Mean waiting time estimation of steady state queues

Coen Lenting *(11332484, coen.lenting@student.uva.nl)*
Sjoerd Terpstra *(11251980, sjoerd.terpstra@student.uva.nl)*

1 December 2020

Abstract

In this paper we looked at a special type of Discrete Event Simulation; queuing theory. We examined the mean waiting time of a M/M/n queue with exponential (with shortest task priority and first in first out principle) and hyperexponential service time and $M/D/n$ simple queues under varying load characteristics. It was mathematically proven that the $M/M/1$ queue has a higher expected waiting time that the $M/M/2$ queue. The $M/M/1$ queue with shortest task priority was found to have the lowest mean waiting time among the different systems. $M/D/n$ was found to be the second most efficient system in terms of mean waiting time, followed by a $M/M/n$ queue with exponentially distributed service time and the hyperexponential distributed service time.

1 Introduction

There exists multiple frameworks for the modelling of systems. They can be divided in continuous time, discrete time and discrete event models. We will look at the latter; Discrete Event Simulation (DES). This framework is largely used in complex systems or in network protocol modelling (Larocque and Lipoff, 1996). It is also used in for example health-care services (Fone et al., 2003; Katsaliaki and Mustafee, 2011; Navonil Mustafee, 2010).

In real life, one can imagine many situations where queues play an important role. Think for example about customers waiting in line in grocery stores, peoples sending requests to servers with limited capacity or people waiting in line for the bus. Queuing theory is the field that describes the mathematical process of customers in a queue and their distributions over servers with limited capacity, where they can perform their desired task for a certain length of time (Erlang, 1909).

In this paper we look at a DES of simple queues under varying load for single- and multiple-server systems. We will also give a mathematical derivation and explanation of the mean waiting times of $M/M/1$ versus $M/M/n$ queues. Furthermore, we will look at a deterministic and a hyperexponential distribution for the service time.

2 Theory & Methods

2.1 Discrete Event Simulation

DES is a wide-spread way of modeling complex systems. Instead of basing a system on discrete or continuous time intervals, whereby for each subsystem at each time interval their state is calculated, the DES system only updates the system when an event happens.

2.2 Queuing Theory

The queuing theory was created by Erlang while working at a telephone company (Erlang, 1909). His aim was to reduce the average waiting time of costumers to enter the telephone service by optimizing the amount of operators and stations used at specific times. He did this by describing a model of a queue and the process of people from that queue moving to available servers.

In most real life situations the filling of queues and the time people use for services are non-deterministic. Thereby it is common in queuing theory to make use of stochasticity in determining the influx of arrivals and their service time.

In 1953 Kendall introduced a standard notation for queues, referred to as the Kendall's notation (Kendall, 1953). For a simple queue model it is described by $A/S/n$, with A the arrival process, S the service time distribution and n the number of servers. We will use M to denote a Markovian or memoryless distribution and D for a deterministic system. The arrival rate in the system is denoted by λ.

These systems will always reach a steady state, where the size of the queue is roughly constant, for each initial state of a configuration. The time it takes to reach the steady state from the initial state is called the "burn-in" period. How long this period is, depends on the configuration of the queuing system.

2.3 Pseudo-random number generator

Almost all queuing systems make use of random numbers, except the fully deterministic D/D/n systems. Therefore, we employ a pseudo-random number generator (PRNG). We make use of the CG64 (PCG XSL RR 128/64) PRNG developed by (O'Neill, 2014) This PRNG is incorporated in the random module of the Python NumPy library. PCG stands for Permuted Congruential Generator, which is based on a Linear Congruential Generator (LCG). Permutation functions are applied on the output from the LCG to improve the randomness and statistics of the numbers. For the interested readers, O'Neill (2014) gives an excellent overview of this and similar PRNG's and their (dis)advantages.

2.4 SimPy

In order to simulate the DES queue, the Python SimPy library was used. This is an intuitive process-based DES framework based on standard Python. Processes in SimPy are simulated by generator functions from Python and they can be used to model active elements like arrivals in the case of a DES queue.

2.5 $M/M/n$ vs $M/M/1$ average waiting time

Suppose there are two systems $M/M/n$ and $M/M/1$ which are denoted by α and β respectively. Both have a system load $\rho = \lambda_i/(n_i\mu)$, with λ_i the arrival rate of the system, n_i the amount of servers of the system and μ the capacity of each single server. β has an arrival rate n-fold lower than α, thus $\lambda_\beta = \lambda_\alpha/n$ with $n = n_\alpha$ and then it is trivial that $\rho = \rho_\alpha = \rho_\beta$ (since $n_\beta = 1$).

We will derive a constraint on ρ for which we are sure that for $n > 1$ the average waiting time in the α queue is lower than the average waiting time in the β queue. We will also show that for $n = 2$, the average waiting time of the α queue is lower than the average waiting time in the β queue for all ρ. For this we will first show that

$$E(W_\alpha) < E(W_\beta), \tag{1}$$

with $E(W)$ the expected value of the average waiting time of the system. According to Little's law (Little, 1961) the expected value of the mean waiting time W is given by

$$E(W) = \frac{\lambda}{E(L_q)}, \tag{2}$$

with L_q the length of the queue.

The $E(W_\alpha)$ of a $M/M/n$ system, proposed by Willig, 1999, is given by

$$E(W_\alpha) = \Pi_W \cdot \frac{1}{(1-\rho)n\mu}. \tag{3}$$

The M/M/1 system is a special case where $n = 1$ in Equation 3:

$$E(W_\beta) = \frac{\rho}{(1-\rho)\mu}. \tag{4}$$

Substitute Equation 3 and Equation 4 in Equation 1:

$$\Pi_W \cdot \frac{1}{(1-\rho)n\mu} < \frac{\rho}{(1-\rho)\mu} \tag{5}$$

$$\Pi_W < n\rho \tag{6}$$

Π_W is given by (Willig, 1999):

$$\Pi_W = \frac{\frac{(n\rho)^n}{n!}}{\left((1-\rho)\sum_{m=0}^{n-1}\left[\frac{(n\rho)^m}{m!}\right] + \frac{(n\rho)^n}{n!}\right)}. \tag{7}$$

We are interested in the domain that $0 \leq \rho \leq 1$. In the case that $\rho = 1$ we see that the first term in the denominator disappears in Equation 7

$$\Pi_W = \frac{\frac{(n\rho)^n}{n!}}{\frac{(n\rho)^n}{n!}} = 1. \tag{8}$$

We can also observe that for $0 \leq \rho < 1$, that $(1-\rho)$ is always positive. The summation

$$\sum_{m=0}^{n-1}\left[\frac{(n\rho)^m}{m!}\right] \tag{9}$$

is also always positive. Combining these observations we see that

$$(1-\rho)\sum_{m=0}^{n-1}\left[\frac{(n\rho)^m}{m!}\right] > 0 \tag{10}$$

$$(1-\rho)\sum_{m=0}^{n-1}\left[\frac{(n\rho)^m}{m!}\right] + \frac{(n\rho)^n}{n!} > \frac{(n\rho)^n}{n!} \tag{11}$$

Then it follows from Equation 7 that

$$\Pi_W \leq 1 \text{ for } 0 \leq \rho \leq 1, n \geq 1. \tag{12}$$

So, looking at Equation 6, we see that the average waiting time of α will always be lower than β if the following constraint is obeyed

$$\rho > \frac{1}{n}. \tag{13}$$

There might be n for which all values of ρ result in $E(W_\alpha) < E(W_\beta)$, the above mentioned constraint is only for all proven values of ρ for all $n > 1$ as we derived in this section.

In the specific case that $n = 2$, we get

$$\Pi_W = \frac{2\rho^2}{4\rho^2 + \rho + 1}. \tag{14}$$

Then Equation 6 becomes

$$\frac{2\rho^2}{4\rho^2+\rho+1} < 2\rho, \tag{15}$$

$$\frac{\rho}{4\rho^2+\rho+1} < 1. \tag{16}$$

The left hand side of Equation 16 is smaller than 1, thus the average waiting time of a $M/M/2$ system is always lower than for a $M/M/1$ system with the same load characteristics.

This mathematical derivation can be supplemented by a more intuitive approach. Lets consider the system to consist of a queue and one or multiple servers. In a single server system, whenever the system is full, new arrivals have to enter the queue, which increases the mean waiting time.However, when, with the same load characteristics, an arrival occurs in a four server system, there are five possible states, e.g. all occupied and 4/3/2/1 free servers. Even though the average load is the same, the stochasticity in the system allows for more states where a new arrival can immediately join the server in comparison to the one server system. Based on this, one can reason that a single server system will have a higher mean waiting time than a multiple server system with the same load characteristics.

3 Results

3.1 Experiments

In order to compute a reliable and accurate estimate of the average waiting time in a queuing system, good simulation parameters had to be selected. This was done by considering the distribution of the mean waiting time as a function of the number of arrivals $\max_p$ for $\rho = 0.95$ and $n = 4$. In Figure 1 it can be observed that when the number of arrivals is set too low, the system has not yet reached the steady state, or just entered it. This can be seen by examining the shape of the distribution for the maximum number of arrivals in the figure. For a $\max_p$ of 1000, the distribution of the mean waiting time is very dependent on the random numbers drawn from the exponential distribution, resulting in a transient distribution. For a $\max_p$ of 10000 the distribution is already much more stable. However, for a $\max_p$ of 20000, the distribution is relatively well defined and stable, which is desired as it is the steady state of the system we are interested in.

As seen in Figure 1 the "burn-in" period results in a large variability of the mean waiting time, so it is important to determine at which point a system enters the steady state to lower the variability in the average mean waiting time of the steady state. It is difficult to numerically verify exactly when the system enters the steady state i.e. when $\max_p$ is high enough. The system appeared to be in a steady state around $\max_p > 6000$. This value was selected by considering the least well defined system, $\rho = 0.95$. This value can then be used to truncate the first 6000 elements of the data, thereby removing the waiting times during the "burn-in" period. The maximum number of arrivals for the simulation was set at 20000. The mean waiting time resulting from a simulation is thus the average of the waiting times of the last 14000 customers.

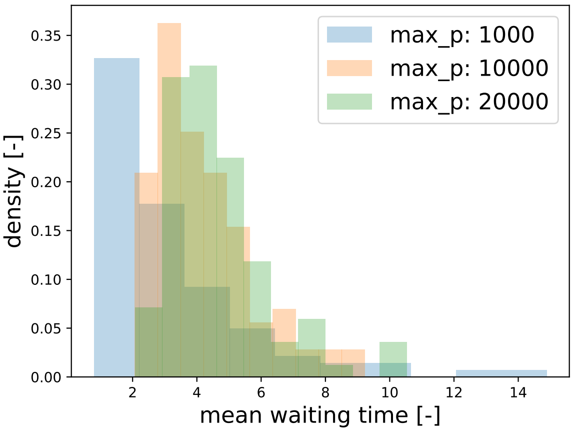

Figure 1: The distribution of the mean waiting time for different values $\max_p$ over 100 simulations, computed for $\rho = 0.95$.

3.1.1 $M/M/n$ vs $M/M/1$ waiting time numerical verification

To verify Equation 16 $M/M/2$ and $M/M/1$ queues were simulated. Also a $M/M/4$ queue was simulated to check our expectations for the average waiting times as described in section 2.4. All were simulated for a range of ρ between 0.1 and 0.95. They were executed with the same load characteristics, with 20000 number of arrivals and $\mu = 1$. The first 6000 data points of the waiting times were truncated and the experiment was repeated 100 times.

3.1.2 $M/M/1$ shortest job priority scheduling

The effect of giving tasks with a lower time to complete a higher priority was tested by implementing this prioritisation in a $M/M/1$ queue, with identical parameters as the other simulations. The difference in mean waiting time can then reliably be compared to the $M/M/1$ queue without priority.

3.1.3 Experiments with different service rate distributions

The effect of different service rate distributions on the mean waiting time was tested by considering a $M/D/n$ queue, with a deterministic task time of 1, and a $M/M/n$ queue, where the service rate distribution was a hyperexponential distribution. The hyperexponential distribution was created by generating a service time by an exponential distribution with mean $\frac{1}{2}$ 75% of the time and mean $\frac{5}{2}$ 25% of the time.

3.2 Results

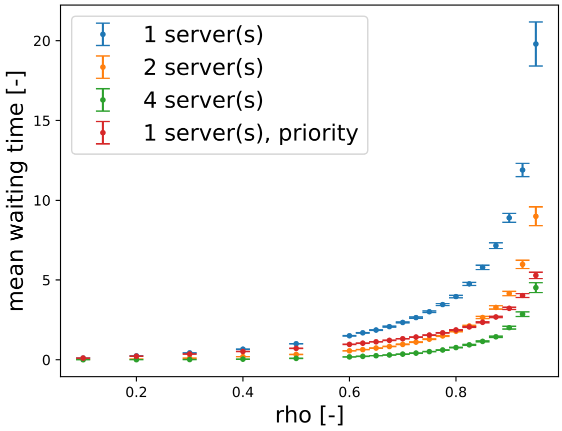

Figure 2: The mean waiting time with the 95% confidence interval as a function of ρ for the $M/M/n$ system and the $M/M/1$ with short task priority. The simulations were repeated for 100 times until $\max_p = 20000$.

The results from the simulations on the $M/M/n$ queues combined with the $M/M/1$ with priority are shown in Figure 2. Here it can be observed that for a system with the same load characteristics and more servers, the average waiting time decreases, corresponding to the theory. The 95% confidence interval of the mean waiting time increases for an increasing value of ρ, becoming more prominent from $\rho = 0.85$ onward. Another interesting finding in the figure is that by prioritising the shortest tasks, the average waiting time for one server is decreased significantly.

The results from the different service rate distributions are shown in the Figures 3 and 4 for the deterministic distribution and the hyperexponential distribution. It can be observed that mean waiting times for the deterministic distribution are relatively shorter compared to the normal $M/M/n$ simulations, which can be explained by the deterministic nature of the task time as no tasks significantly longer than the mean have to be completed, thereby holding up the queue.

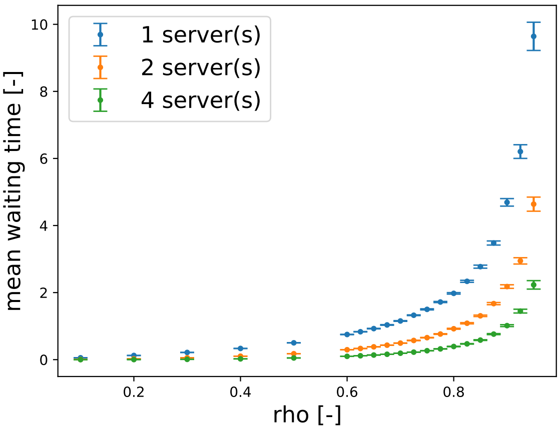

Figure 3: The mean waiting time with the 95% confidence interval as a function of ρ for the $M/D/n$ system. The simulations were repeated for 100 times with $\max_p = 20000$.

The waiting times for the hyperexponential distribution however are much longer than the standard $M/M/n$ simulations. The hyperexponential distribution allows more frequently longer tasks than the normal exponential distribution. Therefore it is more likely that the queue with the hyperexponential service time will be held up by these tasks, increasing the mean waiting time of this system compared to the system with an exponential service distribution.

4 Conclusion

Taking the results above into consideration it can be concluded that four variations of a DES queue were successfully simulated: a normal $M/M/n$ queue, a $M/M/1$ queue with priority, a $M/D/1$ queue and a $M/M/n$ queue with a hyperexponential task time distribution.

The analytical proof stating that the mean waiting time for a $M/M/2$ queue is always shorter than for a $M/M/1$ queue with an equal server load was numerically confirmed by simulating the systems.

With an equal server load, the $M/M/1$ queue was found to be the most efficient in terms of mean waiting time among the variations with 1 server. For $n \in \{2, 4\}$ the most efficient variation was the queue with a deterministic task time, followed by the normal $M/M/n$ queue with a roughly two times longer mean waiting time. The most inefficient queue was the $M/M/n$ queue with a hyperexponential task time distribution, having the longest mean waiting time among all server values.

These results are practically relevant when designing systems with queue-like structures. If it is possible, it would be the most efficient to give all the costumers

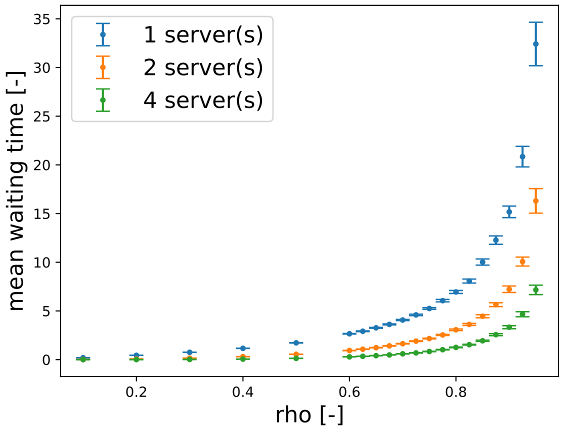

Figure 4: The mean waiting time with the 95% confidence interval as a function of ρ for the $M/M/n$ system with a hyperexponential distribution. The simulations were repeated for 100 times with $\max_p = 20000$.

a fixed task time. However, when the system can not function with only fixed task times, the best alternative would be a queue with priority if one is trying to keep the average waiting time as low as possible. An important downside of this variation is that a costumer with a long task time, could be forced to wait indefinitely when costumers with a shorter task time keep entering the queue and being given priority.

For further research it might be interesting to look in the $M/M/n$ priority queues with more than 1 server and their relation with the standard $M/M/n$ queues.

References

Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik*, *20*(B), 33. https://web.archive.org/web/20111001212934/http://oldwww.com.dtu.dk/teletraffic/erlangbook/pps131-137.pdf

Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G., & Palmer, S. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health*, *25*(4), 325–335. https://doi.org/10.1093/pubmed/fdg075

Katsaliaki, K., & Mustafee, N. (2011). Applications of simulation within the healthcare context. *Journal of the Operational Research Society*, *62*(8), 1431–1451. https://doi.org/10.1057/jors.2010.20

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*, *24*(3), 338–354. http://www.jstor.org/stable/2236285

Larocque, G. R., & Lipoff, S. J. (1996). Application of discrete event simulation to network protocol modeling. *Proceedings of ICUPC - 5th International Conference on Universal Personal Communications*, *2*, 508–512 vol.2. https://doi.org/10.1109/ICUPC.1996.562625

Little, J. D. C. (1961). A proof for the queuing formula: $L = \lambda W$. *Operations Research*, *9*(3), 383–387. https://doi.org/10.1287/opre.9.3.383

Navonil Mustafee, S. J. T., Korina Katsaliaki. (2010). Profiling literature in healthcare simulation. *SIMULATION*, *86*(8-9), 543–558. https://doi.org/10.1177/0037549709359090

O'Neill, M. E. (2014). *Pcg: A family of simple fast space-efficient statistically good algorithms for random number generation* (technical report HMC-CS-2014-0905). Harvey Mudd College. Claremont, CA.

Willig, A. (1999). A short introduction to queueing theory.