University of Amsterdam

Stochastic simulation, Assignment 2

Comparison of different queuing systems

06-12-2021

Florian Tiggeloven, Lars Grim
11872802, 13511157
f.tiggeloven@student.uva.nl
lars.grim@student.uva.nl

Abstract

Due to its many applications, queuing theory remains a concept that is computed frequently. In this report several queuing styles are analysed to investigate which style is most suitable to maintain a low mean waiting time in queue. It was found that for a M/M/n queue, selecting members in queue with a priority based on service time reduces the mean waiting time of the system significantly with a P-value of $p = 1.46e^{-3}$. Out of M/M/n, M/D/n, and M/L/n queues, the M/D/n was found to be the most efficient queue style. However, these results were deemed insubstantial due to P-values only reaching down to a value of $p = 0.46$ which is contradictory to theoretical derived expectations.

1 Introduction

Queuing theory is a branch of mathematics and applied probability theory examining every aspect of a person or abject arriving in a line, to being served and leaving the line (Willig, 1999). It is applied in many different fields, e.g. the line at the supermarket cash register, data requests at websites or waiting at the traffic light. The queuing theory is most often applied to find a balance between the efficiency and costs of a particular system. In this research different queuing systems will be compared. After some theoretical background, hypotheses will be given (see chapter 2.4). After which some applications of the different queuing models will be explained.

2 Basic System, definitions and hypotheses

2.1 The Kendall Notation

The basic system of a queue is visualized in figure 1 obtained from Willig, 1999. It can be seen that a customer from the population joins the queue (if there is one). Customers are being served by one server from a set of servers containing a minimum of 1 server, after which the customer leaves the system. Different styles of queuing systems exist. A well known and used notation of representing the different queuing systems is the Kendall Notation (Willig, 1999) and is described as followed:

$$A/B/n/K/S \tag{1}$$

A represents the distribution for the arrival process and B represents the service time distribution. Both A and B are most often exponential distributions (M) or deterministic distributions (D) while other distributions also exist. n represents the number of servers. K represents the maximum number of customers in the queue. S represents the Queuing Discipline which is most often one of the following:

- FIFO: First In First Out, customers are served in order they arrived.
- LIFO: Last In First Out, the customers are served opposite to FIFO. The last customer that entered the queue is the first to being served.
- SIRO: Service In Random Order, the customers are served in random order.
- Priority Queuing: customers are served with respect to a given priority weight. The higher the priority is the quicker customer will be served.

The system used is dependent on the application it represents. In the case of answering phones at a mobile service centre at a big company, a M/M/n/FIFO system would most probably yield the best results. While in the case of a hospital priority must be given to the more injured people. In this research different systems will be compared.

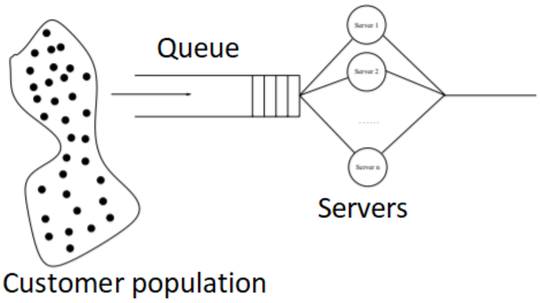

Figure 1: Visualization of a basic queuing system, obtained from Willig, 1999

2.2 Little's Law

Little's Law is one of the most important and useful relationships in the queuing theory (Shanmugasundaram and Umarani, 2015). It is often represented in one of the three following forms:

$$E(N) = \lambda E(S) \tag{2}$$

$$E(L) = \lambda E(W) \tag{3}$$

$$\rho = \frac{\lambda E(B)}{n} \tag{4}$$

For Little's Law applied on the system (eq:2), λ being the arrival rate, $E(N)$ being the expected number of customers in the system and $E(S)$ being the expected sojourn time (waiting time + service time) in the system. For Little's Law applied on the queue plus server (eq:3), λ being the arrival rate, $E(L)$ being the expected queue length under steady state and $E(W)$ being the expected waiting time under steady state. For Little's Law applied on the server (eq:4), ρ representing system load, λ being the arrival rate, n the number of servers and $E(B)$ representing the mean service time.

These relations lead to two important aspects. First, it allows us to derive one of the variables when two of the three are known. Second, it allows us to understand the impact of variable changes to the system. Furthermore, from equation 4 it can be derived that ρ must be smaller than the number of servers in order to prevent the queue length becoming infinitely long. Therefore it is desired to keep $\frac{\lambda}{n}$ smaller than $E(B)$.

2.3 Mean queue length and waiting time for M/M/n system

Multiple servers should in theory decrease the mean queue waiting time. Logically, the more servers there are the quicker the customer will get served. The mean queue length can be described as:

$$E(L) = \frac{(n\rho)^n}{n!}\left((1-\rho)\sum_{m=0}^{n-1}\frac{(n\rho)^m}{m!} + \frac{(n\rho)^n}{n!}\right)^{-1}\frac{\rho}{1-\rho} \tag{5}$$

In this equation (Adan and Resing, 2002) the mean queue length is represented as $E(L)$, the number of customers is represented as m, the number of parallel identical servers is represented as n and system load is represented as ρ. The derivation of this equation is described in the appendix.

Equation 5 can be extended to the mean queue waiting time with μ representing the service time:

$$E(W) = \frac{(n\rho)^n}{n!}\left((1-\rho)\sum_{m=0}^{n-1}\frac{(n\rho)^m}{m!} + \frac{(n\rho)^n}{n!}\right)^{-1}\frac{\rho}{1-\rho}\frac{1}{n\mu} \tag{6}$$

The equations suggest that as the number of servers (n) increases, $E(L)$ and $E(W)$ decreases. As described in the appendix, this is the case for all $\rho < 1$

2.4 Hypotheses

As suggested in chapter 2.3, the mean queue length and mean waiting time shortens and decreases respectively due to an increase in identical parallel servers for a M/M/n system.

As discussed in chapter 2 it is of interest to see how the mean waiting time differs between systems M/M/n/FIFO, M/D/n/FIFO, M/L/n/FIFO, M/M/1/prio and M/M/n/prio. Due to the characteristics of each system it is expected that M/M/n/FIFO and M/M/n/prio yield the shortest mean waiting time and of these two, M/M/n/FIFO will have the lowest variance.

3 Method

To test the hypotheses, different queuing systems were implemented in `Simpy` built upon `Python 3.8`. For all different queuing systems 120 simulations have been executed to decrease stochasticity and ensure reliable results. In this way the mean and standard deviation for different variables of each system could be calculated. For all queuing systems ρ will be constant at the value $\rho = \frac{\lambda}{n\mu} = 0.9$ to ensure equal system load across different systems. To observe the statistical implications of the different queuing systems, independent T-tests were made between various queuing systems to determine the significance.

3.1 M/M/n/FIFO

For the M/M/n/FIFO queuing system, sampling is done with the exponential distribution for both the inter arrival time and the service time. The mean value of the exponential distribution for arrivals is set to the inverse of the arrival rate λ^{-1} with $\lambda = 4.5$. The mean value of the exponential distribution for arrivals is set to the inverse of the arrival rate μ^{-1} with $\mu = 5$. From the exponential distribution with a given mean, random values will be drawn for the inter arrival and service time. The amount of servers (n) will be set on one, two and four servers. The arrival time of the customers will also increase linearly with n to accurately compare the different queuing systems. To compare these systems, boxplots of the average waiting time are made in figure 2. Independent T-tests for this system were also done, which are shown in the upcoming results section.

3.2 M/M/n/prio

The M/M/n/FIFO systems will also be compared to M/M/n/prio systems. The mean value of the exponential distribution for arrivals is set to the inverse of the arrival rate λ^{-1} with $\lambda = 4.5$. The mean value of the exponential distribution for arrivals is set to the inverse of the arrival rate

μ^{-1} with $\mu = 5$. Priority is given to the jobs which take the shortest time. Therefore, a caller that only needs a small amount service time is helped before a caller that will need more time. To compare these systems, boxplots of the average waiting time are made for M/M/n/FIFO and M/M/n/prio for values of n=1,2,4 in figure 3. To test the significance of the difference between these methods, independent T-tests were executed between the two datasets which are shown in the results section.

3.3 M/D/n/FIFO

For the M/D/n/FIFO system the inter arrival time is sampled from an exponential distribution with mean $\mu = 4.5$ and the service time will be deterministically set to $\lambda = 5$ to reach a value of $\rho = 0.9$ which was desired (see beginning chapter 3). The amount of servers (n) will be set on one, two and four servers. The arrival time of the customers will also increase linearly with n to accurately compare the different queuing systems. To compare these systems boxplots of the average waiting time are made, see figure 4. To test the significance of the difference between these methods, independent T-tests were executed between the two datasets and shown in the results section.

3.4 M/L/n/FIFO

The longtail queuing system uses two different means for the exponential distribution of the service time. 75% of the time a mean of $\mu = 1$ will be used, the other 25% of the time a mean of $\mu = 5$ will be used. To keep a constant ρ of 0.9, the mean of the exponential distribution of the inter arrival time is set to the inverse of the arrival rate = 4.5, resulting in a mean of $\frac{1}{4.5}$. This value will be scaled accordingly to the amount of servers, since more servers with an equal arrival rate will simply return faster waiting times. The amount of servers will be set on one, two and four. To compare these systems, boxplots of the average waiting time are made in figure 4 and T-tests were executed to test their significance. All of which can be found in the results section.

4 Results

4.1 M/M/n FIFO

The mean waiting time of each collection of simulations is shown in figure 2. Although the median of each queuing system stays similar, around a value of $E(W) = 40$, the spread of decreases drastically when increasing the number of servers. For n=1 the spread of the boxplot arms reaches a distance of $\Delta E \approx 80$, where this is lowered to about $\Delta E \approx 25$ for a server size of n=4. Through an independent t-test between the M/M/1 and M/M/2 a T-value of $T = -0.678$ and a p-value of $p = 0.498$ were found. Secondly for a t-test between M/M/2 and M/M/4 a T-value of $T = 1.686$ and a p-value of $p = 0.0925$ were found. Lastly, for a t-test between M/M/1 and M/M/4, a T-value of $T = 0.595$ and a p-value of $p = 0.552$ were found.

4.2 M/M/n Priority & M/M/n FIFO

To visualize the process of priority a line plot was made for the values of n=1,2,4. A clear difference between the two techniques can be distinguished. The line portraying the priority queue style contains a significantly larger value of standard deviation for all points, spreading as far as $\Delta E \approx 80$. The line plot for a queue style without priority ending on a value of $E(W) = 60$. The priority queue style follows a similar curve shape around a significantly lower value. The queue starts at a mean waiting time of $E(W) \approx 20$ for $n = 1$ and eventually develops to a value of $E(W) \approx 10$. Mean waiting times for a queue with priority selection is therefore about 5 times faster than just first come first serve selection. Between these two methods an independent T-test was made. This test returned T-value of T= 7.79 and a p-value of $p = 1.46e^{-3}$.

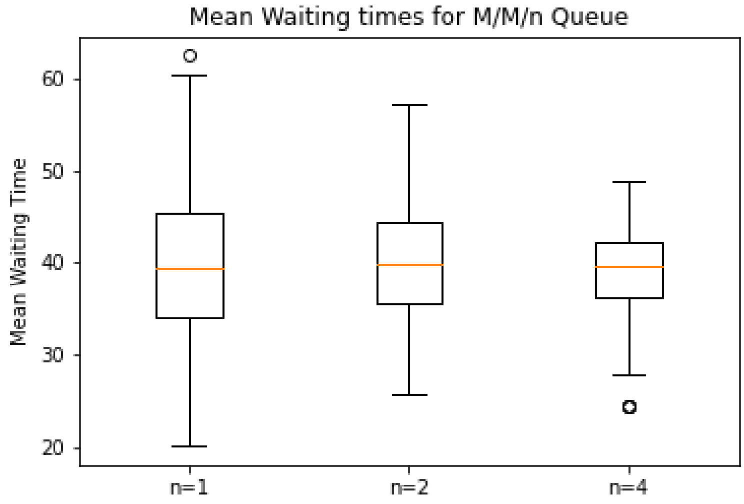

Figure 2: The mean waiting times of each M/M/n queue is shown through 3 boxplots in order of magnitude. Outliers are shown with the dots whereas max and minimal points are shown through the whiskers of the boxplot. The median being represented by the orange line in each box.

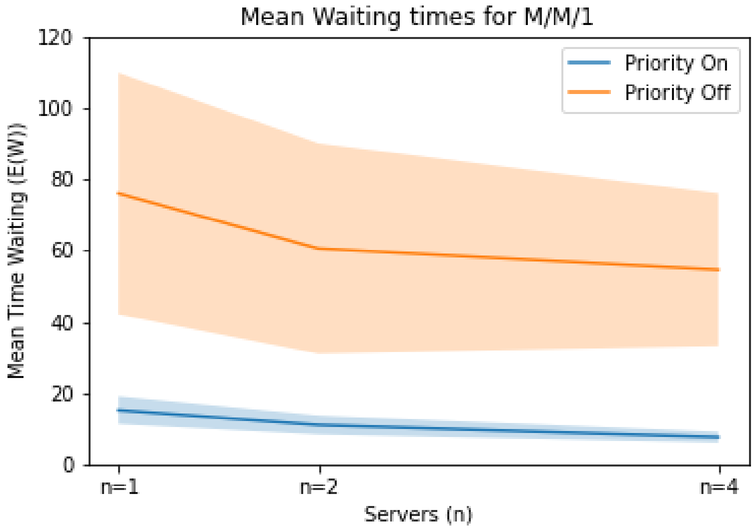

Figure 3: An M/M/1 simulation plotted in a line plot with priority turned on and off. Each line representing the mean with standard deviation in a lighter shade around it. Priority included is shown in blue, whereas priority turned off is shown in orange.

4.3 M/M/n, M/D/n, M/L/n Queue Styles

Line plots were chosen to display the values of mean waiting times along with the standard deviation of each queue style. Once again this was done over server capacity values of $n = 1, 2, 4$. All queue styles have a similar range of standard deviation around them. For a value of $n = 1$ the M/M/n queue has the highest mean waiting time, followed shortly by the M/L/n queue and finally the M/D/n queue. Out of all styles a similar pattern is continued over the different values of server size, with the M/D/n keeping its place as lowest waiting time. A significant difference over the progression of servers is that even though the M/L/n queue started under the M/M/n queue, it ends up being the queue with the highest mean waiting time at the final value $n = 4$. Between each of the queues, independent T-tests were conducted in order to analyse the correlation. Between the M/M/n queue and M/D/n queue, a T-value of $T = 0.535$ and a p-value of $p = 0.621$ were found. Between the M/D/n queue and M/L/n a T-value of $T = -0.806$ and a p-value of $p = 0.466$ were found. And lastly, between the M/M/n and the M/L/n a T-value of $T = -0.100$ and a p-value of $p = 0.925$ were found.

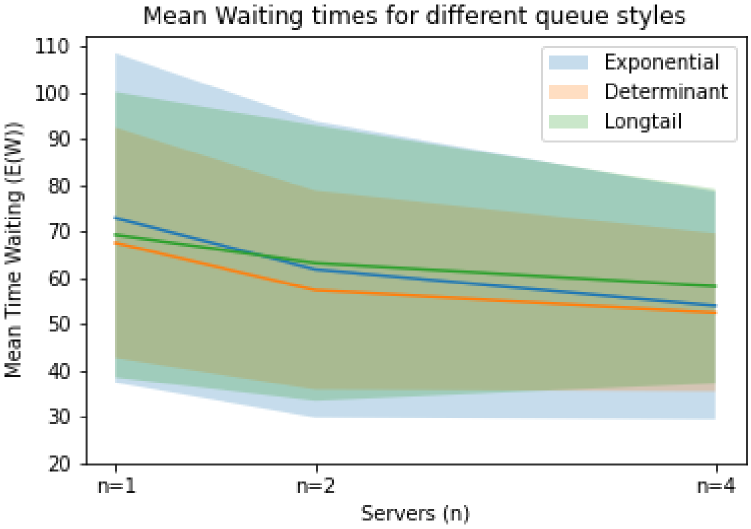

Figure 4: All different queue styles plotted in a line plot. The mean waiting time of each style is shown by the line, where the standard deviation is shown in the lighter shade around each mean line. The exponential is shown in blue, the deterministic in red, and the longtail style in green.

5 Discussion

Looking at the results of the M/M/n FIFO queue (see 4.1), several important correlations can be noted. In the simulation of the M/M/n FIFO queue, the mean waiting time for a customer remained partially the same over the increasing of the factor of servers available (n). This is not as expected, since theoretical derivations of $E(W)$ and $E(L)$ showed a decrease of mean waiting time and mean queue length (see appendix). The differences found between the simulation and theoretical results could be due to the amount of simulation per queuing style. As can be seen in the simulated results, the standard deviation is high compared to the mean value which could indicate that the number of simulations per queuing style is too low. .

By simulating the mean waiting times of an M/M/1 queue with a constrained value of $\rho = 0.9$, the effect of specific priority can be severe. If selection is based on priority (with a higher priority given to a shorter time needed to solve the problem) the mean waiting time can be decreased drastically. For a server value of n=1 a difference of almost 60 minutes in waiting time can be observed between the two methods. This finding was expected, as a M/M/n/prio queue deals with customers in a certain order making the chance for callers unnecessarily waiting for a quick task less.

Based upon the results, the following general statement can be made. The M/D/n queue seemed to return the lowest mean waiting time, followed by the M/M/n queue which surpassed the M/L/n queue in mean waiting time over the course of the values of n. However, if these waiting times are analysed using an independent T-test, no significant differences between the methods are found. P-values between each of the styles reach no lower than a value of $p \approx 0.4$, signalling that selection of a specific style does not result in significant change for the current selection of parameters. This is different than expected from the theoretical derivations. Possible explanations could be in the selection of the parameter ρ, which was set at 0.9 for all simulations. The effect of this parameter as it approaches closer to 1 might return different results, as lowering it away from 1 would only reduce the waiting time. Another possibility is using the parameter of the average number of callers in each system. This value is only mentioned in this work through the theory, but has not been subjected to any modelling or visualization. Doing so might reveal different elements about the system currently not shown through the mean waiting time.

Taking the found results into considerations, it is clear that through even a slight adjustment in queue type, significant improvements can be made for a system. Many different situations exist for the application of these specific types of queues and therefore the correct model should always be carefully selected in order to reduce minimize unnecessary waiting time. Lastly, since the result found for the M/M/n/prio queue is the most significant one found in this work, we advise researchers planning to adapt to this work to further investigate the effect of priority queuing on the M/D/n and M/L/n queues as well.

6 Conclusion

In conclusion, there has been found that an M/M/n style queue with multiple parallel servers returns a smaller spread of mean waiting times. However, independent T-tests returned poor results signalling to insignificant correlation between datasets. Furthermore, prioritizing tasks based on the time needed to assist callers decreases the mean waiting time of a simulation drastically. Lastly, a deterministic queue style has been found to return the smallest value of mean waiting times. This result is however not deemed as significant due to independent T-tests returning insignificant values.

References

Willig, A. (1999). A short introduction to queueing theory. *Technical University Berlin, Telecommunication Networks Group, 21.*

Shanmugasundaram, S., & Umarani, P. (2015). Queuing theory applied in our day to day life. *International Journal of Scientific & Engineering Research*, *6*(4), 533–541.

Adan, I., & Resing, J. (2002). Queueing theory.

Derivation of $E(L)$ and $E(W)$

In this section the equations for the mean length of the queue ($E(L)$) and mean waiting time in the queue ($E(W)$) will be derived as in the book of Adan and Resing, 2002. The occupation rate per server always (ρ) is smaller than 1 and is defined as:

$$\rho = \frac{\lambda}{n\mu} < 1 \tag{7}$$

Let p_m denote the probability that there are m customers in steady state. Iterating through different set of through states $[0, 1, ..., m-1]$ yields:

$$p_m = \frac{(n\rho)^m}{m!} p_0 \text{ for } m = 0 \text{ , ... , } n$$

and

$$p_{n+m} = \rho^n \frac{(n\rho)^n}{n!} p_0 \text{ for } m = 0 \text{ , ... , } n$$

.

p_0 comes from normalization:

$$p_0 = \left(\sum_{m=0}^{n-1} \frac{(n\rho)^m}{m!} + \frac{(n\rho)^n}{n!} \frac{1}{1-\rho} \right)^{-1}$$

The PASTA property states that for a given system with Poisson arrivals, the expected value of any parameter of the queue at the instant of a Poisson arrival is equivalent to the long-run average value of that parameter. Therefore, the fraction of customers that find the system in some state S on arrival is exactly the fraction of time the system spends in state S. This than leads to the probability a job has to wait (Π_w):

$$\begin{aligned} \Pi_w &= p_n + P_{n+1} + p_{n+2} + ... = \frac{p_n}{1-\rho} \\ &= \frac{(n\rho)^c}{c!} \left((1-\rho) \sum_{m=0}^{n-1} \frac{(n\rho)^m}{m!} + \frac{(n\rho)^n}{n!} \right)^{-1} \end{aligned} \tag{8}$$

This can be extended to the mean queue length $E(L)$ by:

$$\begin{aligned} E(L) &= \sum_{m=0}^{\infty} m p_{n+m} \\ &= \frac{p_n}{1-\rho} \sum_{m=0}^{\infty} m(1-\rho)\rho^m \\ &= \Pi_w \frac{\rho}{1-\rho} \\ &= \frac{(n\rho)^n}{n!} \left((1-\rho) \sum_{m=0}^{n-1} \frac{(n\rho)^m}{m!} + \frac{(n\rho)^n}{n!} \right)^{-1} \frac{\rho}{1-\rho} \end{aligned} \tag{9}$$

This can be extended to the mean waiting time $E(W)$ with Little's Law by:

$$\begin{aligned} E(W) &= \Pi_w \frac{\rho}{1-\rho} \frac{1}{n\mu} \\ &= \frac{(n\rho)^n}{n!} \left((1-\rho) \sum_{m=0}^{n-1} \frac{(c\rho)^m}{m!} + \frac{(n\rho)^n}{n!} \right)^{-1} \frac{\rho}{1-\rho} \frac{1}{n\mu} \end{aligned} \tag{10}$$

As can be seen in table 1 and table 2, the mean queue length and mean waiting time decreases when the number of servers (n) is increased. This is for all $\rho < 1$, with ρ kept constant by changing μ accordingly.

	$\rho = 0.3$	$\rho = 0.45$	$\rho = 0.75$	$\rho = 0.9$
n=1	0.43	0.81	3	9
n=2	0.25	0.52	2.4	8.18
n=4	0.14	0.31	1.7	6.92

Table 1: $E(L)$: $\lambda = n * 4.5$ and μ in $[5, 6, 10, 15]$

	$\rho = 0.3$	$\rho = 0.45$	$\rho = 0.75$	$\rho = 0.9$
n=1	0.03	0.08	0.5	1.8
n=2	0.008	0.03	0.2	0.82
n=4	0.002	0.008	0.07	0.35

Table 2: $E(W)$: $\lambda = 4.5$ and μ in $[4, 6, 10, 15]$