Universiteit van Amsterdam

Stochastic Simulation

Assignment 2: Discrete Event Simulation - Simple Queues

Caterina Buranelli (13169408)[a]
Ignas Krikštaponis (13250868)[b]

[a]caterina.buranelli@student.uva.nl
[b]ignas.krikstaponis@student.uva.nl

1 Dec. 2020

Abstract

In queueing models, costumers enter the system, wait for service, get served and exit the system. This setting varies accordingly to three main items: the waiting and service time distribution and the number of servers. In this paper the waiting time distribution was always assumed exponential, and we investigated how the average waiting time changed for different service time distributions (exponential, deterministic or hyper-exponential), for different number of servers (1, 2 and 4) and for two priority rules (FIFO and SPFS). Those discussed are Discrete Event systems and lay on the fundamental assumption that they are stable between the events. The number of costumers needed in order to achieve stability was investigated for different values of server utilization; for server utilization approaching 1, the system needed more costumers to achieve stability. It was found that more servers decreased the average waiting time; the best model with lowest average waiting time had deterministic service times and 4 servers, while for 1 server models the exponential model with SPSF priority was found to have the lowest average waiting time.

1 Introduction

The queuing theory is concerned with studying situations where customers must wait for service. The quality of the queuing system can be measured in many different ways but it is usually done by evaluating the time customers had to wait for service or the percentage of customers that left before being served. It is obvious that a high number of servers is beneficial to improve the quality of the service, however this is not possible as it is economically unsustainable. Exploring different queuing methods and establishing the relationship between number of resources and utilisation is the main purpose of the queuing theory. Queues are a common social phenomena found everywhere from shops to airports. Furthermore, manufacturing processes such as assembly lines can be modelled as queues (Wah Chun [2014]). Understanding this phenomena is paramount for appropriately managing available resources/workforce and therefore sustainability of many businesses. The basic queue model can be seen in Figure 1 below.

The research objectives of this paper are as follows:

- investigate the relationship between resource utilisation and system's stability
- experimentally confirm the theoretical mean waiting times of the system.

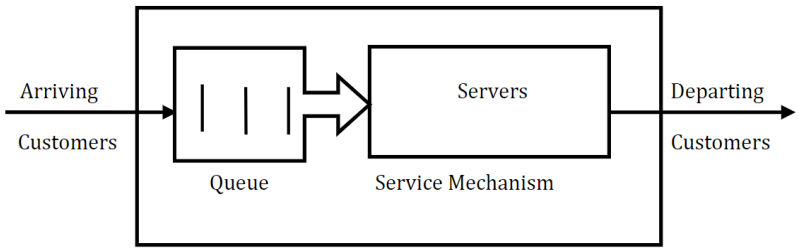


Figure 1: Basic queue model (Wah Chun [2014]).

2 Methods

In Discrete Event systems the focus is on the time at which the event of interest happens, while the behaviour of the system between one event and the next one is not modelled; that is justifiable when it is assumed that the system is steady between the events.

One method for modelling discrete events is by using the queuing theory: costumers enter the system, they get served and they exit accordingly to the server capacity and to the service times. From what was just stated, the main characteristics of a queuing model are highlighted: the distribution of the arrivals, the distribution of the serving time and the service capacity.

The Kendall's notation (Kendall [1951]) for queuing models encapsulates the three main characteristics with the following abbreviation: $a/b/n$, where a refers to the arrival distribution, b to the service time distribution and n to the number of servers. Usually the arrival rate is indicated with λ, while the processor capacity is indicated with μ.

An interesting parameter is the server utilization which is found as

$$\rho = \frac{\text{arrival rate} \times \text{mean service time}}{\text{number of servers}} = \frac{\lambda}{n\mu}$$

In order to prevent the queue to grow to infinity, ρ is required to be smaller than 1.

Arrival and service distribution For arrival and service times several distributions can be assumed; in this paper we will investigate the following: exponential (M), deterministic (D) and hyper-exponential (H).

For the $M/M/n$ model, $\frac{1}{\lambda}$ is the inter-arrival mean, while $\frac{1}{\mu}$ is the service time mean.

The $M/D/n$ model differs from the $M/M/n$ one because the service time is constant for all the costumers; it will be set equal to the average service time.

Finally, a $M/H/n$ model will be investigated, where 75% of the service times have an exponential distribution with mean $\frac{1}{\mu_1}$, while the remaining 25% follows an exponential distribution with mean $\frac{1}{\mu_2}$. μ_1 and μ_2 were chosen so the mean service time is the same as for M/M/n and M/D/n.

2.1 Waiting time

The average waiting time, denoted by $E(W)$, is the parameter taken into consideration while comparing queue models with different number of servers; under the assumption of stability, the theoretical value can be calculated by applying the Little's law and the PASTA propriety. (Adan and Resing [2001])

Little's law yields a relation between the number of customers L, the sojourn time S and λ:

$$E(L) = \lambda E(S)$$

When applied only to the queue length L^q, the following relation is found: $E(L^q) = E(W)$.

For Poisson distributed arrival times, the fraction of arrivals finding the system in a state S is exactly the same as the fraction of time the system is in state S. That is known as the PASTA property: Poisson Arrivals See Time Averages.

In this paper mathematical derivations will not be deepened further, but it can be demonstrated (Levy [1983]) that for a $M/M/n$ queue model, by using the two laws explained above, the average waiting is

$$E(W) = \Pi_W \cdot \frac{1}{1-\rho} \cdot \frac{1}{n\mu}$$

where

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

is the so called *delay probability*. For $n = 1$, the basic model, $E(W_1) = \frac{\rho}{(1-\rho)\mu}$, while for $n = 2$, $E(W_2) = \frac{\rho^2}{(1-\rho^2)\mu}$; as $\rho < 0$, $E(W_1) > E(W_2)$. This result will be investigated in the experiments.

For the $M/D/1$ model, it is easy to calculate the average waiting time, which is found as (Levy [1983]):

$$E(W) = \frac{\rho}{2\mu(1-\rho)}$$

From the formulation above, it is clear that when the service time is deterministic, the mean waiting time is halved than when the service time is exponentially distributed.

Priorities There are different types of queue, depending on the costumer priority; in this paper two variations will be taken into consideration. The two rules will be compared for a $M/M/1$ system only, to see which one is better in order to have the smaller average waiting time.

When the priority of being served is given by the order of arriving, the queue is said to follow the *First In, First Out* (FIFO) rule. That is used as the default rule in this paper.

The *Short Processing Time First* (SPTF) rule gives a higher priority to customers with a lower service time needed. It can be demonstrated (Kleinrock [1975]) that the average waiting time in this situation is calculated as:

$$E(W) = \rho \int_{x=0}^{\infty} \frac{e^{-x}dx}{(1-\rho(1-e^{-x}-xe^{-x}))^2}$$

In this paper, experiments are carried out to check which of the two priorities leads to a smaller mean waiting time.

2.2 Stability

Running simulations on the computer entails that the experiments start with an empty system, so the first costumers will be highly dependent on the random sample drawn from the distribution for the arrival times. Therefore only after a certain amount of time (or costumers), the assumption of a stable system can be realistic. This is why one of the main concern of this paper will be to investigate after how many arrivals the system can be considered stable, as only under that assumption specific properties of the system are valid.

In the paper, the behavior in reaching the stability is examined for four values of ρ, and for the following queue models: $M/M/1$, $M/D/1$ and $M/H/1$. Once found the best number of costumers after which the system is stable, it will be possible to study more in deep the models and compare model with different number of servers in terms of average waiting time.

3 Results

In order to find the steady state in $M/M/1$, $M/D/1$ and $M/H/1$, 50 simulations were performed for a number of costumer c equals to (100, 500, 1000, 5000, 10 000, 50 000, 100 000, 250 000, 500 000). The average waiting time ($\bar{W}$) based on the 50 simulations was calculated for each c. This procedure was performed for ρ equal to

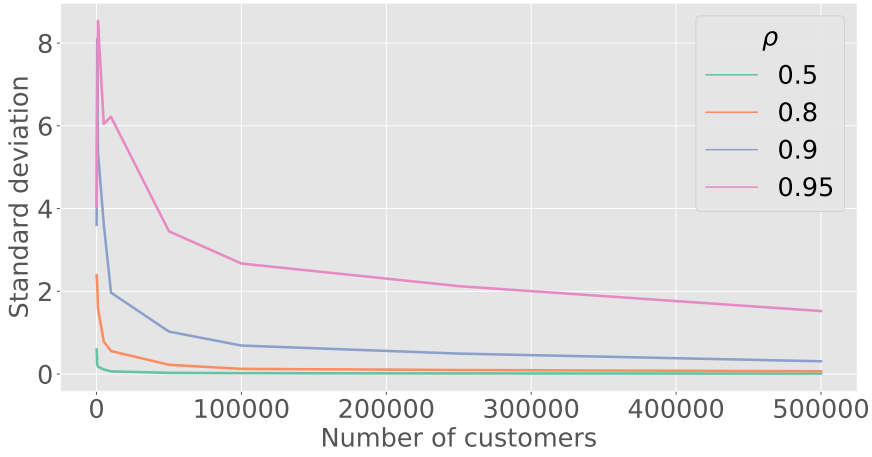


Figure 2: Progression of the standard deviation of $\bar{W}$ in an $M/M/1$ model, for different values of ρ and c; $\mu = 1/1.25$.

(0.5, 0.8, 0.9, 0.95), in order to investigate how ρ influences the number of costumers needed to reach stability. To quantify this influence we chose the difference between the final ($c = 500\ 000$) standard deviation measures ($\gamma = SD_{\rho=0.95} - SD_{\rho=0.9}$) as a benchmark. The best c was found accordingly to the standard deviation of the average waiting time. The behaviour of the system for $\rho = 0.95$ was used as final decision. This is justified because for the comparisons between same models, but different number of servers, ρ is set to that value. Once chosen c, the normality of the distribution was checked both through the Shapiro-Wilks test and graphically with the fitting of the data (in Appendix B the graphs for each combination of number of servers and ρ are reported).

For the chosen c and for $\rho = 0.95$, 100 simulations were performed for each model, for a number of servers equal to 1, 2 and 4. Within the same model, the average waiting time for different number of servers was compared by observing the 95% confidence interval.

3.1 $M/M/n$

In Figure 2 the progression of the standard deviation of the average waiting time is plotted. As the number of costumers increases, the standard deviation decreases; for higher values of ρ, the system requires exponentially more costumers to reach stability. For instance, for $\rho = 0.95$, the SD is around 1.5 at $c = 500\ 000$, while for

Table 1: Results for different number of servers in $M/M/\cdot$; $\rho = 0.95; \mu = 1/1.25$.

Model	$\bar{W}$	$E(W)$	CI-95%	p-normality
M/M/1	23.583	23.750	(23.305, 23.861)	0.285
M/M/2	11.538	11.571	(11.408, 11.667)	0.609
M/M/4	5.566	5.571	(5.507, 5.624)	0.164
M/M/1-SPTF	6.564	6.580	(6.515, 6.612)	0.325

$\rho = 0.9$ that level of precision is reached at $c = 10\ 000$.

c is chosen to be 500 000, as for that value the standard deviation is considerably smaller than the other c. Here $\gamma = 1.2$. Moreover, the goodness of the results is given by the fact that the distribution is Normal for the chosen c and $n = 1, 2, 4$ (Table 1, Appendix A).

Comparisons with different number of servers For higher number of servers, the average waiting time decreases. As the confidence intervals of the sample mean waiting time never overlap among different servers, the sample mean are significantly different; the result is established also from the t-test, which has a p-value near to 0 for all the pairs (p-value of the t-test $<<$ 0.001 for 1-2, 2-4, 1-4 (Appendix C)). This empirical result confirms the theoretical one, explained in Section 2.1.

In Table 1 all the details are reported, including the p-value of the Shapiro-Wilks test.

$M/M/1$ **with SPTF priority** In Figure 3 the progression of the standard deviation of the average waiting time is plotted, in the case of a $M/M/1$ model where costumers have higher priority for lower service time needed. In the system, oscillations weaken faster than FIFO M/M/1: even for $\rho = 0.95$, the standard deviation is 0.5 at $c = 100\ 000$; moreover for $c = 500\ 000$ the discrepancy between standard deviations for $\rho = 0.95$ and the other values of ρ is small: as an example, $\gamma = 0.14$.

From Table 1 it is clear that a different organization of priorities significantly (p-value of the t-test $<<$ 0.001 (Appendix D)) reduces the average waiting time: for SPTF priority, $\bar{W}$ is ≈ 3.5 times smaller than FIFO, for one server system and $\rho = 0.95$.

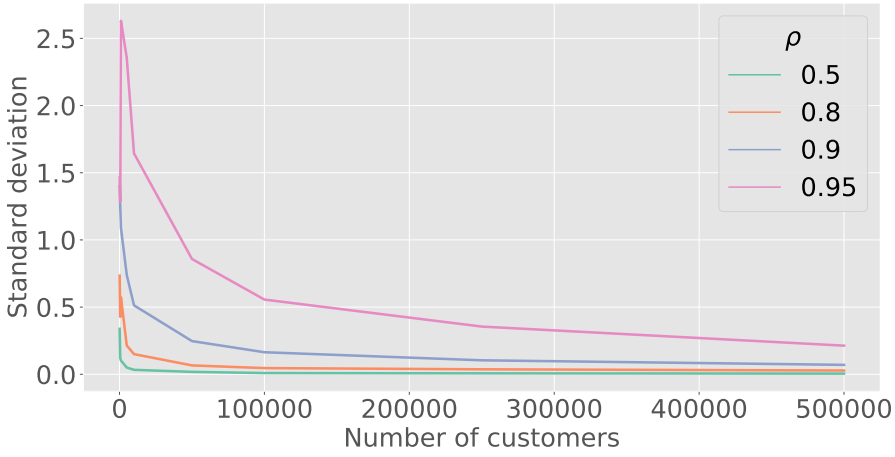


Figure 3: Progression of the standard deviation of $\bar{W}$ in an $M/M/1$ model with SPTF, for different values of ρ and c; $\mu = 1/1.25$.

3.2 $M/D/n$

In Figure 4 the progression of the standard deviation of the average waiting time is plotted. Like it happens with $M/M/1$, as the number of costumers increases, the standard deviation decreases; but in this case the difference between the curves is less stressed and generally the values of the standard deviation are smaller. This is understandable for the $M/D/1$ model is subject to stochasticity only for arrival times. ρ has a weaker influence on the fluctuation of the system.

Even thought the change in standard deviation from $c = 250\ 000$ to $500\ 000$ is not as significant as for previously discussed models, c was still chosen as $500\ 000$ to ensure the highest quality of simulations. Here $\gamma = 0.37$. Moreover, the goodness of the results is guaranteed by the fact that the distribution is Normal for $n = 1, 2, 4$ (Table 2, Appendix B).

Comparisons with different number of servers For higher number of servers, the average waiting time decreases. As the confidence intervals of the sample mean waiting time never overlap among different servers, the sample mean are significantly different; the result is established also from the t-test, which has a p-value near to 0 for all the pairs (p-value of the t-test $<<$ 0.001 for 1-2, 2-4, 1-4 (Appendix C)).

In Table 2 all the details are reported. From there we can see that the theoretical mean is included in the 95% confidence interval for $n = 1, 2$; for $n = 4$, it is not

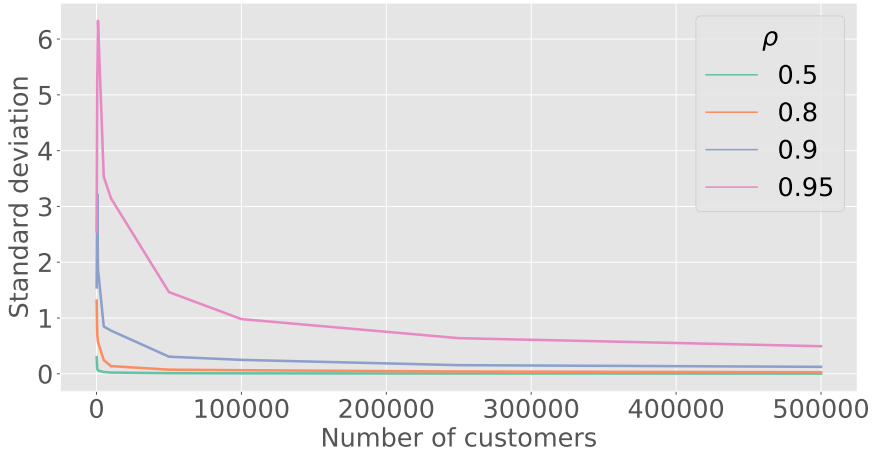


Figure 4: Progression of the standard deviation of $\bar{W}$ in an $M/D/1$ model, for different values of ρ and c; $\mu = 1/1.25$.

Table 2: Results for different number of servers in $M/D/\cdot$; $\rho = 0.95; \mu = 1/1.25$.

Model	$\bar{W}$	$E(W)$	CI-95%	p-normality
M/D/1	11.807	11.875	(11.72, 11.894)	0.303
M/D/2	5.803	5.785	(5.754, 5.852)	0.462
M/D/4	2.828	2.786	(2.803, 2.852)	0.715

included, but it is really near to the same value. This means that, even if the distribution is Normal, in the case of $M/D/4$ it would be better to use more costumers, in order to have more reliable results. Furthermore, by comparing the results of M/M/n and M/D/n we can see that the average waiting time is 2 times lower for M/D/n model even though μ is the same. This agrees with theoretical calculations.

3.3 $M/H/n$

In Figure 5 the progression of the standard deviation of the average waiting time is plotted. As expected, when the number of costumers increases, the standard deviation decreases. The standard deviation assumes high values, because the $M/H/1$ model has three sources of randomness, that implies more stochasticity and therefore more fluctuation.

c is chosen to be 500 000, as for that value the standard deviation is still consid-

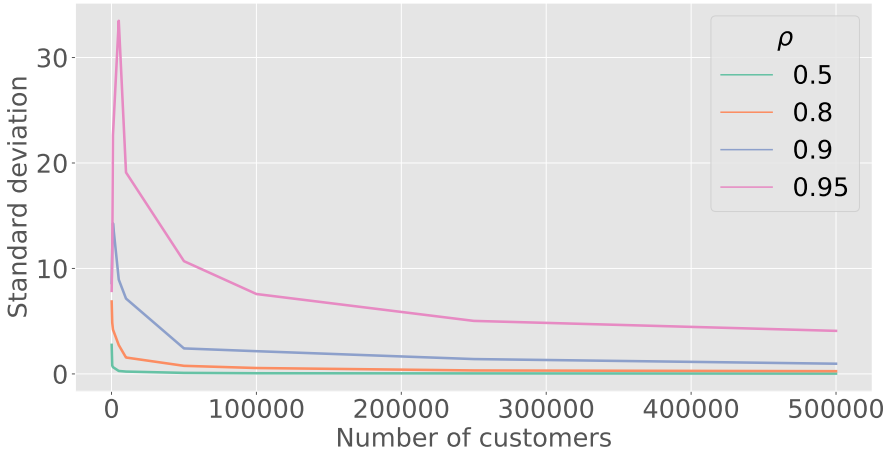


Figure 5: Progression of the standard deviation of $\bar{W}$ in an $M/H/1$ model, for different values of ρ and c;$\mu_1 = 1/0.5$, $\mu_2 = 1/3.5$.

Table 3: Results for different number of servers in $M/H/\cdot$; $\rho = 0.95; \mu = 1/1.25$.

Model	$\bar{W}$	CI-95%	p-normality
M/H/1	49.184	(48.402, 49.966)	0.085
M/H/2	24.126	(23.725, 24.528)	0.202
M/H/4	11.453	(11.276, 11.63)	0.003

erably smaller than the other values. Here $\gamma = 3.13$. The distribution is Normal for $n = 1, 2$ but not for $n = 4$ (Table 3, Appendix B); therefore in order to get a stable system for a $M/H/4$, more than 500 000 costumers are needed.

Comparisons with different number of servers For higher number of servers, the average waiting time decreases. As the confidence intervals of the sample mean waiting time never overlap among different servers, the sample mean are significantly different; the result is established also from the t-test, which has a p-value near to 0 for all the pairs (p-value of the t-test $<<$ 0.001 for 1-2, 2-4, 1-4 (Appendix C)).

In Table 3 all the details are reported. For this model the theoretical mean is not mentioned, because its derivation goes further the purpose of this paper. By comparing the results with M/M/n and M/D/n we can see that the average waiting time is the highest of the 3 models even though μ is the same, which is due to the extra stochasticity added to to the model.

4 Discussion

The behavior of the system in reaching a steady state is highly dependent on the load of the system, ρ. As ρ approaches 1, the system is more subject to oscillations and needs higher number of costumers in order to reach stability. Between $\rho = 0.5, 0.8, 0.9$ the difference in the standard deviation is generally small, but it becomes meaningful when comparing $\rho = 0.95$. This is valid for all the queue models analyzed, but with some differences. As expected, the model with more sources of randomness and therefore more variable, $M/H/1$, showed the highest discrepancy between $\rho = 0.95$ and the other values. ρ had the weaker influence on $M/M/1$ with priority SPTF, even less than when the service time are assumed constant.

By doubling the number of servers of models with same load characteristics and same priority rule, the average waiting time is usually halved; this was observed for each model taken into consideration in the analysis.

The model with the lowest average waiting time for 1 server was the one which assumed an exponential distribution for both the arrival and service times, and with the Short Processing Time First priority. In reality, applying this ordering of the costumers is not always possible: if the purpose is to model a call centre, it is not feasible to previously know how much time one costumer will need to be satisfied; on the other hand, an example where the service time could be calculated before the actual serving is the modelling of an automatized process in a factory, for instance.

Considering the FIFO rule, the best model is $M/D/1$. Here the time that a server dedicates to a costumer is fixed; this is useful in factory processes. In a more *social* context, the model would not be realistic without the option that a costumer who had expired its time but not completed the service, would join the queue again.

The experiments carried out in this paper could have been improved by using a higher amount of costumers in the simulations; moreover, different distribution could have been assumed both for the service and the arrival times.

References

Chan Wah Chun. *Elementary Introduction To Queueing Systems, An.* World Scientific, 2014. ISBN 9789814612005. URL `https://search-ebscohost-com.proxy.uba.uva.nl:2443/login.aspx?direct=true&db=e000xww&AN=810382&site=ehost-live&scope=site`.

David G. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):151–173, 1951. doi: https://

doi.org/10.1111/j.2517-6161.1951.tb00080.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1951.tb00080.x.

I. Adan and J. Resing. *Queueing Theory: Ivo Adan and Jacques Resing*. Eindhoven University of Technology. Department of Mathematics and Computing Science, 2001. URL https://books.google.nl/books?id=dAViMwEACAAJ.

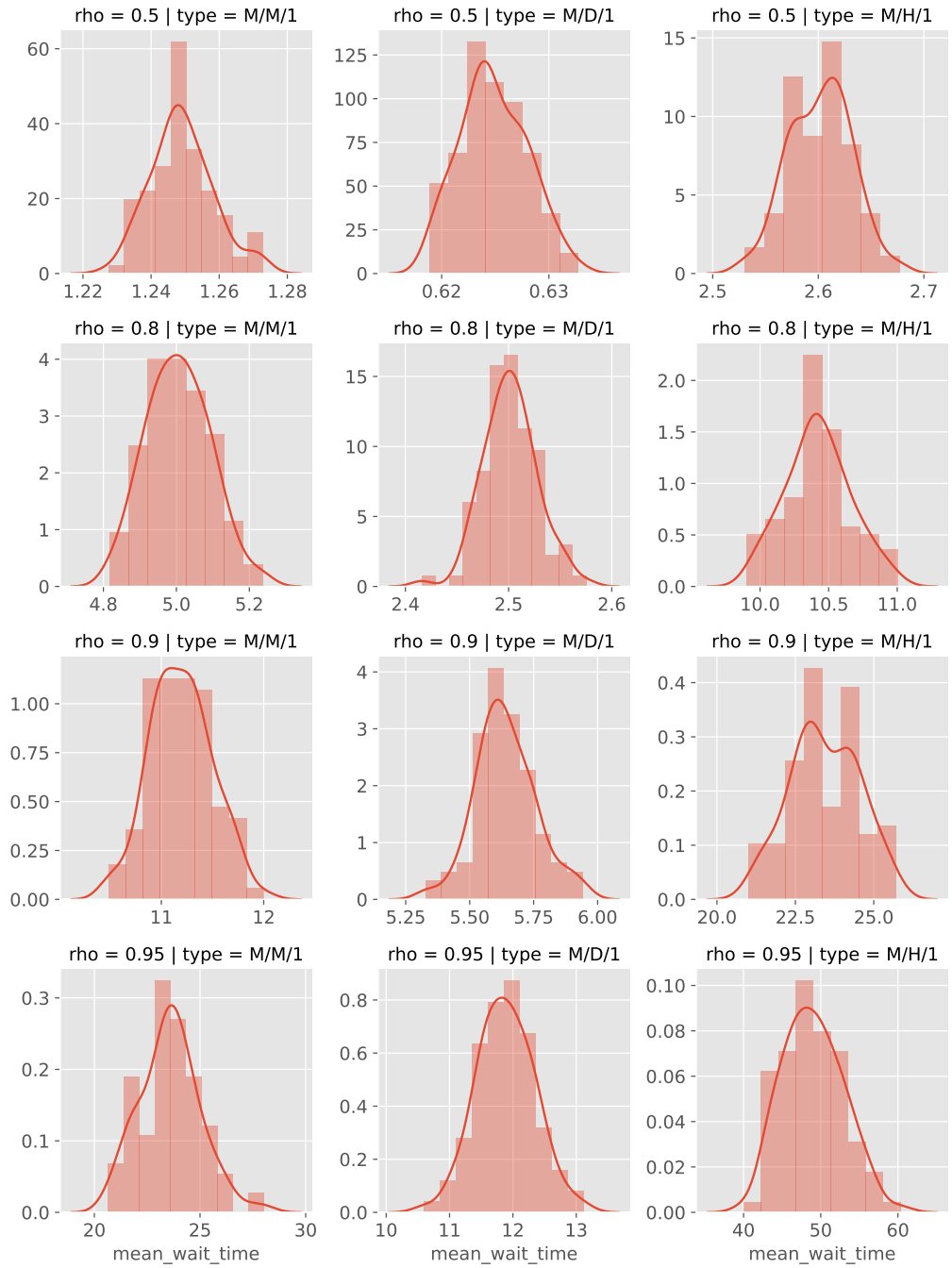
Yonatan Levy. Introduction to queueing theory, 2nd ed., by robert b. cooper, elsevier north holland, new york, 1981, 347 pp. price: $24.95. *Networks*, 13 (1):155–156, 1983. doi: https://doi.org/10.1002/net.3230130112. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230130112.

Leonard Kleinrock. *Theory, Volume 1, Queueing Systems*. Wiley-Interscience, USA, 1975. ISBN 0471491101.

A Detailed table for $M/M/n$

Model	ρ	$\bar{W}$	$E(W)$	CI-95%	p-normality
M/M/1	0.50	1.249	1.250	(1.247, 1.251)	0.221
M/M/2	0.50	0.417	0.417	(0.416, 0.418)	0.630
M/M/4	0.50	0.109	0.109	(0.109, 0.109)	0.918
M/M/1	0.80	5.005	5.000	(4.988, 5.022)	0.908
M/M/2	0.80	2.223	2.222	(2.216, 2.231)	0.163
M/M/4	0.80	0.931	0.932	(0.927, 0.934)	0.141
M/M/1	0.90	11.194	11.250	(11.135, 11.252)	0.844
M/M/2	0.90	5.324	5.329	(5.291, 5.357)	0.480
M/M/4	0.90	2.469	2.462	(2.456, 2.483)	0.040
M/M/1	0.95	23.583	23.750	(23.305, 23.861)	0.285
M/M/2	0.95	11.538	11.571	(11.408, 11.667)	0.609
M/M/4	0.95	5.566	5.571	(5.507, 5.624)	0.164
M/M/1-SPTF	0.50	0.890	0.891	(0.889, 0.891)	0.388
M/M/1-SPTF	0.80	2.351	2.353	(2.347, 2.355)	0.275
M/M/1-SPTF	0.90	4.002	3.996	(3.988, 4.015)	0.842
M/M/1-SPTF	0.95	6.564	6.580	(6.515, 6.612)	0.325

B Normality fitting



C Welch's t-test for M/M/n, M/D/n and M/H/n

Model	Servers	Servers	p-value
M/D/n	1	2	5.461e-154
M/D/n	2	4	1.980e-139
M/D/n	1	4	3.922e-146
M/H/n	1	2	5.322e-101
M/H/n	2	4	3.732e-96
M/H/n	1	4	3.886e-105
M/M/n	1	2	3.441e-116
M/M/n	2	4	9.384e-119
M/M/n	1	4	5.662e-118

D Welch's t-test between FIFO M/M/1 and SPTF M/M/1

ρ	p-value
0.50	3.749e-234
0.80	2.419e-163
0.90	6.477e-149
0.95	2.992e-113