UNIVERSITY OF OKLAHOMA

GENERAL EXAM REPORT

A Study on Performance Analysis of Queuing System with Multiple Heterogeneous Servers

Prepared by

HUSNU SANER NARMAN

husnu@ou.edu

based on the papers

1) F. S. Q. Alves, H. C. Yehia, L. A. C. Pedrosa, F. R. B. Cruz, and L. Kerbache, Upper bounds on performance measures of heterogeneous M/M/c queues, Mathematical Problems in Engineering, vol. 2011, p. 18, May 2011.

2) C. Misra and P. K. Swain, Performance analysis of finite buffer queueing system with multiple heterogeneous servers, in 6th international conference on Distributed Computing and Internet Technology, ser. ICDCIT10, Bhubaneswar, India, Feb 2010, pp. 180 183.

October 8, 2012

Abstract

Most of the real life multi server queuing systems have heterogeneous servers which means each service rates are different than each other. Analysis of such queuing system plays important role to improve performance. In this report, I have summarized analysis of heterogeneous multi server queuing system which has a finite and an infinite buffer. Average waiting time and average queue length are formalized for both cases. Blocking probability of the system is also formalized when the buffer size of the system is finite. The formalized metrics approximation also tested with implemented simulation by using different allocation methods when the system has infinite buffer. Such analyzing methods and results can help us to understand multi heterogeneous queuing system.

CONTENTS

I. Introduction

Most of the real life multi server queuing systems have heterogeneous servers which means each service rates are different than each other. For example, each worker in Papa John's Pizzas has different service times because of distinct work speed of human beings. There would be many reason to have in heterogeneous servers. One of common reasons is that any failed or misbehaved component of a multi server system are replaced by more powerful one that causes system to be heterogeneous [1]. As a result, heterogeneous queuing system can be seen in every field of life. Heterogeneity of system arises a question which arrived job should be distributed to which servers, namely allocation policy, to have high throughputs or increase performance [1]. The question getting complex while different class of jobs has been considered. For instance, business and economy class customers waiting in the boarding pass queue are two type of customers and check-in officers can be considered as servers. This kind of system are called multi class multi server system,(MCMS). Flexibility of each class and server can be added to multi server system to make it more complex. Like economy class customers of some airline can be controlled by two check-in officers while one check-in officer controls business class customers or vice verse.

There have been voluminous research about multi heterogeneous servers system in the literature. These research subjects can be classified under four questions. (1) How many servers are needed? (2) What should be the allocation policies? (3) What should be flexibility level of each server? (4) What should be flexibility level of each class? A review of detail related literature can be found in [2].

In this report, I have reviewed performance of single-class heterogeneous multi servers queuing system which has an infinite and a finite buffer as described in [3] and [4]. The objectives of this report are:

- Analyzing of performance of heterogeneous multi server single queuing system under single class arrival.
- Reviewing upper bounds approximation for performance measurement of the system which has an infinite buffer and testing these approximations with different allocation policies.
- Formulating lower bounds approximation for performance measurement of the system which has an finite buffer.

The rest of the report is organized as follows: Firstly, the detail information about models will be given in Section II. Then, analysis of model I and II are explained in Sections III and IV, respectively. After analysis part, results are given in Section III-C and IV-C. Finally, the report is concluded in Section V.

II. Description of Models

Two different models of heterogeneous queuing system have been studied. Model I and Model II are described in subsections II-A and II-B, respectively.

A. Model I

First model is an $M/M_i/c$ queuing model with an infinite buffer as it is showed on Figure 1. As a general rule of M/M queues, the jobs arrive to servers according to Poisson distribution with rate λ. It is assumed that service times of jobs follows exponential distribution with rate μ_i where i = 1, 2, ... , c

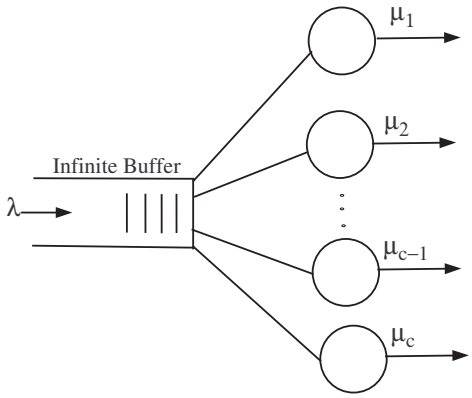

Fig. 1: Heterogeneous Queuing System with Infinite Buffer.

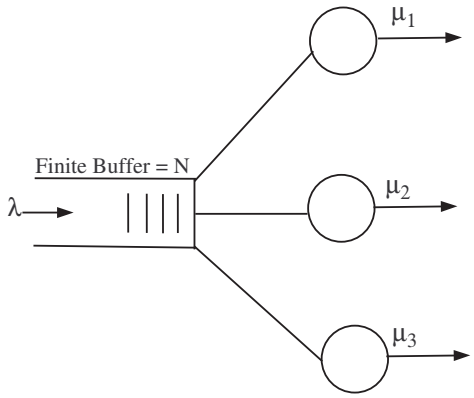

Fig. 2: Heterogeneous Three Queuing System with Finite Buffer.

and c is the number of the servers. Service rates can be different than each other and assume that;

$$\mu_1 \leq \mu_2 \leq \mu_3 \leq ..... \leq \mu_c \tag{1}$$

When a job arrives to server, it can be forwarded to one of idle servers based on allocation policy. Three more common allocation policies, which are (i) the fastest server first (FSF), allocating a job to the fastest available server, (ii) randomly chosen server (RCS), allocating a job to randomly chosen server, and (iii) the slowest server first (SSF), allocating a job to the slowest server are analyzed in the first model. If all servers are busy by serving some other jobs then the new arrived job is queued. If a server finishes its job, queued job is served according to FCFS rule which means first arrived job is served first based on the aforementioned allocation policies.

B. Model II

Second model is an $M/M_i/3/N$ queuing model with buffer size N as it is showed on Figure 2. Similar to previous model, as a general rule of $M/M_i/3/N$ queues, the jobs arrive to servers according to Poisson distribution with rate λ. It is assumed that service times of jobs follows exponential distribution with rate μ_i where i = 1, 2, and 3. Service rates can be different than each other and assume that;

$$\mu_1 \geq \mu_2 \geq \mu_3 \tag{2}$$

When a job arrives to server, it can be forwarded to an idle server based on allocation policy. In this model, we will not consider allocation policy while giving results for the queuing system but we will use FSF allocation policy to find lower blocking probability bound. If all three servers are busy to serve for other jobs then the new arrived job is queued. If a server finishes its job, queued job is served according to FCFS rule. The difference between this model with previous model is that this model has three servers and has a finite buffer. Therefore; an arrived job can be dropped if the buffer of the system is full.

III. Analysis of Model I

Average waiting time and average queue length are two main metrics to analyze the performance of queuing system with infinite queue size. To find average queue length and average waiting time, state probability, which is the probability of number of jobs in the system, need to be measured. All equations are mentioned for model I are derived in [3]. To simplify explanation of analysis of first model, the notations used in analysis are listed below:

λ	Job arrival rate.
μ_i	Service rate of i^{th} server.
ρ	Utilization of the system.
p_i	State probability of i^{th} state.
E_n^u	Upper bound for average queue length of the system.
E_T^u	Upper bound for average waiting time of the system.
μ_{ti}	Total service rates until i^{th} server. Or

$$\mu_{ti} = \sum_{j=1}^{i} \mu_j \qquad 1 \leq i \leq c \tag{3}$$

A. State Probability

It is assumed that queuing system is under heavy traffic flows. If multi jobs arrives, the job which is served by the slowest server is more likely to stay on system longer. This means that probability of finding a job on the slowest server is higher than the fastest server. The server rate of the system actually state dependent. When one job in the system, the server rate is μ_1 and when two jobs in the system, the server rate is $\mu_1 + \mu_2$. Server rate of the system is increasing until total number of the servers, c. Then the server rate of the system is fixed with μ_{tc}. Therefore, by using assumption of equation (1), state diagram of the system is formed as Figure 3. Based on Markov Chain on Figure 3, state probabilities can be formulated.

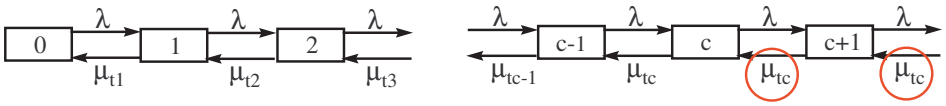

Fig. 3: State transition diagram for $M/M_i/c$ model

Indeed, it is SSF allocation policy and worst case scenario. Thus, by using state diagram of Figure 3, upper bound for two main metrics, E_n^u and E_T^u, are be computed for heterogeneous multi server system.

For better understanding, state probability equations until c state can be written as follows:

$$\begin{aligned}
&\lambda p_0 = \mu_{t1} p_1 \iff p_1 = p_0 \frac{\lambda}{\mu_{t1}} \\
&\lambda p_1 = \mu_{t2} p_2 \iff p_2 = p_0 \frac{\lambda^2}{\mu_{t1}\mu_{t2}} \\
&\lambda p_2 = \mu_{t3} p_3 \iff p_3 = p_0 \frac{\lambda^3}{\mu_{t1}\mu_{t2}\mu_{t3}} \\
&\quad * \\
&\quad * \\
&\lambda p_{c-1} = \mu_{tc} p_c \iff p_c = p_0 \frac{\lambda^c}{\mu_{t1}\mu_{t2}\mu_{t3} * * * \mu_{tc}}
\end{aligned} \tag{4}$$

State probability equations after c state are different because the system has only c servers. Thus; state probability equations can be written as follows:

$$\lambda p_{i-1} = \mu_{tc} p_i \iff p_i = p_0 \frac{\lambda^i}{\mu_{t1}\mu_{t2}\mu_{t3} * * * \mu_{tc-1}\mu_{tc}\mu_{tc}^{i-c}} \qquad where \qquad i > c \tag{5}$$

Or shortly,

$$p_i = p_0 \frac{\lambda^i}{\left(\mu_{tc}^{i-c}\right) \prod_{j=1}^{c} \mu_{tj}} \qquad where \qquad i > c \tag{6}$$

and we have

$$\sum_{j=0}^{\infty} p_j = 1 \tag{7}$$

In order to find state probabilities, we need to measure p_0 by using equation (7).

$$1 = \sum_{j=0}^{\infty} p_j = \sum_{j=0}^{c} p_j + \sum_{j=c+1}^{\infty} p_j \tag{8}$$

p_0 can be written as

$$p_0^{-1} = 1 + \sum_{j=1}^{c} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \sum_{j=c+1}^{\infty} \left(\frac{\lambda^j}{\left(\prod_{i=1}^{c} \mu_{ti} \right) \left(\mu_{tc}^{j-c} \right)} \right) \tag{9}$$

Equation (9) can be written as

$$p_0^{-1} = 1 + \sum_{j=1}^{c-1} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \sum_{j=c}^{\infty} \left(\frac{\lambda^j}{\left(\prod_{i=1}^{c} \mu_{ti} \right) \left(\mu_{tc}^{j-c} \right)} \right) \tag{10}$$

To force the system to be on steady state utilization, $\rho < 1$. Therefore;

$$\rho = \frac{\lambda}{\sum_{i=1}^{c} \mu_i} = \frac{\lambda}{\mu_{tc}} < 1 \tag{11}$$

After substituting equation (11) into equation (10), we will get

$$p_0^{-1} = 1 + \sum_{j=1}^{c-1} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \left(\frac{\mu_{tc}^c}{\prod_{i=1}^{c} \mu_{ti}} \right) \sum_{j=c}^{\infty} \rho^j \tag{12}$$

From geometric series and $\rho < 1$,

$$\sum_{j=c}^{\infty} \rho^j = \frac{\rho^c}{1-\rho} \tag{13}$$

By using equations (12) and (13), we get finally p_0 as

$$p_0 = \frac{1}{1 + \sum_{j=1}^{c-1} \left(\frac{\lambda^j}{\prod_{i=1}^{j} \mu_{ti}} \right) + \left(\frac{\lambda^c}{(1-\rho) \prod_{i=1}^{c} \mu_{ti}} \right)} \tag{14}$$

B. Average Queue Length and Waiting Time

Average queue length and average waiting time can be formulated by using state probability. Average queue length, E_n for $M/M/1$ queue can be formulated as

$$E_n = \sum_{j=1}^{\infty} j p_j \tag{15}$$

However, $M/M_i/c$ queue system has c servers and model uses SSF, E_n^u for $M/M_i/c$ will be

$$E_n^u = \sum_{j=c+1}^{\infty} (j-c) p_j \tag{16}$$

By using equation (6) and (16), we obtain

$$\begin{aligned} E_n^u &= \sum_{j=c+1}^{\infty} p_0 \frac{(j-c)\lambda^j}{\left(\mu_{tc}^{j-c}\right) \prod_{i=1}^{c} \mu_{ti}} \\ &= p_0 \frac{\mu_{tc}^c}{\left(\prod_{i=1}^{c} \mu_{ti} \right)} \sum_{j=c+1}^{\infty} (j-c)\rho^j \quad where \quad \rho = \frac{\lambda}{\mu_{tc}} \end{aligned} \tag{17}$$

After simplification of (17), finally we will have

$$E_n^u = p_0 \frac{\mu_{tc}^c}{\left(\prod_{i=1}^{c} \mu_{ti}\right)} \frac{\rho^{c+1}}{(1-\rho)^2} \tag{18}$$

By using Little's law, average waiting time can be formulate as

$$E_T^u = p_0 \frac{\mu_{tc}^c}{\left(\prod_{i=1}^{c} \mu_{ti}\right)} \frac{\rho^{c+1}}{(1-\rho)^2} \frac{1}{\lambda} \tag{19}$$

Equation (18) and (19) are upper bound approximations for average queue length E_n and average waiting time E_T, respectively.

C. Results of Model I

In this section, we will briefly explain results by using figures in [3]. Alves $et\ al$ simulate a $M/M_i/c$ model in order to find E_T for FSF, RCF, and SSF allocation policies. Then they compare their upper bound approximation formulas (19) and (18) and traditional homogeneous $M/M/c$ formulas approximation with simulation results in order to validate that proposed equations (19) and (18) are better approximation than traditional homogeneous $M/M/c$ approximation for $M/M_i/c$ model. They have only compare E_T because E_n can easily be obtained from E_T by Little's Law. In order to understand efficiency of proposed approximation, at least one of the following parameters needs to be changed:

λ : arrival rate
c : number of servers
Gini Index : heterogeneity of servers.

Gini Index, which of values change between zero and one, is a metric to measure level of heterogeneity. A Gini Index of zero means perfect equality where all server rates are equal and Gini Index of one means maximal inequality among servers rate. For example, if the system has two servers, and the server rates are distributed %50-%50 among two server then Gini Index of zero and the server rates are distributed%98-%2 among two servers then Gini Index of one. The main purpose using Gini Index is to be able to compare homogenity with heteronegty cases. Figure 4 and 5 show E_T in queue by changing aforementioned parameters. Each figure shows E_T for SSF, FSF, and RCF with proposed and homogeneous approximation by using the number of servers, $c = 2$, 3, 6 and 12. Figure 4 and 5 represent heavy loaded, $\rho = 0.9$ and less loaded, $\rho = 0.6$, respectively.

- Multi heterogeneous system performance by using FSF can be better than homogeneous system performance on some cases. These places are showed on Figure 4 as optimal regions. Optimal regions depends on both heterogeneity of the system and the number of the servers.
- While increasing heterogeneity, system performance decreases for all allocations except some cases for FSF.

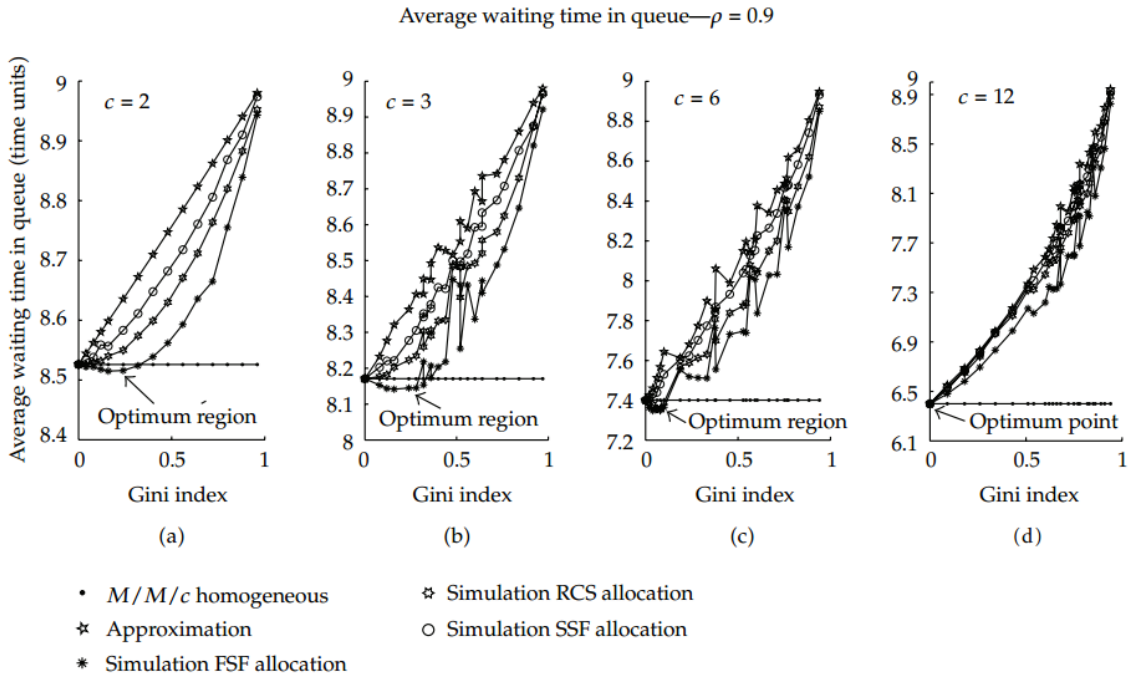

Fig. 4: Comparison of average waiting time between different allocation policies and approximations while $\rho = 0.9$ [3]

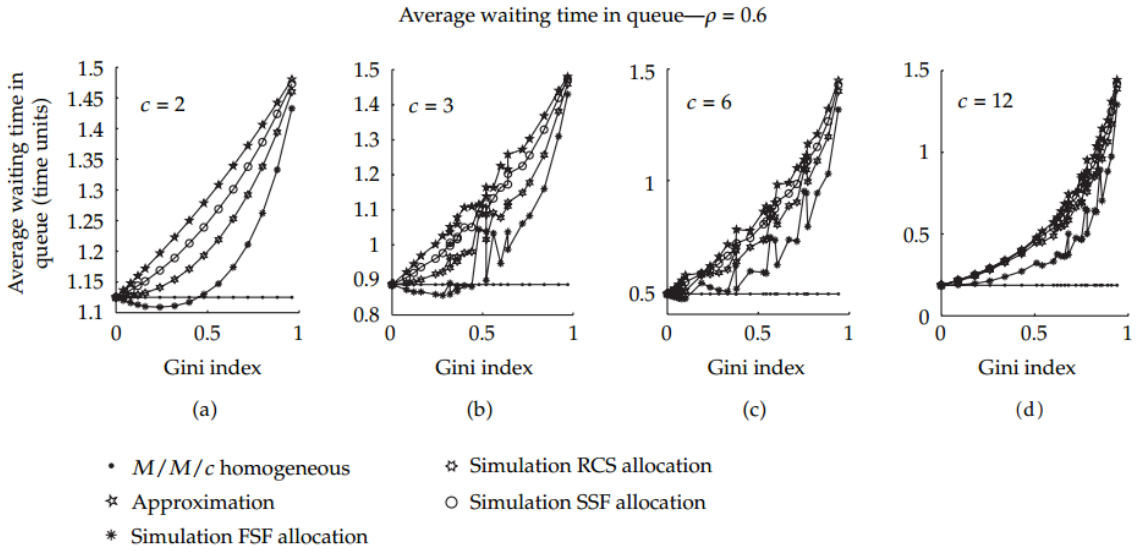

Fig. 5: Comparison of average waiting time between different allocation policies and approximations while $\rho = 0.6$ [3]

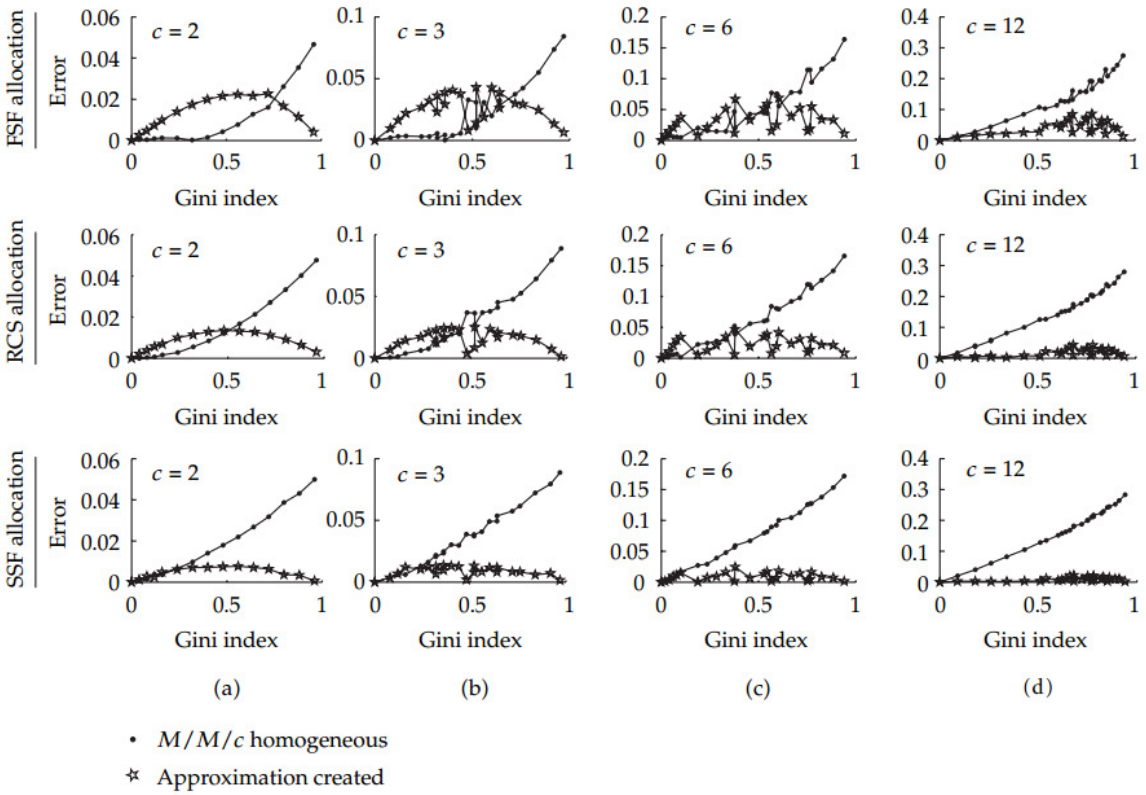

Fig. 6: Comparison of error of average waiting time between homogeneous and proposed approximation methods for different allocation policies while $\rho = 0.9$ [3]

- As it is expected, performance order of the system from greater to less is FSF, RCF, and SSF.

In order to compare the proposed approximation with homogeneous approximation, well known normalized error formula, equation (20) is used.

$$error = \frac{\|Simulation - Formula\|}{Simulation} \tag{20}$$

Figure 6 and Figure 7 shows normalized errors of proposed and homogeneous approximation for different allocations while system is under heavy loaded, $\rho = 0.9$ and less loaded, $\rho = 0.6$, respectively.

Some observations has been done from Figures 6 and 7 are listed below:

- When number of servers is low, like 2 and 3, homogeneous approximation for FSF and RCS is better than proposed one for less heterogeneity of the system.
- While increasing heterogeneity, proposed approximation is better than homogeneous approximation for all allocation policies.
- While increasing number of servers, proposed approximation is better than homogeneous approximation for all allocation policies except some cases for FSF.

Overall proposed upper bound approximations for multi heterogeneous system are better than traditional homogeneous system approximations.

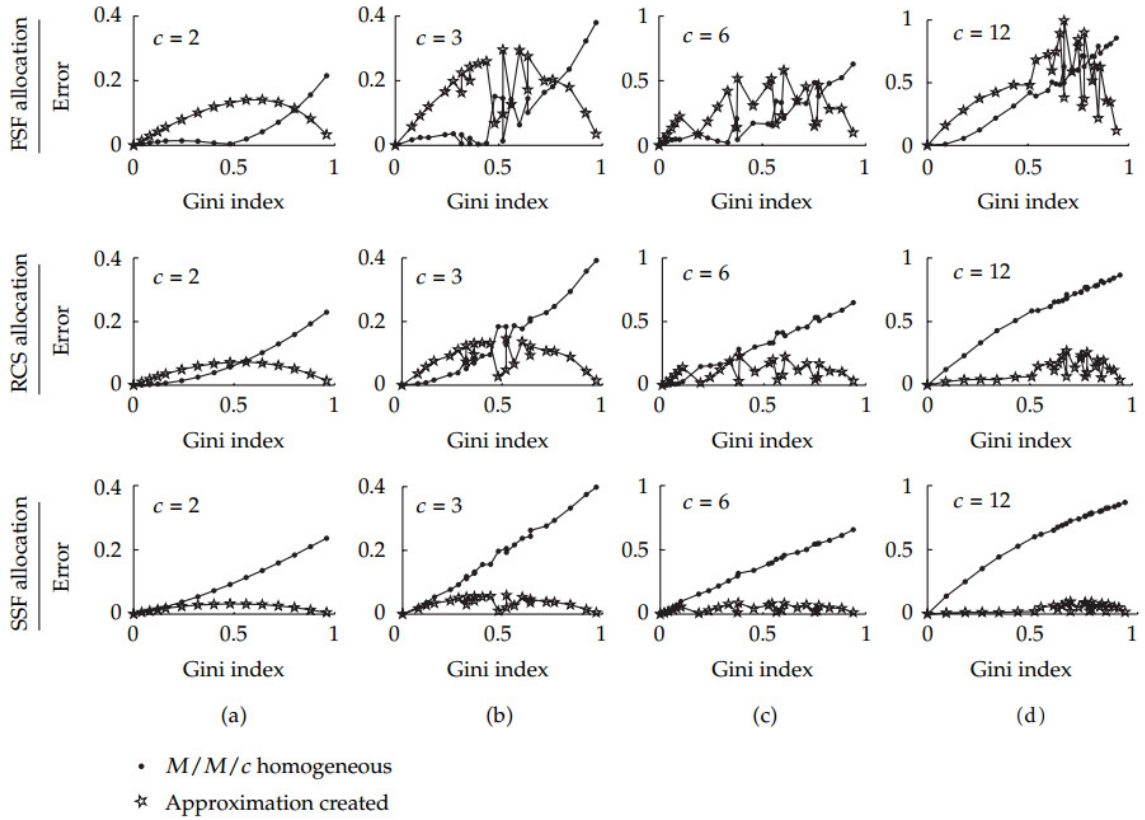

Fig. 7: Comparison of error of average waiting time between homogeneous and proposed approximation methods for different allocation policies while $\rho = 0.6$ [3]

IV. Analysis of Model II

In addition to average waiting time and average queue length, drop rate of the system is important metric to analyze the performance of queuing system which has a finite buffer. To find mentioned metrics, servers conditions, whether is idle or busy, and state probability, which is the probability of number of jobs in the system, need to be measured. To simplify explanation of analysis, the notations used in analysis are listed below.

- λ Job arrival rate.
- μ_i Service rate of i^{th} server where i = 1, 2, and 3. It is assumed that $\mu_1 \geq \mu_2 \geq \mu_3$.
- μ_{t3} Total service rate. Or $\mu_{t3} = \mu_1 + \mu_2 + \mu_3$
- ρ Utilization of the system.
- $\pi_{i,j,k}$ States of three servers. i, j, and k represents first, second and third servers and can be 0 or 1. 0 means server is idle and 1 means server is busy.
- $\pi_{n,3}$ probability of n^{th} state which means all three servers are busy and there are n customer in the queue. $0 \leq n \leq N$
- N Size of the buffer.

A. State Probability

Similar to previous model, it is assumed that queuing system is under heavy traffic flows. If a job arrives, the job is allocated to the fastest idle server. This means that the system allocation policy is FSF. If multi jobs arrive, jobs assigned to servers according to μ_1, μ_2, and μ_3 rate and state of servers (Whether servers are busy or idle). The buffer size of the system is finite. Thus; every jobs will be dropped if the buffer is full. While we are analyzing the system, we will focus on three main metrics. They are average queue length, average waiting time, and drop rate because of finite buffer. It is worth to mention that FSF policy is the best case scenario for heterogeneous multi server system. Hence, three main formalized metrics will represent lower bound for this model. Although it is easily understandable that the fastest server has the highest influence on the system, it is important to see how each server affects the performance of the system. By using aforementioned assumptions, state diagram of the system are formed as Figure 8 and 9 by taking busyness of servers in consideration. Job transition state diagram has only N states

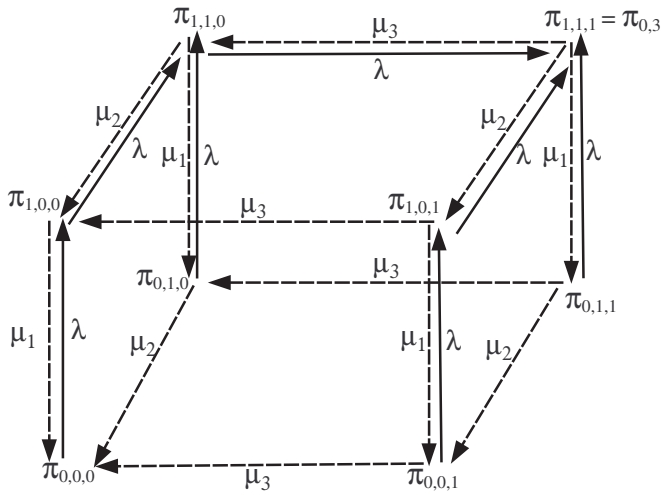

Fig. 8: Server state transition diagram for $M/M_i/3/N$ model

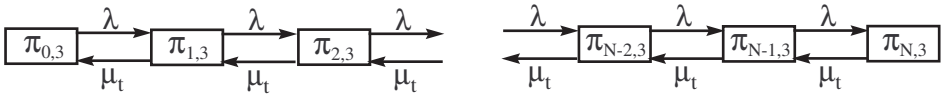

Fig. 9: Job state transition diagram for $M/M_i/3/N$ model

because of finite buffer. Based on Markov Chain on Figure 8 and 9, state probabilities are formulated by [4]. However, an error is recognized in [4] which affect all solutions while formulate state probabilities because of a forgotten value. Detail information about the error and alternative solutions are explained next section. By using server state transition diagram on Figure 8, following formulas are obtained in [4]:

$$\lambda\pi_{0,0,0} = \mu_1\pi_{1,0,0} + \mu_2\pi_{0,1,0} + \mu_3\pi_{0,0,1} \tag{21}$$

$$\lambda\pi_{1,0,0} + \mu_1\pi_{1,0,0} = \lambda\pi_{0,0,0} + \mu_2\pi_{1,1,0} + \mu_3\pi_{1,0,1} \tag{22}$$

$$\lambda\pi_{0,1,0} + \mu_2\pi_{0,1,0} = \mu_1\pi_{1,1,0} + \mu_3\pi_{0,1,1} \tag{23}$$

$$\lambda\pi_{0,0,1} + \mu_3\pi_{0,0,1} = \mu_1\pi_{1,0,1} + \mu_2\pi_{0,1,1} \tag{24}$$

$$\lambda\pi_{1,1,0} + \mu_1\pi_{1,1,0} + \mu_2\pi_{1,1,0} = \lambda\pi_{1,0,0} + \lambda\pi_{0,1,0} + \mu_3\pi_{1,1,1} \tag{25}$$

$$\lambda\pi_{0,1,1} + \mu_2\pi_{0,1,1} + \mu_3\pi_{0,1,1} = \mu_1\pi_{1,1,1} \tag{26}$$

$$\lambda\pi_{1,0,1} + \mu_1\pi_{1,0,1} + \mu_3\pi_{1,0,1} = \mu_2\pi_{1,1,1} \tag{27}$$

By using server state transition diagram on Figure 8 and job state transition diagram on Figure 9, following formula is obtained

$$\lambda\pi_{0,3} + \mu_1\pi_{0,3} + \mu_2\pi_{0,3} + \mu_3\pi_{0,3} = \lambda\pi_{0,1,1} + \lambda\pi_{1,1,0} + \lambda\pi_{1,0,1} + \mu_1\pi_{1,3} + \mu_2\pi_{1,3} + \mu_3\pi_{1,3} \tag{28}$$

By using job state transition diagram on Figure 9, following formula is obtained

$$\lambda\pi_{n,3} + \mu_1\pi_{n,3} + \mu_2\pi_{n,3} + \mu_3\pi_{n,3} = \lambda\pi_{n-1,3} + \mu_1\pi_{n+1,3} + \mu_2\pi_{n+1,3} + \mu_3\pi_{n+1,3} \quad where \quad 1 \leq n \leq N \tag{29}$$

Or equation (29) can be written as:

$$\lambda\pi_{n,3} = \mu_t\pi_{n+1,3} \iff \pi_{n,3} = \frac{\mu_t}{\lambda}\pi_{n+1,3} = \rho^{-1}\pi_{n+1,3} \quad where \quad 0 \leq n \leq N \quad and \quad \rho = \frac{\lambda}{\mu_t} \tag{30}$$

$$\mu_1\pi_{N,3} + \mu_2\pi_{N,3} + \mu_3\pi_{N,3} = \lambda\pi_{N-1,3} \tag{31}$$

From equation (29) and (31),

$$\pi_{N-1,3} = \pi_{N,3}\frac{\mu_t}{\lambda} \iff \pi_{n,3} = \rho^{n-N}\pi_{N,3} \quad where \quad 0 \leq n \leq N \tag{32}$$

However, equation (27) should be

$$\lambda\pi_{1,0,1} + \mu_1\pi_{1,0,1} + \mu_3\pi_{1,0,1} = \mu_2\pi_{1,1,1} + \lambda\pi_{0,0,1} \tag{33}$$

because as it is showed Figure 10, $\lambda\pi_{0,0,1}$ is ignored (It is also confirmed by the authors of [4]). Therefore, all solutions which was obtained in [4] using equation (27), are not correct. Therefore, we have ignored all generated formulas for average queue length, average waiting time, and probability of blocking in [4]. By following similar method as it was followed in [4] to find state probabilities by using corrected equation (33) will be very hard. However, in order to analyze the model, state probabilities are needed. Therefore, We have used similar method which is used in [3] in order to find state probabilities for second model.

State transition diagram has only $N+3$ possibilities because buffer size is N and the system has three servers which means $c = 3$. By using above assumptions, state transition diagram for second model can be formed as in Figure 11. To simplify explanation of analysis of second model, the notations used in analysis are listed below:

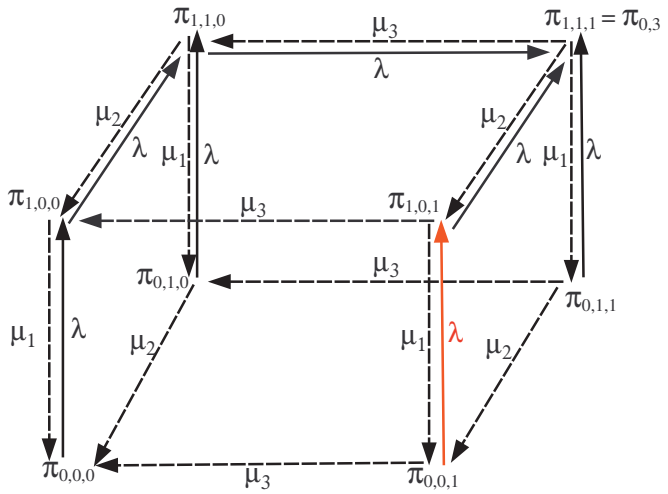

Fig. 10: Server state transition diagram for $M/M_i/3/N$ model

Fig. 11: State transition diagram for $M/M_i/3/N$ model

λ	Job arrival rate.
p_i	State probability of i^{th} state where $0 \leq i \leq N+3$.
μ_{t1}	= μ_1.
μ_{t2}	= μ_1 + μ_2.
μ_{t3}	= μ_1 + μ_2 + μ_3.
E_n^l	Lower bound for average queue length of the system.
E_T^l	Lower bound for average waiting time of the system.
P_B^l	Lower bound drop rate of the system
γ^u	Upper bound throughput of the system

For better understanding, state probability equations can be written as follows:

$$\begin{aligned}
\lambda p_0 = \mu_{t1} p_1 \iff p_1 &= p_0 \frac{\lambda}{\mu_{t1}} \\
\lambda p_1 = \mu_{t2} p_2 \iff p_2 &= p_0 \frac{\lambda^2}{\mu_{t1}\mu_{t2}} \\
\lambda p_{i-1} = \mu_{t3} p_i \iff p_i &= p_0 \frac{\lambda^i}{\mu_{t1}\mu_{t2}\mu_{t3}^{i-2}} = p_0 \frac{\mu_{t3}^2 \rho^i}{\mu_{t1}\mu_{t2}} \quad where \quad 3 \leq i \leq N+3 \quad and \quad \rho = \frac{\lambda}{\mu_{t3}}
\end{aligned} \tag{34}$$

In order to find state probabilities, we need to measure p_0 by using equation (35).

$$\sum_{j=0}^{N+3} p_j = 1 \tag{35}$$

Substituting equations (34) to equation (35), we get

$$1 = \sum_{j=0}^{N+3} p_j = p_0 + p_0\frac{\lambda}{\mu_{t1}} + p_0\frac{\lambda^2}{\mu_{t1}\mu_{t2}} + \sum_{j=3}^{N+3} p_0\frac{\mu_{t3}^2\rho^j}{\mu_{t1}\mu_{t2}} \tag{36}$$

After simplifying equation (36), we get

$$p_0^{-1} = 1 + \frac{\lambda}{\mu_{t1}} + \frac{\lambda^2}{\mu_{t1}\mu_{t2}} + \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}\sum_{j=3}^{N+3}\rho^j \tag{37}$$

or

$$p_0 = \frac{1}{1 + \frac{\lambda}{\mu_{t1}} + \frac{\lambda^2}{\mu_{t1}\mu_{t2}} + \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}\sum\limits_{j=3}^{N+3}\rho^j} \tag{38}$$

ρ can be any positive real number because the system has finite buffer. If $\rho = 1$ then

$$p_0 = \frac{1}{1 + \frac{\lambda}{\mu_{t1}} + \frac{\lambda^2}{\mu_{t1}\mu_{t2}} + \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}\frac{(N+3)(N+4)-12}{2}} \tag{39}$$

If $\rho \neq 1$, geometric series can be used to simplify equation (40) as

$$\sum_{j=3}^{N+3}\rho^j = \frac{\rho^3 - \rho^{N+4}}{1-\rho} \tag{40}$$

Finally p_0 is obtained as

$$p_0 = \begin{cases} \frac{1}{1+\frac{\lambda}{\mu_{t1}}+\frac{\lambda^2}{\mu_{t1}\mu_{t2}}+\frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}\frac{\rho^3-\rho^{N+4}}{1-\rho}} & \rho \neq 1 \\ \frac{1}{1+\frac{\lambda}{\mu_{t1}}+\frac{\lambda^2}{\mu_{t1}\mu_{t2}}+\frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}}\frac{((N+3)(N+4)-12)}{2}} & \rho = 1 \end{cases} \tag{41}$$

Drop probability of the second model is actually the final state probability which is p_{N+3}. Therefore; drop rate and throughput of the second model can be formulated as

$$P_B^l = p_0\frac{\mu_{t3}^2\rho^{N+3}}{\mu_{t1}\mu_{t2}} \tag{42}$$

$$\gamma^u = \lambda(1 - P_B^l) \tag{43}$$

B. Average Queue Length and Waiting Time

Average queue length and average waiting time can be formulated by using state probability. Average queue length, E_n for $M/M/1/N$ queue is

$$E_n = \sum_{j=1}^{N} jp_j \tag{44}$$

However, $M/M_i/3/N$ queue system has three servers and the model uses FSF, lower bound average queue length, E_n^l for $M/M_i/3/N$ will be

$$E_n^l = \sum_{j=4}^{N+3} (j-3)p_j \tag{45}$$

If $\rho = 1$ then

$$\begin{aligned} E_n^l &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \sum_{j=4}^{N+3} (j-3)\rho^j \\ &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} (1 + 2 + ... + N) \\ &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \frac{N(N+1)}{2} \end{aligned} \tag{46}$$

if $\rho \neq 1$ then

$$\begin{aligned} E_n^l &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \sum_{j=4}^{N+3} (j-3)\rho^j \\ &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \left(\rho^4 + 2\rho^5 + ... + N\rho^{N+3}\right) \\ &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \rho^4 \left(1 + 2\rho + 3\rho^2 + ... + N\rho^{N-1}\right) \\ &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \rho^4 \frac{d}{d\rho} \left(\rho + \rho^2 + \rho^3 + ... + \rho^N\right) \\ &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \rho^4 \frac{d}{d\rho} \left(\frac{\rho - \rho^{N+1}}{1-\rho}\right) \\ &= p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \rho^4 \left(\frac{1 - (N+1)\rho^N + N\rho^{N+1}}{(1-\rho)^2}\right) \end{aligned} \tag{47}$$

From equation (47) and (46), E_n^l will be

$$E_n^l = \begin{cases} p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \rho^4 \left(\frac{1-(N+1)\rho^N + N\rho^{N+1}}{(1-\rho)^2}\right) & \rho \neq 1 \\ p_0 \frac{\mu_{t3}^2}{\mu_{t1}\mu_{t2}} \frac{N(N+1)}{2} & \rho = 1 \end{cases} \tag{48}$$

From Little's law and by using equation (43) and (48), average waiting time can be formulate as

$$E_T^l = \frac{E_n^l}{\gamma^u} \tag{49}$$

It is also worth to mention that the used method for second model with three servers to find E_n^l, E_T^l, and P_B^l, can be easily be extended to n servers but it would be impossible with the method which is followed by [4].

C. Results of Model II

In this section, the quality of approximation formulas are tested with a implemented simulation. Simulation is implemented in Matlab environment based on district time simulation model. Figure 12 and 13 show E_n and E_T in queue obtained from approximation and simulation. The graphs present three different μ_2 values while $\lambda = 32$, $N = 20$, $\mu_3 = 1$, and varying μ_1 values. There are some small differences between simulation and approximations. Figure 14 and 15 show blocking probability, P_B and throughput,

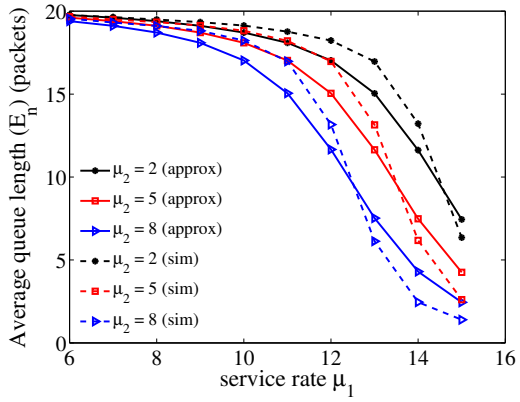

Fig. 12: Average queue length comparison between analytical and simulation.

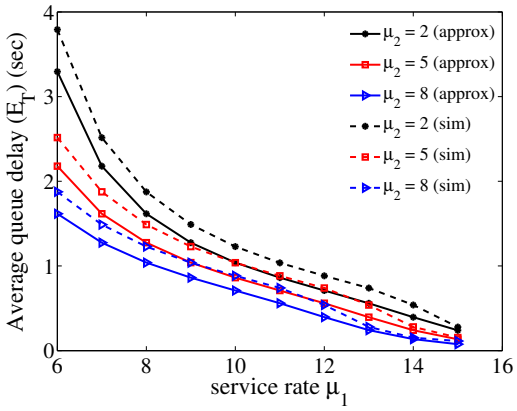

Fig. 13: Average waiting time comparison between analytical and simulation.

γ in queue obtained from simulation and approximation. The same parameters which were used to test E_n and E_T are used to test P_B and γ. Approximations results for P_B and γ almost exactly match with simulation results. The obtained results from both simulation and approximation for E_n, E_T, P_B, and γ verify the correctness of the approximation.

We also would like to see how changing in a parameter like μ_1 and N affects queue performance. The obtained results from these experiments are displayed on Figure 16, 17, 18 and 19. Figure 16 and 17 show E_n and E_T in queue for three different μ_2 values while $\lambda = 32$, $N = 20$, $m_3 = 1$, and varying μ_1 values. Because of assumptions which made for second model, displayed E_n and E_T values are lower bound results (Section II). Figure 17 follows similar path with Figure 17 with different rate because E_T depends on E_n. While μ_1 service rate increasing and higher μ_2 values, the performance of the system is better. Figure 19 shows affects of ρ on E_n. When ρ increases, E_n is getting exponentially higher.

Figure 18 shows P_B in queue while $\lambda = 32$, $\mu_1 = 1$, $\mu_2 = 5$, $m_3 = 20$, and varying N values. Displayed P_B values are lower bound results again because of assumptions which made for second

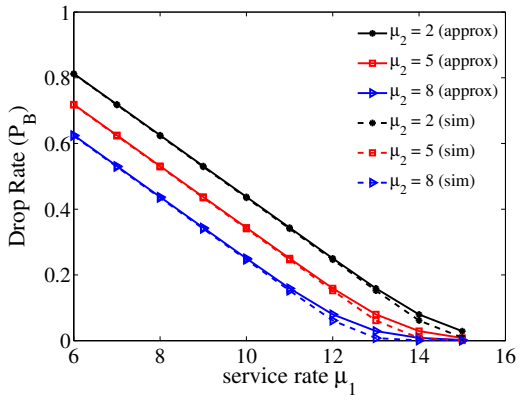

Fig. 14: Drop rate comparison between analytical and simulation.

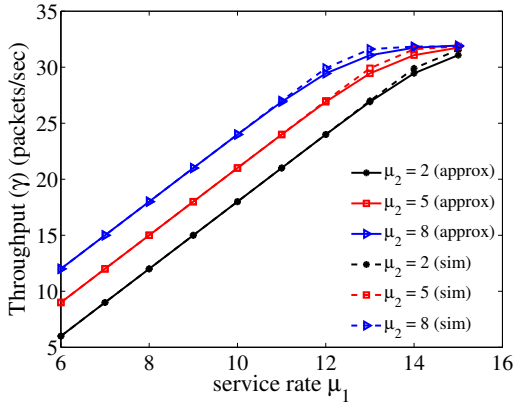

Fig. 15: Throughput comparison between analytical and simulation.

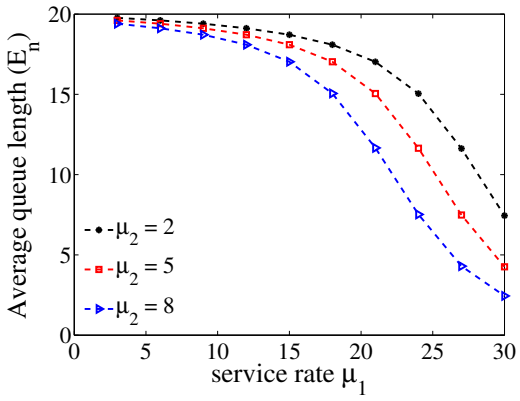

Fig. 16: Effect of μ_1 with varying μ_2 on E_n.

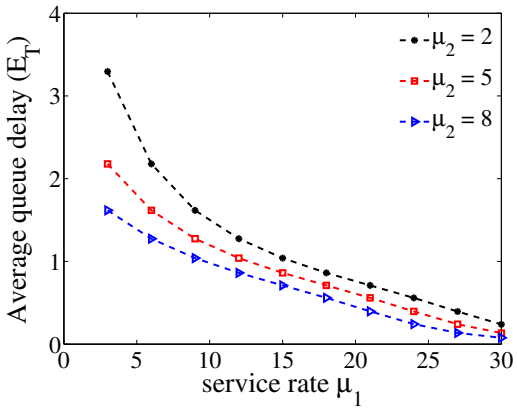

Fig. 17: Effect of μ_1 with varying μ_2 on E_T.

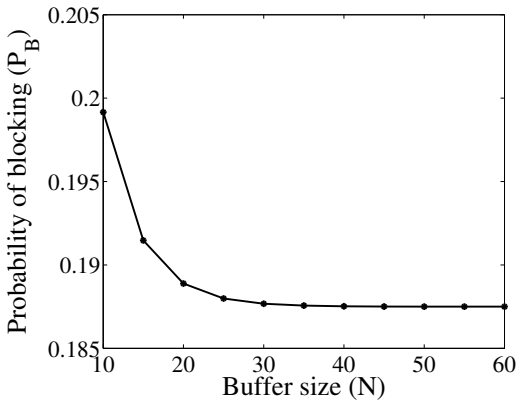

Fig. 18: Effect of buffer size on P_B.

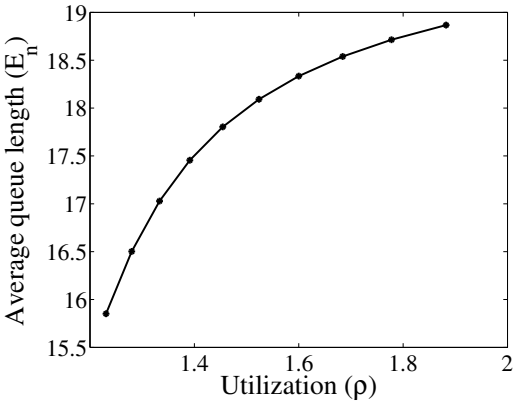

Fig. 19: Effect of ρ on E_n.

model. If the system buffer size is higher, P_B rates are getting lower but after $N = 30$, the buffer size does not significantly affect P_B. This result verifies that large buffer size is not needed in order to increase throughput [5].

V. Conclusion

Multi heterogeneous servers queuing system is encountered in every corner of real life. Therefore; it is needed to be find better bounds approximation to design efficient system. By designing the worst case approximation for $M/M_i/c$, the upper bounds approximations for average queue length and average waiting time are developed and verified by simulated results and also by designing the best case approximation for $M/M_i/3/N$, the lower bounds approximations for average queue length, average waiting time, and blocking probability are developed and verified by simulation. These analysis helps us to understand and design more efficient multi server queuing systems.

References

[1] T. Heath, B. Diniz, E. V. Carrera, W. M. Jr., and R. Bianchini, "Energy conservation in heterogeneous server clusters," in *Principles and Practice of Parallel Programming*, Chicago, Illinois, June 2005, pp. 186–195.

[2] S. Gurumurthi and S. Benjaafar, "Modeling and analysis of flexible queueing systems," *Naval Research Logistics*, vol. 51, pp. 755–782, June 2004.

[3] F. S. Q. Alves, H. C. Yehia, L. A. C. Pedrosa, F. R. B. Cruz, and L. Kerbache, "Upper bounds on performance measures of heterogeneous M/M/c queues," *Mathematical Problems in Engineering*, vol. 2011, p. 18, May 2011.

[4] C. Misra and P. K. Swain, "Performance analysis of finite buffer queueing system with multiple heterogeneous servers," in *6th international conference on Distributed Computing and Internet Technology*, ser. ICDCIT'10, Bhubaneswar, India, Feb 2010, pp. 180–183.

[5] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," *Computer Communication Review*, vol. 34, pp. 281–292, Oct 2004.