QingheGao
12589896
Maud Bremer
11347872

Assignment 2
Queue formation for four different queuing systems

2019/12/01

1 Abstract

Understanding how queues form is useful, so large waiting times can be prevented. In this report the waiting times for four different methods ($M/M/n$, $M/M/n - priority$, $M/D/n$ and $M/LT/n$) are compared while having the same work load. With increasing work load the waiting times also increased, but by having more servers in the system this can be reduced. The $M/M/n$ system with priority scheduling and the $M/D/n$method had shorter mean waiting times and had the best performance out of all the methods.

2 Introduction

Queues are unavoidable in modern society. Not a day comes by without standing in a queue, which could be in a supermarket, or an airport or at the elevator. These queues usually form because a server, n, cannot process all the customers at a given time. This leads to a customer having to wait until the server is free again. Queuing theory describes the dynamics of this process and helps understand how a queue is formed. This way queues and long waiting times can be prevented.
The process of queuing depends on the arrival and service of a customer. A queuing system can be modeled in the following manner: a customer arrives at a random time, after which it will enter the queue for the service desk. When there are no customers in front of them, the customer will be served and leaves after that. However, if there are customers already being served, the new customer will have to wait for them to be finished; the customer will have to form a queue. The arrival and service times are random in a real system, which is why these are usually drawn from a distribution in a computational model. This way the queues can be modeled.
In this report different amount of servers and different service time distributions are used to understand how the mean waiting time can be minimised, while keeping the work load, ρ the same. This work load, and also the waiting time, was shown to be dependent on how many customers are present in the systems. Lastly, different methods were compared, FIFO (first-in-first-out) and priority handling and different distributions of the service time are used, exponential, deterministic and long-tail distribution. These methods shows decreasing waiting times with increasing n. Of all the methods the $M/M/n - priority$ and $M/D/n$ showed smallest waiting times, for a constant work load.

3 Theory

3.1 Queuing theory model and Kendall notation

Since queues are common in the daily life, understanding the queues and waiting time can help stores and companies to improve the customers' service.

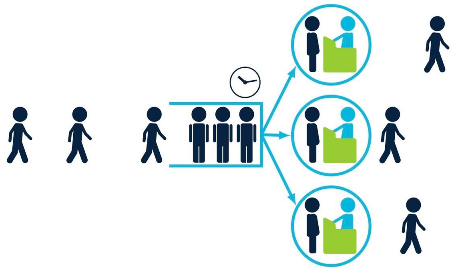

Figure 1: Queues in the daily life, adapted from [1]

In the queuing model used in this report, a customer arrives with arrival rate, λ. Afterwards it tries to access a server, n. This server can only serve one customer at a time, with a service time given by the service rate μ. If the server is not available, a queue is formed, until the server is free again. The waiting time is thus the time between arrival and the customer being served. After the service time is completed the customer leaves again.

These events do not follow in order, the interarrival and service times have a random nature. Additionally, the customers can be helped in different orders. FIFO (first in first out) will help the customer in order that they entered independent of their job size. While priority handling will help the customer with the smallest jobs first.
Different versions of this model are characterised by the Kendall notation: $A/B/n/N-S$. [2]

1. A is the distribution of the interarrival time.
2. B the distribution of service time
3. n the number of servers
4. N the maximum amount of customers that can enter the waiting line
5. S the service discipline. FIFO, Priority and so on

The distributions used in this report for A and B are the following:

- Markov, M, uses an exponential distribution: $A(t) = 1 - e^{-\lambda t}$
- Deterministic, D, uses a constant value for A and B
- Long-tail distribution, LT, is similar to an exponential distribution, but there are more samples drawn further away from the expected value.

3.2 Little's Law

Little's law describes how the average waiting time is affected by the number of customers in a system and its arrival rate. [3] The average of customers arriving at time t is given by:

$$\bar{A}(t) = \frac{A(t)}{t} \tag{1}$$

Each customer will spend time inside the system. This is the time from arrival to being helped at a server and is given by T.

$$\bar{T}(t) = \frac{1}{A(t)} \sum_{i=1}^{A(t)} T_i \tag{2}$$

The function $F(t)$ gives the space between arrival and departure of a customer, which can be given by substraction of the time of arrival and departure. But it can also be given by the summation of all customer times minus a small error, which is due to begin and end values of the response time.

$$F(t) = \int_0^t (A(u) - D(u))du = \int_0^t N(u)du = \sum_i^{A(t)} T_i - E(t) \tag{3}$$

Using Equations 3.2 and 3.2, Little's law can be derived: Equation 3.2. This law shows the relations between the mean number of customers in the system, the arrival rate and the mean time a customers spends in the system.

$$\int_0^t N(u)du = \sum_i^{A(t)} T_i - E(t) \tag{4}$$

$$\frac{1}{t}\int_0^t N(u)du = \frac{A(t)}{A(t)t} \sum_i^{A(t)} T_i - E(t) \tag{5}$$

$$\bar{N} = \lambda \bar{T} \tag{6}$$

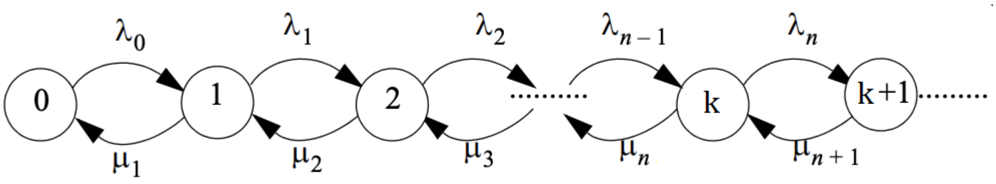

Figure 2: CTMC for a birth-death process, where λ_i , i=1,2,$\cdots$ n and μ_i , i=1,2,$\cdots$ n. [4]

3.3 Performance measurements for the $M/M/n$ system

Steady state:
In order to study the performance measures, a special case continuous time Markov chain (CTMC) for a birth-death process is constructed, see Figure 2. In this way all $M/M/n$ systems can be described by this Markov chain.
In order to get the steady-state, first calculate the probabilities over k customers:

$$p_k = \lim_{t\to\infty} P_k(t) \tag{7}$$

$$\sum_{k=0}^{n} p_k = 1 \tag{8}$$

$P_k(t)$ denotes at the probability at time t with k customers. The steady state probability, in which there are k customers in system, is given by p_k where t goes to infinity. Thus, for $\lim_{t\to\infty} \frac{dP_k(t)}{dt} = 0$, the system will be stable.
From Figure 2 the steady state flow equations are derived:

$$\lambda_0 p_0 + \mu_2 p_2 - \lambda_1 p_1 - \mu_1 p_1 = 0 \tag{9}$$

$$\lambda_1 p_1 + \mu_3 p_3 - \lambda_2 p_2 - \mu_2 p_2 = 0 \tag{10}$$

$$\cdots \tag{11}$$

$$\lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1} - \lambda_k p_k - \mu_k p_k = 0 \tag{12}$$

From these steady state equations the following expression is derived of the steady state probability for k, given by p_0 the arrival rate, λ, and service rate, μ.

$$p_k = \frac{\lambda_{k-1}\lambda_{k-2}\cdots\lambda_0}{\mu_k\mu_{k-1}\cdots\mu_0} p_0 = p_0 \prod_{i=1}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \quad for \quad k \geq 2 \tag{13}$$

3.3.1 M/M/1

For M/M/1, the arrival rates and service rates are constant ($\forall\lambda_i = \lambda, \forall\mu_i = \mu$). Thus:

$$p_k = (\frac{\lambda}{\mu})^k p_0 \tag{14}$$

The system load, ρ, depends on the arrival rate and service and is defined by: $\frac{\lambda}{\mu}$. There can only be a steady state when the work load is smaller than 1, $\rho < 1$. Following from Equation 7, p_0 for all customers will be:

$$p_0 = \frac{1}{1 + \sum_{j\geq 1}^{k} \prod_{i=1}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} = 1 - \frac{\lambda}{\mu} = 1 - \rho \tag{15}$$

The mean number of customers will be:

$$\bar{N} = E(n) = \sum_{k=0}^{n} np_k = \sum_{k=0}^{n} n\rho^n(1-\rho) = (1-\rho)\sum_{k=0}^{n} n\rho^n = (1-\rho)\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho} \tag{16}$$

Where these geometrical series have been used:

$$\sum_{k=0}^{\infty} kx^k = \frac{x}{(1-x)^2} for\ |x| < 1 \tag{17}$$

Using Little's law (Equation 3.2) the mean response time can be given by:

$$W = E(r) = \frac{E(n)}{\lambda} = \frac{1}{\mu - \lambda} \tag{18}$$

And the mean waiting time in the queue is the response time minus the service time, which is given by:

$$W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)} \tag{19}$$

And the mean number of customers in the queue for one server is given by: [4]

$$\bar{L_q} = \lambda W_q = \frac{\rho^2}{1-\rho} \tag{20}$$

3.3.2 Performance in an M/M/n system

For a M/M/n queuing system, the arrival rates are constant ($\forall \lambda_k = \lambda$), but the service rate, μ depends on the number of servers and how many customers are present in the system.

$$\mu_k = \begin{cases} k\mu & \text{if } k < n \\ n\mu & \text{if } k \geq n \end{cases} \tag{21}$$

Using Equation 13, the following equation for p_k is derived.

$$p_k = \begin{cases} p_0(\frac{\lambda}{\mu})^k(\frac{1}{k!}) & \text{if } k < n \\ p_0(\frac{\lambda}{\mu})^k(\frac{1}{n!n^{k-n}}) & \text{if } k \geq n \end{cases} \tag{22}$$

Again using Equation 7, 22 and $\rho = \frac{\lambda}{n\mu}$ p_0 will be given by:

$$p_0 = \left[\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!}\frac{1}{(1-\rho)}\right]^{-1} \tag{23}$$

Thus the mean number of customers in the system is:

$$\bar{N} = E(n) = \sum_{k=0}^{n} np_k = \sum_{k=0}^{n} n\rho^n(1-\rho) = n\rho + \rho\frac{(n\rho)^n}{n!}\frac{p_0}{(1-\rho)^2} \tag{24}$$

And the probability of all the servers being busy (Erlang-C formula) is given by :

$$P_Q = \frac{(n\rho)^n}{n!}\frac{p_0}{1-\rho} \tag{25}$$

And the mean number of customers in the queue is:

$$\bar{N_q} = P_Q\frac{\rho}{1-\rho} \tag{26}$$

And the mean waiting time in the queue time and mean response is:

$$W_q = P_Q\frac{\rho}{\lambda(1-\rho)} \tag{27}$$

$$W = \frac{1}{\mu} + \frac{P_Q}{n\mu - \lambda} \tag{28}$$

Take M/M/1 , M/M/2 and M/M/4 as examples.
For $M/M/1$the mean waiting time is given by Equation 19.

For $M/M/2$, $n = 2$.

$$
\begin{aligned}
W_{q2} &= P_Q \frac{\rho}{\lambda_2(1-\rho)} && (29)\\
&= \frac{(2\rho)^2}{2!} \frac{p_0}{1-\rho} \frac{\rho}{\lambda_2(1-\rho)} && (30)\\
&= \frac{(2\rho)^2}{2!} \frac{1}{1-\rho} \frac{\rho}{\lambda_2(1-\rho)} \left[\sum_{k=0}^{1} \frac{(2\rho)^k}{k!} + \frac{(2\rho)^2}{2!} \frac{1}{(1-\rho)} \right]^{-1} && (31)\\
&= \frac{\rho^2}{\mu(1-\rho^2)} && (32)
\end{aligned}
$$

And for $M/M/4$:

$$W_{q4} = P_Q \frac{32\rho^5}{3\lambda_4(1-\rho)^2} \tag{33}$$

$$P_Q = \left[1 + 4\rho + 8\rho^2 + \frac{32}{3}\rho^3 + \frac{32\rho^4}{3(1-\rho)} \right]^{-1} \tag{34}$$

4 Method

4.1 Statistical significance

The significance test is a method used in statistics to determine whether the conclusions drawn from a sample can be inferred to the population. Confidence interval is used to describe the statistical significance. Besides, according to the central limit theorem (CLT), the distribution of the sample mean approximates a normal distribution regardless of the distribution of population.
Thus, the central limit theorem was used to get the error bars. Equation 35 shows the confidence interval. $\bar{x}$ is mean of samples, σ is standard deviation and n is sample number.
And in this report, all the confidence interval are 95%

$$\bar{x} \pm \frac{1.96\sigma}{\sqrt{n}} \tag{35}$$

4.2 Determination of the distribution of the mean waiting time

And in this report, the Shapiro-Wilk test is used to determine if the average waiting time data is normally distributed. The test statistics are given by Equation 4.2, this is then used to get the p-values. If these values were larger than 0.05 the data was normally distributed.

$$W = \frac{(\sum a_i x_i)^2}{\sum(x_i - \bar{x})} \tag{36}$$

4.3 Determination of the number of customers

Using Simpy, Numpy and Matplotlib the queuing system was programmed in Python 3. [5] [6] [7] [8] Queues and thus waiting times will only form if there are enough customers in the system. Thus, the number of customers for $\rho = 0.8$ is varied to find how the number of customers influences the waiting time. Since this report focuses on $n = 1, 2, 4$, $n = 4$ was chosen to discuss how the number of customers influences the waiting time. If queues are forming for $n = 4$, thus the mean waiting time will not be zero, queues will definitely form for systems with lower amount of servers.
The total number of customers is ranged from 100 to 20000 with $\rho = 0.8$. The minimal number of customers that gave a waiting time was calculated using 100 simulations.

4.4 Waiting times for different service distributions and server numbers

The effect of system load, ρ, on the waiting time was investigated. The ρ was ranged from 0.1 to 1, with a total customer of 10000 for server number of 1,2 and 4.
Finally, four different methods were compared. First $M/M/N - FIFO$, with an exponential arrival and service time distribution ,and FIFO ordering. Then the same $M/M/N - Priority$ but instead of FIFO ordering it

uses priority. Here priority is given to customers with a large job. Followed is the $M/D/N$ which instead of a random service time, has a fixed service and is thus deterministic, while the arrival is drawn from an exponential distribution. Finally, the $M/LT/N$ is discussed, where the service time is drawn from a long tail distribution instead of an exponential distribution. For all of these methods the waiting times are compared while the system work load, ρ, had a constant value of 0.8.

5 Results and Discussion

5.1 Queuing dynamics in the $M/M/n$ model

$M/M/N$ is a basic model of queuing theory, where the arrival and service time are taken from an exponential distribution and the jobs are serviced via FIFO. Figure 3 shows these dynamics for $n = 1$. First a customer arrives, after which it will go to a server if it is free. However, sometimes a server is not available and a queue is formed. The customer will have to wait and there is now a waiting time. The Figure also shows that the interarrival times and service times are random. A queue is formed when the interarrival time of a new customer was short or the service time of the previous customer was longer. The average waiting time is time on average a customer will spend into the system. Thus, it is important to explore the waiting time of queuing system.

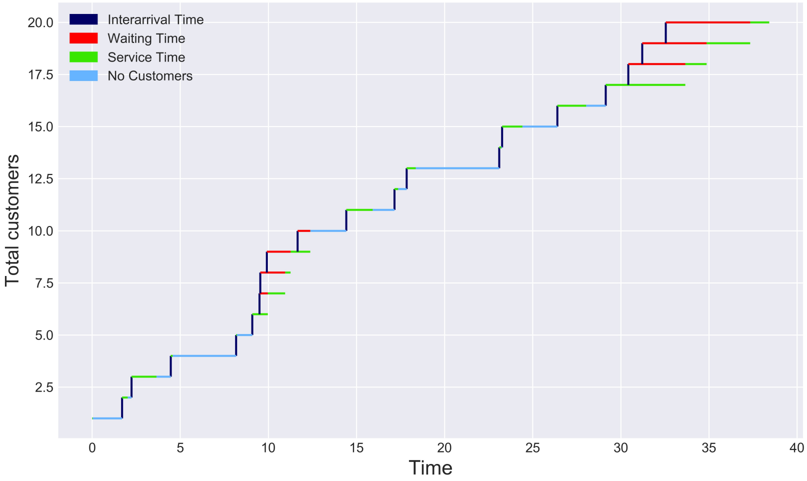

Figure 3: Customer flow of $M/M/1$ system, with $\lambda = 0.6$, $\mu = 0.8$, $\rho = 0.75$ and total customers = 20. In dark blue is the interarrival time, red the waiting time of a customer, green the service time of a customer and in light blue the time that there are no customers in the system.

5.2 Study on mean waiting time

5.2.1 The mean waiting time is normally distributed

In order to ensure the statistical significance, the distribution of the mean waiting time is explored. Normal distribution has been tested first. As Figure 4 and Table 1 show, the P values for the different methods are bigger than 0.05. So these can be accepted and thus the mean waiting times are normally distributed. So, central limit theory can be used to get the confidence interval to ensure the statistical significance for the following simulations.

Table 1: Normal test for four methods, $\rho = 0.9$, number of customers=1000.

Methods	P value	H_0
M/M/N	0.31109	Accept
M/M/N-P	0.49056	Accept
M/D/N	0.39529	Accept
M/LT/N	0.02235	Accept

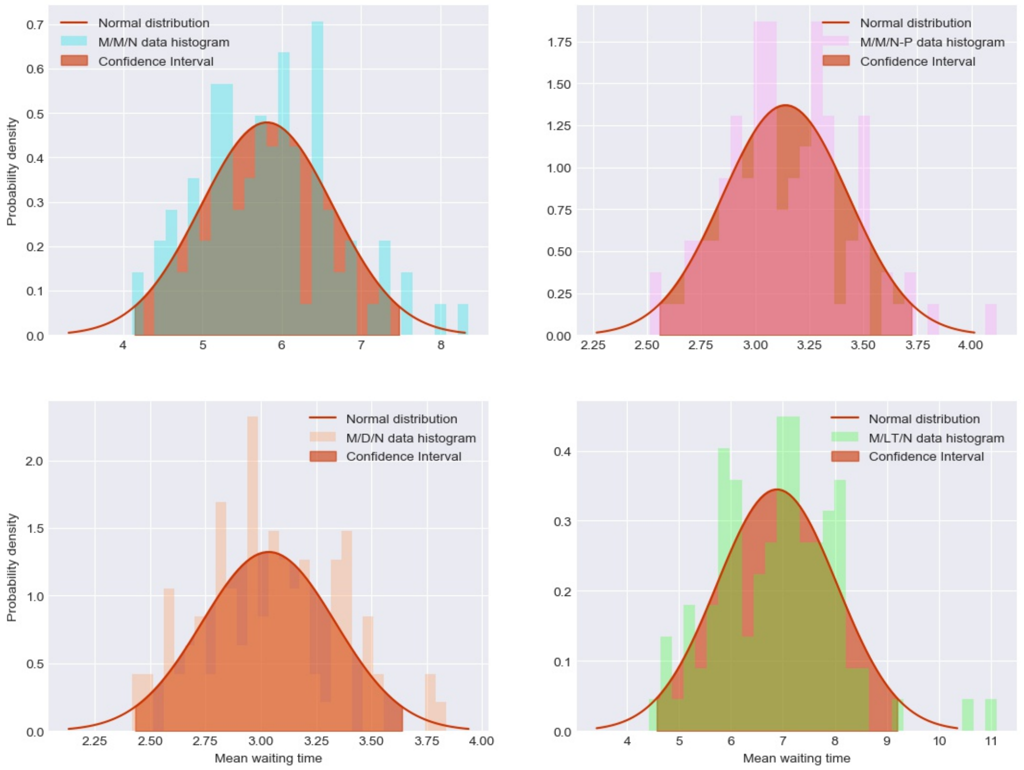

Figure 4: Normal distribution test for four methods. In the top left is $M/M/N$, top right is $M/M/1/P$, bottom left is $M/D/N$ and bottom right is $M/LT/N$. And the ρ=0.8, number of customers=10000 and simulations=100.

5.2.2 Determination of number of customers

A system's mean waiting time is dependent on the system load, ρ, based on Equation27. However, this is only true for a system where t goes to infinity or in this case a system with a large number of customers. As Figure 5 shows, the mean waiting time first increases slightly when the total customers increases. Also the standard deviation decreases; the distribution gets sharper with increasing number of customers. This Figure shows the waiting time for a M/M/4 system, so in system with lower amount of servers the waiting time will definitely never be zero. In other words, for a customer number of 10000 a $M/M/n$ system with 1,2 or 4 servers and $\rho = 0.8$ will always have queues forming.

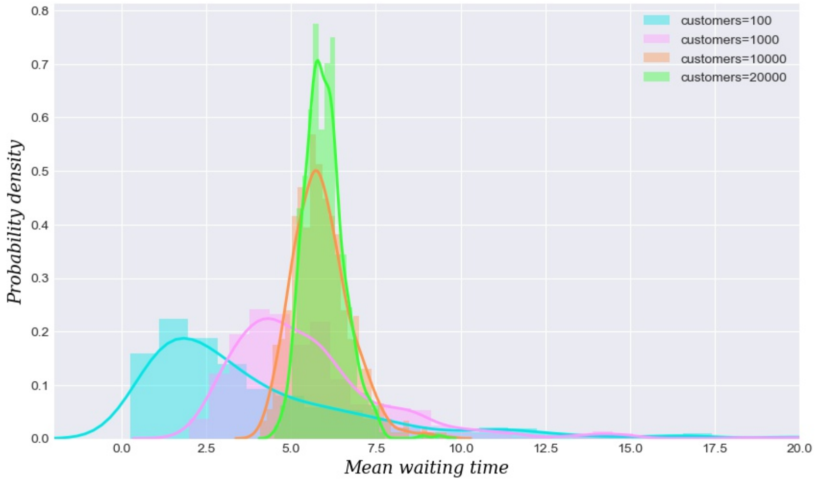

Figure 5: Determination of number customers in an $M/M/4$ system with $\rho = 0.8$.

5.2.3 Mean waiting time increases with more work load

Based on the determination of the number of customers, the effect of ρ on the mean waiting time for four methods has been explored. Figure 6 shows that when ρ increases, the mean waiting time also increases. This is seen for $n = 1, 2, 4$. Especially when ρ approaches to the 1, the mean waiting time increases significantly. When the system load is 1, the arrival times will be equal to the server time ($\rho = \frac{\lambda}{n\mu}$). This means that the arrival rate will be too high for the servers to handle all the incoming customers, which will increase the mean waiting time for each customer.

The Figure 6 also shows that the average waiting time will be smaller for high server numbers. More servers always have a shorter mean waiting time than systems with low servers, even though each system has the same system work load.. Especially when ρ is around 0.9-1.0, the mean waiting time increases fast. Apparently, having more servers available will increase the run through of all the customers. Even when a customer has a large service time, the other servers can still receive the incoming customers and prevent the formation of a large queue.

Additionally, Figure 7 shows Figure 6 again, but for the region of the workload from 0.85 to 1. It it clear to see that the confidence intervals(shade part) for the waiting time for each ρ in all methods are quite small and do not overlap between the different number of servers.

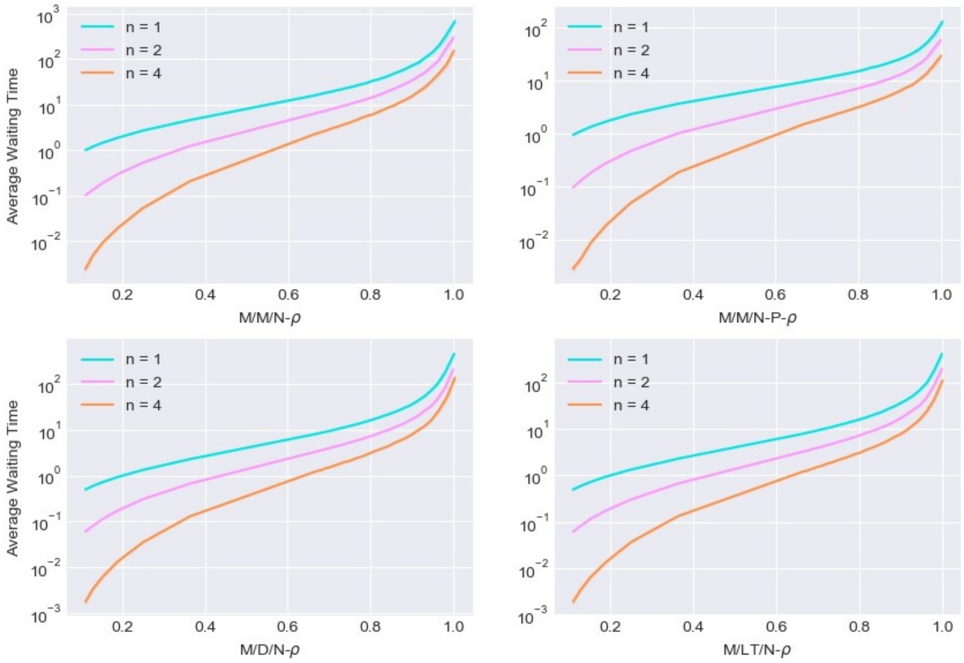

Figure 6: The correlation of changing ρ and mean waiting time for four methods. ρ is ranged from 0.1-1 for server numbers, $n = 1, 2, 4$. In the top left is $M/M/N$, top right is $M/M/1/P$, bottom left is $M/D/N$ and bottom right is $M/LT/N$.

5.3 Performance for different number of servers for different methods

Figure 8 shows the average waiting times for different servers ($n = 1, 2, 4$). With increasing of the number of servers, the waiting times decreases. This can be expected, since customers can be helped more quickly when there are more servers available. The difference between the waiting time $n = 1$ and $n = 2$ is larger, than the waiting time from $n = 2$ to $n = 4$. So it seems that for increasing the number of servers, the waiting time does not change drastically anymore, while keeping the same workload.

The same figure also shows the average waiting times for the different methods: $M/M/n$, $/M/n$ with priority to the largest jobs first, $M/D/n$, and $M/LT/n$ (with Long-Tail distribution). $M/M/n$ and $M/LT/n$ shows similar performance, while priority and $M/D/n$ show significant improvement. Especially for $n = 1$. However, for larger n the differences between the methods decreases.

The priority scheduling performs better because customers who had shorter service time do not have to wait if a previous customer had a longer service time. Thus, in this way the mean waiting time will be shorter. A deterministic service time will result in short queues, because every customer has the same service time. Resulting, queues will only form if the interarrival time was shorter, but not because of a customer with a longer service time compared to the other customers.

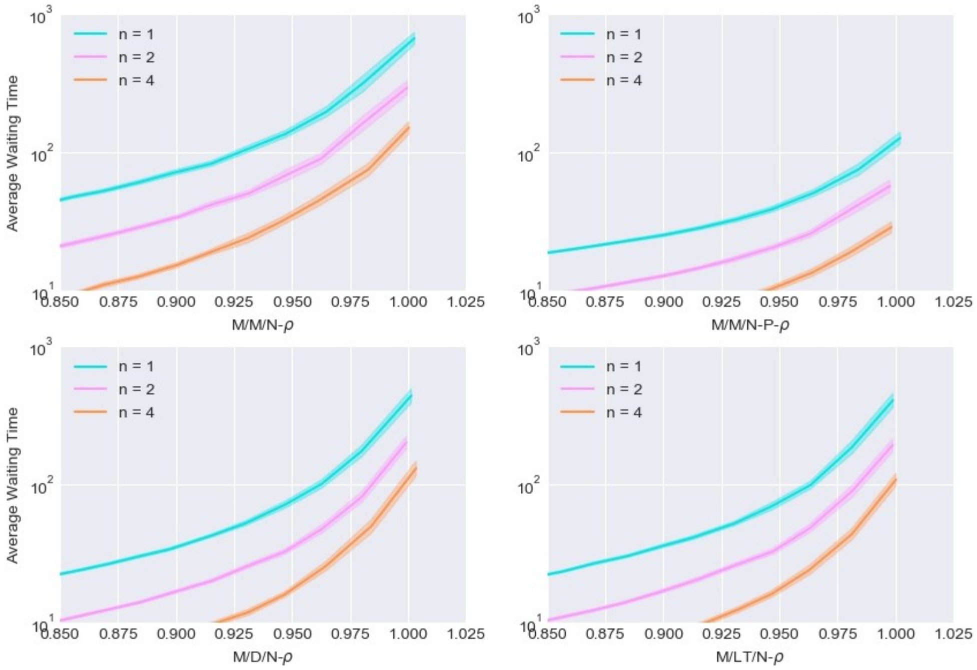

Figure 7: Zoomed in of the correlation of changing ρ and mean waiting time for four methods. ρ is ranged from 0.85-1 and servers numbers, $n = 1, 2, 4$. The shade part is confidence intervals for each ρ.In the top left is $M/M/N$, top right is $M/M/1/P$, bottom left is $M/D/N$ and bottom right is $M/LT/N$.

$M/LT/n$ has the longest mean waiting time compared to the other three methods, because the distribution of the service time is more scattered. It will have more customers with very long waiting times, which results in customers having to queue. Thus, the mean waiting time becomes larger than the waiting time for the $M/M/n$ system.

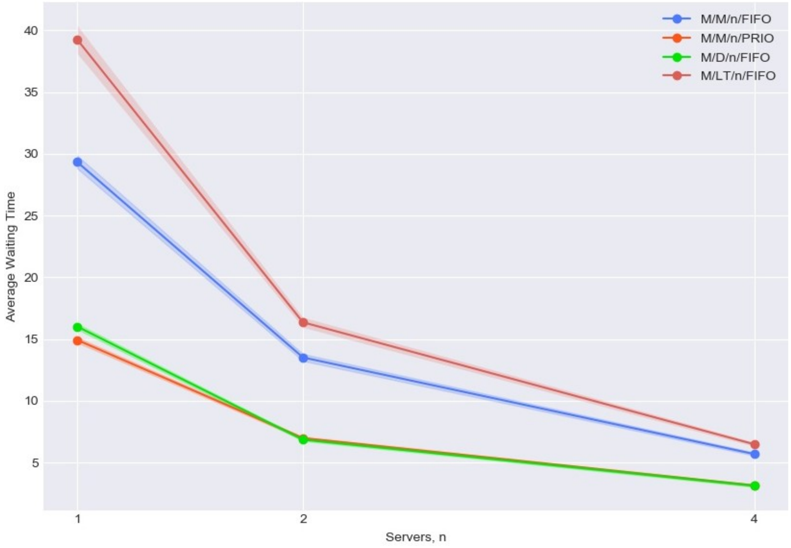

Figure 8: The mean waiting time as a function of total servers ($n = 1, 2, 4$) with 10000 customers, $\rho = 0.8$ and 100 simulations. The colours indicate the different methods. Shade area is confidence interval.

6 Conclusion

A queuing system was implemented for the four different methods: $M/M/n$, $M/M/n$ with priority scheduling, $M/D/n$ and $M/LT/n$. For small work load with 4 servers when the number of customers was 10000 the average waiting time was shown to always be larger than zero. For this amount of customers, it still had a mean waiting time, but was not too expensive to use in simulations considering the available computing power. Additionally, the average waiting times were normally distributed. The different systems showed for increasing ρ an increasing of the waiting time. Also the mean waiting time decreased with more servers. Finally, comparison of the waiting time for the different methods showed that the priority scheduling and deterministic service time performed the best. These methods had the shortest mean waiting time.

References

[1] MOOC de l'IMT. Queuing theory - mooc - free online course. `https://www.youtube.com/watch?v=QqVDNPb5faY`, 12 2017. Accessed: 1-12-2019.

[2] Andreas Willig. *A Short Introduction to Queuing Theory*. Technical University Berlin, Telecommunication Networks Group, 7 1999.

[3] Randolph Nelson. *Probability, Stochastic Processes, and Queueing Theory – The Mathematics of Computer Performance Modeling*. Springer Verlag, New York, 1995.

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

[5] Oliphant Travis E. *A guide to NumPy*. USA: Trelgol Publishing, 2006.

[6] S. Chris Colbert Stéfan van der Walt and Gaël Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13:22–30, 2011.

[7] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9:90–95, 2007.

[8] Simpy discrete event simulation for python. `https://simpy.readthedocs.io/en/latest/index.html`, 2002–2019. Accessed: 29-11-2019.