How does the arrival rate and service rate affect the average waiting time in a single-server queue using the M/M/1 model?*

Ruiyi Liu

March 5, 2025

1 Introduction

Waiting in queues is a universal experience, whether at supermarkets, banks, hospitals, or public transport systems. Long wait times can lead to frustration, inefficiency, and financial losses for businesses and organizations. Understanding how queues function and what factors influence waiting times is crucial for optimizing service operations and improving customer satisfaction.

Queuing Theory is a branch of mathematics that models and analyzes waiting lines. It is widely applied in industries such as telecommunications, traffic management, and healthcare to enhance efficiency. One of the simplest and most commonly studied models in queuing theory is the M/M/1 queue, which represents a system with a single server, exponential interarrival times, and exponential service times.

This essay investigates the effect of two key parameters—arrival rate (λ) and service rate (μ)—on the average waiting time in an M/M/1 queue. The research question is:

How does the arrival rate (λ) and service rate (μ) affect the average waiting time in a single-server queue using the M/M/1 model?

By exploring this question, I aim to analyze the mathematical relationships governing queue dynamics and provide insights into how businesses and service providers can reduce waiting times through strategic decisions. This study combines both theoretical analysis and simulated data to demonstrate how queuing systems respond to different inputs.

*Code and data are available at: https://github.com/zora0131/MATH_IA_FINAL.git.

2 Mathematical Background

2.1 Queueing Theory

Queueing theory is the mathematical study of waiting lines, or queues. A queueing model is designed to predict queue lengths and waiting times (Wikipedia contributors 2025b).

Queues are a type of data structure with specific rules for adding and removing elements. Queues operate on a first-in, first-out (FIFO) basis, meaning the first element added is the first one removed. A simple way to understand a queue is by imagining a line at a cafeteria: the person at the front is served first, while new arrivals join at the back. This ensures that the first person in line is served before those who arrived later, maintaining the order in which elements were added (Brilliant.org n.d.).

2.2 M/M/1 queue

In queueing theory, a branch of probability theory, an M/M/1 queue models the queue length in a system with a single server, where arrivals follow a Poisson process and service times are exponentially distributed (Wikipedia contributors 2025a).

The M/M/1 queue is a standard model in queuing theory with the following assumptions:

- Arrivals follow a Poisson process with rate λ (customers per unit time).
- Service times follow an exponential distribution with rate μ (customers served per unit time).
- One server serves customers in a first-come, first-served manner (FIFO).

In our essay, as our research question is "How does the arrival rate and service rate affect the average waiting time in a single-server queue using the M/M/1 model?", we are mostly focus on the formulae about the average time and the indicators that can represents the busyness in the system. Here are the formualea we mainly used in our essay:

- Utilization factor (ρ):

$$\rho = \frac{\lambda}{\mu},$$

  with $0 \leq \rho < 1$, ρ represents the proportion of time the service busy. This is the fundamental result from Markovian queuing model. (ScienceDirect n.d.)

- Average Waiting Time in Queue (W_q):

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

This is comes from the **Little's Theorem** and the **Poisson/exponential distribution** assumptions in a M/M/1 queue. (ISSSP n.d.)

- Average Time in Queue (W):
$$W = \frac{1}{\mu - \lambda}$$
- Little's Law:
$$L = \lambda W,$$
where L is the average number of customer in the system. (ISSSP n.d.)

3 Apporach

Since we are only focus on how does the arrival rate and service rate affect the average waiting time, we do not have to use the real life dataset. Instead, we can simulate a dataset which follow our assumptions in the M/M/1 model, and use our simulated data to indicate our research question. We used the`R`programming language (R Core Team 2022), the `dplyr` packages (Wickham et al. 2023) to simulate our data, and the `knitr` package (Xie 2023) to plot the table.

4 Data Simulation and Analysis

To test these relationships, a simulation in `R` was used to model different scenarios.

4.1 Simulation Methodology

- 100 customers were simulated. The choice of 100 customers was made to balance computational efficiency and statistical significance. A smaller sample size (e.g., 10 or 20) might result in too much variability, making it harder to identify clear trends, while a significantly larger sample (e.g., 1000 or more) would require additional computational resources without substantially improving the insights gained.
- Arrival and service times were generated using an exponential distribution, since they are exponentially distributed in the assumption of M/M/1 model.
- Different values of λ (arrival rate) and μ (service rate) were tested.

4.2 R Code Used

Here are the `R` code we used to simulate the data.

```r
set.seed(123)
library(dplyr)

simulate_queue <- function(lambda, mu, num_customers) {
  interarrival_times <- rexp(num_customers, rate = lambda)
  service_times <- rexp(num_customers, rate = mu)
  arrival_times <- cumsum(interarrival_times)
  service_start_times <- numeric(num_customers)
  service_end_times <- numeric(num_customers)

  service_start_times[1] <- arrival_times[1]
  service_end_times[1] <- service_start_times[1] + service_times[1]

  for (i in 2:num_customers) {
    service_start_times[i] <- max(arrival_times[i], service_end_times[i - 1])
    service_end_times[i] <- service_start_times[i] + service_times[i]
  }

  waiting_times <- service_start_times - arrival_times
  time_in_system <- waiting_times + service_times

  data.frame(
    Customer = 1:num_customers,
    Arrival_Time = arrival_times,
    Waiting_Time = waiting_times,
    Service_Time = service_times,
    Time_In_System = time_in_system
  )
}
```

4.3 Results and Interpretation

The simulation was run for different values of λ and μ, and the following results were obtained:

```r
library(knitr)
lambda_values <- c(3, 4, 5, 5.5, 5.9)
mu_values <- c(5, 6, 6, 6, 6)
num_customers <- 100
```

```r
results <- data.frame(
  Lambda = lambda_values,
  Mu = mu_values,
  W_q = sapply(1:length(lambda_values), function(i)
    mean(simulate_queue(lambda_values[i], mu_values[i], num_customers)$Waiting_Time))
)

kable(results, col.names = c("Arrivals per unit time",
                             "Service rate",
                             "Avg Waiting Time"),
      align = 'c')
```

Table 1: Simulate results on different lambda and mu

Arrivals per unit time	Service rate	Avg Waiting Time
3.0	5	0.2035607
4.0	6	0.1353617
5.0	6	0.2482000
5.5	6	1.2279302
5.9	6	1.9734553

From the Table 1, we can see that as λ increase and approaches μ, the waiting time increase significantly, This comfirms the formula

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

which shows that as $\lambda \to \mu$, W_q approaches infinity, indicating long wait times and system congestion.

5 Discussion

The findings of this study reveal a strong relationship between the arrival rate (λ) and the service rate (μ) in determining queue efficiency. A higher arrival rate leads to longer waiting times, especially as it nears the service rate, confirming theoretical expectations. Conversely, increasing the service rate reduces average waiting time, making the system more efficient. When λ approaches μ, the queue becomes unstable, resulting in exponentially increasing wait times. These findings highlight the importance of optimizing service capacity to prevent excessive congestion.

From a practical perspective, businesses and organizations such as banks, hospitals, and supermarkets can utilize these insights to enhance customer experience. Increasing the service rate during peak hours or adding more servers (M/M/c model) can significantly reduce wait times. For industries like call centers, ensuring that the ratio $\lambda/\mu < 0.85$ helps maintain manageable waiting times and operational efficiency.

While the study successfully models queue dynamics, certain limitations exist. The assumption of exponential distributions may not perfectly reflect real-world scenarios, where service times might follow different distributions. Future research can explore multi-server queue models (M/M/c) or prioritize certain customers based on different queueing disciplines.

Overall, this IA demonstrates the power of queuing theory in understanding and optimizing service operations. The results confirm mathematical predictions and provide valuable recommendations for real-world applications.

Reference

Brilliant.org. n.d. "Queues - Basic." https://brilliant.org/wiki/queues-basic/.

ISSSP. n.d. "Finding the Average Wait Time – Little's Law." https://isssp.org/finding-the-average-wait-time-littles-law/.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

ScienceDirect. n.d. "Traffic Intensity in Queueing Systems." https://www.sciencedirect.com/topics/mathematics/traffic-intensity.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wikipedia contributors. 2025a. "M/m/1 Queue." *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/M/M/1_queue.

———. 2025b. "Queueing Theory." *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/Queueing_theory.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.