UNIVERSITY OF AMSTERDAM

Queuing Theory

STOCHASTIC SIMULATION - ASSIGNMENT 2

Authors:
Nadav Levi, 11806990,
nadav.levi@student.uva.nl
Yunxiang Li, 13189387,
yunxiang.li@student.uva.nl

Abstract

In this paper we will be introducing Queuing Theory, a branch of applied probability with origins in the early 20th century. Queuing theory attempts to model the relation between a service centre and a population entering the service. We begin by a brief introduction of the topic, followed by laying out the theoretical groundwork. Then, we continue by investigating the average waiting times for M/M/n, M/D/n, M/H/1 models and with different scheduling disciplines. The average waiting times are shorter for an M/M/n queue than for a single M/M/1 queue with the same system load ρ is proved.

December 6, 2021

1 Introduction

Queuing theory is a branch of applied probability theory that was first described by Erlang [1] in the 20th century. The subject is particularly interesting due to its relevance in a big number of applications, in areas such as communication networks, computer systems, machine plants, and call centres to name a few[2]. The process of a Queuing model can be described as following: Consider a service centre with a certain capacity, and a population which engages with the service centre at a point in time. However, the service centre can serve only a limited number of customers at a time, based on the finite capacity. Therefore, if a new customers arrives but the service is at full capacity, the customer can not be serviced and will have to enter a waiting line. In the waiting line, the customer will wait until the the service becomes available again. Queuing Theory, in essence, is an attempt to quantify and predict this behaviour based on probability theory.

In the following section, we will briefly introduce all the relevant background information necessary to discuss queuing models, before we can start described the experiment which we have followed.

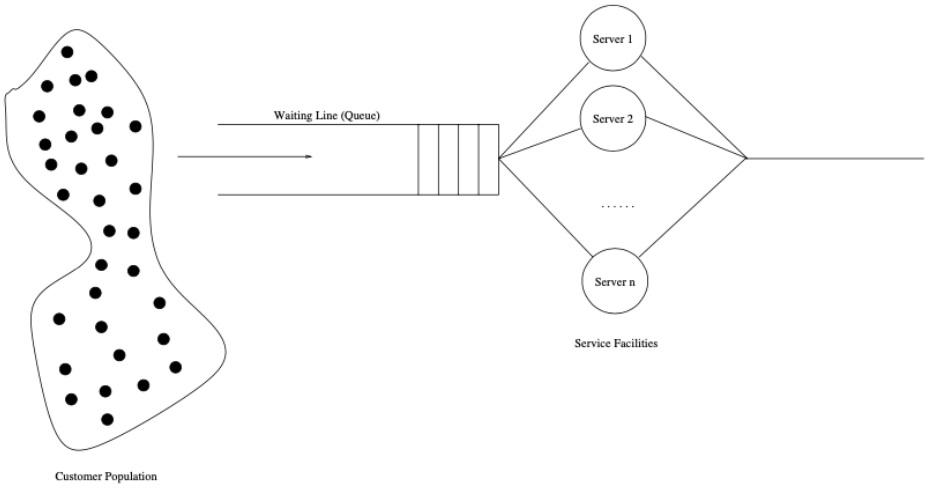

Figure 1: Model of a Service Center. From *A Short Introduction to Queueing Theory*[2]

2 Theoretical background

A queuing model has various characteristics. Some of which are:
Characteristics of a queuing model:

- *Arrival process of customers* Usually denoted by $A(t)$ is the distribution of inter-arrival times of customers into the queue. Random variables are usually i.i.d and commonly the Poisson distribution is chosen.
- *Behaviour of customers.* Are they willing to wait in order to be served, or will they grow impatient and leave?

- *Service time*, denoted by $B(t)$. Similar to the interarrival rate, the random variables are i.i.d. but are independent to $A(t)$
- *Capacity of the queue*: One server, or multiple servers serving customers?
- *Service Discipline*: The rules in which we serve customers. One by one or in batches? Some examples are
  - FIFO(First In, First Out) Customers are served in the orders in which they arrive.
  - LIFO(Last In, First Out) Customers are served in a reversed order
  - Random, where customers are served randomly

In this paper, we will be adopting the *Kendall Notation* : $A/B/n/N-S$ where A is the arrival rate, B is the service time, n is the number of servers, N is the maximum queue capacity and S is the service discipline.

For example, in the general case of $M/M/n$ we have the following:

- A and B are replaced by M which denotes the Markovian property. Hence, $A(t) = 1 - \exp(-\lambda t)$
- Since the discipline of service is FIFO, $N - S$ is omitted
- $\mathbb{E}[B]$ is the mean service time of one server.

Furthermore, we let

- $\lambda :=$ Arrival rate into the system as a whole.
- $\mu := \frac{1}{\mathbb{E}[B]}$ Capacity of each of n equal servers.
- $\rho := \frac{\lambda \mathbb{E}[B]}{n} = \frac{\lambda}{n\mu}$ For a queue with $n = 1, 2, 3, ...$ equal servers is the server utilization rate.
- And lastly, we want $\rho < 1$ - Otherwise the queue will grow

2.1 Equilibrium Probabilities

In this subsection we will touch upon Equilibrium Probabilities, such as the Delay Probability, Mean Queue Length, and Mean Waiting Time.

Definition 2.1. (Delay Probability) First we denote the probability that there are n customers in the system by p_n. Then, the Delay Probability is given by

$$\begin{aligned} \Pi_W &= p_n + p_{n+1} + p_{n+2} + \dots \\ &= \frac{p_n}{1-\rho} \\ &= \frac{(n\rho)^n}{n!}\Big[(1-\rho)\sum_{i=0}^{n-1}\frac{(n\rho)^i}{i!} + \frac{(n\rho)^n}{n!}\Big]^{-1} \end{aligned} \tag{1}$$

Before we continue, we will introduce *Little's Law*, which describes the relationship between various probabilities in a queuing system.

Proposition 2.2. ***Little's Law****: There exists a relationship between $\mathbb{E}[L]$, the mean number of customers entering the system, $\mathbb{E}[S]$, the mean sojourn time and λ, the average number of customers entering the queuing system per unit time.*

Namely, the following relation holds:

$$\mathbb{E}[L] = \lambda \mathbb{E}[s] \tag{2}$$

Furthermore, applying Little's Law to the queue yields a relation between the mean queue length $\mathbb{E}[L_q]$ and $\mathbb{E}[W]$, the mean waiting time:

$$\mathbb{E}[L_q] = \lambda \mathbb{E}[W] \tag{3}$$

For the proof of Little's Law, refer to [3].

Now, we can continue by formally introducing the *Mean Queue Length* $\mathbb{E}[L_q]$ and *Mean Waiting Time* $\mathbb{E}[W]$
From the Delay Probability we can directly obtain the Mean Queue Length

$$\begin{aligned}\mathbb{E}[L_q] &= \sum_{i=0}^{\infty} i p_{n+i} \\ &= \frac{p_c}{1-\rho} \sum_{i=0}^{\infty} i(1-\rho)\rho^i \\ &= \Pi \frac{\rho}{1-\rho}\end{aligned} \tag{4}$$

And then from *Little's Law* we obtain the *Mean Waiting Time*

$$\mathbb{E}[W] = \Pi \frac{\rho}{1-\rho} \frac{1}{n\mu} \tag{5}$$

Proposition 2.3. *Denote the Mean waiting time of a FIFO $M/M/2$ queue with two servers as $\mathbb{E}[W]$ and the mean waiting time for a $M/M/1$ queue by $\mathbb{E}[W^*]$*
Then, the following condition holds for $\rho < 1$

$$\mathbb{E}[W^*] > \mathbb{E}[W].$$

Proof. We begin by calculating the Delay Probability Π_W^* for the $M/M/1$ queue.
Taking $n = 1$, and substituting directly into (1) yields

$$\begin{aligned}\Pi_W^* &= \frac{(\rho)^1}{1!}\Big[(1-\rho)\sum_{i=0}^{1-1}\frac{(1\rho)^i}{i!} + \frac{(\rho)^1}{1!}\Big]^{-1} \\ &= \rho\Big[(1-\rho)+\rho\Big]^{-1} \\ &= \rho\end{aligned}$$

Then, substituting that into (4) and (5) we obtain:

$$\begin{aligned}\mathbb{E}[L_q] &= \rho\frac{\rho}{1-\rho} \\ &= \frac{\rho^2}{1-\rho}\end{aligned}$$

And subsequently,

$$\begin{aligned}\mathbb{E}[W^*] &= \frac{\rho^2}{1-\rho}\frac{1}{\lambda} \\ &= \frac{\rho}{1-\rho}\frac{1}{\mu}\end{aligned}$$

Similarly for $n = 2$

$$\begin{aligned}\Pi_W &= \frac{(2\rho)^2}{2!}\Big[(1-\rho)\sum_{i=0}^{2-1}\frac{(2\rho)^i}{i!} + \frac{(2\rho)^2}{2!}\Big]^{-1} \\ &= 2\rho^2\Big[(1-\rho) + (1-\rho)2\rho + 2\rho^2\Big]^{-1} \\ &= \frac{2\rho^2}{1+\rho}\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}[W] &= \frac{2\rho^2}{1-\rho}\frac{1}{1-\rho}\frac{1}{2\mu} \\ &= \frac{\rho^2}{(1-\rho^2)\mu}\end{aligned}$$

Putting it all together, we have that

$$\frac{\rho^2}{(1-\rho^2)\mu} < \frac{\rho}{(1-\rho)\mu} \implies \mathbb{E}[W] < \mathbb{E}[W^*]$$

which hold for every $0 < \rho < 1$. □

This result of course makes sense intuitively. Ceteris paribus, a queue with more servers *should* have lower waiting times as more customer can be engaged with at the same time, whereas in the case of $n = 1$, a customer needs to completely finish the process in order for the next customer to enter. Then, for $n = 2$, the first two customers can enter the queue immediately, while subsequent customers will have to wait less because now there are two possible servers that they can enter into.

Note that this result also generalises to queues where $n > 2$ but for the sake of brevity we will not include a proof. For a deeper dive on the topic, refer to [4]. However, in the following section we will demonstrate this property by means of an empirical experiment.

3 Methods

The following DES programs were completed to investigate the average waiting time for different service time distributions and service disciplines. We set exponential distributed inter-arrival times and variate the service time with exponential(M), deterministic(D) and hyper-exponential(H) distributions. A cooperation between FIFO scheduling and shortest job first scheduling was also carried out.

As stated above, λ is the overall arrival rate into the system, μ is the capacity of each server, and $\rho = \frac{\lambda}{n\mu}$ represents the system load. In terms of the M/M/n model, $\frac{1}{\lambda}$ is the inter-arrival mean, $\frac{1}{\mu}$ is the average service time. For M/D/n models, The service time is equal to the constant $\frac{1}{\mu}$ for all measurements. In the end, a hyperexponential distribution is used where 75% of the jobs have an exponential distribution with an average service time of $\frac{1}{\mu_1} = 1$ and the remaining 25% an exponential distribution with an average service time of $\frac{1}{\mu_2} = 5$. The corresponding μ is $\frac{1}{0.75\times1+0.25\times5} = 0.5$.

4 Results

For all experiments implemented, 500 simulations were performed to obtain the average waiting time. The number of measurements were quantified as the number of customers generated. Thus the average waiting time was calculated based on different duration of the simulation as the number of customers was equal to [100 to 1000 in the steps of 100, 1000 to 10000 in the steps of 1000 and 10000 to 100 000 in the steps of 10000].

4.1 M/M/1 and M/M/n

First, we implemented the program to simulate the Markovian Systems for different number of servers $n = 1$,$n = 2$ and $n = 4$. The distribution of the inter arrival times and the service times are exponential and therefore exhibit the memoryless property. FIFO scheduling was approached by default. The avarage waiting time was computed for $\mu = 1$ and ρ is equal to [0.7,0.8,0.9,0.95] to investigate how ρ influence the stability and statistical significance. The corresponding λ equals $\rho n \mu$. Each combination of parameters chose was simulated for 500 times.

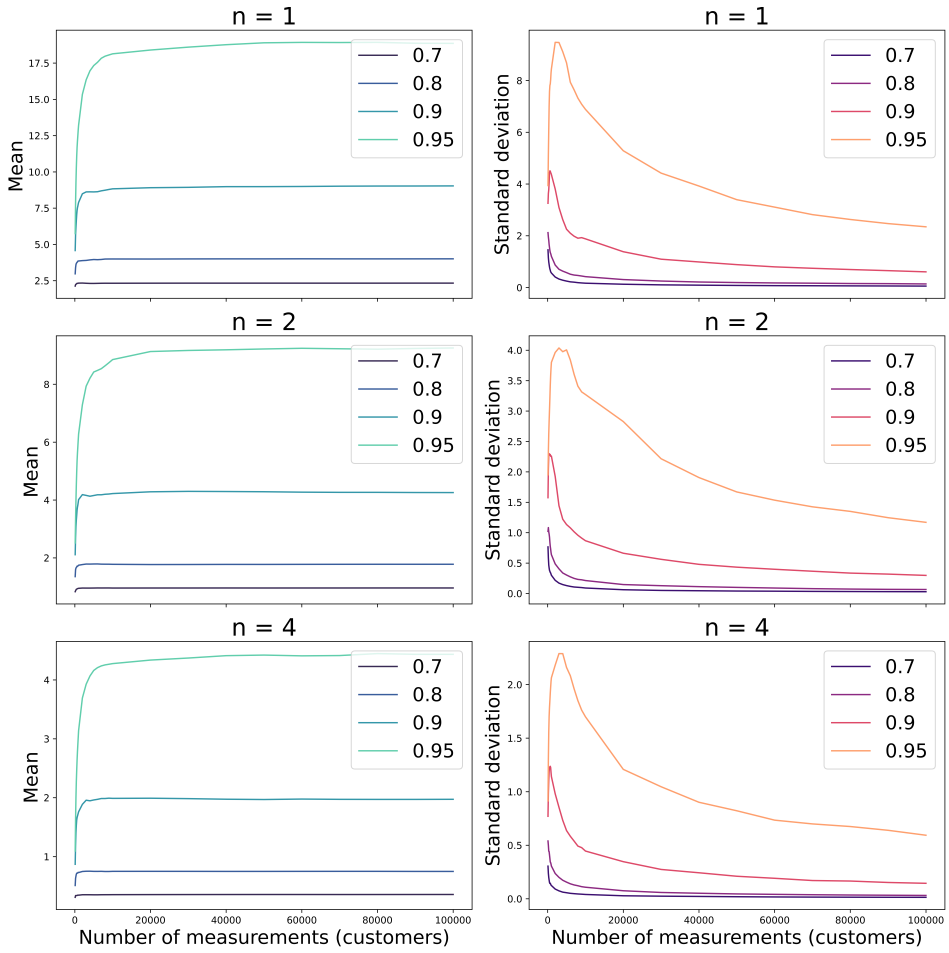

Figure 2: Investigating the average waiting time for multi-server Markovian System with $n = 1, 2, 4$ and different choices of $\rho = 0, 7, 0.8, 0.9, 0.95$.

As the mean values of average waiting time over 500 runs demonstrated in Figure

2, the average waiting times for an M/M/n queue are significantly shorter than for a single M/M/1 queue with n-fold lower arrival rate. It was reasonable since multi servers could compensate each other when there is a rush hour. With a larger system load ρ, the average waiting times have larger standard deviations and take longer time to reach stability. The results are showed in details in Table 1. The results distribution was also checked by Shapiro-Wilk test for normality. The histogram for maximum and minimum ρ seleted was ploted as Figure 3 indicates. The normality of distribution is less significant for larger ρ. This phenomenon could be due to the crowed warming stage as ρ increases.

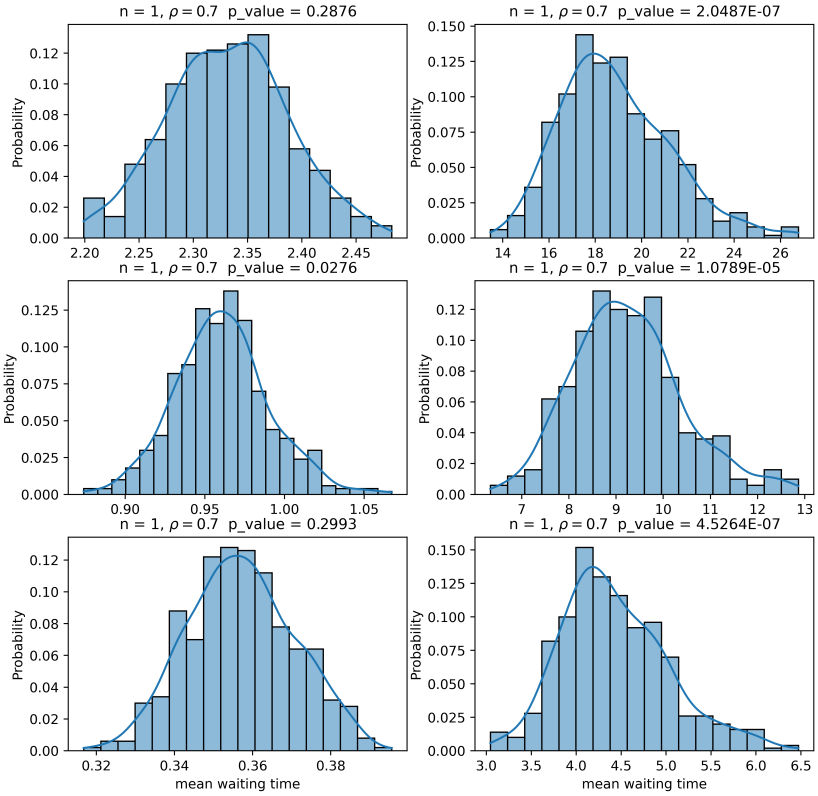

Figure 3: The normality fitting and Shapiro-Wilk test performed for minimum $rho = 0.7$ and maximum $rho = 0.95$.

Table 1: Summary of results (based on maximum measurements 100 000).

Model	μ	ρ	Mean	SD	CI 95%
M/M/1	1	0.7	2.3319	0.0553	(2.2793, 2.3844)
M/M/1	1	0.8	4.0064	0.1385	(3.8748, 4.1381)
M/M/1	1	0.9	9.0331	0.6057	(8.4578, 9.6085)
M/M/1	1	0.95	18.8571	2.3462	(16.6282, 21.0859)
M/M/2	1	0.7	0.9609	0.0294	(0.9329, 0.9888)
M/M/2	1	0.8	1.7803	0.0659	(1.7176, 1.8429)
M/M/2	1	0.9	4.2572	0.2976	(3.9745, 4.5399)
M/M/2	1	0.95	9.2541	1.1703	(8.1423, 10.3659)
M/M/4	1	0.7	0.3571	0.0136	(0.3442, 0.3700)
M/M/4	1	0.8	0.7484	0.0313	(0.7186, 0.7782)
M/M/4	1	0.9	1.9730	0.1446	(1.8356, 2.1104)
M/M/4	1	0.95	4.4343	0.5933	(3.8707, 4.9980)
M/M/1 (shortest job priority)	1	0.9	3.1959	0.1338	(3.0688, 3.3230)
M/M/2 (shortest job priority)	1	0.9	3.2687	0.2864	(3.9967, 4.5408)
M/M/4 (shortest job priority)	1	0.9	1.9495	0.1329	(1.8233, 2.0758)
M/D/1	1	0.9	4.5018	0.1890	(4.3223, 4.6814)
M/D/2	1	0.9	2.1457	0.1089	(2.0423, 2.2492)
M/D/4	1	0.9	0.9996	0.0534	(0.9489, 1.0503)
M/H/1 ($1/\mu_1 = 1$, $1/\mu_2 = 5$)	0.5	0.9	31.4549	2.7473	(28.8450, 34.0648)
M/H/2 ($1/\mu_1 = 1$, $1/\mu_2 = 5$)	0.5	0.9	14.5493	1.2865	(13.3272, 15.7715)
M/H/4 ($1/\mu_1 = 1$, $1/\mu_2 = 5$)	0.5	0.9	6.8034	0.7692	(6.0727, 7.5341)
M/M/1	0.5	0.9	18.0855	3.4483	(14.8096, 21.3613)

4.2 M/M/n with Shortest Job Priority

The M/M/n model with priority scheduling discipline was simulated. The customer who has lower service time needed is first served. Intuitively, the average waiting time should be lower than FIFO scheduling since the total waiting time must decreases. The comparison of M/M/1 models between FIFO and shortest job priority scheduling is demonstrated in Figure 4, which proves our point. The results for n=1,2,4 in Table 1 proves the statement that the average waiting times are shorter for an M/M/n queue than for a single M/M/1 queue with the same system load works for shortest job first scheduling as well.

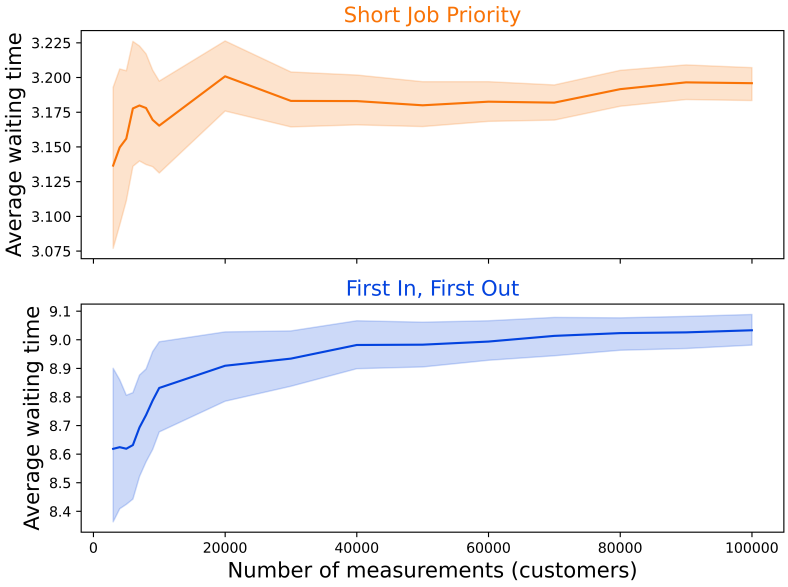

Figure 4: Comparison of M/M/1 models between First In First Out and Shortest job priority scheduling.$rho = 0.9$, number of customers from 3000 to 100 000. The average waiting time for two disciplines over 500 simulations.

4.3 Different service rate distributions

The M/D/n model with FIFO discipline was simulated. The service time is deterministic and set to the mean service time $\mu = 1$. Thus the system is less random and less rush hour occurs. The average waiting time should be lower than a pure Markovian system. The comparison between M/M/1 and M/D/1 models is showed in Figure 5. Again, the results for n=1,2,4 in Table 1 proves the statement that the average waiting times are shorter for an M/D/n queue than for a single M/D/1 queue with the same system load. Although it's less obvious compared with the M/M/n models.

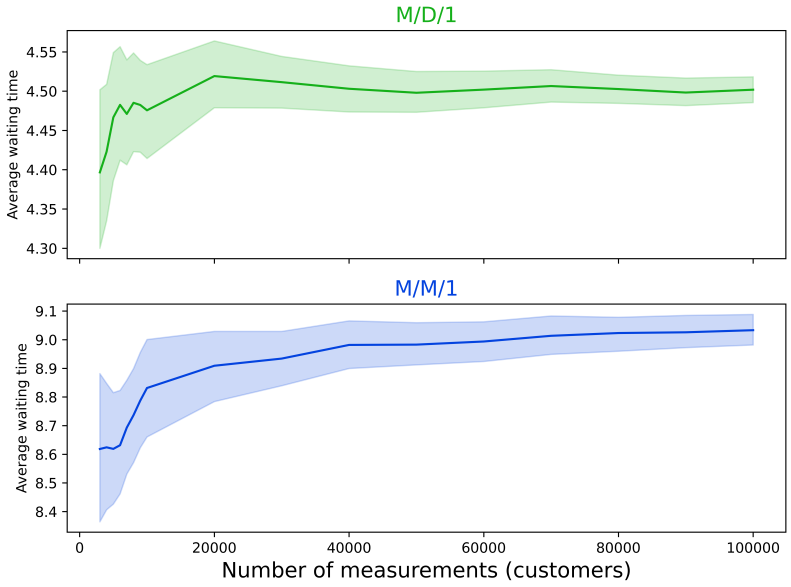

Figure 5: Comparison between M/D/1 and M/M/1 models. $rho = 0.9$**, number of customers from 3000 to 100 000. The average waiting time for two disciplines over 500 simulations.**

The M/H/n model with FIFO discipline was simulated. The service time is determined by two exponential distribution combined. The comparison between M/M/1 and M/H/1 models with same overall $\mu = 0.5$ is showed in Figure 5. The average waiting time for M/H/1 system is significantly larger than M/M/1. The reason might be its increased randomness involved. The results for n=1,2,4 in Table 1 proves the statement that the average waiting times are shorter for an M/H/n queue than for a single M/H/1 queue with the same system load.

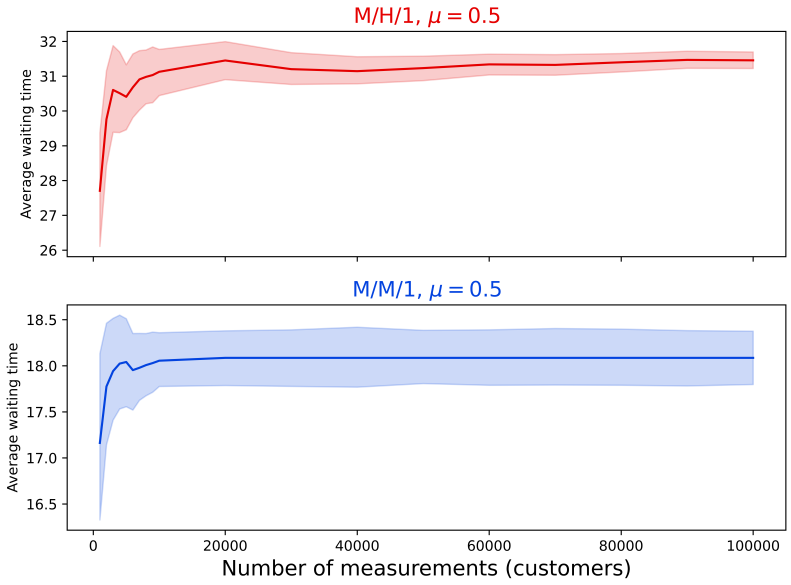

Figure 6: Comparison between M/H/1 and M/M/1 models. $\mu = 0.5$**,** $75\% of \mu_1 = 1$ **and** $75\% of \mu_2 = 1/5$**,** $rho = 0.9$**, number of customers from 1000 to 100 000. The average waiting time for two disciplines over 500 simulations.**

5 Discussion

Based on the multiple experiments above, the average waiting times are shorter for an M/M/n queue than for a single M/M/1 queue with the same load characteristics was verified. And this conclusion still hold for some other service distributions stated above, deterministic and hyperexponential.

By investigating the average waiting time, we found the system load ρ could heavily influence the state of the system. With a larger ρ selected, the system takes longer time approaching the steady state. In terms of 100 000 customers over 500 simulations, the normality of some distributions can be proven through Shapiro-Wilk test when $\rho = 0.7 or 0.8$. However, for an even larger ρ, the statistical significance is trivial and its distribution is unlikely normal. This may be the result of including the warming up stage since the distribution is right skewed. A further study and manipulation on truncating the initial part of the simulation should be carried on.

References

[1] A. Erlang, *The Theory of Probabilities and Telephone Conversations.* Nyt Tidsskrift for Matematik B, 20, 33, 1909.

[2] A. Willig, "A short introduction to queueing theory," 1999.

[3] R. B. Cooper, *Queueing Theory.* Association for Computing Machinery, 1981.

[4] I. Adan and J. Resing, *Queueing Theory.* Eindhoven University of Technology. Department of Mathematics and Computing Science, 2001.