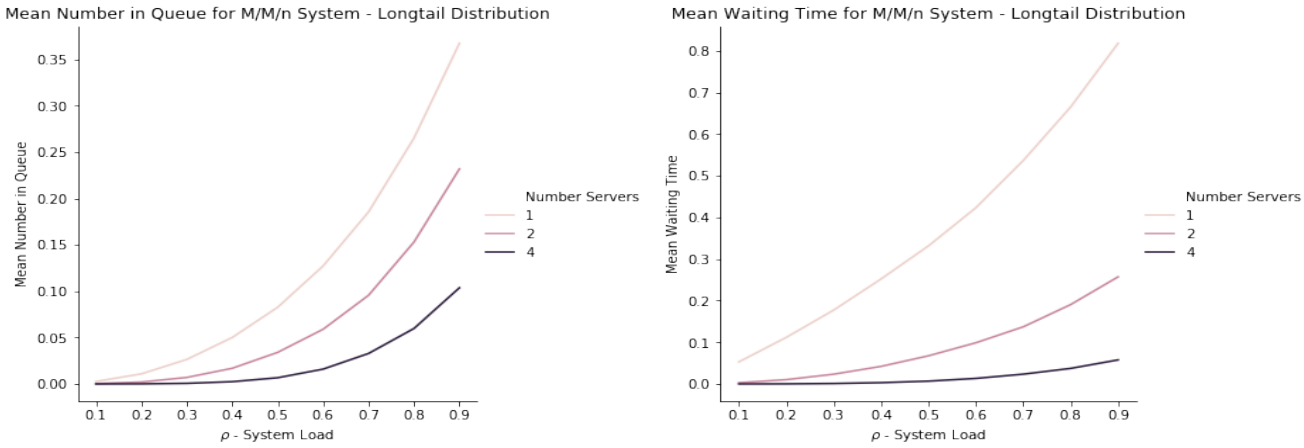
servers were increased.




**(a)** **(b)**

**Figure 5:** M/D/n FIFO Queuing System simulated and theoretical average queue length and wait times for various system loads.

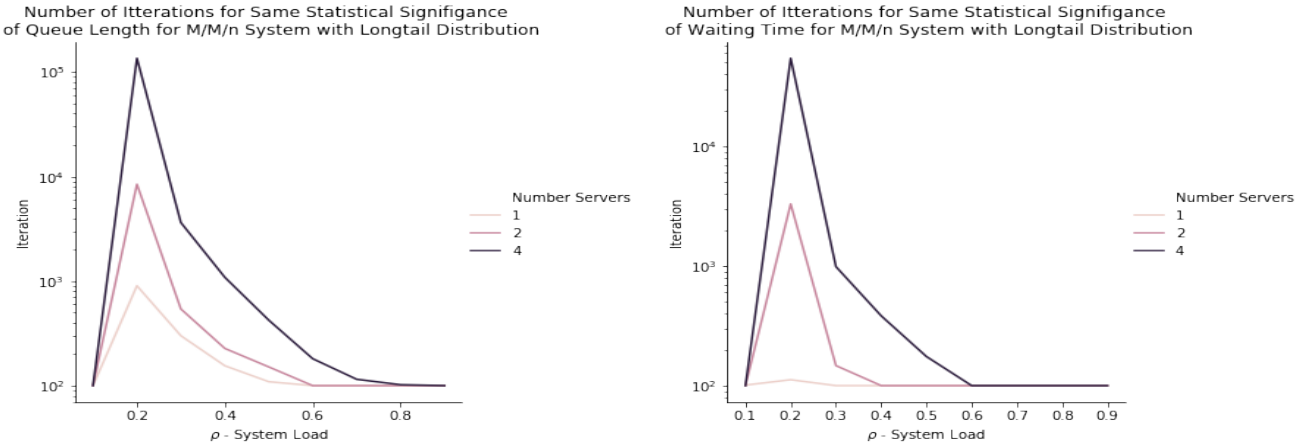
Finally the distribution of samples was altered to allow for a longtail distribution with 75% of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25% an exponential distribution with an average service time of 5.0 queuing system. The simulation was plotted against various values of system load $\rho$ for different numbers of servers. The results can be seen in Figure 6. Comparing Figure 6 to Figure 2 it can be seen that the longtail distribution caused a much lower number in the queue and lower waiting time.




**(a)** **(b)**

**Figure 6:** M/M/n FIFO with long-tail distribution (75% of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25% an exponential distribution with an average service time of 5.0 Queuing System) simulated average queue length and wait times for various system loads.

It was noted during the experiment though, that for low numbers of the system load, the simulation took an increasing amount of iterations to achieve the same statistical significance and range. In order to see how the system was being effected, the number of iterations required to achieve the same statistical significance in the FIFO part of the experiment (the confidence interval used was 99%, and the length of the range (tolerance bands) sought was $0.1 \cdot (mean)$ except for $\rho = 0.1$ where the range was 0.01 due to computational expense). This can be seen in Figure 7. The number of iterations can be seen to be rising at a higher system load than in the FIFO experiment (Figure 3). The reason for this could be because the service time is draw from 2 different distributions, causing the variance to increase between runs. This meant that more runs were needed to achieve the same statistical significance as the original FIFO run.




**(a)** **(b)**

**Figure 7:** M/M/n FIFO with long-tail distribution 75% of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25% an exponential distribution with an average service time of 5.0 Queuing System) number of iterations required to achieve the same statistical significance (l = 0.1) average queue length and wait times for various system loads.

# 6 Conclusion

In this experiment, the dynamics of a queue system was analyzed from a theoretical and practical standpoint. First the theoretical system was defined and the theoretical system dynamics of a FIFO M/M/n system were compared to the theoretical system dynamics of a FIFO M/M/1 system. It was found that the M/M/n system should be equal to or faster than the M/M/1 system. Once this was theoretically proven, the system was simulated and the theoretical results were compared to the experimental results. The system load was varied with a fixed service rate, and the systems were analyzed. The experimental results were confirmed by the theoretical results and it was found that for small values of the system load, that the number of iterations in the experiment needed to achieve the same statistical significance was higher than those with higher system loads.

Next the M/M/1 system was modified to observe the effects of a shortest job first scheduling method. The dynamics of the system were observed and it was found that for the shortest jobs first, the number of people in the queue and the average wait time was shorter, but it is possible that those in the queue had been there for a long period of time due to the jobs never being the shortest job in the system.

Finally, different service rate distributions were altered. At first the system was changed to a deterministic service rate (M/D/n) and the effects were observed. It was found that the systems perform slightly worse than those with a Markov service rate. This is likely because without the stochastic distribution, the M/D/n model did not have a chance to receive the shorter service times that the M/M/n system experienced. The final distribution modeled was a longtail service distribution, which was an example of a hyper-exponential distribution. This longtail distribution caused a much lower number in the queue and lower waiting time than for the original FIFO experiment. There was an increase in the number of runs needed to achieve the same statistical significance, which could be in part due to the increased variance in the system since the distribution was being selected from 2 peaks.

# References

[1] Gely P. Basharin, Amy N. Langville, and Valeriy A. Naumov. The life and work of A.A. Markov. In *Linear Algebra and Its Applications*, 2004.

[2] Anthony Ralston, Edwin D. Reilly, and David Hemmendinger, editors. *Encyclopedia of Computer Science*. John Wiley and Sons Ltd., Chichester, UK, 4th edition, 2003.

[3] Sheldon M. Ross. *Simulation*. 2013.

[4] M. Veeraraghavan. M/m/1 and m/m/m queueing systems. 2004.

[5] Andreas Willig. A short introduction to queueing theory. 1999.