The Average Waiting Time in Queueing Systems

Investigating Effects of the Number of Servers and the Serving Time Distribution by Discrete Event Simulation

December 1, 2020

Isabelle Brakenhoff
isabelle.brakenhoff@student.uva.nl
11283912
Universiteit van Amsterdam

Jorrim Prins
jorrimprins.prins@student.uva.nl
11038934
Universiteit van Amsterdam

***Abstract*—We examine the effect of increasing the number of servers in a queueing system on the average waiting time. Initially we derive a theoretical relation between these variables, we thereafter use Discrete Event Simulation (DES) to perform experiments that verify this relation. The experiments provide empirical evidence for a significantly decreasing waiting time with the number of servers, under exponential distribution of the service times. Prioritised queueing is also shown to shorten waiting times compared to FIFO scheduling and subsequently, additional simulations are executed for other distributions of the service time. Similar conclusions about the increase of the number of servers can be made under these alternative assumptions. The parameter setting for the system load shows to be of major importance for the stabilisation of the system, additional observations are necessary for successful statistical analysis when the system load is increased towards 1.**

I. Introduction

Mathematical investigation of the behavior of waiting lines and queues is usually referred to as queueing theory. The theory enables mathematical analysis of an abundance of queue-related processes, such as arrival at a queue, waiting in a queue and eventually being processed to leave the queue. Performance measures of queueing systems can be derived and calculated and are often defined as the average waiting time in a queue, the expected number of items waiting or receiving service or the probability of encountering the system in a certain state. Simulation of these systems is generally done by Discrete Event Simulation (DES), providing a popular framework for experiments and analysis.

The mathematical set-up and methods for analysis in queuing theory arose as a result of telephone network expansion (Stordahl, 2007). It can help improve the efficiency of waiting line processing, e.g. in telephone centers of stores, but queueing theory also plays an important role in the examination of data processing. Efficient queueing can improve performance of machines that perform tasks, but also reduce the machine's power consumption while maintaining high quality. Other uses of queueing theory are presented by Bailey (1954) and Mendelson (1985) that explore the effective use of hospital resources and the quantification of waiting time costs respectively.

An important trade-off can be found between the decision to increase the number of servers or increase the service speed of an existing server. This research will explore the effect of an increased number of servers on average waiting times under varying model specifications. We hypothesise a higher number of servers in the system to result in lower waiting times, under equal settings for the service load. Results are presented under multiple settings for the service load, so the effect of this variable will also become visible. To provide insight into the model specification, its underlying assumptions and its consequences, an alternative to FIFO queueing and alternatives for the widely used exponential distribution of the service rate are investigated. Adan and Resing (2002) provide thorough mathematical analysis of queueing processes and help define a theoretical expression for the average waiting time under M/M/$\cdot$ processes. We will compare our estimations to these theoretical values to confirm the results of our simulations.

Section II provides a theoretical background for the queueing mechanisms used in this research, as well as proof for the effect of increasing the number of servers. The set up of our experiments is described in Section III and the corresponding results are presented in Section IV. We conclude in Section V, reflecting on our hypotheses and highlighting the principal findings of our research.

II. Theoretical background

Queueing theory generally studies the specification of a service centre with one or more servers, where customers (or items, but we will consider customers for clarity,) arrive, wait until a service slot is available for service and receive service to leave the system afterwards. The arrival rate λ and the capacity of the servers μ are usually specified as parameters to generate random variables, these subsequently determine the resulting model dynamics. When $\lambda > \mu$, the system encounters more arrivals than it can process and the waiting time for customers increases without stabilising. Analysis is

predominantly interesting in stable situations, so we specify the system load $\rho = \frac{\lambda}{c\mu}$, with c the number of servers, to be smaller than 1.

Kendall (1953)'s early work on queueing processes results in his widely known $A/S/c$ notation, respectively standing for the specifications of the arrival process, the service time distribution and the number of servers. The notation has later been extended to contain the queues capacity K, the number of customers to be served N and the queueing discipline D (Adan and Resing, 2002). The theoretical part of this research will leave these additional parameters out of the notation and assume K and N to be ∞ and D as First In First Out (FIFO), this is standard practice (Willig, 1999). The empirical segment will mostly consider the former three parameters as well, there will additionally be one model with Priority Service (PQ) as its queueing discipline.

While analysis of queueing systems can be done for various performance measures, we decide to focus on the average waiting time. This quantity can be determined by subtracting the average service time from the average system time:

$$E(W) = E(S) - \frac{1}{\mu} \tag{1}$$

Quantification of our performance measure can be done theoretically in a sagacious manner, using knowledge on the underlying service distribution, the Poisson Arrivals See Time Average (PASTA) property (Wolff, 1982) and Little's law (Little, 1961). We describe these aspects to subsequently give theoretical values for the average waiting time in M/M/1 and M/M/c systems, of which the comparison is our principal interest.

Exponentially distributed arrival and service times are commonly used and will initially be assumed in this research as well, as it has advantageous properties. The probability and cumulative distribution functions of the service time[1] can be respectively expressed as follows:

$$f(t) = \mu e^{-\mu t} \quad F(t) = 1 - e^{-\mu t} \tag{2}$$

The most important property of the exponential distribution is arguably its memorylessness: occurrence of a following event is independent from past or future information. It can be mathematically expressed for $\Delta t \geq 0$ as (Adan and Resing, 2002):

$$\begin{aligned} P(X > t + \Delta t \mid X > t) = \frac{P(X > t + \Delta t, X > t)}{P(X > t)} = \\ \frac{P(X > t + \Delta t)}{P(X > t)} = \frac{e^{-\mu(t+\Delta t)}}{e^{-\mu t}} = e^{-\mu \Delta t} \end{aligned} \tag{3}$$

Memorylessness of the exponential distribution gives name to its letter M in Kendall's notation. It also allows for the use of the PASTA property when the system is M/·/·, as exponentially distributed times between events are the result of a Poisson process by definition. The property states that on average, arriving customers find the system in the same situation that an outsider would observe from the system at a random point in time. More precisely, the fraction of arriving customers finding the system in some state A is exactly the same as the fraction of time the system is in state A.

Little's law (Little, 1961) can be used to alternate between the average number of customers in the system $E(L)$ and the average system time $E(S)$. It is given by:

$$E(L) = \lambda E(S) \tag{4}$$

with λ as specified before.

To derive an analytical expression of the average waiting time in an M/M/1 or M/M/c system we can rewrite the memoryless property for $\Delta t \to 0$:

$$P(X < t + \Delta t \mid X > t) = 1 - e^{-\mu \Delta t} = \mu \Delta t + o(\Delta t) \tag{5}$$

with $o(\Delta t)$ denoting a function for which $\frac{o(\Delta t)}{\Delta t} \to 0$ when $\Delta t \to 0$. If we denote the the probability that at time t there are n customers in the system as $p_n(t)$ we can simply use Equation 5 to find:

$$\begin{aligned} p_n(t + \Delta t) = \lambda \Delta t p_{n-1}(t) + (1 - (\lambda + \mu)\Delta t) p_n(t) \\ + \mu \Delta t p_{n+1}(t) + o(\Delta t) \end{aligned} \tag{6}$$

Resulting in the following differential equation for $\Delta t \to 0$:

$$p_n'(t) = \lambda p_{n-1}(t) - (\lambda + \mu) p_n(t) + \mu p_{n+1}(t) \tag{7}$$

Cohen and Browne (1982) show that $p_n'(t) \to 0$ and $p_n(t) \to p_n$ for $t \to \infty$. Setting the differential from Equation 7 equal to zero gives a second order recurrence relation that has solutions of the form $p_n = c_1 + c_2 \rho^n$. Using $\sum_{n=0}^{\infty} p_n = 1$ shows that c_1 should be zero, c_2 should be $1 - \rho$ and the relation looks as follows:

$$p_n = (1 - \rho)\rho^n \tag{8}$$

The equilibrium distribution apparently depends on the arrival and service rates through their ratio, the system load ρ (Adan and Resing, 2002).

We use this probability to find $E(L) = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho}$ by the convergence of a geometric series. The average waiting time can now be found by applying Little's law and using Equation 1:

$$\begin{aligned} E(W_{M/M/1}) = E(S) - \frac{1}{\mu} = \frac{E(L)}{\lambda} - \frac{1}{\mu} \\ = \frac{1/\mu}{1 - \rho} - \frac{1}{\mu} = \frac{\rho/\mu}{1 - \rho} \end{aligned} \tag{9}$$

An alternative to Equation 8 can be specified by a recursive relation as $p_n = \rho^n p_0$. We use this to evaluate the average waiting time in an M/M/c system with equal ρ, leading to a system load for individual servers of $\frac{\rho}{c}$. The next specifications all have ρ as the individual service load and multiply this by c to get the system service load. The recursive relation can now be specified in two ways, one representing the case when

[1] Similarly, the following proof holds for arrival times when μ is replaced by λ.

$n \leq c$ and one representing the case when $n > c$ and one will be placed in queue:

$$\begin{aligned} p_n &= \frac{(c\rho)^n}{n!} p_0, \quad n = 0, \ldots, c \\ p_{c+n} &= \rho^n p_c = \rho^n \frac{(c\rho)^c}{c!} p_0, \quad n = 0, 1, 2, \ldots \end{aligned} \tag{10}$$

p_0 then follows from combining $\sum_{n=0}^{\infty} p_n = 1$ and the two equalities above (Adan and Resing, 2002): $p_0 = \left(\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \cdot \frac{1}{1-\rho} \right)^{-1}$. We use this to specify the *delay probability* as the probability that a customer finds a completely occupied system and has to wait:

$$\begin{aligned} \Pi_W &= \sum_{n=0}^{\infty} p_{c+n} = \frac{p_c}{1-\rho} \\ &= \frac{(c\rho)^c}{c!} \left((1-\rho) \sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \right)^{-1} \end{aligned} \tag{11}$$

This helps to find the average number of customers in the queue $E(L^q)$, which finds $E(W_{M/M/c})$ by Little's law:

$$\begin{aligned} E\left(W_{M/M/c}\right) &= \frac{1}{\lambda} \sum_{n=0}^{\infty} n p_{c+n} \\ &= \frac{1}{\lambda} \frac{p_c}{1-\rho} \sum_{n=0}^{\infty} n(1-\rho)\rho^n \\ &= \frac{1}{\lambda} \Pi_W \cdot \frac{\rho}{1-\rho} \\ &= \Pi_W \cdot \frac{1/c\mu}{1-\rho} \end{aligned} \tag{12}$$

which is a more general expression of Equation 9 (they are equal for $c = 1$).

We ultimately want to show that the average waiting time for a M/M/c system is less than for a M/M/1 system, so we use $c = 2$ and use this in the equation above:

$$\begin{aligned} E(W_{M/M/2}) &= \frac{2\rho^2}{1+\rho} \cdot \frac{1/\mu}{1-\rho} \cdot \frac{1}{2} = \frac{\rho}{1+\rho} \cdot \frac{\rho/\mu}{1-\rho} \\ &< \frac{\rho/\mu}{1-\rho} = E(W_{M/M/1}) \end{aligned} \tag{13}$$

where the final step holds because $\frac{\rho}{1+\rho} < 1$.

This is an intuitive result, as a single server will be heavily affected by instances with large service time. Waiting times will automatically increase for all the arriving customers after this instance, whereas this effect is way smaller in a setting where other servers can still provide service to the arriving customers.

III. Experiments

Although we have a theoretical quantification of the average waiting time for M/M/· systems, we want to provide empirical examples of this measure and extend our research by exploring different service distributions and queueing disciplines. We specify experiments under exponential arrival and service rates with $\mu = 1$ and varying ρ. We start by exploring the relation between the average waiting time and number of servers, which we have theoretically described under exponential service assumptions in Section II. We explore the distributions and test the effect of increasing the number of servers from 1 to 2, 4 and 8. In addition, we provide insight into the effect of ρ on the statistical properties of the models and on hypothesis tests we perform.

The shortened model notation M/M/· that we have used until now implies a FIFO queue, but we want to explore the effect of different queueing disciplines as well. PQ, as defined in Section II, orders the queue by the service length of customers in this queue and provides service to the shortest service lengths first. The experiments we perform for the increasing number of servers, as specified in the paragraph above, are also performed on a M/M/1 model with PQ as its queueing discipline (M/M/1/∞/∞/PQ).

Furthermore, we want to relax the assumption of an exponentially distributed service rate and explore models with alternative distributions. Again, similar experiments are performed as before, looking at the difference between $c = 1$ and $c = 2$. We do this for a hyperexponential, Erlang-5 and deterministic service rate. We specify a hyperexponetial distribution where 25% of customers has a service rate that is $\frac{\mu}{5}$, the rest has the regular service rate of μ. Our Erlang-5 specification is simply the average of 5 exponentially distributed service rates with μ and the deterministic set-up always has μ as its service time. The difference in average waiting time between a model with 1 or 2 servers can now be analysed, as well as the difference between the various distributions we specify.

Implementation of the DES structure of our experiments is done by using the SimPy package. We explore ρ values between 0.5 and 0.9 to see its effect and simulate until $t = 20000$. Every simulation is repeated 300 times to allow statistical properties to be quantified by the Law of Large Numbers.

IV. Results

Figure 1 presents the distribution of the average waiting time for various model specifications under $\rho = 0.7$. We decide to use this value for ρ as it is the value closest to 1 that still satisfies the hypothesis of normality from the omnibus test by d'Agostino (1971) for all model specifications. A more detailed analysis of normality is discussed later. Our most basic M/M/1 model clearly shows its distribution to be located in the most rightward location, representing the highest values for $E(W)$ as expected. Adding additional servers so c becomes 2, 4 and 8 results in waiting times lowering towards zero.

We show clear statistical evidence for a decreasing $E(W)$ with the number of servers by calculating t-statistics, comparing an M/M/1 model with the other specifications for various ρ values in Figure 2. The black line, representing the critical value of a one-sided test with $\alpha = 0.01$, shows to be very low compared to the calculated values of the statistics. This is true for all ρ and all comparisons. Test statistics increase when an M/M/1 model is compared with an M/M/c model with higher c. To be more precise, the critical value for a t-statistic with

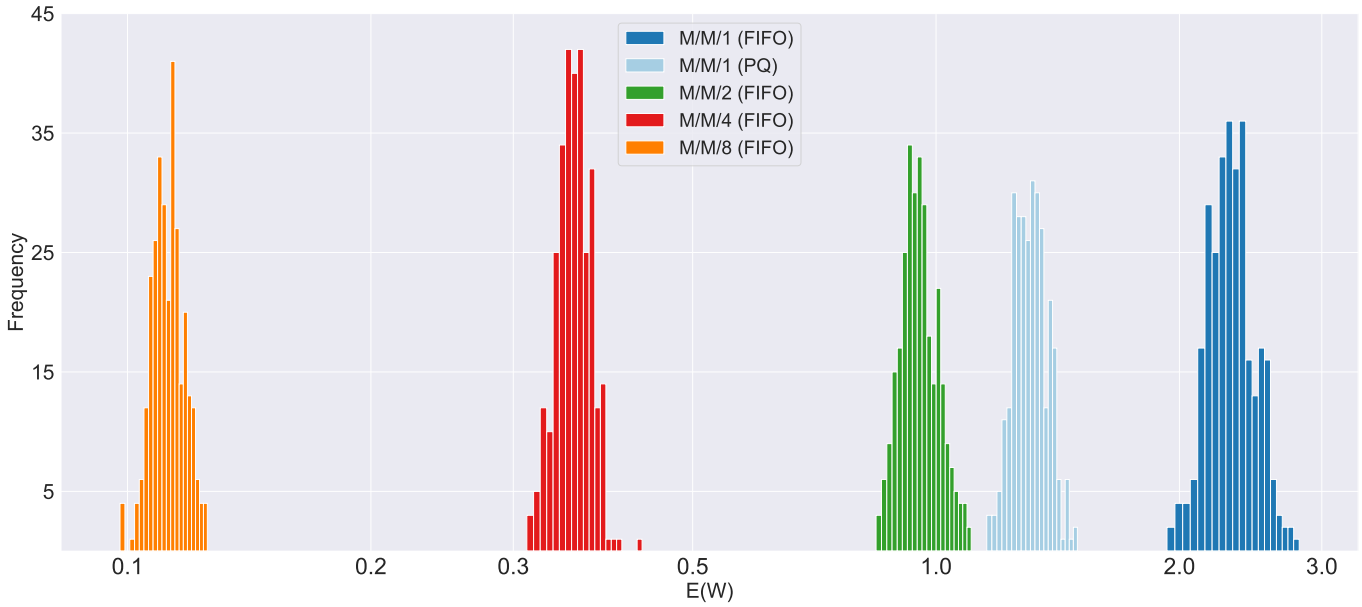

Fig. 1. Distribution of the average waiting time for $\mu = 1$, $\rho = 0.7$ and $N = 300$. Comparing a varying number of servers in the system and a varying specification of the queueing discipline. A decreasing average waiting time can clearly be observed when the number of servers increases, the same thing holds for switching to priority scheduling. Note that a logarithmic scale is used on the x-axis and that distributions further left appear wider than they are.

$(300 - 1 =)299$ degrees of freedom is equal to 1.97 while the t-statistics for M/M/1 compared to M/M/2, M/M/4 and M/M/8 respectively are 145.31, 218.95 and 247.61 for $\rho = 0.7$. For higher ρ, the test statistics show to decrease towards the critical value, but even at $\rho = 0.9$ they vastly exceed 1.97.

We previously noted that not all of our models satisfy normality for $\rho > 0.7$. Testing with $\alpha = 0.01$ only provides statistical evidence for normality of M/M/4 under $\rho = 0.9$ for example, the rest of the normality hypotheses are rejected (p-values are very low). Lowering ρ delivers an increasing number of models that can not be rejected to be normally distributed, and at $\rho = 0.7$ we find a situation where all our specified models satisfy normality. Therefore we present the actual statistical values of our estimates for $\rho = 0.7$ like we did for the t-statistics above. This rejection of normality is likely the cause of a simulation time in our experiments that is set too low to stabilise the queueing model, we however did not have the computational resources to increase this parameter and find normally distributed models for higher ρ.

Apart from the increasing number of servers in the model, we also look at alternative queueing discipline PQ as described in II. We specify this model with 1 server and this alternative discipline, it is tested to be normally distributed under the test by d'Agostino (1971) with a p-value of 0.23. Figure 1 shows an obvious decrease in $E(W)$ when the alternative queueing discipline is used, although the effect is less than adding an additional server into the model. The t-statistic for comparing the M/M/1 model to an M/M/1 (PQ) model is 105.21 and a significant decrease in $E(W)$ is therefore found at $\alpha = 0.01$.

Our experiments regarding the underlying service time distributions are done under similar parameter settings, as the

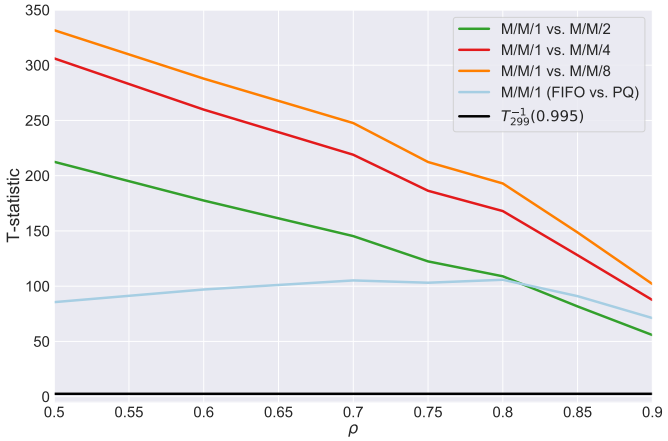

Fig. 2. Values of the t-statistic for $\mu = 1$, varying ρ and $N = 300$. Black line at the bottom presents the critical value for a one-sided test at $\alpha = 0.01$ as we want to investigate if $E(W_{M/M/c}) < E(W_{M/M/1})$. Note that for values $\rho > 0.7$, normality can not be shown for all models and t-statistics might be off.

test conclusions about normality are similar to the conclusions before. Histograms of $E(W)$ under varying service time distributions can be found in Figure 3. The M/M/1 specification again shows to perform the worse in terms of $E(W)$, but it is remarkable that our M/M/· specifications perform worse than all the other specifications for both values of c. A hyperexponential distribution as specified in III, as well as the Erlang-5 distribution, clearly decrease the average waiting time for $c = 1$ and provide fairly close results. A deterministic set-up of the service time shows an even lower $E(W)$. It can also be observed from Figure 3 that all models with $c = 1$

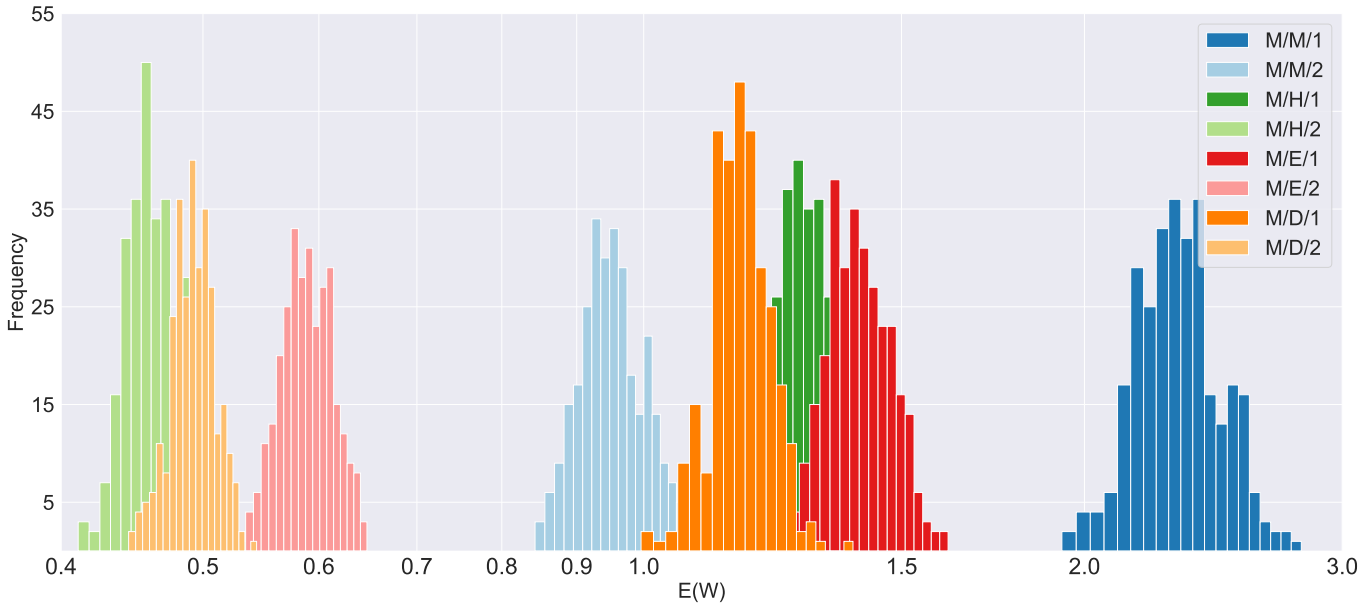

Fig. 3. Distribution of the average waiting time for $\mu = 1$, $\rho = 0.7$ and $N = 300$. Comparing varying service time distributions in the system and models with 1 or 2 servers. Of course, a decreasing average waiting time can clearly be observed when the number of servers becomes 2 for the same distribution. Furthermore, the hyperexponential and Erlang-5 distributions result in lower waiting times than the exponential distributions. Deterministic service times cause the lowest $E(W)$ for $c = 1$. Note that a logarithmic scale is used on the x-axis and that distributions further left appear wider than they are.

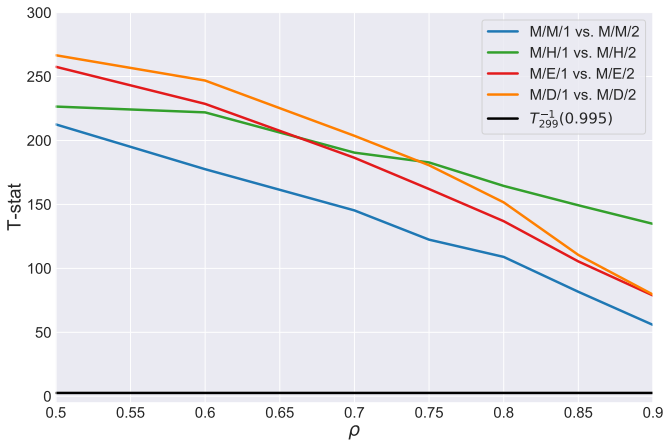

Fig. 4. Values of the t-statistic for $\mu = 1$, varying ρ and $N = 300$. Black line at the bottom presents the critical value for a one-sided test at $\alpha = 0.01$ as we want to investigate if $E(W_{M/\cdot/2}) < E(W_{M/\cdot/1})$. Note that for values $\rho > 0.7$, normality can not be shown for all models and t-statistics might be off.

have higher average waiting times than all models for $c = 2$, independent of the underlying distribution used. For $c = 2$, the M/M/2 model performs worst again, but the lowest $E(W)$ is now reached by a model with hyperexponentially distributed service times. The latter observation might be true because our specification of the hyperexponential distribution has a lower expectation than the original exponential distribution, of which we use the expectation in our deterministic specification.

Figure 4 again shows a relatively low critical value compared to the t-statistics of our specified distributions. We compare models with 1 and 2 servers under the four distributional assumptions an find very high t-statistics. We can therefore conclude that an extra server will significantly decrease the average waiting time of the system. Like before, we see a decreasing t-statistic with increasing ρ for all distributional assumptions, meaning that a higher system load lowers that statistical evidence for $E(W_{M/\cdot/2}) < E(W_{M/\cdot/1})$.

At $\rho = 0.7$, the point at which we can still statistically assume normality, t-statistics for the four distributional assumptions are 145.31, 190.48, 186.46 and 203.64 for exponential, hyperexponetial, Erlang-5 and deterministic specifications respectively. All values are vastly higher than the critical value of 1.97. We also perform some additional tests for peaks that appear to be close in Figure 3. We test for the difference between the hyperexponential and Erlang-5 and the difference between hyperexponential and deterministic estimates under $c = 1$. The test statistics for these tests are 21.18 and 22.47 respectively. The difference between hyperexponential and deterministic estimates under $c = 2$ has a t-statistic of 16.97. All values again under $\rho = 0.7$.

To present a concise and clear overview of our estimations and their similarities with theoretical values from Adan and Resing (2002), we provide Table I. The table contains estimations for all our model specifications under $\mu = 1$ and $\rho = 0.7$ and theoretical values that can be calculated through Equation 12.

The estimates in Table I show that the average waiting times from our simulations are equal to the theoretical values that have been derived from Equation 12. It shows that the simulation has stabilised and the limiting behaviour of the process

is reached. We see that for the other models, with increased c, priority scheduling or different service time distribution, none of the confidence intervals include the theoretical value for $E(W_{M/M/c})$ and they are significantly different.

TABLE I
AVERAGE WAITING TIMES AND 95 PERCENT CONFIDENCE INTERVALS, COMPARED TO THE THEORETICAL VALUE UNDER M/M/·. $\mu = 1$, $\rho = 0.7$ AND $N = 300$

c	1	2	4	8
M/M/c	2.33 (2.02, 2.64)	0.96 (0.85, 1.06)	0.36 (0.32, 0.39)	0.11 (0.10, 0.12)
M/M/c (PQ)	1.31 (1.19, 1.44)			
M/H/c	1.28 (1.14, 1.43)	0.47 (0.42, 0.51)		
M/E/c	1.41 (1.26, 1.55)	0.59 (0.54, 0.64)		
M/D/c	1.17 (1.06, 1.28)	0.49 (0.46, 0.53)		
Theory	2.33	0.96	0.36	0.11

V. CONCLUSION

We estimate the average waiting time and its distribution in a queueing process for varying model specifications. We initially use an exponential service time distribution and an increasing number of servers in the system. Subsequently, we include priority scheduling and different service time distributions to explore the effects on the average waiting time. These effects are all monitored under an increasing value for the service load ρ.

Simulations determine the average waiting time for our models to be normally distributed up to a ρ value of 0.7, as the normality test by d'Agostino (1971) can not be rejected. Although experiments become more interesting when we further increase ρ to 1, the current computational power does not allow for increments in computational time. This would theoretically lead to enough statistical evidence to accept normally distributed waiting times.

Under ρ0.7, we find strong statistical evidence that increasing the number of servers leads to a lower average waiting time. Similar statistical evidence is found for the implementation of priority scheduling as opposed to FIFO scheduling in a M/M/1 model. Not only will an increased number of servers cause lower waiting times under exponentially distributed service time, it is also shown to hold under hyperexponential, Erlang-5 and deterministic service time.

Furthermore, the t-statistics of our inequality tests for increasing c show to decrease with ρ. This implies that further increasing ρ leads to the need of additional number of observations to be able to statistically proof decreasing waiting times (or other performance measures).

The theoretical quantification of the average waiting time for specific μ, ρ and c in M/M/c processes from Adan and Resing (2002) matches the estimations from our simulations. However, we have made the early conclusion that increasing ρ above 0.7 leads to a lack of evidence for normally distributed waiting times and statistical analysis of the simulation becomes infeasible. Ideally, ρ is increased towards 1 further (e.g. increased to 0.95 or 0.995) to make the test statistics closer to the critical value. This would make the comparison between models more interesting, as there would likely be equality tests between models that could not be rejected. The different behaviour between models with varying specifications could be presented more clearly.

In an optimal situation, our available computational power would be higher, we would increase the simulation time for our models and find evidence for normality even under $\rho = 0.995$. Further research is encouraged to make this increased simulation time feasible and improve on our experiments. Additionally, the theoretical derivations we provide for M/M/c models could be extended to show similar waiting time expressions under different distributional assumptions. This way, all estimations could be compared to their theoretical value.

REFERENCES

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

Bailey, N. T. (1954). Queueing for medical care. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 3(3):137–145.

Cohen, J. W. and Browne, A. (1982). *The single server queue*, volume 8. North-Holland Amsterdam.

d'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, 58(2):341–348.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, pages 338–354.

Little, J. D. (1961). A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387.

Mendelson, H. (1985). Pricing computer services: queueing effects. *Communications of the ACM*, 28(3):312–321.

Stordahl, K. (2007). The history behind the probability theory and the queuing theory. *Telektronikk*, 103(2):123.

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

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operations Research*, 30(2):223–231.