Performance Comparison of Different Queueing Systems

Burak Bozdağ
Computer Engineering Department
Istanbul Technical University
Istanbul, Türkiye
bozdagb17@itu.edu.tr

***Abstract*—In this paper, different queueing systems are compared by their performance. By determining arrival rate as constant, varying service rates are considered when comparing these systems. In this way, performances of these systems are compared equally.**

***Keywords*—queueing theory, performance, queue, queueing system, arrival, service, Poisson, delay, loss**

I. Introduction

Queueing theory is widely used in various fields such as computer science, transportation systems, manufacturing processes, telecommunications, etc. Queueing systems consist of service facilities (servers) and waiting lines (queues) that require a service from service facilities.

In this paper, different queueing systems are analyzed and compared in terms of response time, throughput, utilization, waiting time and queue length. These queueing systems include:

- M/M/1,
- M/M/2,
- M/M/3,
- Two M/M/1 queues in parallel,
- Three M/M/1 queues in parallel,
- M/M/1/m,
- M/M/2/m
- and M/M/3/m.

II. Queueing Systems Studied

A. M/M/1 Queue

This queueing system consists of a single queue and server. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed.

The utilization of the system is calculated as:

$$\rho = \lambda / \mu \tag{1}$$

The expected waiting time in queue is calculated as:

$$W_q = \rho / (\mu - \lambda) \tag{2}$$

The queue length is calculated as:

$$L_q = \rho^2 / (1 - \rho) \tag{3}$$

The expected throughput is equal to arrival rate (λ), while the expected response time is equal to $W_q + (1 / \mu)$.

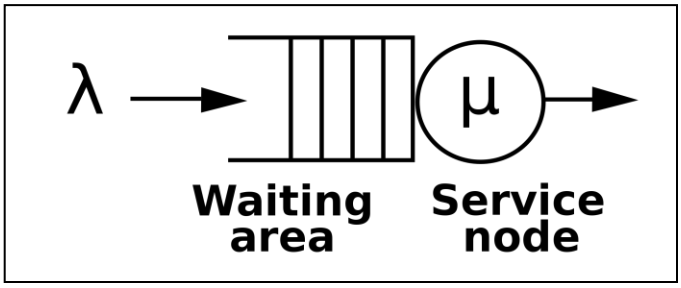

Figure 1. M/M/1 Queue.

B. M/M/2 Queue

This queueing system consists of single queue and two servers in parallel. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed.

The utilization of the system is calculated as:

$$\rho = \lambda / (2\mu) \tag{4}$$

The expected waiting time in queue is calculated as:

$$W_q = P_1 / (2\mu - \lambda) \tag{5}$$

P_1 in the formula (5) is calculated with the Erlang C formula where c is number of processes and a means the workload (λ/μ):

$$P_1 = \frac{a^c c}{c!(c-a)} \sum_{k=0}^{c-1} \left(\frac{a^k}{k!} + \frac{a^c c}{c!(c-a)} \right)^{-1} \tag{6}$$

The queue length is calculated as:

$$L_q = P_1\lambda / (2\mu - \lambda) \tag{7}$$

The expected throughput is equal to double arrival rate (2λ), while the expected response time is equal to $W_q + (1 / \mu)$.

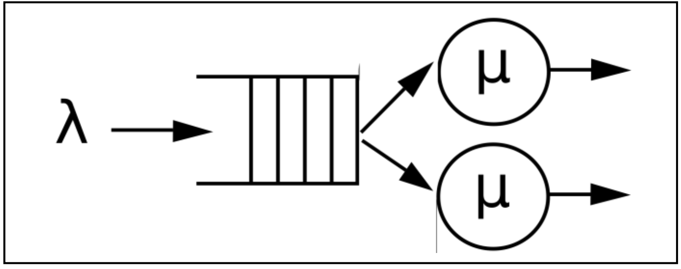

Figure 2. M/M/2 Queue.

C. M/M/3 Queue

This queueing system consists of single queue and three servers in parallel. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed.

The utilization of the system is calculated as:

$$\rho = \lambda / (3\mu) \tag{8}$$

The expected waiting time in queue is calculated as:

$$W_q = P_1 / (3\mu - \lambda) \tag{9}$$

P_1 is calculated with Erlang C formula (6). The queue length is calculated as:

$$L_q = P_1\lambda / (3\mu - \lambda) \tag{10}$$

The expected throughput is equal to triple arrival rate (3λ), while the expected response time is equal to $W_q + (1 / \mu)$.

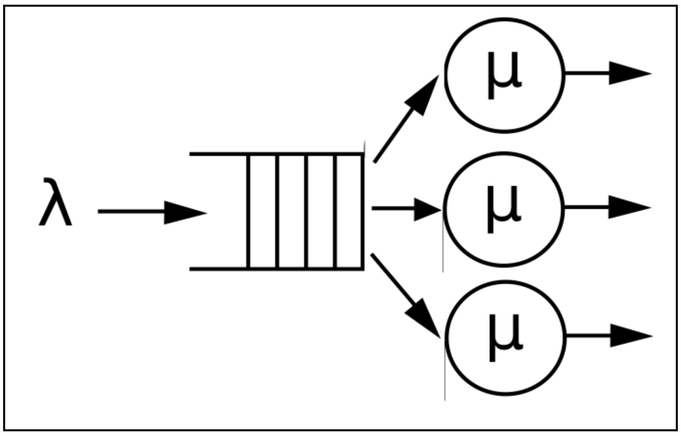

Figure 3. M/M/3 Queue.

D. Two Parallel M/M/1 Queues

This queueing system consists of two queues and two servers in parallel. It can be represented as two independent M/M/1 queues in parallel. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed.

The formulations can be considered same as a single M/M/1 queue except the arrival rate is divided by two.

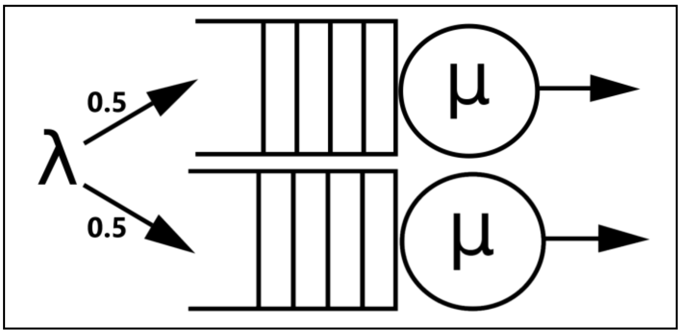

Figure 4. Two M/M/1 Queues in Parallel.

E. Three Parallel M/M/1 Queues

This queueing system consists of three queues and three servers in parallel. It can be represented as three independent M/M/1 queues in parallel. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed.

The formulations can be considered same as a single M/M/1 queue except the arrival rate is divided by three.

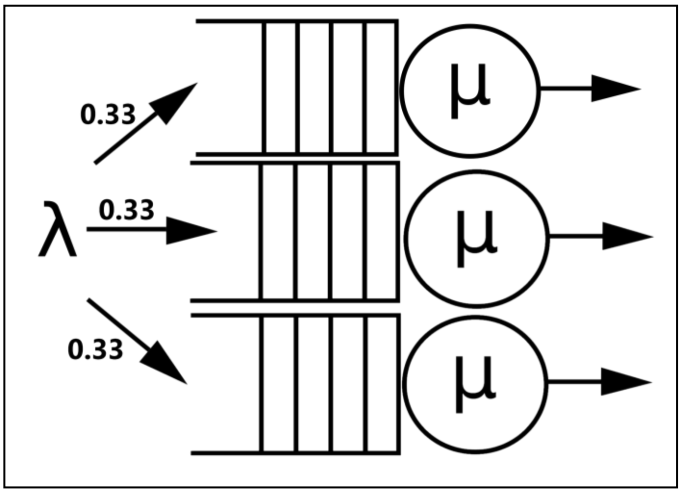

Figure 5. Three M/M/1 Queues in Parallel.

F. M/M/1/m Queue

This queueing system consists of a single queue and server. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed. In addition to the M/M/1 queue, the parameter "m" means the maximum number of customers (packets) that can be in the queueing system at any given time.

The formulations can be considered same as a single M/M/1 queue but with the limitation of $L_q < m$.

G. M/M/2/m Queue

This queueing system consists of a single queue and two servers. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed. In addition to the M/M/2 queue, the parameter "m" means the maximum number of customers (packets) that can be in the queueing system at any given time.

The formulations can be considered same as a single M/M/2 queue but with the limitation of $L_q < m$.

H. M/M/3/m Queue

This queueing system consists of a single queue and three servers. Arrival rate (λ) is determined with a Poisson process and service time (1/μ) is exponentially distributed. In addition to the M/M/3 queue, the parameter "m" means the maximum number of customers (packets) that can be in the queueing system at any given time.

The formulations can be considered same as a single M/M/3 queue but with the limitation of $L_q < m$.

III. Performance Evaluation

When evaluating the performance, the arrival rate is determined constantly for all queueing systems. Varying parameters for the performance evaluation include:

- The service rate (μ)
- The buffer space meaning that the queue is limited with a specified number of packets

With changing these values, the evaluation of performance between different queueing systems is performed.

A. Delay Comparison

In terms of response time, ones can be compared with utilization-delay charts.

Charts for queueing systems are given below.

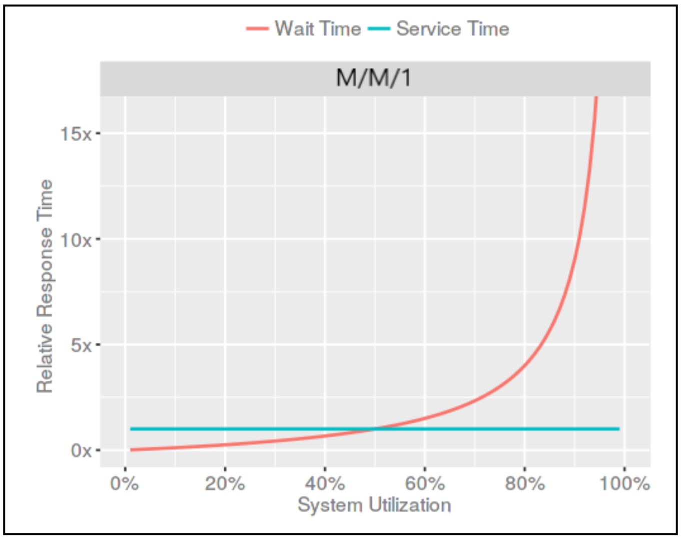

Figure 6. M/M/1 Utilization-Delay Chart.

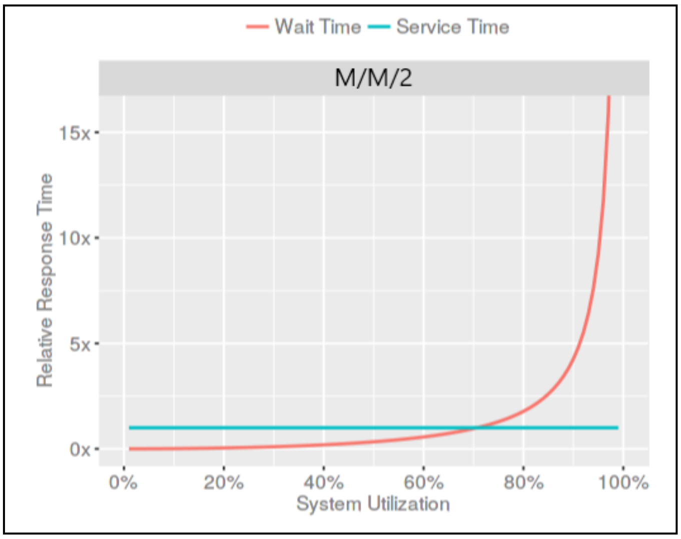

Figure 7. M/M/2 Utilization-Delay Chart.

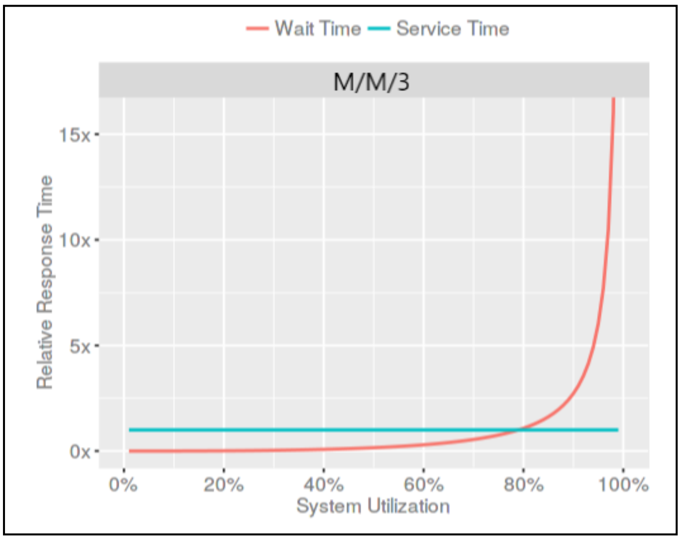

Figure 8. M/M/3 Utilization-Delay Chart.

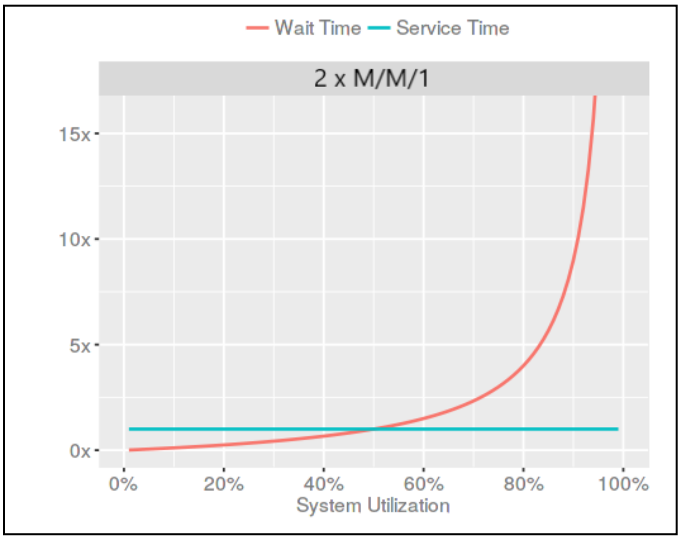

Figure 9. Two Parallel M/M/1Utilization-Delay Chart.

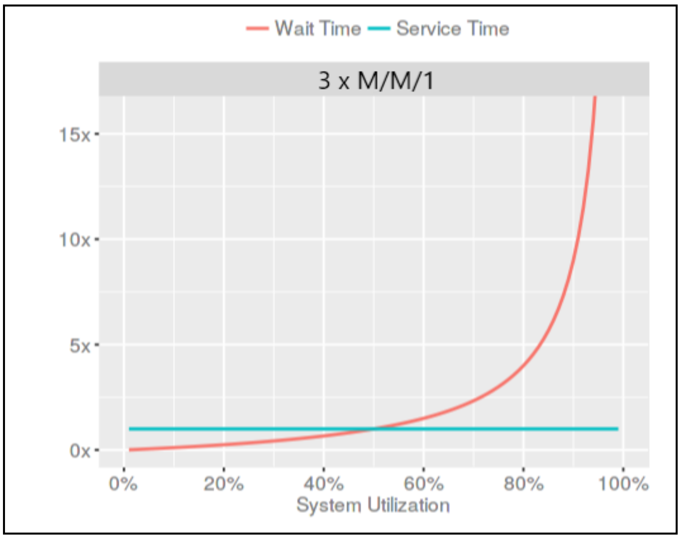

Figure 10. Three M/M/1 Utilization-Delay Chart.

For more comparative perspective, we can compare ones with the same service rate.

The delay comparison between M/M/2, two parallel M/M/1 and M/M/1 is given below. Arrival and service rates are assumed equally as they are black box systems.

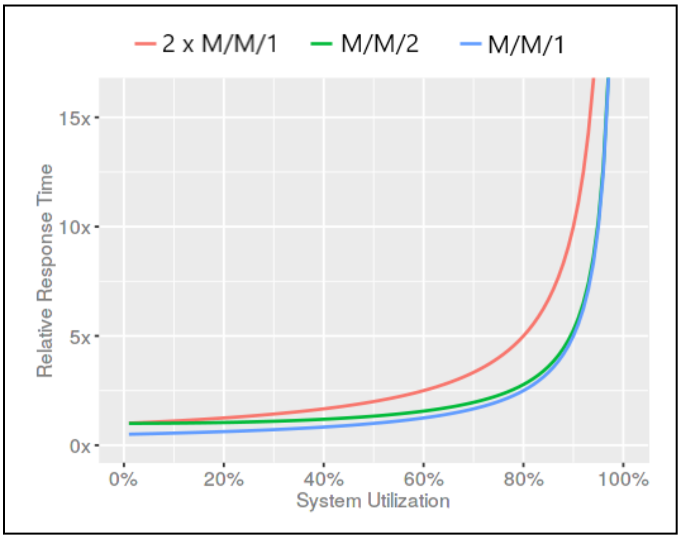

Figure 11. Comparison of M/M/1 and M/M/2.

The delay comparison between M/M/3, three parallel M/M/1 and M/M/1 is given below. Arrival and service rates are assumed equally as they are black box systems.

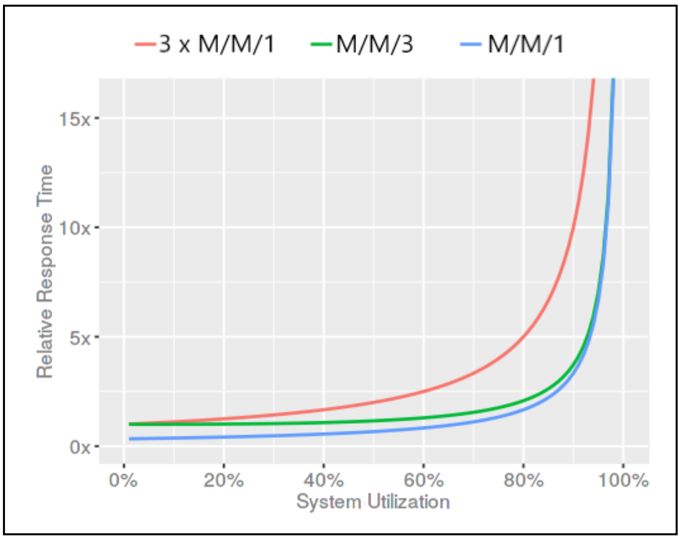

Figure 12. Comparison of M/M/1 and M/M/3.

B. *Loss Comparison*

As for loss comparison, since M/M/1, M/M/2 and M/M/3 queues have no limit on their queues, there will not be any packet loss for these systems theoretically.

The ones to compare is M/M/1/m, M/M/2/m and M/M/3/m queueing systems. Relative response time comparisons are also valid for these systems.

With high queue limits, the loss values reach to zero. As the parameter “m” decreases, the advantage of M/M/3 over M/M/2 and M/M/2 over M/M/1 becomes more visible.

All systems have a single queue but different number of servers. Ones with more servers will be able to process packets from queue faster.

The loss rate for these queues can be written as M/M/1 > M/M/2 > M/M/3.

IV. Results

As a result, the performance between different queueing systems is shown. Usually, when the number of servers is high, the delay and loss rate get lower.

For delay comparison between M/M/1, M/M/2 and 2xM/M/1 queues:

- Fast (low utility): M/M/1 < M/M/2 = 2xM/M/1
- Medium (mid utility): M/M/1 < M/M/2 < 2xM/M/1
- Slow (high utility): M/M/2 < M/M/1 < 2xM/M/1

For delay comparison between M/M/1, M/M/3 and 3xM/M/1 queues:

- Fast (low utility): M/M/1 < M/M/3 = 3xM/M/1
- Medium (mid utility): M/M/1 < M/M/3 < 3xM/M/1
- Slow (high utility): M/M/3 < M/M/1 < 3xM/M/1

The loss rate for queueing systems with limits can be compared as M/M/1 > M/M/2 > M/M/3.

References

[1] A. Herzog, (2021). Simulation mit dem warteschlangensimulator: Mathematische Modellierung und simulation von... produktions- und logistikprozessen. GABLER.

[2] Wikimedia Foundation. (2023, April 15). M/M/C queue. Wikipedia. Retrieved April 25, 2023, from https://en.wikipedia.org/wiki/M/M/c_queue.