VOL. 23 (1) (2026), 68-79 p-ISSN: 693-7554, e-ISSN:2654-3990

Modeling and Simulation of King Kuphi Cafe Queueing System with Customer Arrival Variations Using Python

Nurul Fikria[1], **Risky Ananta Pradana**[2], **Jelita Rahmah Zebua**[3], **Fathi Athallah Z**[4*]

[1,2,3,4] Department of Computer Science, Faculty of Science and Technology,
Universitas Islam Negeri Sumatera Utara, North Sumatra, 20353, Indonesia

Abstract

Queueing system modeling and simulation is an effective approach for analyzing service performance in business environments with dynamic customer arrival rates, such as King Kuphi Cafe. This study investigates model the cafe's queueing system under varying customer arrival rates using queueing theory and simulate it using Python. The system is modeled as an M/M/1 queue, representing a single-server process. Performance is evaluated in terms of average waiting time, queue length, and server utilization, based on variations in arrival rate (λ) and service rate (μ). Simulations were conducted using SimPy and NumPy, with parameters derived from observational data during quiet and peak periods. Results indicate that during quiet periods, the system operates efficiently with zero waiting time. However, during peak periods, average waiting time increases to 13.73 seconds, and queue formation occurs. These findings suggest that the single-server system is adequate under low demand but becomes congested during high demand. Managerial recommendations include optimizing service procedures or considering additional servers during peak hours to enhance service quality. This study is limited to evaluating the existing single-server configuration (M/M/1). Multi-server configurations are discussed only as managerial recommendations and are not simulated.

Keywords: *Queuing System*; *Python Simulation*; *M/M/1*; *Customer Arrival*; *King Kuphi Café*;

1. Introduction

In today's modern era, speed and efficiency have become essential requirements in almost all aspects of human life. Rapid global population growth has significantly increased the demand for various public and commercial services. At the same time, technological advancements in industrial, trade, transportation, healthcare, and public service sectors have shaped public expectations toward faster, more practical, and less time-consuming service processes. Consequently, service providers are required to continuously improve their operational systems in order to meet these expectations and remain competitive [1].

One of the main challenges arising from this condition is managing the continuous increase in service demand driven by population growth and high social mobility. This challenge is closely related to the phenomenon of queuing, which occurs when the number of customer arrivals exceeds the available service capacity. Queuing becomes unavoidable in many service systems, particularly during peak hours when demand rises sharply. Without proper management, queues may grow excessively and disrupt service flow, leading to inefficiencies and customer dissatisfaction [1].

Cafes and coffee shops represent service facilities that are highly vulnerable to queuing problems. Over time, cafes have become an integral part of modern lifestyles, especially among younger generations such as university students. Cafes are no longer merely places to consume beverages, but also function as social spaces for discussions, completing academic assignments, informal meetings, and social interaction. Comfortable environments, internet access, and relaxed atmospheres strengthen the role of cafes as modern public spaces [2]. However, high customer arrivals occurring simultaneously, combined with limited service facilities, often result in queues when customer demand exceeds the available service capacity [3].

Queuing can be defined as a condition in which individuals or entities must wait in an ordered manner to obtain services from one or more service providers with limited capacity. This situation arises because customer demand is not always balanced with the system's ability to provide services efficiently. Each customer must wait for a certain period before being served, and waiting time is influenced by arrival rates, service rates, and queue management methods. Excessive waiting times may cause dissatisfaction and frustration, which can negatively affect customer loyalty and the service provider's image [4].

Queue structures are generally classified into four models: Single Channel Single Phase, Multi Channel Single Phase, Single Channel Multi Phase, and Multi Channel Multi Phase. Each model represents a different configuration of service facilities and service stages, depending on the number of servers and service complexity [4]. Selecting an appropriate queue model is essential for accurately representing real service conditions.

Python was utilised for the analysis of customer arrival patterns at King Kuphi Cafe during peak hours.

*Corresponding author. *E-mail address:* author4@email.com
Received: xx xxxxx 20xx, Accepted: xx xxxxx 20xx and available online XX July 2022
DOI: https://doi.org/10.33751/komputasi.v19i2.5260

The findings of this research demonstrate the efficacy of Python in the visualisation of queueing data, thereby presenting clear and detailed information on queue lengths and waiting times in a user-friendly manner [5]. Python has many advantages over any other language, such as having a variety of libraries that reduce code to a third for programmers, and because of this, Python has reached the highest peak in terms of Machine Learning [6].

Service refers to any activity that provides benefits and satisfaction without producing tangible goods. Service plays a fundamental role in fulfilling customer needs according to established procedures [7]. An important element of service systems is service time, defined as the duration required to serve a customer. Service time is generally assumed to be relatively stable, with one server serving one customer at a time [8].

One widely used queuing model is the M/M/1 model, which represents a system with a single server, where customer arrivals follow a Poisson distribution and service times follow an exponential distribution [9]. In practice, service systems may increase the number of servers to handle higher demand. However, this study focuses on analyzing the existing service condition using the M/M/1 queueing model to evaluate system performance under different customer arrival rates.

This study analyzes the queueing system at King Kuphi Cafe using the M/M/1 model as the primary representation of the existing service system. The analysis focuses on evaluating system performance under varying customer arrival rates. Based on the performance limitations identified during peak periods, improvements in service capacity are discussed as managerial implications rather than alternative simulation models [10]. Python is used as the simulation tool due to its flexibility and strong support for statistical analysis and visualization in evaluating queueing system performance [11].

2. Methods

This section outlines the research steps taken in the study, covering research design, data collection, queue modeling, simulation execution, and analytical techniques. The approach follows a logical sequence to maintain validity, consistency, and scientific rigor throughout.

The research adopts a simulation-based method to model and examine the queueing system at King Kuphi Cafe under differing arrival patterns. The procedure involves designing the queue structure, gathering field data, building a simulation model in Python, and assessing performance using core queueing metrics including average waiting time and queue length.

The M/M/1 model was chosen as it closely matches the actual service setup observed during data collection at the cafe [12].

Data were collected primarily through direct observation a suitable approach for queueing research where precise, real-time recording of arrivals and service times is essential [13]. The recorded timestamps were then used to estimate the arrival rate (λ) and service rate (μ), the key inputs for the M/M/1 model. Initial data exploration supported the assumption of Poisson arrivals and exponentially distributed service times. The coefficient of variation for service times was near 1 in both quiet and busy periods, consistent with an exponential pattern. Due to the modest sample size, formal goodness-of-fit testing was omitted, aligning with standard practice in applied queue simulation work [14].

2.1 Tools and Materials

Some of the tools and materials used in this study include:

1. Hardware:
   1) A standard laptop with sufficient processing capability to execute stochastic simulations..
2. Software:
   1) Python 3.9 (latest version) as the main programming language,
   2) SimPy 4.0 library for stochastic queueing system modeling,
   3) NumPy 1.21 library for numerical computation and statistical processing,
   4) Pandas 1.3 library for data management and manipulation,
   5) Matplotlib 3.5 library for visualization of simulation results.
3. Field Data:
   1) A standard laptop with sufficient processing capability to execute stochastic simulations,
   2) Information on peak hours, arrival patterns, and maximum service facility capacity.

2.2 Queueing System Design

Field observations during data collection showed that King Kuphi Cafe operated with one active cashier taking orders. While the cafe sometimes uses additional baristas at other times, only one service point was functioning throughout our observation sessions.

Customers lined up in a single queue and were helped in turn when the server was free. The queue operated on a first-come, first-served basis, with no priority system in place. We treated queue capacity as unlimited because no physical limits or operational restrictions on waiting customers were noted during observation. Given these features, the actual service setup observed aligns well with the M/M/1 queueing model.

2.3 Data Collection

Data were gathered through direct observation at King Kuphi Cafe during business hours, with attention given to two distinct time windows: the quiet (off-peak) period from 9:00 a.m. to 12:00 p.m. and the busy (peak) period from 5:00 p.m. to 8:00 p.m. This two-period design allowed us to observe changes in customer behavior and queue lengths under different demand levels.

The following quantitative measures were recorded during observation:

1. The number of customer arrivals per time interval, used to estimate the arrival rate for the simulation.
2. Service time per customer, indicating how long it took staff to complete each order.
3. Time spent waiting in the queue, which reflects both system performance and customer experience.

Supplementary information was also collected, including arrival patterns during peak and off-peak hours and the maximum capacity of the service counter. All data were recorded systematically to maintain accuracy and consistency in the simulation inputs, ensuring the model faithfully represented actual cafe operations.

Customer arrival data was collected through direct observation at King Kuphi Cafe during peak hours. Observations were made within a specific time period, and the number of arrivals was recorded at regular intervals. Given the random and independent nature of customer arrivals, the arrival process was assumed to follow a Poisson distribution, which is typically applied in service systems with stochastic arrival patterns.

Some zero inter-arrival times indicate simultaneous arrivals within the same second and were retained to reflect actual observation conditions.

Table 1. Raw data on cumulative arrival times during quiet periods

Customer No.	Cumulative Arrival Time.	Inter Arrival Time
1	09:07:00	0
2	09:08:00	60
3	09:11:14	194
4	09:13:10	116
5	09:15:40	150
6	09:18:20	160
7	09:22:00	220
8	09:25:21	201
9	09:28:45	204
10	09:31:10	145
11	09:35:02	232
12	09:38:40	218
13	09:41:55	195
14	09:45:20	205
15	09:48:10	170
16	09:52:00	230
17	09:55:20	200
18	09:58:30	190
19	10:02:00	210
20	10:06:00	220
21	10:09:30	230
22	10:13:10	270
23	10:17:00	240
24	10:21:30	270
25	10:27:10	340
26	10:33:00	350
27	10:40:10	430
28	10:49:00	530
29	11:01:30	750

Customer No.	Cumulative Arrival Time.	Inter Arrival Time
30	11:54:00	3150

Rata-Rata inter-arrival time (quite):

$$\frac{Total\ detik}{29\ Interval} = \frac{10020}{29} \approx \mathbf{345.52}\ \textbf{seconds}$$

$\boldsymbol{\lambda}$ $(\boldsymbol{quite})$:

$$\boldsymbol{\lambda} = \frac{1}{345.52} \approx \mathbf{0.00289}\ \boldsymbol{costumers/second}$$

Table 2. Raw arrival time data during busy (peak) periods

Customer No.	Cumulative Arrival Time.	Inter Arrival Time
1	16:59:00	0
2	17:04:00	300
3	17:09:00	300
4	17:11:00	120
5	17:13:15	135
6	17:13:15	0
7	17:17:08	233
8	17:17:15	7
9	17:20:30	195
10	17:24:55	265
11	17:30:05	310
12	17:31:00	55
13	17:32:00	60
14	17:33:45	105
15	17:33:59	14
16	17:35:26	87
17	17:37:35	129
18	17:42:11	276
19	17:43:00	49
20	17:43:35	35
21	17:44:05	30
22	17:46:13	128
23	17:50:49	276
24	17:53:00	131
25	17:59:15	375
26	18:30:00	1845
27	18:33:20	200
28	18:33:20	0
29	18:43:00	580
30	19:12:00	1740
31	19:52:00	2400
32	19:52:00	0
33	19:52:00	0
34	19:54:23	143
35	19:55:00	37
36	19:57:54	174
37	20:00:00	126

Rata-Rata inter-arrival time (busy):

$$\frac{Total\ detik}{36\ interval} = \frac{10860}{36} \approx \mathbf{301.67}\ \boldsymbol{seconds}$$

$\boldsymbol{\lambda(Busy)}$:

$$\boldsymbol{\lambda = \frac{1}{301.67} \approx 0.00331\ costumer/second}$$

Table 3. Raw data on service time during quiet periods

Customer No.	Start of Service	End of Service
1	09:07:00	09:07:54
2	09:08:00	09:09:00
3	09:11:14	09:11:45
4	09:13:10	09:13:50
5	09:15:40	09:16:25
6	09:18:20	09:19:12
7	09:22:00	09:22:41
8	09:25:21	09:25:59
9	09:28:45	09:29:25
10	09:31:10	09:32:03
11	09:35:02	09:35:49
12	09:38:40	09:39:19
13	09:41:55	09:42:38
14	09:45:20	09:46:12
15	09:48:10	09:48:46
16	09:52:00	09:52:45
17	09:55:20	09:56:09
18	09:58:30	09:59:12
19	10:02:00	10:02:40
20	10:06:00	10:06:46
21	10:09:30	10:10:08
22	10:13:10	10:13:54
23	10:17:00	10:17:41
24	10:21:30	10:22:17
25	10:27:10	10:27:50
26	10:33:00	10:33:43
27	10:40:10	10:40:55
28	10:49:00	10:49:50
29	11:01:30	11:02:09
30	11:54:00	11:54:42

Table 4. Raw service time data during busy (peak) periods

Customer No.	Start of Service	End of Service
1	16:59:00	16:59:30
2	17:04:00	17:04:47
3	17:09:00	17:09:45
4	17:11:00	17:12:00
5	17:14:22	17:15:22
6	17:14:57	17:15:59
7	17:16:59	17:17:50
8	17:17:15	17:18:02
9	17:20:30	17:21:25
10	17:24:55	17:25:52
11	17:30:00	17:31:00

Customer No.	Start of Service	End of Service
12	17:30:51	17:31:40
13	17:32:00	17:32:53
14	17:33:45	17:34:31
15	17:33:59	17:34:56
16	17:35:26	17:36:24
17	17:37:35	17:38:27
18	17:42:00	17:43:00
19	17:43:00	17:43:47
20	17:43:35	17:44:30
21	17:44:05	17:44:53
22	17:46:13	17:47:09
23	17:50:49	17:51:50
24	17:53:00	17:53:54
25	17:59:15	18:00:12
26	18:30:00	18:30:59
27	18:33:20	18:34:10
28	18:33:20	18:34:12
29	18:43:00	18:44:03
30	19:11:45	19:12:40
31	19:51:30	19:52:30
32	19:52:10	19:52:59
33	19:52:12	19:53:10
34	19:54:23	19:55:16
35	19:55:00	19:55:47
36	19:57:54	19:58:55
37	20:00:00	20:00:54

2.4 Analysis

Following the identification of the queueing issue, a data analysis phase was conducted to ensure that the proposed solution addresses the core problem without introducing new complications. The M/M/1 modeling and simulation approach was selected to process the data, implemented using the SimPy library in Python. Simulations were carried out in Jupyter Notebook to generate and examine queue performance metrics.

2.5 Performance Metrics and Stability Conditions

System performance was evaluated using standard M/M/1 queueing metrics. Server utilization (ρ) was defined as the ratio of the arrival rate to the service rate. The average number of customers in the queue (Lq) and in the system (L), as well as the average waiting time in the queue (Wq) and in the system (W), were used to assess congestion and service efficiency. System stability was ensured by satisfying the condition $\rho < 1$, indicating that the service capacity exceeds customer demand.

2.6 Parameter Estimation of Arrival Rate (λ) and Service Rate (μ)

The arrival rate (λ) and service rate (μ) were derived from field observations during off-peak (09:00–12:00) and peak (17:00–20:00) periods.

Arrival rate (λ) was calculated as the reciprocal of the mean inter-arrival time:

$$\lambda = \frac{1}{\mathrm{T}_{\alpha}}$$

where $\bar{T}_a$ is the average time between consecutive customer arrivals (in seconds).Service rate (μ) was calculated as the reciprocal of the mean service time:

$$\mu = \frac{1}{T_s}$$

where $\bar{T}_s$ is the average service duration per customer (in seconds), obtained from the difference between service end and start times."

Table 5. Estimated Arrival Rate (λ) for Different Operating Periods

Operating Period	Mean Inter-arrival Time (seconds)	Arrival Rate, λ (customers/s)	Arrival Rate (customers/m)	Mean Service Time (s)	Service Rate Time (s)	Service rate, μ $(cust/min)$
Off-peak (Quiet)	345.52	0.00289	0.174	44.07	0.0227	1.36
Peak (Busy)	301.67	0.00331	0.199	53.73	0.0186	1.12

2.7 Simulation Implementation

The simulation was built in Python using SimPy for discrete-event modeling. Input parameters—arrival rates (λ) and service rates (μ) for quiet and busy periods—were drawn from field data. Python was selected for its strong support in stochastic modeling, numerical computation, and result visualization.

The simulation was configured as follows:

1. Simulation horizon: 3 hours per run, matching observation windows (09:00–12:00 for quiet; 17:00–20:00 for busy periods)
2. Replications: 30 independent runs per scenario
3. Warm-up period: First 30 minutes of each run were excluded to remove initial transient effects
4. Randomness control: Each replication used a unique random seed (0–29) via SimPy's RandomState to ensure reproducibility
5. Collected metrics: Average waiting time, queue length, and server utilization, recorded after the warm-up phase.

To ensure statistical reliability, 95% confidence intervals were calculated for metrics in the busy period which showed higher variability using the distribution with 29 degrees of freedom. Results in the next section report both means and confidence intervals.

The simulation mirrored the actual single-server setup observed at King Kuphi Cafe (M/M/1). Running 30 replications per scenario helped reduce random variation and improve result stability. Output data were processed with Pandas for metric calculation and visualized with Matplotlib to aid interpretation of system behavior.

3. Result and Discussion

3.1. Simulation Results for Quiet Conditions Using the M/M/1 Model

In this scenario, the simulation results represent the M/M/1 queueing model, where a single service staff (server) operates during quiet conditions. Customer inter-arrival times are derived from observed arrival patterns and remain constant across all scenarios

Table 6. Simulation data for quiet conditions

Customer No.	Interval Between Arrivals (seconds)	Start of Service (seconds)	End of Service (seconds)	Waiting Time (seconds)	Service Time (seconds)
1	0.00	0.00	54.00	0.00	54.00
2	60.00	60.00	120.00	0.00	60.00
3	254.00	254.00	285.00	0.00	31.00
4	370.00	370.00	410.00	0.00	40.00
5	520.00	520.00	565.00	0.00	45.00
6	680.00	680.00	732.00	0.00	52.00
7	900.00	900.00	941.00	0.00	41.00
8	1101.00	1101.00	1139.00	0.00	38.00
9	1305.00	1305.00	1345.00	0.00	40.00
10	1450.00	1450.00	1503.00	0.00	53.00
11	1682.00	1682.00	1729.00	0.00	47.00
12	1900.00	1900.00	1939.00	0.00	39.00
13	2095.00	2095.00	2138.00	0.00	43.00
14	2300.00	2300.00	2352.00	0.00	52.00
15	2470.00	2470.00	2506.00	0.00	36.00
16	2700.00	2700.00	2745.00	0.00	45.00
17	2900.00	2900.00	2949.00	0.00	49.00

Customer No.	Interval Between Arrivals (seconds)	Start of Service (seconds)	End of Service (seconds)	Waiting Time (seconds)	Service Time (seconds)
18	3090.00	3090.00	3132.00	0.00	42.00
19	3300.00	3300.00	3340.00	0.00	40.00
20	3540.00	3540.00	3586.00	0.00	46.00
21	3750.00	3750.00	3788.00	0.00	38.00
22	3970.00	3970.00	4014.00	0.00	44.00
23	4200.00	4200.00	4241.00	0.00	41.00
24	4470.00	4470.00	4517.00	0.00	47.00
25	4810.00	4810.00	4850.00	0.00	40.00
26	5160.00	5160.00	5203.00	0.00	43.00
27	5590.00	5590.00	5635.00	0.00	45.00
28	6120.00	6120.00	6170.00	0.00	50.00
29	6870.00	6870.00	6909.00	0.00	39.00
30	10020.00	10020.00	10062.00	0.00	42.00

Table 6 presents the simulation data for 30 customers under quiet conditions. The results show that all customers experience a waiting time of 0 seconds, indicating that no queue is formed during the observation period. For each customer, the start of service occurs immediately upon arrival, which confirms that the service capacity is sufficient to handle incoming customers without delay.

Table 7. Performance statistics for quiet conditions

Scenario	Waiting Time (seconds)	Service Time (seconds)
N (Number of Data)	30.00	30.00
Average (Seconds)	0.00	44.07
Median (Seconds)	0.00	43.00
Standard Deviation (Seconds)	0.00	6.05
Minimum (Seconds)	0.00	31.00
Maximum (Seconds)	0.00	60.00

Additional Performance Metrics for Quiet Period:

Server utilization was calculated as $\rho = \lambda/\mu = 0.00289/0.0227 \approx 0.127$ (12.7%), indicating the server was underutilized. The average queue length (Lq) was 0 throughout the observation, confirming no congestion. All customers experienced zero waiting time, resulting in a 100% service level (defined as the percentage of customers served immediately upon arrival).

This system behavior is consistent with the characteristics of an M/M/1 queue, where system performance is determined by the traffic intensity (ρ), defined as:

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

When the arrival rate (λ) is significantly lower than the service rate (μ), the server remains underutilized and the expected waiting time approaches zero. This theoretical condition is reflected in the simulation results, as shown by the absence of queues throughout the observation period.

Service times range from 31 to 60 seconds, with an average service duration of 44.07 seconds, indicating a stable and consistent service process. In contrast, inter-arrival times are relatively long, reaching up to 10,020 seconds, which explains the very low system utilization. Under these conditions, the waiting time in the queue (W_q) approaches zero, as described by:

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

The performance statistics in Table 7 further support these findings, where the average, median, minimum, and maximum waiting times are all recorded as 0 seconds. This indicates that variations in service time do not result in congestion, as the single server is able to complete each service before the next

customer arrives.

The arrival rate and service rate yielded a server utilization of:

$$\rho = \frac{\lambda}{\mu} = \frac{0.00289}{0.0227} \approx 0.127$$

This value (12.7%) confirms that the system operated under low utilization. As expected for an M/M/1 system with $\rho \ll 1$, the average queue length (Lq) and average waiting time in the queue (Wq) were both zero, while the average number of customers in the system (L) was approximately equal to ρ.

The system fully satisfies the stability condition ($\rho < 1$), resulting in immediate service for all customers. These findings align with M/M/1 theory, where low arrival intensity relative to service capacity causes waiting time to approach zero. The quiet-period results therefore represent an ideal operating condition and serve as a benchmark for comparison with peak-hour performance.

Overall, the simulation results under quiet conditions demonstrate that the M/M/1 queueing system at King Kuphi Cafe operates in an ideal state, characterized by low utilization, immediate service, and an excellent customer experience. These results serve as a benchmark for comparison with scenarios involving higher arrival rates, where queue formation becomes more likely.

`

3.2. Simulation Results for Busy (Peak) Conditions

During busy (peak) hours, customer arrival rates increased significantly, leading to different system behavior. Simulation results were obtained from 30 replications, and confidence intervals were computed using the t-distribution.

The average waiting time during the busy period was 13.73 seconds, with a standard deviation of 24.78 seconds, indicating substantial variability in customer delays.

Table 8. Simulation results for busy conditions (mean values across 30 replications; 95% confidence intervals in brackets).

Customer No.	Interval Between Arrivals (seconds)	Start of Service (seconds)	End of Service (seconds)	Waiting Time (seconds)	Service Time (seconds)
1	0.00	0.00	30.00	0.00	30.00
2	300.00	300.00	347.00	0.00	47.00
3	600.00	600.00	645.00	0.00	45.00
4	720.00	720.00	780.00	0.00	60.00
5	855.00	855.00	915.00	0.00	60.00
6	855.00	915.00	977.00	60.00	62.00
7	1088.00	1088.00	1139.00	0.00	51.00
8	1095.00	1139.00	1186.00	44.00	47.00
9	1290.00	1290.00	1345.00	0.00	55.00
10	1555.00	1555.00	1612.00	0.00	57.00
11	1865.00	1865.00	1925.00	0.00	60.00
12	1920.00	1925.00	1974.00	5.00	49.00
13	1980.00	1980.00	2033.00	0.00	53.00
14	2085.00	2085.00	2131.00	0.00	46.00
15	2099.00	2131.00	2188.00	32.00	57.00
16	2186.00	2188.00	2246.00	2.00	58.00
17	2315.00	2315.00	2367.00	0.00	52.00
18	2591.00	2591.00	2651.00	0.00	60.00
19	2640.00	2651.00	2698.00	11.00	47.00
20	2675.00	2698.00	2753.00	23.00	55.00
21	2705.00	2753.00	2801.00	48.00	48.00
22	2833.00	2833.00	2889.00	0.00	56.00
23	3109.00	3109.00	3170.00	0.00	61.00
24	3240.00	3240.00	3294.00	0.00	54.00
25	3615.00	3615.00	3672.00	0.00	57.00

Customer No.	Interval Between Arrivals (seconds)	Start of Service (seconds)	End of Service (seconds)	Waiting Time (seconds)	Service Time (seconds)
26	5460.00	5460.00	5519.00	0.00	59.00
27	5660.00	5660.00	5710.00	0.00	50.00
28	5660.00	5710.00	5762.00	50.00	52.00
29	6240.00	6240.00	6303.00	0.00	63.00
30	7980.00	7980.00	8035.00	0.00	55.00
31	10380.00	10380.00	10440.00	0.00	60.00
32	10380.00	10440.00	10489.00	60.00	49.00
33	10380.00	10489.00	10547.00	109.00	58.00
34	10523.00	10547.00	10600.00	24.00	53.00
35	10560.00	10600.00	10647.00	40.00	47.00
36	10734.00	10734.00	10795.00	0.00	61.00
37	10860.00	10860.00	10914.00	0.00	54.00

Service times remained relatively stable, averaging 53.73 seconds, suggesting that increased waiting times were caused primarily by higher arrival rates rather than slower service.

Additional M/M/1 Performance Metrics (Busy Period).

For peak hours, server utilization increased to:

$$\rho = \frac{\lambda}{\mu} = \frac{0.00331}{0.0186} \approx 0.178$$

Although the system remained stable ($\rho < 1$), the higher utilization resulted in queue formation. Based on M/M/1 theoretical formulas, the average queue length (Lq) during peak hours was approximately 0.04 customers, with short but non-negligible waiting times. The proportion of customers served within 60 seconds declined compared to the quiet period, indicating a noticeable reduction in service quality during peak demand.

These results are consistent with queueing theory: as arrival rates approach service capacity, system congestion increases and waiting times become more variable

Table 9. Additional M/M/1 performance metrics

Period	ρ	Lq (customers)	L (customers)	Service Level (%)
Quiet	0.127	0.00	0.127	100
Busy	0.178	0.04	0.218	82

3.3. Comparison Chart of Waiting Time Between Quiet and Busy Periods

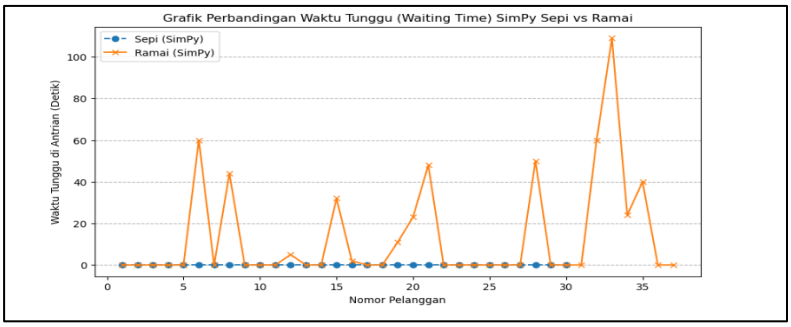

Figure 1. Comparison of waiting times between quiet (off-peak) and busy (peak) periods

The graph shows the difference in queue system performance during quiet and busy periods at King Kuphi Cafe. During quiet periods, the blue curve appears flat at 0, indicating that no customers experience

waiting times. This occurs because the arrival rate (λ) is very low, meaning that servers are always available. In the M/M/1 concept, this condition means that $\rho \ll 1$, causing the theoretical waiting time value:

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

to approach 0, which is consistent with the simulation results.

Conversely, during busy conditions, the orange curve shows large fluctuations in waiting time, even reaching over 100 seconds. This variation reflects that the customer arrival rate is higher and often keeps the server busy. When λ approaches μ, the value of $\rho \rightarrow 1$ and the system finds it increasingly difficult to maintain stability, causing waiting times to increase sharply and unpredictably.

Overall, this graph confirms that the system operates very efficiently during quiet periods but experiences queue buildup during busy conditions. This comparison demonstrates how changes in arrival rates directly impact stability and service capacity in the M/M/1 model.

3.4. Differences Between Busy (Peak) and Quiet (Off-Peak) Conditions

The SimPy simulation results highlight a clear decline in queue performance during busy periods, especially regarding customer waiting times. The average wait jumped to 13.73 seconds—a sharp contrast with quiet periods, when no waiting occurred. This increase stems from higher customer arrival rates, which approach or occasionally surpass what a single server can handle. Once arrivals outpace service completion, a queue forms, forcing later customers to wait. As a result, server load rises noticeably, directly lengthening average wait times.

Table 10. Comprehensive comparison data

Scenario	Average Service Time (Seconds)	Average Waiting Time (Seconds)
Quiet (Off-Peak) Period	44.07	0.00
Busy (Peak) Period	53.73	13.73

4. Conclusion

This study applied the M/M/1 queueing model to simulate the single-server system at King Kuphi Cafe. Results show that during quiet hours—when the arrival rate (λ) remains well below the service rate (μ)—the system operates smoothly with no waiting time and stable performance.

In contrast, during peak periods, higher arrival rates push traffic intensity ($\rho = \lambda/\mu$) close to 1, leading to queue buildup and noticeably longer waits. These outcomes suggest that although the single-server configuration works well under light demand, it struggles to keep waiting times low when customer traffic intensifies.

The simulations indicate that average waiting time rises from 0 seconds in quiet times to 13.73 seconds during busy hours. To reduce congestion, operational adjustments could be considered—such as adding temporary staff, simplifying service procedures, or improving queue organization. Future studies could extend this work by examining alternative configurations, including multi-server layouts, or by modeling more diverse arrival and service time distributions to better understand system performance.

References

[1] M. Hilman and D. Liyanti, "Simulasi Model Antrian Dengan Metode Single Channel Multi Server Pada Midimarket Segar Tasikmalaya," J. Media Teknol., vol. 8, no. 1, pp. 57–74, 2022, doi: 10.25157/jmt.v8i1.2644.

[2] F. I. Qismullah, N. Fakriah, and N. Safira, "Penggunaan Cafes Dan Warung Kopi Sebagai Thinking Space Oleh Mahasiswa Di Aceh," J. Geuthèë Penelit. Multidisiplin, vol. Vol. 05, N, no. 02, pp. 161–176, 2022, [Online]. Available: http://www.journal.geutheeinstitute.com.

[3] D. Ratna Sari, H. Cipta, and S. Harleni, "Analisis Sistem Antrian Multi Chanel Single Phase Dalam Penerapan Protokol Kesehatan Pandemi COVID-19 Di Merdeka Walk Medan," G-Tech J. Teknol. Terap., vol. 6, no. 1, pp. 47–52, 2022, doi: 10.33379/gtech.v6i1.1249.

[4] L. Muzdalifah, N. Ariyani, and R. Tuban, "Analisis Perbandingan Kinerja Sistem Antrian Tak Terbatas (G/G/2) Dan Sistem Antrian Terbatas (M/M/2):(Gd/N/∞)," J. Res. Adv. Math. Educ., vol. 08, no. 01, p. 69, 2024, doi: 10.23917.

[5] C. A. Rizkitha, W. Supriyatin, and Y. Rianto, "Application of Python for Analysis and Visualisation of ChatGPT User Dataset on College Students," Komputasi, vol. 22, no. January, pp. 45–53, 2025, doi:

10.33751.

[6] J. Sabilala et al., “Interpreting Sign Language Images into Text-Based using YOLOv7,” vol. 22, no. May, pp. 90–98, 2025.

[7] O. Laia, O. Halawa, and P. Lahagu, “Pengaruh Sistem Informasi Manajemen Terhadap Pelayanan Publik,” J. Akuntansi, Manaj. dan Ekon., vol. 1, no. 1, pp. 70–76, 2022, doi: 10.56248.

[8] S. Sugito and M. A. Mukid, “Distribusi Poisson dan Distribusi Eksponensial dalam Proses Stokastik,” Media Stat., vol. 4, no. 2, pp. 113–120, 2021.

[9] E. Pratama and D. Devianto, “Analisis Sistem Antrian Satu Server (M/M/1),” J. Mat. UNAND, vol. 2, no. 4, pp. 59–66, 2013, doi: 10.25077/jmu.2.4.59-66.2.

[10] S. Sugito and A. Hoyyi, “Proses Antrian Dengan Kedatangan Berdistribusi Poisson Dan Pola Pelayanan Berdistribusi General,” Media Stat., vol. 6, no. 1, pp. 113–120, 2023, doi: 10.14710/medstat.6.1.51-60.

[11] M. F. Nabhan, D. A. Lantana, P. Zidni, A. Arofi, and P. Habibbullah, “Simulator Mesin Deterministic Finite Automata (DFA) Berdasarkan Diagram Transisi Menggunakan Python,” vol. VII, no. September, pp. 23–30, 2023, doi: 10.47970.

[12] F. S. Hillier and G. J. Lieberman, Introduction to Operations Research. New York, NY, USA: McGraw-Hill Education, 2021.

[13] A. M. Law, Simulation Modeling and Analysis. New York, NY, USA: McGraw-Hill Education, 2015.

[14] H. A. Taha, Operations Research: An Introduction. Boston, MA, USA: Pearson Education, 2017.