Discrete Event Simulation with Queueing Theory

Meifang Li
13043390
meifang.li@student.uva.nl
30/11/2020

Yuhao Qian
13011456
yuhao.qian@student.uva.nl
30/11/2020

ABSTRACT

The study of queuing theory can help provide better service and reduce waiting time for customers. In this report, we model four different methods with DES simulation (M/M/n, M/M/n with priority, M/D/n, M/LT/n) and compare the mean waiting time given different system loads. The number of customers the system needs to reach the steady state is larger with a higher system load. The mean waiting time of customers declined when increasing the number of servers in each model. The M/M/n with priority scheduling and the M/D/n model achieved better performance than the M/M/n model under the same system load.

1 INTRODUCTION

Queuing theory can be viewed as one of the most extensive theories of stochastic models[7]. The study and development of queuing theory both in theoretical fields and areas of application are still thriving. Understanding this process can truly help the service center to achieve better performance and reduce customers' waiting time.

Agner Krarup Erlang who worked for Copenhagen Telephone Exchange first published a paper about queuing theory in 1909[3]. About forty years later, David George Kendall introduced the modern notation for queuing theory, known as Kendall's notation now[5]. Although many mathematicians made contributions to solve the formulas of different models for the mean waiting time, the classical unsolved problems like the queue M/G/k are still open in this field[6].

Intuitively, the queues would not form unless the servers could not handle all customers at the same time. Thus the customers have to wait to be served until at least one server is free. Since the process of queue relies on the arrival rate and service time of each customer, we could model the system in a way that is similar to the real world, given by discrete event simulation with queuing theory.

In this report, the service time would follow different distributions, like exponential, deterministic, and long-tail distribution. We also use multiple servers and different service disciplines, like FIFO (First In, First Out) and priority with shortest jobs to study the effects on the queues. Given different arrival intervals, we could get different system load-ρ and the number of customers the system needs to reach the steady state are larger with a higher system load. The mean waiting time of customers is declining when increasing the number of servers in each model. The mean waiting time of the model with priority service discipline is the lowest, followed by the model with deterministic service time and the M/M/n model at last, given the same system load.

2 THEORY

2.1 Queuing theory and Kendall's notation

The queuing phenomenon can be widely seen in our daily life so the study of queuing theory can help the service center improve the service and reduce mean waiting time.

In this report, we used a similar model like the example of a queuing system shown in Figure1. A customer arrives at a certain rate-λ into the system, then tries to access a server-n. In this system, one server can only deal with one customer at a time. If at least one server is available at that time, the customer would access this server and leave after a certain service time given by the service rate-μ. If all servers are busy with customers, the arriving customers would form a queue until at least one server is available again. Thus there will be a queue waiting for available servers and the waiting time of each customer will be the time between arrival and the customer start being served.

The arrivals of customers could follow a random nature and so do the service time of the service center. Additionally, the system could have one or several servers and the customers could be served with different orders, for example, the shortest job priority. Different methods of this model can be categorized by Kendall's notation: $A/B/n/N-S$[1].

- A denotes the inter-arrival time distribution.
- B denotes the service time distribution.
- n denotes the number of servers.
- N denotes the maximum size of the queue.
- S denotes the service discipline.(FIFO: First In, First Out./Priority: Every customer has a predefined priority and the servers always select the one with highest priority/etc.)

The distribution of A and B used in this report are the following:

- M(Markov): M for memoryless, as this is the exponential distribution $A(t) = 1 - e^{-\lambda t}$.
- D(Deterministic): deterministic distribution, for example, constant value(no randomness)
- H(Hyper-exponential): summation of exponential distribution each weighted with a probability, for example, 75% of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25% have an exponential distribution with an average service time of 5.0(also viewed as a long-tail distribution).

2.2 Little's Law

Little's Law is a general result that holds for every distribution type and every queue discipline[8]. It describes a relationship between the mean number of customers in the system-$E(L)$, the arrival rate-λ (average number of customers entering the system per unit time) and the mean sojourn time-$E(S)$ (time spent in the system: waiting

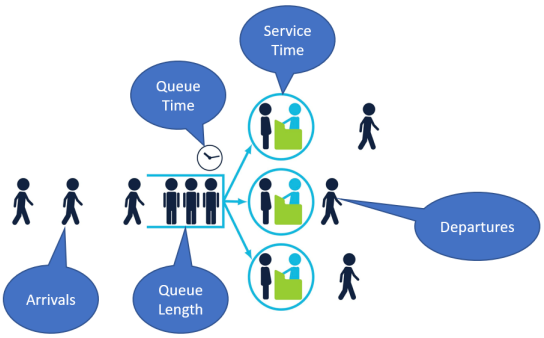

Figure 1: Queuing explanation[4]

time plus service time) in the steady state. We assume that the queue does not grow to infinity and obtain:

$$E(L) = \lambda E(S) \tag{1}$$

2.3 Performance measurement

2.3.1 Warming-up period. There will be a warming-up period[2] in the initial stage of one simulation before the system could be steady when the flows into the system equal the flows out of the system. At the beginning of one queuing simulation, the servers are not busy and some customers would not need to wait to be served so the mean waiting time of this period is lower than the steady-state period, which leads to the bias of the estimated resulting from incorrect initial conditions.

From the Law of Large Numbers, we can tell that the longer we run a simulation, the estimated value would be closer to the true value. However, the cost of computing would be extremely expensive if we continuously increase simulation to infinite. We have to balance between the bias and our computing power. If we have to take the bias into account, several methods could reduce the impact of the bias, including discarding the results of the warming-up period.

2.3.2 Mean performance measures. If we let ρ represent the system load, then in a single server system we get $\rho = \lambda/\mu$, in a multi-server system (one queue with n equal servers, each with capacity μ) we have $\rho = \lambda/(n\mu)$. Queuing theory tells us that for FIFO scheduling, the mean waiting times are shorter for an M/M/n queue than for a single M/M/1 queue with the same load ρ, which could be derived from the following proof.

M/M/1 queue. Since the exponential distribution is memoryless, we assume $p_k(t)$ denotes the probability that there are k customers in the system at time t, $k = 0, 1, \cdots$. We have

$$p_0(t + \Delta t) = (1 - \lambda\Delta t)p_0(t) + \mu\Delta t p_1(t) + o(\Delta t) \tag{2}$$

$$p_k(t + \Delta t) = \lambda\Delta t p_{k-1}(t) + (1 - (\lambda + \mu)\Delta t)p_k(t) + \mu\Delta t p_{k+1}(t) + o(\Delta t) \tag{3}$$

When $\Delta t \to 0$, we could get

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \tag{4}$$

$$p_k'(t) = \lambda p_{k-1}(t) - (\lambda + \mu)p_k(t) + \mu p_{k+1}(t) \tag{5}$$

In the limit($t \to \infty$), we require $p_k'(t) = 0$ so we arrive at the following steady-state equations:

$$0 = -\lambda p_0 + \mu p_1 \tag{6}$$

$$0 = \lambda p_{k-1} - (\lambda + \mu)p_k + \mu p_{k+1} \tag{7}$$

We know from Equation 6 that $p_1 = \lambda/\mu p_k = \rho p_0$. Taking this into Equation 7 for $k = 1$ and recursively express all probabilities in terms of p_0, we could get $p_k = \rho^k p_0$. Since the p_k are probabilities, they must follow from the normalization equation $\sum_{k=0}^{\infty} p_k = 1$. We then could obtain $p_k = (1-\rho)\rho^k$, $k = 0, 1, 2, \cdots$, thus the mean number of customers in the system-$E(L)$ is given by

$$E(L) = \sum_{k=0}^{\infty} k p_k = \sum_{k=0}^{\infty} k(1-\rho)\rho^k = \frac{\rho}{1-\rho} \tag{8}$$

and by applying Little's Law Equation 1, the mean sojourn time-$E(S)$ is given by:

$$E(S) = \frac{1/\mu}{1-\rho} \tag{9}$$

The mean number of customers in the queue-$E(L^q)$ can be obtained from $E(L)$ by subtracting the mean number of customers in service, so

$$E(L^q) = E(L) - \rho = \frac{\rho^2}{1-\rho} \tag{10}$$

The mean waiting time-$E(W)$ can be obtained from $E(S)$ by subtracting the mean service time, thus

$$E(W) = E(S) - 1/\mu = \frac{\rho/\mu}{1-\rho} \tag{11}$$

M/M/n queue. Similar as for the M/M/1, we could obtain the steady-state equations for the probabilities p_k by equating the flow between the two neighboring states $k-1$ and k:

$$\lambda p_{k-1} = min(k, n)\mu p_k, k = 1, 2, \cdots \tag{12}$$

through iterations, we can get

$$p_k = \frac{(n\rho)^k}{k!} p_0, k = 0, \cdots, n \tag{13}$$

$$p_{n+k} = \rho^k p_n = \rho^k \frac{(n\rho)^n}{n!} p_0, k = 0, 1, 2, \cdots \tag{14}$$

The probability p_0 follows from normalization, yielding

$$p_0 = \left(\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!} \cdot \frac{1}{1-\rho}\right)^{-1} \tag{15}$$

Let Π_W denote the probability that a job has to wait, it follows that

$$\Pi_W = p_n + p_{n+1} + p_{n+2} + \cdots = \frac{p_n}{1-\rho} = \frac{(n\rho)^n}{n!}\left((1-\rho)\sum_{k=0}^{n-1}\frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!}\right)^{-1} \tag{16}$$

From the steady-state probabilities we could directly obtain the mean queue length

$$E(L^q) = \sum_{k=0}^{\infty} k p_{n+k} = \frac{p_n}{1-\rho}\sum_{k=0}^{\infty} k(1-\rho)\rho^k = \Pi_W \cdot \frac{\rho}{1-\rho} \tag{17}$$

and then from Little's Law Equation 1,

$$E(W) = \Pi_W \cdot \frac{1}{1-\rho} \cdot \frac{1}{n\mu} \tag{18}$$

comparison between M/M/1 and M/M/2. We can obtain the mean waiting time of M/M/2 system by taking n=2 into Equation 18 and Equation 16,

$$E(W_2) = \frac{(2\rho)^2}{2!} \cdot \frac{1}{(1-\rho)^2} \cdot \frac{1}{2\mu}\Big(1 + 2\rho + \frac{(2\rho)^2}{2!} \cdot \frac{1}{1-\rho}\Big)^{-1}$$
$$= \frac{\rho^2}{\mu(1-\rho^2)} \tag{19}$$

We also know the mean waiting time of M/M/1 system from Equation 11, so we divide $E(W_2)$ by $E(W_1)$ and get

$$\frac{E(W_2)}{E(W_1)} = \frac{\rho^2}{\mu(1-\rho^2)} \cdot \frac{\mu(1-\rho)}{\rho} = \frac{\rho}{1+\rho} < 1 \tag{20}$$

The mean waiting time of M/M/2 system is always lower than M/M/1 system with the same system load-ρ.

2.3.3 Confidence interval. When we use simulation to approximate a value, it is always accompanied by confidence interval. If $X_1, \cdots, X_n$ are from the same distribution with mean θ and variance σ, the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ is an effective estimator of θ, but $\bar{X}$ does not exactly equal θ. Note that $E(\bar{X}) = \theta$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$ and thus, from *Central Limit Theorem*, we know that for large n : $\frac{\bar{X}-\theta}{\frac{\sigma}{\sqrt{n}}}$ $N(0,1)$. We also know that if we approximate $\sigma \to S$, then it remains the case and we can get $\frac{\bar{X}-\theta}{\frac{S}{\sqrt{n}}}$ $N(0,1)$. For any α, $0 < \alpha < 1$, let Z_α be such that $P\{Z > Z_\alpha\} = \alpha$, where Z is a standard normal random variable. We could obtain:

$$P\{\bar{X} - Z_{\alpha/2}\frac{S}{\sqrt{n}} < \theta < \bar{X} + Z_{\alpha/2}\frac{S}{\sqrt{n}}\} \approx 1 - \alpha \tag{21}$$

In other words, with probability $1-\alpha$ the population mean θ will lie within the region $\bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}$. In most cases, a 95% confidence level is used[9]. Thus for $\alpha = 0.5$, we could get $Z_{.025} = 1.96$, then $\bar{X} \pm \frac{1.96S}{\sqrt{n}}$ given 95% confidence level.

3 METHODOLOGY

3.1 Determination of the number of customers

If there are not enough customers arriving into the system, a waiting line will not form. Besides, a queue system will not reach equilibrium unless it has an adequate number of customers. Before we compare the mean waiting time of different models, we have to determine the number of customers that could lead the system to reach the steady state. We started our experiment at 100 customers and even added up to 20,000 customers in some cases. Because our report focused on n=1, 2, 4 servers, we used M/M/4 queuing system with different system loads ($\rho = 0.4, 0.8$) to determine the customer number. It is intuitively clear that if a certain number of customers would queue up in M/M/4, the queue would also appear in the other systems with fewer servers. Similarly, if M/M/4 would queue up with a higher system load ρ, the queue would also appear in systems with lower ρ. Our figures were obtained by using *seaborn.distplot*, this function visualized the distribution of mean waiting time by counting the observations that fall in the discrete bins(we have 10 bins in the experiment). Smooth curves, which estimated the probability density function, were added to the histograms.

3.2 Consideration about the bias

We also need to decide whether we should take the bias of warming up period into account in one single simulation when the system could reach the steady state. To complete the comparison, we ran one simulation for full data and discarding the first 20% of full data respectively and observed the distribution of probability density and the mean waiting time.

3.3 Normality test

To make sure we have a high and known statistical significance result, we have to confirm that after reaching the steady state, the mean waiting time of a given customer number is normally distributed. We used *scipy.stats.shapiro()* to perform the *Shapiro-Wilk Test* testing the null hypothesis: the mean waiting time in our experiment is drawn from a normal distribution. In each test, if the p-value we obtain is larger than 0.05, then we could not reject the null hypothesis. Besides, we continuously increased the customer number by 1000 every time until it passed the normality test five times in total.

3.4 Mean waiting time for different models

The discrete event simulation program in our experiment was built based on *Simpy*. We created several functions like *MMn-queueing* to generate a waiting list containing each customer's waiting time. The mean service time for M/M/n system and M/D/n system is 8 where server number $n = 1, 2, 4$. The M/M/n system with priority of shortest job first scheduling also ran in the mean service time of 8. The service time of M/LT/n is a long-tail distribution with mean service time equals to 2 (75% of the jobs have an exponential distribution with an average service time of 1.0 and the remaining 25% have an exponential distribution with an average service time of 5.0). The PDF of our long-tail distribution was shown in Figure 2.

By using *numpy.mean()*, we can calculate the mean waiting time for one simulation. If we repeat the simulation 100 times, we could finally obtain a distribution of mean waiting time. We also tried to compare the mean waiting time with different system loads-ρ(by using different arriving intervals)and multiple servers in each model. Additionally, we compared the mean waiting time of different models with the same system load-ρ.

During investigating the effect of ρ on mean waiting time, we set 19 different ρs, $\rho = 0.05, 0.1, 0.15, ..., 0.9, 0.95$, for each model.

4 RESULTS AND DISCUSSION

4.1 A non-mathematical comparison

Figure 3 has shown that the theoretical mean waiting time of M/M/2 is lower than M/M/1 through all the system loads-ρ, which is in accordance with the proof as mentioned above.

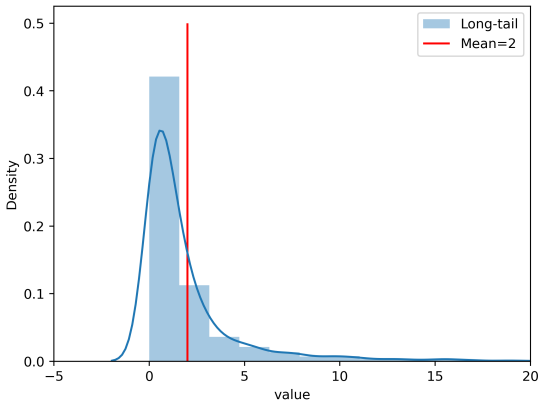

Figure 2: A long-tailed distribution: 75% has an exponential distribution with an average service time of 1.0 and the remaining 25% has an exponential distribution with an average service time of 5.0

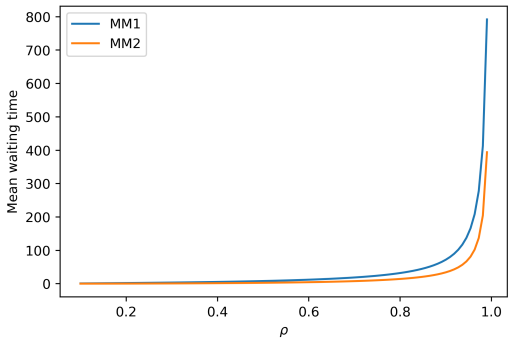

Figure 3: The comparison of the theoretical mean waiting time between M/M/1 and M/M/2 with $\mu = 1/8$

4.2 Number of customers when reaching steady state

Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8 revealed the process of reaching the steady state with different customer numbers. It was clear that, for all scenarios, the number of customers had significant effects on the distribution of mean waiting time. The spread of distribution shrank as the customer number went larger, which meant the system do not have obvious change when the customer number was larger enough. Thus we can say, under this circumstance, this system reached a steady state. Figure 5, for instance, the distribution and the mean waiting time did not change significantly when customer number was larger than 10,000. Referred to our computing power, we chose customer number as 10,000 instead of a enormously large number.

From Figure 4 and Figure 5, we also noticed that the number of customers the system needed to reach the steady state was smaller with low ρ(3000 for $\rho = 0.4$) and larger with high ρ(10000 for $\rho = 0.8$). Additionally, as shown in Figure 6, Figure 7and Figure 8, all the systems could reach the steady state when customer number was larger than 10000.

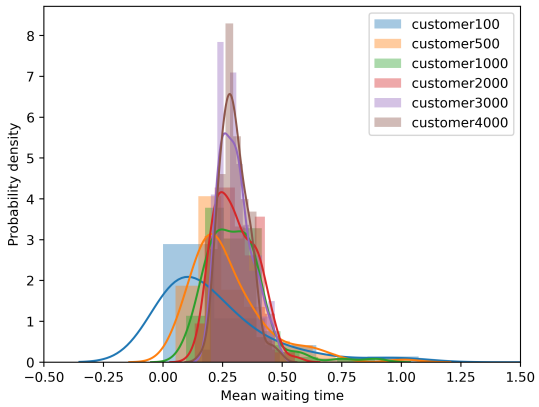

Figure 4: The distribution of mean waiting time with different customer numbers in M/M/4, where $\rho = 0.4$, simulation=100. Because mean waiting time is very small, y-axis has large density (larger than one).

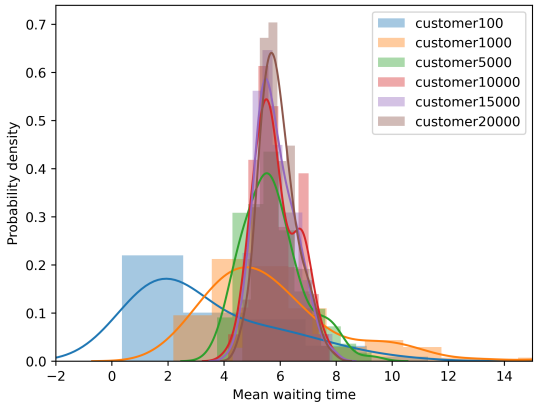

Figure 5: The distribution of mean waiting time with different customer numbers in M/M/4, where $\rho = 0.8$, simulation=100.

4.3 The effect of warming up period

We used a M/M/4 model to verify the effect of the warming up period. As shown in Figure 9, the mean waiting time distribution of M/M/4 without warming up period did not have significant differences with M/M/4 with warming up period. These two models not only had similar distribution but also had very close mean waiting time, which was 5.992867 and 5.965858 respectively. Therefore we concluded that the warming-up period did not have great effects on the distribution of mean waiting time. Then we did not consider its effect in later experiments.

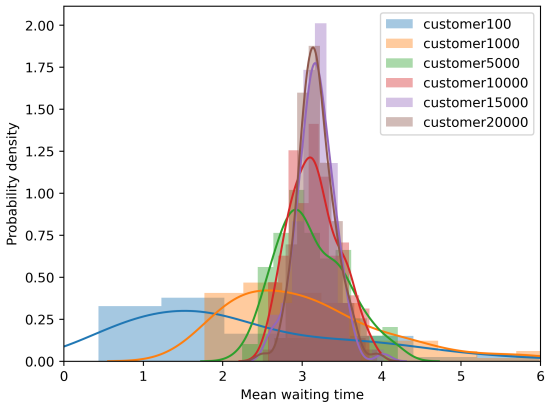

Figure 6: The distribution of mean waiting time with different customer numbers in M/M/4 having shortest job priority, where $\rho = 0.8$, simulation=100.

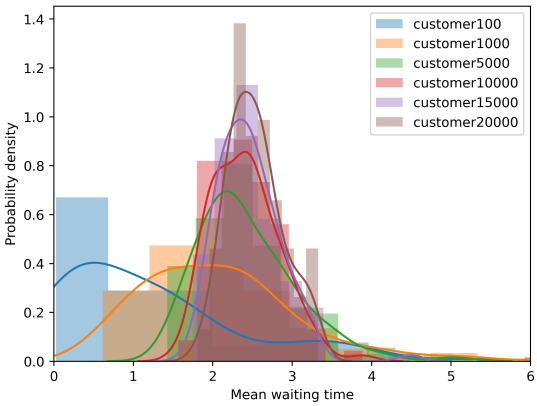

Figure 8: The distribution of mean waiting time with different customer numbers in M/LT/4 having long tail service time distribution , where $\rho = 0.8$, simulation=100.

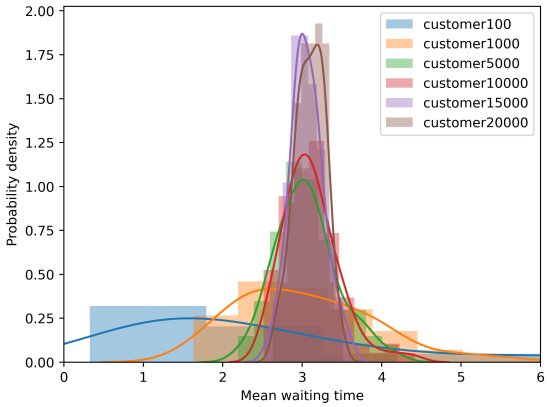

Figure 7: The distribution of mean waiting time with different customer numbers in M/D/4, where $\rho = 0.8$, simulation=100.

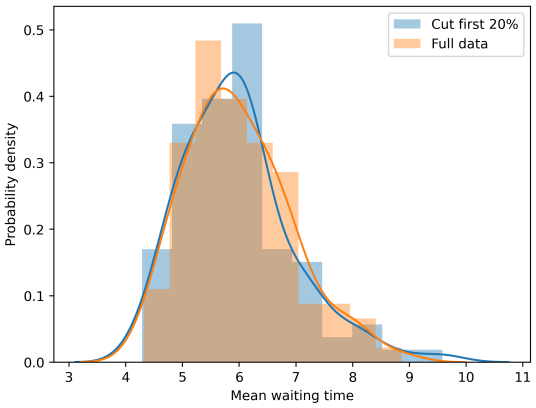

Figure 9: The effect of warming up period, verified by M/M/4, where $\rho = 0.8$, customer number=10000, simulation=100.

4.4 Normality test

Although the system seemed to reach the steady state at relatively large customer numbers, it did not necessarily mean the mean waiting time is normally distributed. For M/M/4 model with $\rho = 0.4, 0.6, 0.8$, We performed a normality test for each customer number and increased the number continuously until it passed the test five times. The results were shown in Table2. It was clear that the larger the customer number is, the easier it could pass the normality test. We also noticed that after reaching the steady state, the mean waiting time of the system did not deviate noticeably. In order to save computing costs, we did not need to have a very large customer number, which meant 10000 customers was enough for all the models to pass the test. Table1 gave test result of four different models with same customer number 10000 and system load $\rho = 0.8$. All the models had normally distributed mean waiting time with given parameters.

Table 1: The normality test of M/M/4, M/M/4 with shortest job priority, M/D/4, M/LT/4 with long-tail service time distribution. $\rho = 0.8$, the customer number is 10000.

Model	M/M/4	M/M/4 P	M/D/4	M/LT/4
p-value(95%)	0.206052	0.423938	0.070286	0.423938
H_0	Accept	Accept	Accept	Accept

4.5 Comparison of mean waiting time

4.5.1 Different server numbers. Figure 10, Figure 12, Figure 13 and Figure 14 revealed the effect of server numbers on the mean waiting time. In general, the more servers a system has, the less its mean waiting time is. Besides, the mean waiting time increased significantly as system load ρ grew. This conclusion applied to all the systems. With 95% confidence level, the mean waiting time of M/M/n at $\rho = 0.2$ is $2.005961 \pm 0.019754, 0.332293 \pm 0.005720, 0.024388 \pm$

Table 2: The number of customers of which the mean waiting time is normally distributed with 95% confidence level.

	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
	Customer	Mean	Customer	Mean	Customer	Mean
1	3000	0.312659	3000	1.44958	10000	5.892968
2	5000	0.308165	6000	1.414327	13000	5.877071
3	8000	0.296645	8000	1.431884	14000	5.942081
4	9000	0.304654	10000	1.426092	16000	6.053509
5	10000	0.301262	11000	1.414722	19000	5.896633

0.001104 for n=1,2,4 respectively. Still, the mean waiting time of M/M/n at $\rho = 0.8$ is 32.387056±0.439310, 14.156140±0.199209, 5.888959± 0.151509 for n=1,2,4 respectively. Detailed data was in Table 3.

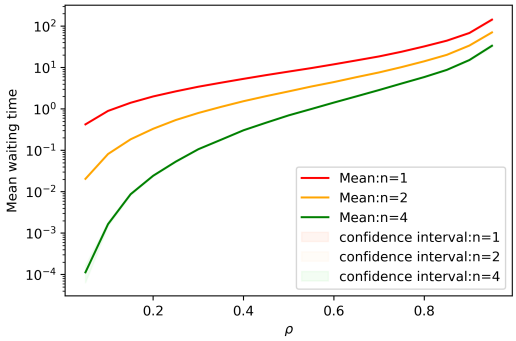

Figure 10: Mean waiting time of M/M/n queueing system with different servers $n = 1, 2, 4$, customer number=10000, simulation=100.

4.5.2 Different service rate distributions. Figure 11 showed a comparison of mean waiting time between M/M/1 and M/M/1 with shortest job priority. When system load ρ was small ($\rho < 0.3$), the difference between these two models' mean waiting time was subtle. The difference became larger as ρ increased close to 1. The mean waiting time of M/M/1 with shortest job priority was lower than M/M/1 throughout the whole experiment.

In Figure 15 we showed the differences between the four models. The difference between M/M/n and the other three models was very large when the server number was 1. This difference shrank obviously when the server number was 4. Among four models, the M/LT/n with long-tail distribution service time had the best performance in mean waiting time. This is because M/LT/n had the shortest mean service time (2 in this case) in this experiment, given others were 8. In addition, the performance of M/D/n was slightly better than that of M/M/n with shortest job priority.

5 CONCLUSION

In general, we investigated the effect of server numbers on the mean waiting time of a queuing system. Before focusing on the server number, we observed the process of reaching the steady state while the number of customers increased. As shown in the experiment, when the customer number was large enough, a system would reach a steady state. A system with a higher load needed

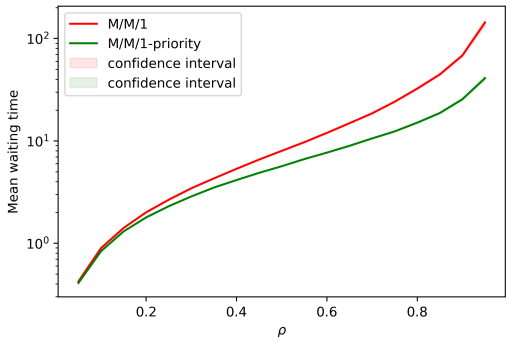

Figure 11: Mean waiting time of M/M/1 and M/M/1 with shortest job priority, customer number=10000, simulation=100.

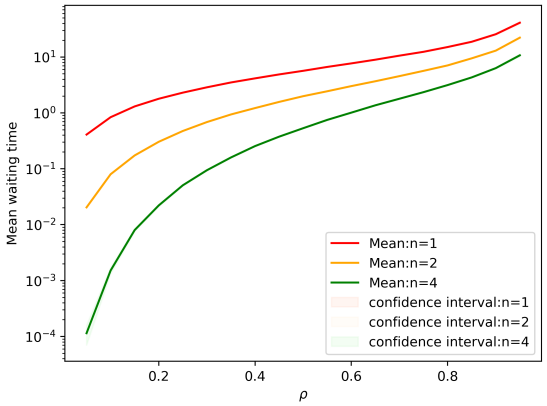

Figure 12: Mean waiting time of M/M/n with shortest job priority having different servers $n = 1, 2, 4$, customer number=10000, simulation=100.

more customers for steady state. Besides, the warming-up period in queuing theory did not have significant effects on the mean waiting time distribution in our experiment. It was clear that systems with more servers always had better performance in waiting time than systems with fewer servers. This conclusion was verified in all the models: M/M/n, M/M/n with shortest job priority, M/D/n and M/LT/n with a long-tail distribution in the service time. Under the same system load, M/M/n with shortest job priority had the lowest

Table 3: The mean waiting time of four queueing systems with different server numbers 1, 2, 4. $\rho = 0.8$, customer number=10000, simulation=100.

Server	M/M/n	M/M/n Priority	M/D/n	M/LT/n
1	32.387056 ± 0.439310	12.371762 ± 0.151813	16.056583 ± 0.149284	13.945829 ± 0.423996
2	14.156140 ± 0.199209	7.0636510 ± 0.118272	7.191170 ± 0.118881	6.103091 ± 0.185883
4	5.888959 ± 0.151509	3.1426864 ± 0.058207	3.078741 ± 0.0522505	2.421921 ± 0.104281

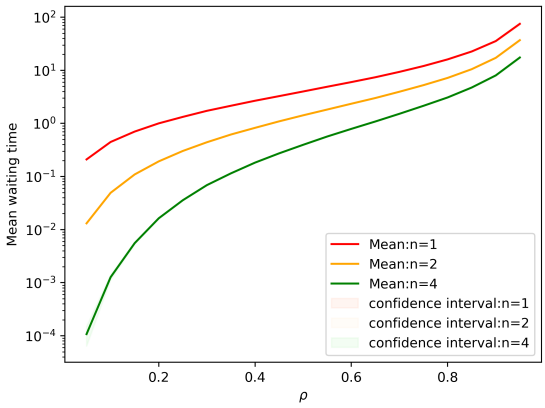

Figure 13: Mean waiting time of M/D/n with different servers $n = 1, 2, 4$, customer number=10000, simulation=100.

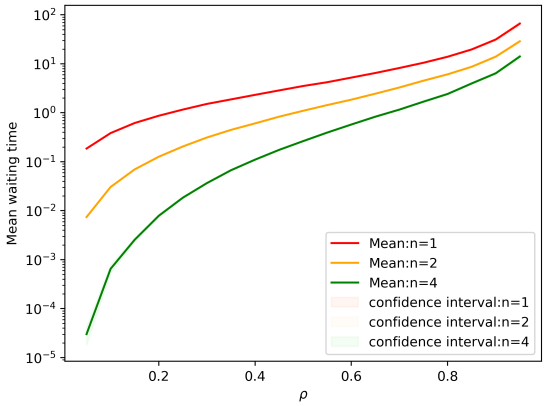

Figure 14: Mean waiting time of M/LT/n with different servers $n = 1, 2, 4$, customer number=10000, simulation=100..

mean waiting time, followed by M/D/n and M/M/n. Since the mean service time of M/LT/n is 2, while the other models are 8, we could not fairly compare their performances. In future, we could compare all the models again with equal mean service time. Additionally, we could use other service rate distributions like E (Erlang-k:the distribution of the sum of k independent exponential variables) and other queue type like LIFO (Last In, First Out)/ Round Robin/ Random to further study the effects on mean waiting time.

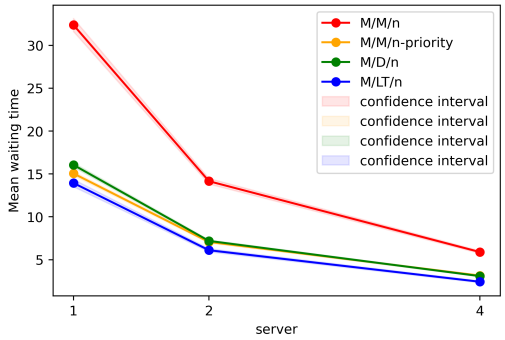

Figure 15: The comparison between four models with same system load $\rho = 0.8$, customer number=10000, simulation=100.

REFERENCES

[1] Ivo Adan and Jacques Resing. 2002. Queueing theory.

[2] Patrick J Delaney. 1995. *Control of Initialization Bias in Queueing Simulations Using Queueing Approximations.* Technical Report. VIRGINIA UNIV CHARLOTTESVILLE SCHOOL OF ENGINEERING AND APPLIED SCIENCE.

[3] Agner Krarup Erlang. 1909. The theory of probabilities and telephone conversations. *Nyt. Tidsskr. Mat. Ser. B* 20 (1909), 33–39.

[4] David Hare. 2019. *Tackling Queued Jobs With Queueing Theory - Part 1.* Community. Retrieved October 22, 2019 from https://community.alteryx.com/t5/Engine-Works/Tackling-Queued-Jobs-With-Queueing-Theory-Part-1/ba-p/475036#

[5] David G Kendall. 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* (1953), 338–354.

[6] JFC Kingman. 2009. The first Erlang century—and the next. *Queueing Systems* 63, 1-4 (2009), 3.

[7] Jyotiprasad Medhi. 2002. *Stochastic models in queueing theory.* Elsevier.

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

[9] Jerrold H Zar. 1999. *Biostatistical analysis.* Upper Saddle River, N.J. : Prentice Hall.