The stochastic analysis of a M/M/C model

Jasper den Duijf
Koen Greuell

Abstract

Queuing theory explains a system with arrivals of customers waiting for service where both the arrival and service rate are Poisson distributed. In this paper in accordance with this theory our simulations showed that M/M/1 queues have significantly longer mean waiting times than M/M/c queues while the system load ρ stays equal. Shortest job first scheduling was found to decrease the mean waiting time and benefit from more servers in an equal ratio as first-in, first-out scheduling. Different service rate distributions can change the queueing times even though the mean service rate is kept constant.

Contents

1 Introduction

People dislike waiting, yet most human being regularly spend time waiting to be served. They could be physically waiting: picking a number at a bakery or standing in line at the bank. One can also call a customer service to listen to a waiting tune and an automatic voice updating the customer on their place in the queue. Also in heath care a person can be in a waiting line, for example for a surgery. In general people dislike this time spent waiting, so companies often search for solutions to minimize the time customers wait for their services in the most economical manner.

Most of the time the cause of queueing is quite simple: the demand for service exceeds the service available. So with an increased number of servers or quicker services the waiting time can be lowered in theory. However, this asks for a study of the waiting time and the length of the queue and the effects of management changes. The queueing theory offers a model to analyze such systems and estimate those properties.

In 1909 the Danish mathematician Erlang published this theory applied to the telephone system in his country (in 1909 you had to wait before the operator put you through) [3]. Erlang showed that both the number of calls per hour and the number of finished conversations per service per hour were Poisson distributed. This property is still fundamental for the queueing theory as it gives the probability of a certain number of customers using a system and makes it possible to model and analyze the system both in a deterministic and a stochastic manner. Later these problems were linked to Markov Chains[4]. The system's states depended on the number of customers in the system and the probabilities of changing to another state were dependent on the earlier mentioned Poisson distributions.

In this paper we will combine the queueing theory with stochastic simulation to explain the behaviour of queueing systems, comparing the theoretical properties with simulations. The results in this paper are all fictive, that means it will not be based on real data, rather we will create input specifically to analyze the output as we implement certain changes. In chapter 2 we will explain the theory behind the used distributions, the central limit theorem, confidence intervals and the M/M/c queueing model. We will also calculate the theoretical properties of a queueing system. In chapter 3 we will digress on the methods used to create the stochastic simulation and the calculations used to create the results, which are given in chapter 4. We'll explain here the differences between the simulations and compare the results. In chapter 5 we'll discuss this paper and evaluate our research.

2 Theory

2.1 Distribution

This paper features Monte Carlo simulation, which means a random input is created and the output is then deterministically calculated. In the first cases input comes from a Poisson distribution, as Erlang[3] has given for both the arrival and the service rates. In later chapters we also examine other distributions and combinations of distributions. In this chapter we briefly discuss those distributions.

For the Poisson distribution, different methods can be used to simulate the arrival of customers and the rate of services. The Poisson distribution is discrete and has the following probability density function (pdf):

$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$

It has the unique property that it only uses one parameter λ that is both the mean and the variation. However in the simulation we are more interested in the time between arrivals and the service time. Fortunately we can derive another distribution for the inter arrival time of an Poisson process: the exponential distribution. This is a continuous distribution with the following pdf where we can use the same parameter as in Poisson distribution:

$f(x, \lambda) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$

and cumulative distribution function (CDF)

$F(x, \lambda) = e^{-\frac{x}{\lambda}}$.

Now given a random $u \sim U(0, 1)$ we can calculate $F^{-1}(u, \lambda) = -\lambda ln(u)$ which gives us a way to create random values in the domain $[0, \infty)$.

Other distributions include a longtail distribution, for which random numbers are picked from two distributions: both exponential distributions with a different mean. Finally it is also possible to pick from multiple distributions which do not all start at one. This makes it possible to describe a system where there are two different kinds of service requests which are inherently different in the time they require. We describe in chapter 3.1 in more detail how these distributions were constructed.

2.2 The Central Limit Theorem

One of the most important rules in stochastic simulation is the Central Limit Theorem. It says that the sample mean of every distribution is normally distributed. The stochastic $\bar{x}$ is the mean of the sample X, μ is the expected value of X and σ^2 is the variation of X. Now we know that $\bar{x}$ follows the distribution of $Y \sim N(\mu, \sigma^2)$.

2.3 Confidence interval

When estimating population means by simulation and the amount of samples taken is large, the central limit theorem applies, leading to $(\bar{x} - \theta)/(\sigma/\sqrt{n})$ being normally distributed [2]. Here $\bar{x}$ is the sample mean, θ is the population mean, σ is the standard deviation and n is the amount of samples taken. Therefore we have:

$$P(|\bar{x} - \theta| > c\sigma/\sqrt{n}) \approx P\bigg(|2(1 - \phi(c)| > c\bigg),$$

where ϕ is the standard normal distribution function. As usually σ is unknown, we assume that $\sigma \approx S$ which is the standard deviation of the sample. We can then find a confidence interval for the true population mean by using the standard the standard normal random variable of the desired confidence (Z_α). For a desired confidence interval of α this is

$$Z_\alpha = 2(\frac{1 - e^{-\frac{\alpha^2}{2}}}{\sqrt{2 * \pi}})$$

Using this standard normal random variable, the true mean lies within $\bar{x} \pm Z_{\frac{\alpha}{2}} * S * \sqrt{n}$ with a confidence of $100(1 - \alpha)\%$.

2.4 The M/M/c queueing model

The $M/M/c$ model is one of most common variation for queueing problems, of which many different variations exist (e.g. $M/D/1$ or $M/M/c/K$). We will start discussing the M/M/c model and describe variations at the end of the chapter.

$M/M/c$ also sometimes named $M/M/c/\infty$ is used to model customers arriving, waiting in a queue, waiting at a server and leaving the system. There are c parallel servers, where c is a positive integer. Both the arrival and service process follow a Poisson distribution with respectively parameters λ and μ. The system load (ρ) is defined as $\rho = \dfrac{\lambda}{c\mu}$. If the system load exceeds 1, the queue length will increase over time. All customers wait in one queue if all c servers are already busy serving customers, the first customer in the queue will be served by any of the c servers as soon as any server will be free from serving previous customer. If all servers are busy, this will give us the service rate μc. If $n < c$ servers are used, the service rate is μn. We can also calculate the utilization factor ρ, which is the mean time the system is used, by $\rho = \frac{\lambda}{\mu c}$.

The entire system state can be described by one parameter n, the number of customers in the system. The change that the system is in a certain state can be calculated as follows:

$$P_n = \begin{cases} \bigg(\frac{1}{n!}(\frac{\lambda}{\mu})^n\bigg)P_0 \text{ if } n < c \\ \bigg(\frac{1}{c!}(\frac{\lambda}{\mu})^c(\frac{\lambda}{\mu c})^{n-c}\bigg)P_0 \text{ if } n > c \end{cases}$$

This will give us a set of equations which we can solve because we know that $\Sigma_{i=o}^{\infty} P_i = 1$. We can also find a general expression.

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \left(\frac{1}{1 - \frac{\rho}{c}} \right) \right]^{-1}$$

With this discrete distribution it becomes possible to express all kinds of average properties of the system. So we can find the average the number of customers in the queue (L_q) in two ways:

$$L_q = \sum_{j=0}^{\infty} j P_{j+c} = \left(\frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} \right) P_0$$

Using this estimation, we can also find an expression for the average waiting time in the queue: $W_q = \frac{L_q}{\lambda}$ and the time spent in the system $W = W_q + \frac{1}{\mu}$. And the number of customers in the service system is calculated $L = \lambda W$.
In the next chapter we will put all those functions to use in an example.

2.5 A theoretical example

We will first derive the theoretical examples for a M/M/1 system. We will set the parameters $\lambda = 1$ and $\mu = \frac{10}{9}$. Since take the number of servers to be 1, we can first calculate $\rho = \frac{\lambda}{n\mu} = \frac{9}{10}$. This is a consciously chosen value, lower than one so the mean queue length will not increase over time, but closely near one so the system is active and there will be less time spent in an empty state.
Next we apply the formulas given in the last chapter.

$$P_0 = \left[\sum_{n=0}^{1-1} \frac{\frac{9}{10}^n}{n!} + \frac{\frac{9}{10}^1}{1!} \left(\frac{1}{1 - \frac{\frac{9}{10}}{1}} \right) \right]^{-1} = \frac{1}{10}$$

$$L_q = \left(\frac{\frac{9}{10}^{1+1}}{(0-1)!(c-\frac{9}{10})^2} \right) \frac{1}{10} = \frac{81}{10}$$

.

$$W_q = \frac{\frac{81}{10}}{1} = \frac{81}{10}, \quad W = \frac{81}{10} + \frac{1}{\frac{10}{9}} = 9, \quad L = 9$$

With these values found, we can compare these theoretical results with the results of our simulation in chapter 4. In table 1 are the values for a different number of servers and arrival rates. We keep the number of servers and arrival rate equal to each other, so even with an increasing number of servers the utilization rate ρ stays the same. This way we can compare different M/M/c systems.

λ&c	ρ	P_0	L_q	W_q	W	L
1	0.9	0.1000	8.1000	8.1000	9.0000	9.0000
2	0.9	0.3793	0.2285	0.1143	1.0142	2.0285
3	0.9	0.4035	0.0300	0.0100	0.9100	2.7300
4	0.9	0.4062	0.0042	0.0010	0.9010	3.6041
5	0.9	0.4065	0.0005	0.0001	0.9001	4.5005
8	0.9	0.4066	6.2e-7	7.7e-8	0.9000	7.2000

Table 1: The probability of zero customers in the system (P_0) increases with an increasing number of servers (c), while the number of customers (L) and waiting time in the queue (W_q) also dramatically decreases. The waiting time in the system (W) seems to converge to μ, since this is equal to the expected service time. The reasoning behind this effect is that queuing time usually is caused by large service time. With multiple servers, there is less chance that people will have to wait for a larger service since other desks will be available.

3 Methods

The model ran various simulations of customer server interaction, always keeping the number of servers equal to the arrival rate of new customers. We ran simulations with first-in first-out priority and shortest job first scheduling. For the arrival rate of new customers exponential, long-tail and bi-modal arrival rates were simulated.

3.1 Input Variation

The simulation ran multiple times using different inputs. The number of servers (c), arrival rate (λ) and system load (ρ) was equal in each simulation and was ran at 1,2,4 and 8. The service rate was picked from an exponential, long-tail and bi-modal distribution for which pseudo random numbers were generated using the random package in Python which generates numbers using Mersenne Twister [1]. The mean service time was set at 0.90 for the exponential distribution (figure 1). For the longtail distribution of the service duration 3/4 of the customers had a duration of $\mu/5$ and the other customers had a mean service duration of 3.06 both picked from an exponential distribution (figure2). For the bimodal distribution 9/10 of all customers had a service request taken from an exponential distribution with a mean of 0.90 and other customers had a service rate taken from a normal distribution with a mean of μ, a minimum of 0.5 and maximum of 1.3 (figure 3). All of those distributions have the same expected inter-arrival time and thus are good to compare.

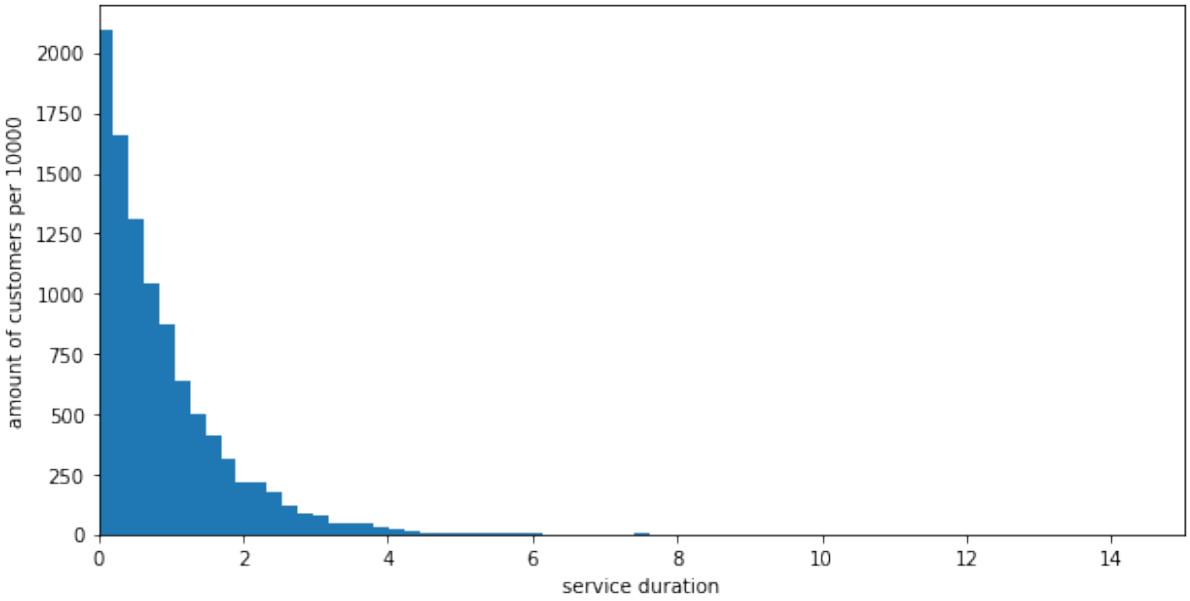

Figure 1: 10000 random drawings from an exponential distribution with $\lambda = 0.9$

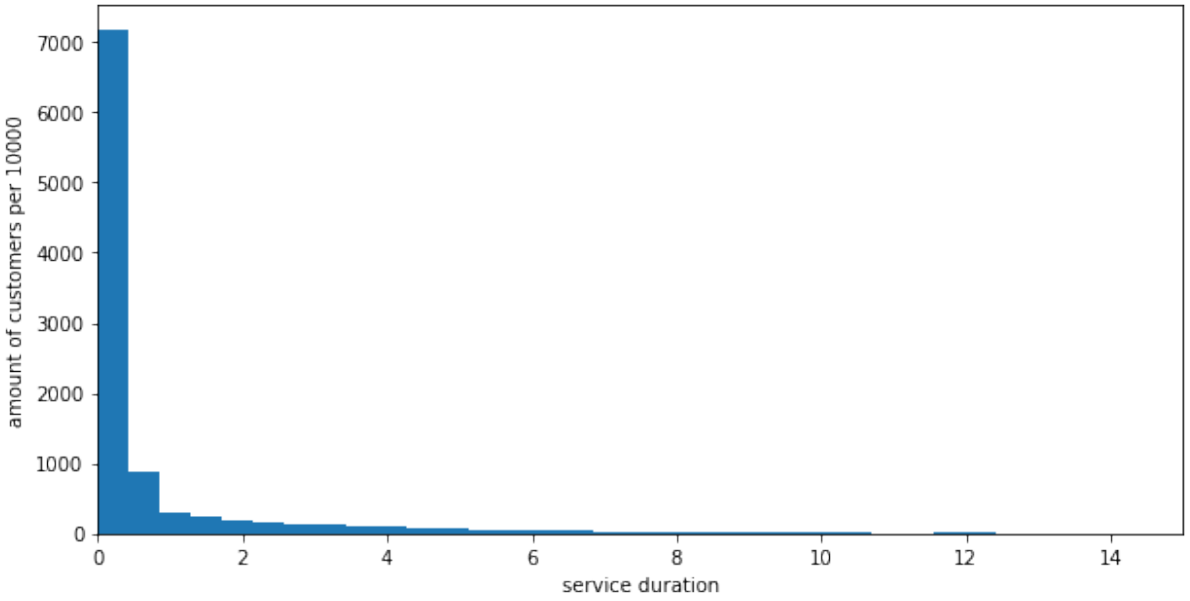

Figure 2: 10000 random drawings from two different exponential distributions with different means. In comparison with 1 we notice that the higher drawings appear more often but also more service times are close to 0.

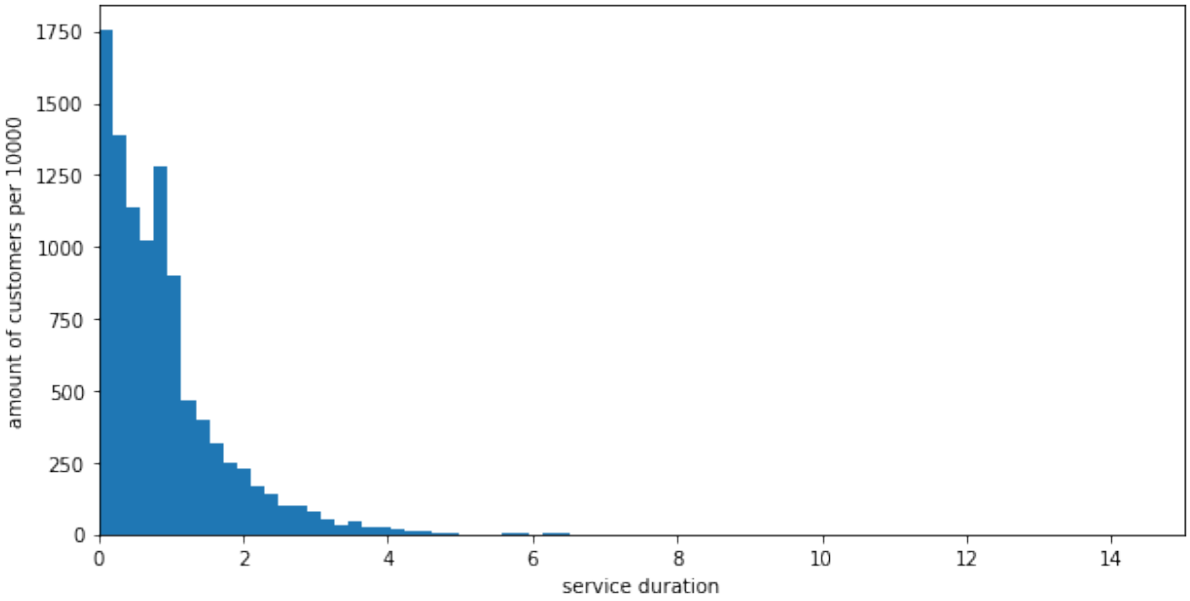

Figure 3: 10000 drawings from a bimodal distribution. Since this is partially a normal distribution, there is a local maximum near 0.9.

3.2 Simulation

The simulation ran 239000 customers for each of the input settings. This number was chosen to fulfill our need for accuracy yet with regard to computing time. For the exponential arrival rate distributions both first-in first-out customer service and shortest job first schedulingwas simulated.

3.3 Sampling Space

The waiting times of customers were analyzed in batches of 1000 consecutive customers. When we stated the simulation we first have a warming-up period, so we disregard the first 100 customers and their waiting times. Batches were made to be independent of each other in each simulation by allowing a window of 200 customers to sit in between each batch that won't be analyzed in the results. In total 200 batches were analyzed.

3.4 Statistics

Of each batch the mean and standard deviation was determined. We have different notation for this. The mean of the batch $< W >$ indicates the mean waiting time of the 1000 customers in one batch. The mean of all the analyzed customers in the simulation is the average of the $< W >$ of all the batches, called $<< W >>$ The standard deviation of the mean $\sigma(< W >)$

indicates how large the variation is of the mean between the batches. The standard deviation of the batches $\sigma(W)$ indicates how much variation in waiting time exists in batch. Then we take all those standard deviations of all batches, we can calculate the mean of the standard deviation $< \sigma(W) >$ standard deviation of the standard deviation $\sigma(\sigma(W))$ indicates how much variation is found between the variation within the batches. For first-in first-out serving with an exponential arrival rate it was determined whether the mean waiting time was significantly shorter for simulations with higher arrival rates and amount of customers using a t-test (see 4.2).

4 Results

4.1 Introduction

In this chapter we will analyze the results for different amounts of servers, different job scheduling methods and different arrival rate distributions.

4.2 Amount of servers

There are significantly shorter queues in first-in, first-out simulations were the arrival rate and amount of servers are increased at equal ratio. The mean and standard deviation of the mean of each sampling batch is indicated in figure 4. The within batch variation in waiting time also decreases when the amount of servers increases (figure 5). 2, 4 And 8 servers all had highly significant shorter service times than a system with 1 server, with t statistics and p values of respectively t stat 12.0020 & p 1.54806e-28, t stat 17.9869 & p 2.28437e-53 and t stat 21.3971 & p 3.69579e-68.

As explained before, we keep the utilization rate or system load ρ in most of the results stable so that the results of different servers are comparable. We see similar effects with different ρ's, but with lower waiting times and quicker converging effects. Since we want to study the waiting times, we remain with the respectively high ρ of 0.9 for the rest of the results, even though a higher ρ can cause more variance and thus will require a higher number of samples to result in statistically significant results.

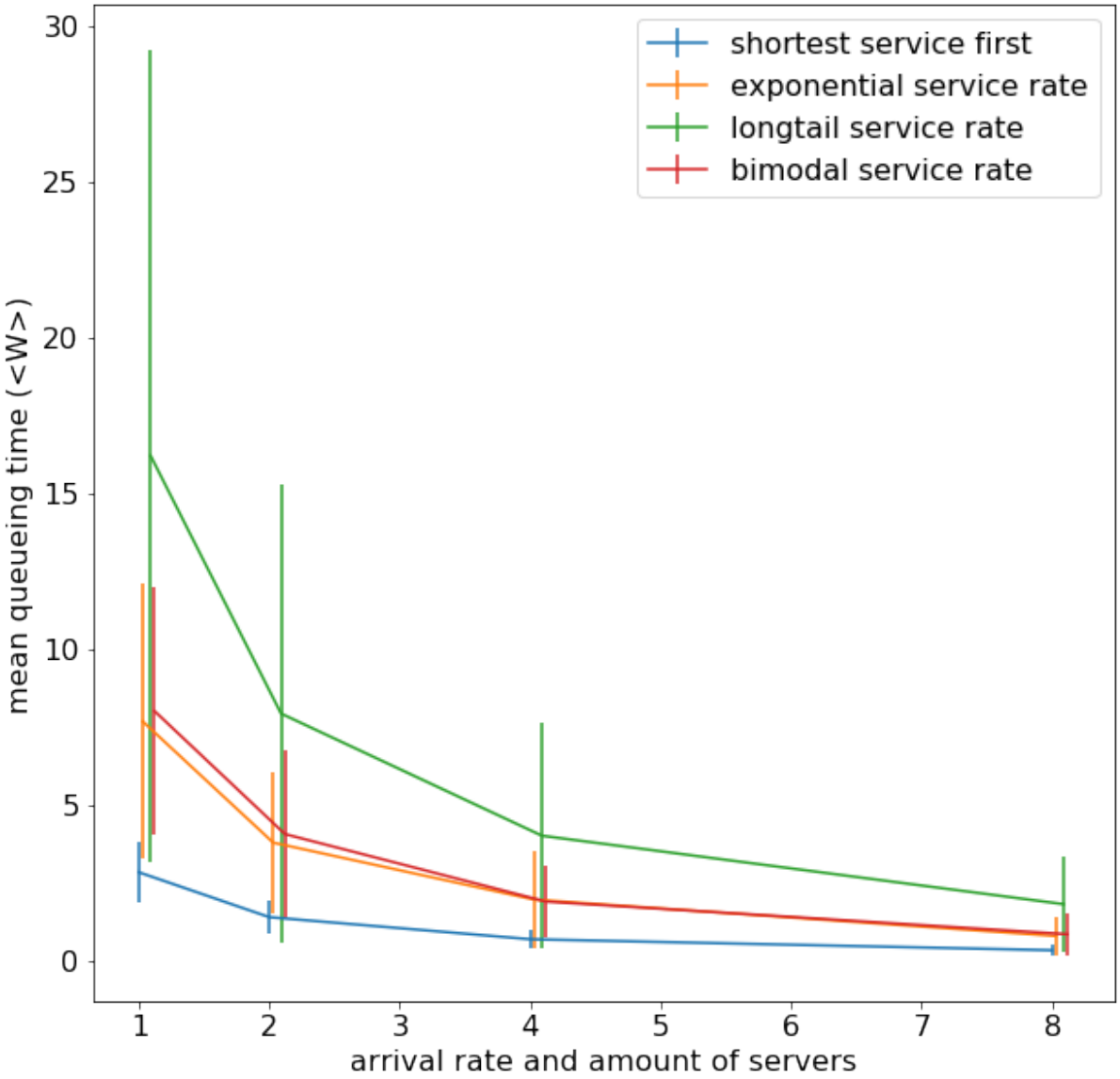

Figure 4: The mean and standard deviation of the mean waiting time per batch. We can see that with the utilization rate ceteris paribus the waiting time in queue L_q drops significantly. All queueing times seem to converge to 0, which makes sense as a system with an infinite amount of servers does not require a row.

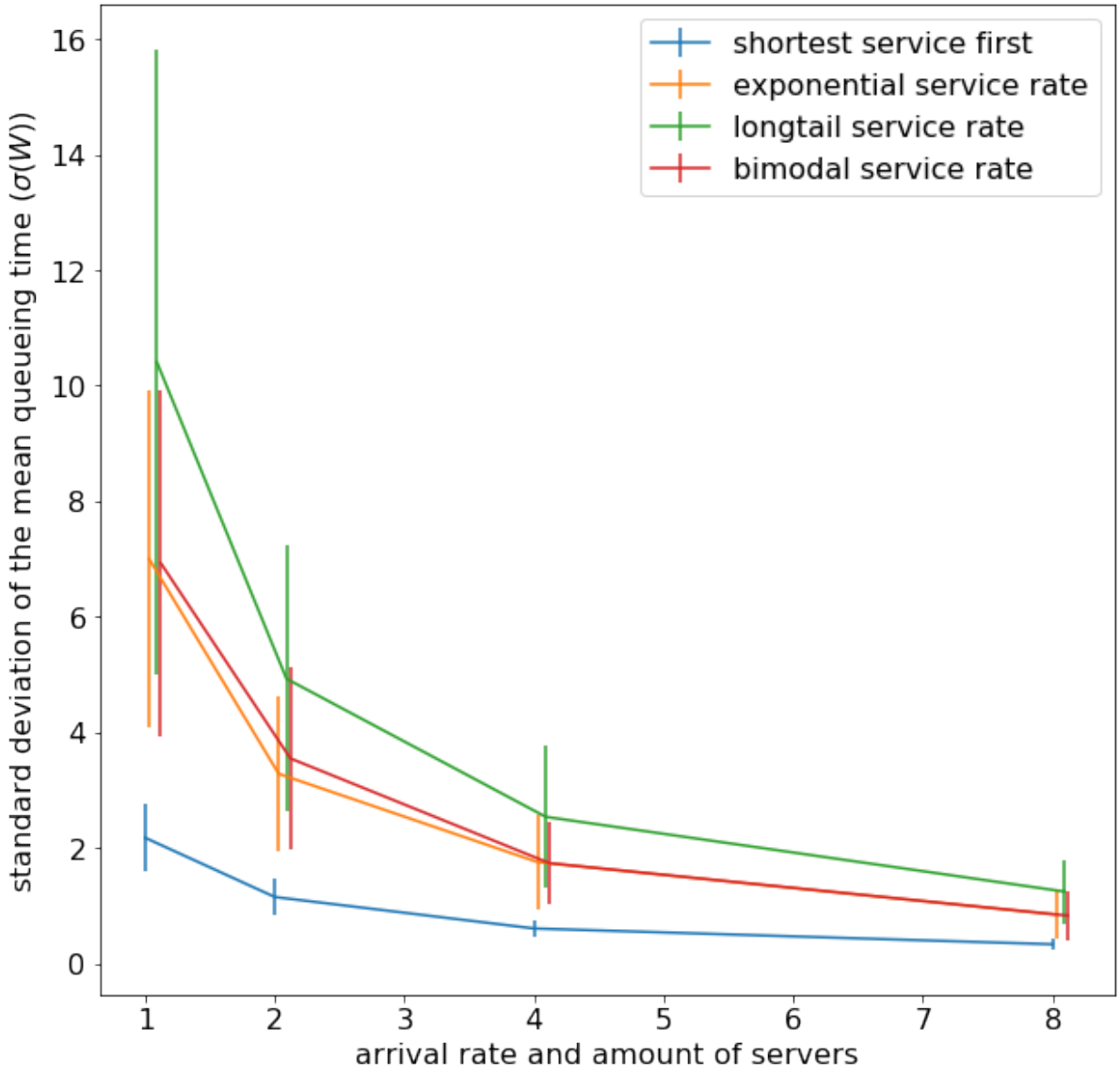

Figure 5: The mean standard deviation of the batches and standard deviation of the standard deviation of the batches plotted as a function of the amount of servers. We see that, as the number of servers grows, the $< \sigma(W) >$ drops.

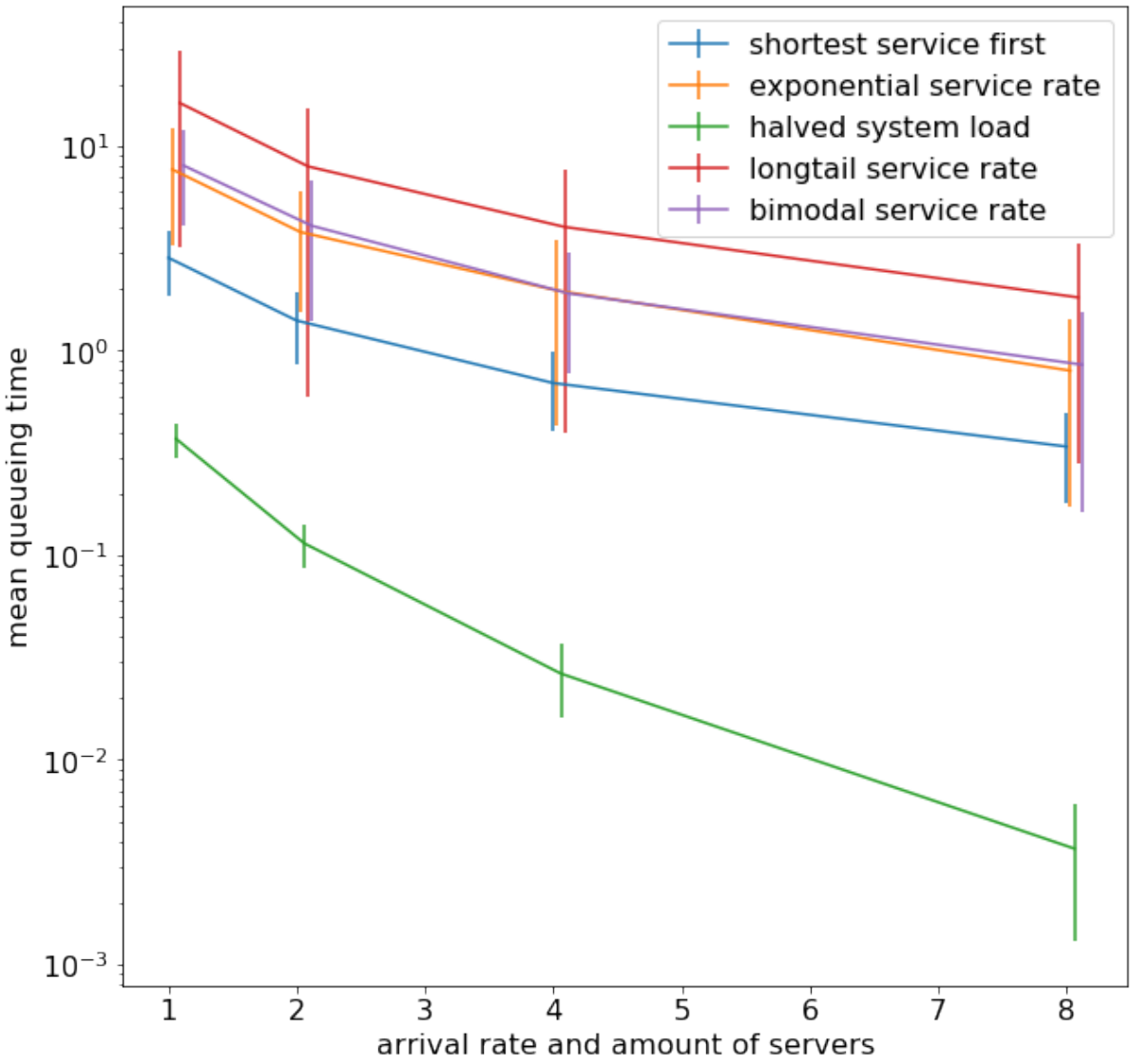

Figure 6: For smaller system loads (ρ) adding multiple servers has a larger negative effect on the mean queueing time. We see that for a ρ twice as small (created by doubling the service rate (μ)) adding more servers makes the mean queueing time converge to zero faster. The halved system load model has a service rate following a Poisson process with first-in, first-out servicing.

4.3 Shortest job first scheduling

The mean waiting time is lower for shortest job first scheduling (figure 7). Shortest job first scheduling decreased the mean waiting time for a one server system from 8.05 to 2.84. This about 2.5 fold decrease in mean queueing time is also visible in simulations with more servers (figure 8) for all types of distributions. Also the variation in waiting time within each batch is lower for shortest job first scheduling then for first-in, first-out (figure 5).

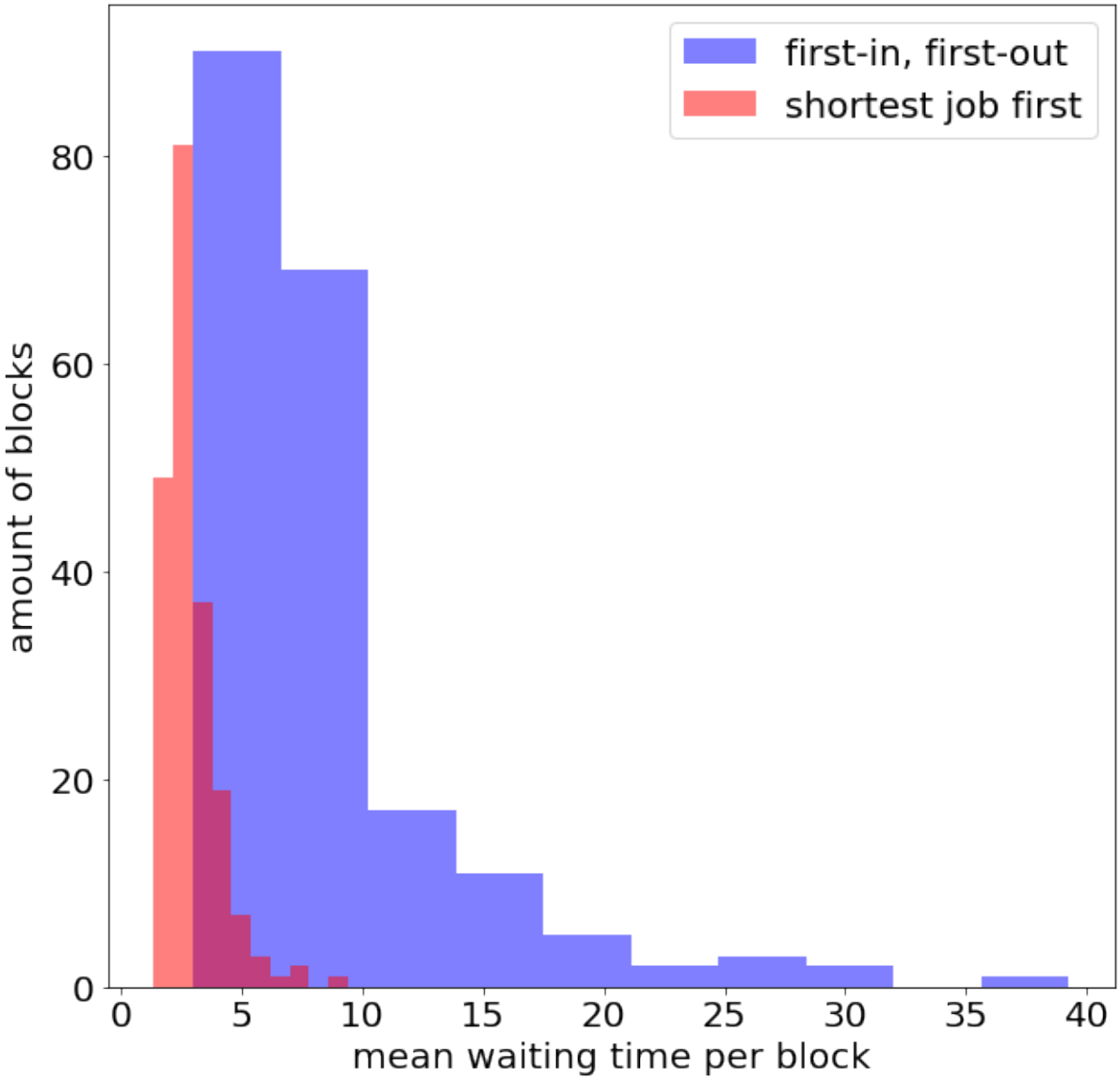

Figure 7: The mean waiting time for shortest job first service is almost always shorter than the mean of first-in, first-out servicing. Shown is the distribution of the mean queueing time of 1000 consecutive customers for the first-in, first-out servicing strategy and shortest job first servicing strategy. 200 batches of 1000 customers were taken per servicing strategy from a system with 1 server.

4.4 Other service rate distributions lead to different waiting times

When keeping the mean service time constant, a long tailed service duration distribution leads to longer mean waiting times (figure 4). There is also more variation in the mean waiting time when the service time has a long tail. This effect of increased waiting time and variation in waiting time is consistent also in simulations with more servers. The bimodal distribution also leads to slightly longer mean waiting times. However this effect is very limited. When correcting

the mean queueing time to 1 at servers = 1, all service rate show the same relative decrease in mean queueing time when the simulation is run with more servers (figure 8).

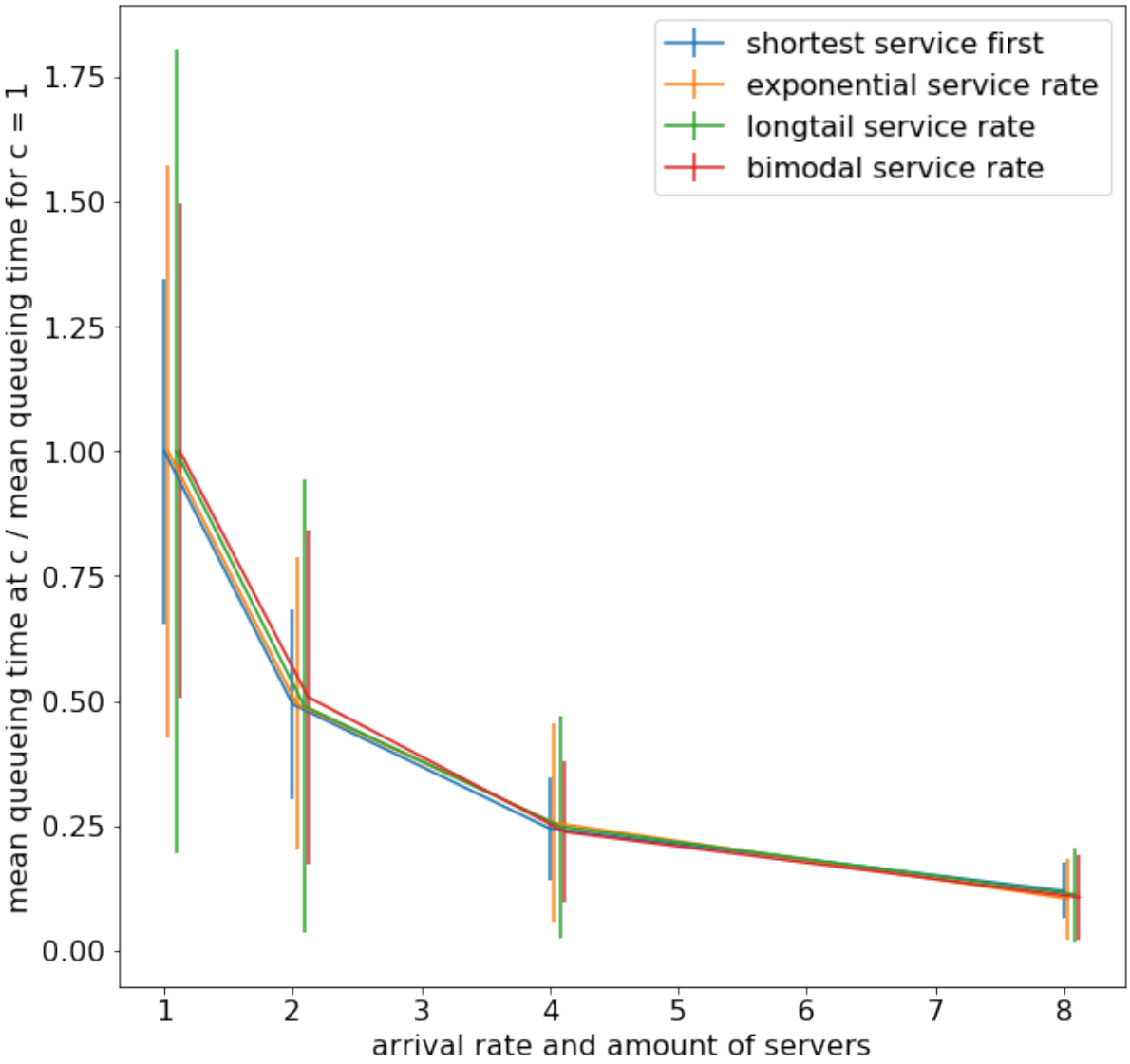

Figure 8: The relative decrease in mean queueing time is equal for all service time distributions and service preferences when the system load is kept constant. The mean queueing time divided by the mean queueing time in the system with 1 server

4.5 Appliance

As discussed in 2.3 we have the ability to create certain confidence intervals. We have no real goal settings given, but assuming we are interested in reducing the waiting time in the queue so that in 99% of the cases the waiting time is less than 5, we have a $\rho = 0.9$ as given in previous results and we start with one server, an exponential distribution and the 'first in first

out'-system, we can calculate an interval that is only upper bound as follows:

$$P\Big(\mu < \bar{x} + Z_{99} * S * \sqrt{n}\Big) = P\Big(\mu < 8,0545 + 2.3263 * 4,1028/\sqrt{200}\Big) = P\Big(\mu < 8.7294\Big) = 99\%$$

Now we can say that or μ is most likely not smaller than 5. Now this paper shows us two solutions, either increase the number of servers (as we have seen, even if the arrival rates grows evenly, this reduces the waiting times) or change to a priority-system, which gives us the following results:

$$P\Big(\mu < \bar{x} + Z_{99} * S * \sqrt{n}\Big) = P\Big(\mu < 2,8412 + 2.3263 * 0,9812/\sqrt{200}\Big) = P\Big(\mu < 3.0026\Big) = 99\%$$

This is an interesting conclusion: by changing the process the servers pick their customers, we can say with 99% certainty that the average waiting time is lower than 3.0026.

5 Discussion

In accordance with theory our simulations show that M/M/1 queues have significantly longer mean waiting times than M/M/c queues while the system load ρ stays equal. Decreasing waiting times for systems with more servers is greater for systems with a lower load. Shortest job first scheduling was found to decrease the mean waiting time and benefit from more servers in an equal ratio as first-in, first-out scheduling. Different service rate distributions can change the queueing times even though the mean service rate is kept constant.

The finding that the relative decrease in queueing time by an increase in the amount of servers is the same for all service rate distributions is supported by the theory, as that suggests only the mean capacity of the servers should matter. However, the decrease in waiting times found in M/M/c systems compared to M/M/1 systems is much lower than the decrease should be according to theory. If we apply queueing theory to real queues, it is more likely to find a free server when the amount of servers is larger.

In the retail industry customer services for a larger amount of customers thus have a scale advantage over systems with less customers. This means customers who visit smaller shops must either accept that they, on average, have to queue longer or smaller shops must employ more servers per customer to achieve the same mean queueing time.

Further research could implement a more realistic model of queues and experiment with more servicing strategies. The model could add a server when queues get too long, add rush hours or model customers queueing in one queue per server with customers choosing the shortest queue. It would also be interesting to investigate the effect of serving short jobs by a special server and to see whether longtail service time distributions profit more from shortest job first service as

opposed to first-in, first-out service then exponentially distributed service times.
In conclusion, when waiting in a queue, we should more often let people go in front of us if they have shorter jobs then us.

References

[1] Matsumoto, M., Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation (TOMACS), 8(1), 3-30.

[2] Sheldon M. Ross, Simulation, Fourth Edition, Academic Press, Inc., Orlando, FL, 2006

[3] Erlang. A.K (1909). The theory of probabilities and telephone conversations. Nyt Tidsskrift Matematika, B.20, pp.33-39

[4] Gaver. D.P (1959) Imbedded Markov chain analysis of a waiting line process, in continuous time. Annals of Mathematical Staistics,Vol.30, No.3, pp.698-720.