A Queuing Model for Outpatient Department to Reduce Unnecessary Waiting Times

Thomas Courtney[1], Akeredolu Damilare[1], Soudabeh Madhkhan Esfahani[1],
Negin Ghasemi[1] , and Haritha Karawitage[1]
[1]University of Ottawa
School of Electrical and Computer Engineering
800 King Edward, Ottawa, ON, Canada, K1N 6N5

Abstract - This report presents a study on the application of queuing theory to reduce unnecessary waiting times in an Outpatient Department (OPD) of a hospital in Sri Lanka. The study focuses on a specific OPD known for its high patient traffic and long waiting times. The research utilizes a Multi-Server (M/M/c) queuing system, treating the entire process as a single queue with multiple stages, rather than separate queues for each service point. The simulation uses data from the Sri Lankan OPD and considers the number of servers in the system. The performance metrics used to determine the optimal number of servers include Server Utilization, Mean Waiting Time in Queue (Wq), and Mean Waiting Time in the System (Ws). The results indicate that increasing the number of servers significantly reduces the Wq value, while the Ws values do not decrease with an increased number of servers. The Server Utilization decreases over time with an increased number of servers. The study concludes that the application of queuing theory and the use of an M/M/c model can effectively reduce waiting times in an OPD setting.

Keywords: Healthcare Management, Multi-Server Model (M/M/c), Outpatient Department (OPD), Queuing Theory, Waiting Times.

1. Introduction

In recent years, due to the growing population and widespread disease, patients stand several hours in long queues to get registered into the system, consult with physicians, obtain drugs from the pharmacies, and finally leave the hospitals. These time-consuming processes would lead patients to get frustrated and dissatisfied [1]. An outpatient department (OPD) is considered as the first and one of the most important departments in the hospitals, and it is defined as the window of external service, connecting directly with the hospital's reputation and profits. In other words, people come in contact with the OPDs first, subsequently being filtered and categorized based on the treatments they would need. Due to the last-minute visit of outpatients and inadequate number of service windows, OPDs experience long waiting lines of patients expecting to get served as fast as possible [2, 3].

Regarding the importance and value of a healthcare system, governments are searching for new ways to reduce long waiting lines, as well as increase patient satisfaction. In recent years, much research has been carried out by researchers to solve this problematic issue. Some solutions including increasing the number of service windows, or decreasing the processing time in the outpatient departments have been proposed [1].

In our project, queuing theory is used to analyze waiting times of patients in the selected OPD and predict the new waiting times with respect to the increase of servers [1]. Queueing theory or stochastic system service theory plays a

role in studying the service system queue random rules and disciplines. In this theory, a number of indicators of statistical law such as waiting time, the length of queues in busy periods, etc. are applied to enhance the system performance in terms of patients' satisfaction and organizations' profits [4].

In the current project, an overcrowded OPD in Sri Lanka is selected. The chosen OPD contains one registration counter, three rooms for physicians, and two pharmacy counters. On weekdays, the working hours for OPD consultations would start from 8.00 a.m. to 12 p.m., and also from 2.00 p.m. to 4.00 p.m. Additionally, this OPD center would work Saturdays from 8.00 a.m. to noon. In this center, doctors do not follow the parallel way to serve the patients. In our OPD center, they always have one doctor and one pharmacy counter available to serve patients. Also, all three doctors and two pharmacy counters are available on clinic days. For OPD patients on clinic days, one doctor and one pharmacy counter are available to serve patients, while other doctors and pharmacy counters are busy serving clinic patients. As a result, in each step, a huge number of patients are waiting in long queues to get their service and leave the outpatient department. Therefore, we mainly focus on the length of the queues in all steps including the registration counter, consultation process, and eventually pharmacy counter [1].

Overall, in the following research, we are concentrating on answering some important questions including 1) The time when the waiting lines occur, 2) The bottlenecks of hospital queuing, 3) Present hospital queuing procedures, 4) The latest methods to reduce the waiting times in the hospital, and 5) The way of decreasing the negative impacts of long queues [1].

2. Literature Review

In hospitals across the globe, particularly in underdeveloped nations, overcrowding causes protracted wait times for outpatient care [5]. Limited funds and resources in developing nations might make this issue worse. Making effective use of hospital resources can lower operating costs while also improving patient satisfaction. Readjusting the resources to use underutilized departments effectively can also be aided by this. In order to find practical methods for cutting waiting times, this study summarizes recent studies on queuing models used in outpatient clinics.

Numerous queuing models, including priority queues, simulations, and single- and multi-server models, have been used in the literature to describe outpatient services. Monte Carlo simulations and Markov decision processes are particularly prevalent in recent studies. For example, Xie Shan *et. al*. study [3] looks at how queuing models—more especially, the single service (M/M/1) and multi-service (M/M/c) models—are used in the analysis of several performance indicators in outpatient systems. These models have been used in a number of empirical research in actual outpatient settings. For instance, Algiriyage N. *et. al*. [6] use a simulation model to assess various appointment scheduling rules and analyze various queues that cause bottlenecks in the Outpatient Department of the National Eye Hospital in Sri Lanka. The goal is to develop a solution that minimizes the overall amount of time patients must wait.

The arrival of patients and the length of the service are both uncertain in outpatient services, which are stochastic service systems. Research [4] has been on the use of DES to simulate complex systems, controlling the flow of events and patient-provider interactions through the use of an event scheduling simulation technique. It implies that by locating inefficiencies and making recommendations for improvements, DES models can help hospitals save a substantial amount of money.

A different strategy is used in the study [2], where the authors suggest a "Waiting Time Estimation System" to use the MQTT protocol to estimate the last waiting time for a new patient on cell phones. The queueing theory-based

algorithm they employ to calculate the waiting time. In order to predict the waiting time for new patients who intend to visit the outpatient department, the paper presents a waiting time estimation technique. Next, they discuss the deployment of an application that provides a real-time estimate of the wait time for incoming patients. Patients can use this information to schedule their arrival time at the outpatient department in order to avoid crowding, which presents a promising solution to wait times, according to this study.

A discrete event simulation is a frequent tool used in studies to simulate patient flow. Data from the Out-Patient Department of an orthopedic hospital was used to create a base model that simulated the patient's travel through many departments [7]. In a similar vein, the National Eye Hospital in Sri Lanka evaluated appointment scheduling systems and examined how lines formed using simulation models [6]. All of the research points to high patient inflow rates and long service times as major causes of issues with waiting in lines. Specifically, because there aren't enough multipurpose servers, the registration and billing counters are identified in the orthopedic hospital's research as the main bottlenecks [7]. On the other hand, the National Eye Hospital study emphasizes the necessity of effective appointment scheduling in order to reduce overall patient waiting times [6].

The studied research emphasizes how well discrete event simulations and queuing models work to alleviate hospital outpatient service inefficiencies. Technological treatments such as the WTE System provide creative ways to manage patients in real-time. These studies offer insightful information about low-cost ways to improve patient satisfaction and service quality without making significant changes to the way hospitals are now set up. Adopting such modifications could result in significant advancements in the provision of healthcare.

3. Methodology

3.1 Area of Study

This research focuses on a specific hospital's outpatient department known for its high patient traffic and long waiting times in the study [1]. As outlined by Weerakoon *et. al*., [1] involves patients going through multiple service points: registration, consultation, and pharmacy. These service points are essentially treated as separate queues, with the study analyzing the waiting times at each point. The queuing theory parameters such as arrival rate and service rate are used to analyze and predict the mean waiting times. However, our study follows a different approach in that we have simulated a model where there is only one waiting line to enter the system, and then patients go through each of the three service points (registration, consultation, pharmacy) in sequence before exiting. Relevant modifications have been made to the existing work done in the study [1]. This modification involved considering the entire process as a single queue, rather than separate queues for each service point.

3.2 Data Collection Methodology

For the development of the model, we have used the data collected by direct observation in the study [1]. Data was collected over a one-week period, capturing a comprehensive view of patient flow. The data points recorded include 1) Patient arrival times to the OPD. 2) Start and end times for registration, consultation, and drug issues. Data from a total of 289 patient visits were recorded. Time periods for staff breaks like tea and lunch were excluded from the service time calculations.

3.3 Application of Queuing Theory

Single-Server Queue Model (M/M/1): This model is used in situations where there is only one server available to provide a service. Examples include a doctor's consultation or a pharmacist at a pharmacy counter. This model aids

in analyzing and optimizing the patient flow and waiting times in services where a single person is responsible for attending to all arrivals.

Multi-Server Queue Model (M/M/c): This model is applied when there are multiple servers available or when the possibility exists to increase the number of servers. This could be relevant in a hospital setting where there are multiple doctors or pharmacists available to attend to patients. This model helps in determining the effect of increasing the number of servers on the overall waiting time and in understanding the potential for improving service efficiency.

In consideration of this study, we have simulated a Multi-Server (M/M/c) queuing system which can be successfully implemented into OPD departments of hospitals as patients must visit each server in a specific sequence, which is a characteristic of a tandem queuing system. The introduced multi-server system in our paper follows one waiting line, and patients sequentially pass through the three service points (registration, consultation, and pharmacy) before exiting the system. This approach is particularly useful for modeling scenarios where multiple service channels are available at different stages. This is typical in healthcare settings like outpatient departments (OPDs), where patients often follow a predefined path through various service points. Therefore, our simulation follows a Multi-Server (M/M/c) Queuing system.

3.4 Detailed Analyses

Within the OPD, patients' arrival times, registration start times, registration end times, consultation start times, consultation end times, drug issuing start times, and drug issuing end times were directly observed where arrivals follow a Poisson Process, service time follows an Exponential Distribution. Arrival rate (λ) and Service rate (μ) are being used to find mean waiting time in the queue (Wq) and mean waiting time in the system (Ws). Calibrating and validating the queuing models was done by collecting historical data on patient arrivals, service times, and waiting times in healthcare facilities by observations and automated data collection systems.

- Arrival Rate (λ): This is the rate at which patients arrive at the OPD. It was calculated by the inverse value of average interarrival time between patient arrivals and multiplying by 60 if considering the average rate per hour. However, the Average rate can still be calculated by taking the average number of arrivals.

$$\lambda = 1/\text{average interarrival time} \qquad (1)$$

- Service Rate (μ): This represents the rate at which service points (the likes of registration, consultation, and issuing of drugs) can serve the patients. It was calculated by taking the inverse of the average time taken to serve a patient.

$$\mu = 1/\text{average service time} \qquad (2)$$

- Mean Waiting Times: These were calculated for both the entire system (Ws) and the queues (Wq).

$$Ws = Ls/\lambda \qquad (3)$$

$$Wq = Lq/\lambda \qquad (4)$$

Note: Ls and Lq represent the total number of patients in the system and the total number of patients in the queue respectively.

3.5 Analyzing and Simulating Different Scenarios

Using the proposed M/M/c queue model, the impact of increasing the number of service providers (like doctors or pharmacists) on waiting times was explored in order to reduce the waiting time and get the most efficient number of servers for the system. The queuing theory models developed can be used to simulate and predict changes in waiting times with different numbers of servers. Once the model is implemented, based on the findings, an optimized system for managing patient flow in the OPD is proposed. This includes recommendations for the number of overall servers that should be implemented in the system to minimize patient wait times. The report's implementation section will go into further depth using the following equations.

$$Ls=Lq+\lambda/\mu \qquad (5)$$

$$Lq = ((\mu/\lambda)^c p/C! (1-p)^2) P_o \qquad (6)$$

$$Ws = Wq+l/\mu \qquad (7)$$

$$Wq = Lq/\lambda \qquad (8)$$

λ = Arrival rate (average number of arrivals per unit time).

μ = Service rate per server (average number of services completed per unit time by a single server).

c = Number of servers.

ρ = Traffic intensity or utilization factor ($(\lambda/(C\mu))$).

P_0 = Probability that there are zero customers in the system.

4. Implementation

In this section, we present details of the implementation of our simulation for an M/M/c queueing system. The goal of the simulation was to simulate the queuing dynamics of an outpatient hospital delivery system, in which patients arrive, form a queue, and receive service from several (c) servers. A popular method in queueing theory is the M/M/c queueing model, which is what our simulation uses. In order to simulate the stochastic nature of client arrivals and service procedures, the model takes into account an exponential distribution for interarrival times and service times.

An Object-Oriented Python script that captures the essential elements of the M/M/c system is used to generate the simulation. Classes that represent clients, service points, the waiting queue, and the entire M/M/c system are used to organize the code. Within the M/M/c queueing system, the Client class represents a single client. Every customer has an ID that identifies them specifically, along with information about their arrival time and the amount of time needed for services. The system's servers, or service points, are modeled by the ServicePoint class. Despite having a unique ID, every service point keeps track of its current status (available or busy). It keeps track of when the service begins and adds up the entire amount of time within the queue. The Waiting Queue class represents a queue where customers wait for assistance, it uses a list structure to control the clients' order. Essential queue methods are also provided by this class including enqueuing a client to the end of the queue, dequeuing a client from the front, and checking if the queue is empty. Overall, the M/M/c class serves as the main simulation component, which is the

fundamental element. It is initialized by taking into consideration the number of servers in that specific simulation while using data from the OPD. Among the many techniques in the class are those for generating random interarrival and service times, based on the exponential distribution, which are stochastic in nature. Additionally, there are methods such as enqueue, which adds customers to the waiting list for a specific day of the week while determining when they will arrive for a later service. Finally, the finish_service method wraps up the service for clients whose service has been completed. The serve_client method enables available servers to attend to clients from the waiting queues. The simulation method orchestrates the simulation process for a specified number of iterations. It captures and returns the mean waiting times and mean service times for each day, offering insights into the system's performance.

The simulation utilizes data from the Sri Lankan OPD used in the paper we modeled. The OPD is open Monday-Friday and therefore the data used for this simulation is 5 separate distinct days, representing each day of the work week. Each occurrence of interarrival times, arrival rates, service rates, and number of patients per day were independent from one another and used to simulate our M/M/c model for Monday-Friday. Since our problem was simulated using an Object-Oriented approach, we can benefit from abstraction which is a common Object-Oriented paradigm that commonly allows for scalability. In our specific situation, the number of servers can be scaled up (increased) and subsequently, the output results can be analyzed to determine the optimal number of servers.

5. Results

When analyzing an M/M/c queueing system, there are several performance metrics to look at for determining the optimal number of servers. Server Utilization represents the ratio of the arrival rate (λ), to the service rate (μ), per server, multiplied by the number of servers (c) [1]. Server Utilization is also known as traffic intensity which can be denoted by ρ, it is said that if $\rho < 1$, the system is stable, if $\rho > 1$, the system is unstable, and finally if $\rho = 1$, the system is in a state of balance, the average number of jobs in the system will remain constant over time [1]. Another metric that is used to judge the performance of an M/M/c is the Mean Waiting Time in Queue (Wq), this represents the average time a job or request spends waiting in the queue before being served [1]. The last and final performance metric that will be considered in our experiment is the Mean Waiting Time in the System (Ws), this represents the average time a job or request spends being serviced by one of the servers, or the time it takes for the server to process a request [1].

Day	Mean Waiting Time	Mean Service Time	Number of Servers	Day	Mean Waiting Time	Mean Service Time	Number of Servers
Monday	71.4377	13.9439	1	Monday	55.735	16.293	4
Tuesday	71.6504	16.5447	1	Tuesday	55.8595	16.8248	4
Wednesday	71.8297	16.226	1	Wednesday	56.045	17.9822	4
Thursday	72.0765	17.881	1	Thursday	56.2351	16.6794	4
Friday	72.325	20.4163	1	Friday	56.4492	17.6154	4
Monday	68.1335	16.836	2	Monday	52.4556	17.0292	5
Tuesday	68.3705	18.1174	2	Tuesday	52.5918	17.6023	5
Wednesday	68.5828	19.3465	2	Wednesday	52.7745	17.8546	5
Thursday	68.7985	18.3627	2	Thursday	52.9692	18.0636	5
Friday	69.0509	17.8125	2	Friday	53.1404	16.9859	5
Monday	60.4033	17.322	3	Monday	52.4018	17.7958	6
Tuesday	60.5557	17.1146	3	Tuesday	52.5067	16.6557	6
Wednesday	60.7042	16.9173	3	Wednesday	52.6462	18.2644	6
Thursday	60.9337	17.6142	3	Thursday	52.8313	16.6418	6
Friday	61.1571	17.2204	3	Friday	52.974	17.0273	6

Table 1: Mean Waiting Time and Mean Service Time for each day of the week, depending on the number of servers for that particular simulation.

According to the results presented in Table 1 from our simulation with c number servers, it can be observed that when increasing the number of servers by 1 in each simulation, the Wq value decreases significantly. However, the shift from 5 servers to 6 servers observed the least amount of difference between waiting times. As we observed the values of Wq decrease while c servers increase, this could be explained by a variety of reasons, including more servers being attended to simultaneously, while handling patients independently which reduces overall waiting time. Additional servers help in distributing patient load more evenly, preventing bottlenecks at service points. Finally, in an OPD setting, there are often peak periods where the demand for certain services can be higher, having a higher number of servers allows the hospital to respond more effectively during those peak hours. During such hours, additional service points can be deployed to handle the increased patient flow, preventing longer waiting times. Table 2 displays the Wq value, as the number of servers increases along the x-axis, the Wq value decreases per unit of time on the y-axis accordingly. Each day of the week is shown for c = 6 servers.

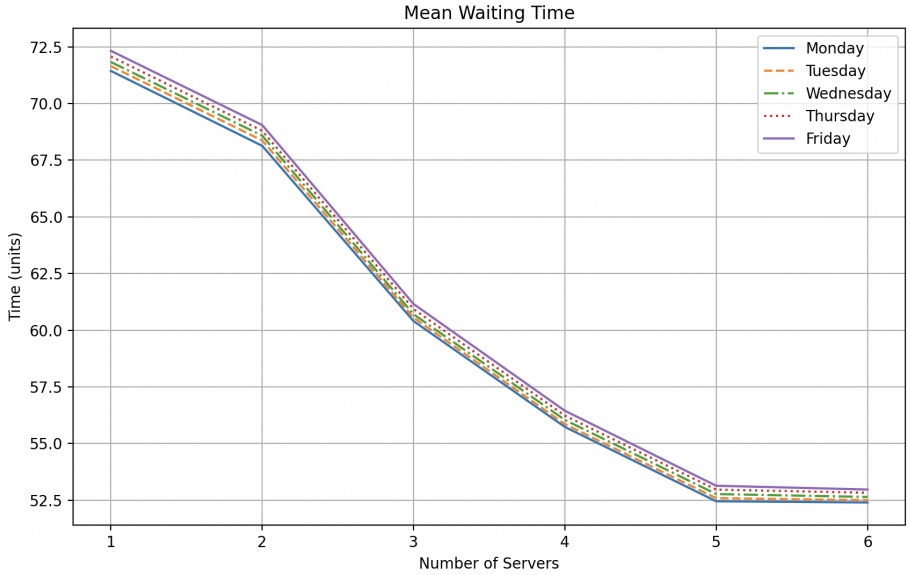

Figure 1:: Mean Waiting Time for each day of the week.

Additionally, when looking at the Mean Waiting time in the System (Ws) which represents the average time a job or request spends being serviced by one of the servers, we can see from Figure 2, that the Ws values do not decrease with an increased number of servers and for even some of the days we can see an increase in the values of Ws. For our particular situation of an OPD, this can be explained by numerous reasons. Increasing the number of servers may introduce complexities in coordination or communication, if servers display these qualities, it can lead to inefficiencies and or delays. Another possible reason why the values of Ws do not decrease with an increased number of servers is that redundancy or overlap in responsibilities might be introduced, while tasks that are not well distributed or if there is duplication of tasks, often unnecessary, this could lead to inefficiencies and longer service times. Finally, when adding more servers to the M/M/c queue, we often disregard increasing the amount of resources proportionally, these resources include technology, equipment, or support staff. This disregard for improved resources could hinder the ability of servers to perform their tasks accordingly.

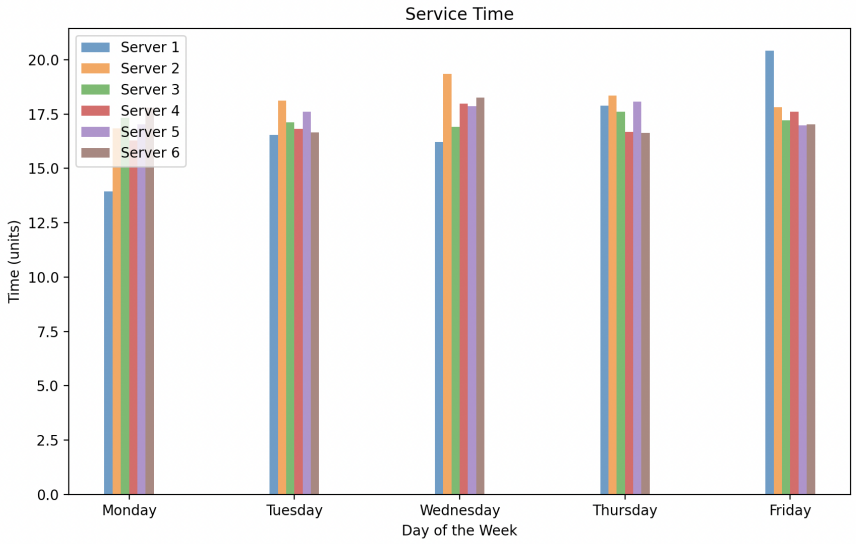

Figure 2: Mean Service Time for each day of the week.

The last performance metric that will be studied in our OPD setting is the Server Utilization for c number of servers. As we can see from Figure 3, our queueing system is stable as it has a value between 0 and 1. It can also be observed that in general, with an increased number of servers, the Server Utilization decreases over time. This can be explained by a variety of reasons, one reason could be the underutilization of new servers. This number of new servers may not be fully utilized compared with a decreased number of servers due to the workload not being evenly distributed among servers, bringing down the overall metric of Server Utilization. Another reason why Server Utilization might decrease in our particular situation is that there could be a mismatch between capacity and demand, an increased number of servers may not align with the particular demand for the OPD at that current time. If the demand is lower than capacity, the servers will not be fully utilized, and thus the Server Utilization metric will decrease. Finally, an explanation for why the Server Utilization metric decreases as the number of servers increases is that there could be an incompatibility with OPD workload patterns. While there exists peak times in an OPD setting if these peak times do not align with server deployment, there could exist times when many servers are being underutilized, adequate monitoring and optimization are necessary so these underutilization issues can be identified and addressed to ensure optimal server utilization.

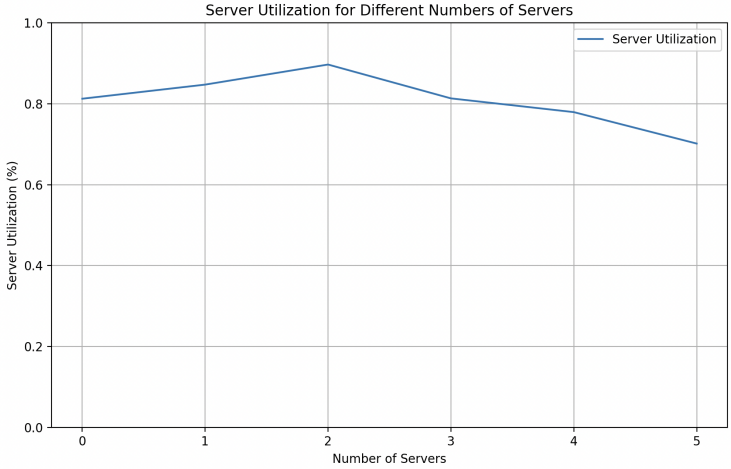

Figure 3: Server Utilization Percentage for c number of servers.

Overall, when analyzing our M/M/c queue simulation, the model performed well, with Mean Waiting time in Queue decreasing with an increased number of servers explained by more resources available, a constant Mean Waiting time in System, explained by our particular situation of an OPD with a possibility of redundancy or overlap and coordination or communication for more optimal results, and finally a decrease in Server Utilization with an increased number of servers, explained by underutilization and or a mismatch between capacity and demand.

6. Conclusion and future proceeding

The application of queuing theory in healthcare management, specifically in Outpatient Departments (OPDs), has proven to be an effective approach to reduce unnecessary waiting times. The study conducted on a specific OPD in Sri Lanka, known for its high patient traffic and long waiting times, demonstrated significant improvements in service efficiency. The use of a Multi-Server (M/M/c) queuing system, treating the entire process as a single queue with multiple stages, allowed for a more streamlined and efficient patient flow.

The results of the study indicated that increasing the number of servers significantly reduced the Mean Waiting Time in Queue (Wq), while the Mean Waiting Time in the System (Ws) remained constant. This suggests that while additional servers can help distribute patient load more evenly and prevent bottlenecks at service points, there may be other factors at play that affect the overall service time. These could include the need for improved coordination and communication among servers, or the need for additional resources such as technology, equipment, or support staff.

In terms of future proceedings, it would be beneficial to explore these factors further. Additional research could focus on optimizing resource allocation in conjunction with increasing the number of servers. This could involve investigating the impact of improved communication systems, the introduction of advanced medical equipment, or the implementation of staff training programs. Furthermore, the study could be expanded to include other departments within the hospital or even other hospitals in different regions. This would provide a more comprehensive understanding of the applicability and effectiveness of the M/M/c queuing model in various healthcare settings.

Another potential area for future research could be the exploration of other queuing models. While the M/M/c model proved effective in this study, other models may offer different insights and could potentially lead to further improvements in service efficiency. For instance, priority queues or simulations could be considered.

In conclusion, the application of queuing theory in healthcare settings presents a promising approach to improving service efficiency and patient satisfaction. With further research and continuous improvements, it is hoped that long waiting times in OPDs will become a thing of the past.

7. Reference

[1] W. M. N. B. Weerakoon, S. Vasanthapriyan, and U. A. P. Ishanka, " A Queuing Model for Outpatient Department to Reduce Unnecessary Waiting Times, "*IEEE Xplore,* Dec. 01, 2019. https://ieeexplore.ieee.org/document/9063348.

[2] N. Tantitharanukul and T. Throngjai, "Waiting time estimation system for outpatient's arrival planning," 2018 International Conference on Digital Arts, Media and Technology (ICDAMT), Phayao, Thailand, 2018, pp. 207-212, doi: 10.1109/ICDAMT.2018.8376525.

[3] X. Shan, L. Jing, L. Zhifeng, Q. Dongjun, and T. Ying, "The Study and Application of Intelligent Queuing in Outpatient Department," 2013 Third International Conference on Intelligent System Design and Engineering Applications, Hong Kong, China, 2013, pp. 1549-1553, doi: 10.1109/ISDEA.2012.372.

[4] Yin Chao, "Outpatient queue business simulation based on acceptable waiting time," 2010 International Conference On Computer Design and Applications, Qinhuangdao, China, 2010, pp. V1-120-V1-123, doi: 10.1109/ICCDA.2010.5541124.

[5] A. Mohammed Ali and A. Kassam, "Optimization of Outpatient Department performance using simulation," ResearchGate,https://www.researchgate.net/publication/322057220_Optimization_of_Outpatient_Department_Performance_Using_Simulation (accessed Nov. 17, 2023).

[6] N. Algiriyage, R. Sampath, C. Pushpakumara and G. Wijayarathna, "A simulation approach for reduced outpatient waiting time," 2014 14th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2014, pp. 128-133, doi: 10.1109/ICTER.2014.7083891.

[7] K. S and S. K. S, "Optimising Waiting Times at a Major Orthopaedic Hospital Using Simulation – A Case Study," 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2023, pp. 542-549, doi: 10.1109/ICCPCT58313.2023.10244941.