Queuing Theory in Modern Technology, Areas Covered, and Challenges

Zhenyu Gui *

Univerisity of Mount Saint Vincent, Riverdale, NY, 10471, United States of America

* Corresponding Author Email: Leo-gzy@outlook.com

Abstract. In the context of accelerated scientific and technological advancement, allocating resources and optimizing services have emerged as significant challenges for numerous industries. This paper aims to examine the potential applications of queuing theory in the context of modern technology. The paper commences with a summary of the fundamental concepts and definitions of queuing theory. Subsequently, the paper illustrates the implementation of queuing theory in a multitude of practical domains and its pivotal function in optimizing the distribution of resources and enhancing efficiency through the examination of illustrative examples. It has been demonstrated that queuing theory significantly improves system performance and resource utilization. However, it also presents a number of challenges, including those associated with modelling optimization, uncertainties and dynamic data. Integrating queuing theory with artificial intelligence and machine learning will facilitate its expanded role in multidisciplinary operations and real-time data processing, thereby providing crucial support for advancing modern science and technology. The research in this paper can provide appreciation for researchers in related fields to better understand or apply queuing theory.

Keywords: Queuing theory; computer optimization; server optimization; resource utilization; efficiency increased.

1. Introduction

In the contemporary era of accelerated technological advancement, the efficient utilization of resources and optimal management have emerged as pivotal concerns for many industrial sectors. In order to achieve optimal performance and utilization, industries such as network communications and transportation are attempting to maximize the efficiency of their resources. As technology and data evolve, the importance and utility of optimization methods continue to grow, as does the need for accurate and effective theoretical tools.

At this time, effective theoretical tools must be capable of optimizing and maximizing resource utilization with accuracy and efficiency under conditions of robustness and scalability. Queuing theory is a mathematical approach that is used to study the phenomenon of queuing and optimization systems waiting. This mathematical approach is based on the development of various mathematical models, including M/M/1, M/M/c and M/G/1, which are used to depict the process of queuing, waiting and processing. It covers probabilities, scheduling algorithms, utilization rates and other key information, and is intended to optimize the system to meet its needs best, providing effective support for all types of queuing systems.

As a pragmatic and contemporary mathematical method, queuing theory is employed in modern technology to address a diverse array of fields. In the field of information technology, queuing theory is employed to optimize performance, improving resource utilization and reducing waiting times. Additionally, queuing models offer numerous methods for anticipating the nature of the problem at hand through a range of computational approaches. For example, understanding how a given architecture behaves with given parameters allows better anticipation of problems such as bottlenecks and performance issues [1]. In the field of communications, queuing theory can be employed to enhance the efficiency of broadband allocation and call center connectivity. For example, when a customer of a communication company encounters a problem and needs to give feedback to the call center, using key performance indicators to determine the optimal number of operators is an important factor in optimizing the communication system. And research has shown that such optimization can

be carried out using queuing theory [2]. In the field of transport, which is becoming increasingly challenging, queuing theory can be employed to enhance flow management and optimize signalization. This entails the utilization of queuing models and stochastic optimality to optimize the existing traffic congestion situation. Similar methodologies can be applied in motorway tolling and airport security domains. Waiting time optimization of waiting queues and resource allocation of services are typical healthcare applications. For example, a queuing system can be used to bring bed occupancy and visits into a mathematical model and plot curves to consider optimization [3]. The application of queuing theory algorithms to manufacturing and supply chain processes enables the optimization of production line scheduling, thereby enhancing overall production line efficiency.

Nevertheless, queuing theory encounters numerous challenges in the context of modern technology, despite its extensive practical applications. These include the complexity of modelling, the impact of external dynamic environmental uncertainty, multi-objective and human factors, and other factors. Furthermore, the synergistic development of multiple fields represents a significant challenge for queuing theory, which must address this issue in order to facilitate more effective advancement.

This paper aims to examine the particular applications of queuing theory in the context of contemporary technology, identify the challenges currently facing this field of study, and suggest future avenues of development. The intention of this study is to provide a reference point and source of inspiration for those engaged in the field of queuing theory and to facilitate the development of modern technology.

2. Overview of Queuing Theory

The field of queuing theory is a mathematical discipline that investigates the dynamics of task arrival, waiting, and service within a queuing system. The objective of queuing theory is to optimize the speed of task processing and achieve more efficient resource utilization through modelling, computation and analysis. This mathematical theory is employed extensively in the domains of computer science, network communications and transport. The following section provides an overview of the fundamental concepts, definitions, components, models, and algorithms of queuing theory.

2.1. Concepts and Basic Definitions of Queuing Theory

The field of queuing theory is founded upon three core concepts: the client, the queue and the server. A customer is typically defined as a task or entity that requires servicing. A queue is the designated area where customers await service and processing. On the other hand, a server represents the location where customers are served or tasks are processed. The field of queuing theory employs mathematical modelling to elucidate the processes of customer arrival, queuing and the acceptance of service. The following section includes a number of fundamental definitions.

Client arrivals are typically expressed in terms of arrival rate (λ), which is the number of clients arriving per unit of time;

Service time usually refers to the amount of time a server is able to serve a customer, and it can also be used to express service rate. The service rate (μ) is usually used to express the quantity of customers that the server is able to serve per unit of time. And in general, $\lambda < \mu$, i.e., the arrival rate in the model is less than the service rate, as a way to avoid overloading;

Waiting time usually refers to the time a customer waits from the start of arrival to receiving service;

Queue length usually refers to the number of customers sorted in the queue.

Queuing theory employs these fundamental definitions to construct mathematical models for the analysis of system optimisation, resource utilisation, and other related processes.

2.2. Basic Components of Queuing Theory

The fundamental elements of a queuing system were previously outlined in broad terms; the following section provides a more detailed account of the underlying components.

In this context, the term "client" refers to an individual who requires a service or assistance. Examples of such entities include consumers in a supermarket queue or tasks in a server.

A queue can be defined as a location where customers await service provision. Examples of such queues include those observed in supermarket checkout lines and task queues within a system.

Servers may provide services to customers, and tasks provided to individuals for processing. Examples include shop assistants in supermarkets, processors in computers, etc. Servers will come in the form of one or more compositions, usually forming but servers with multi-server systems.

The objective of the queuing system is to optimise server performance through the reasonable arrangement of server queues. This includes task processing speed, resource utilisation and reduction of waiting time.

The green arrow in Fig. 1 expresses customer arrival, the order is from the left side into the queuing area; the grey area indicates the queuing area, where customers queue up; the blue area represents the server, which is the area where the customer receives the service; the red arrow represents the completion of the service, and the customer leaves from the right side.

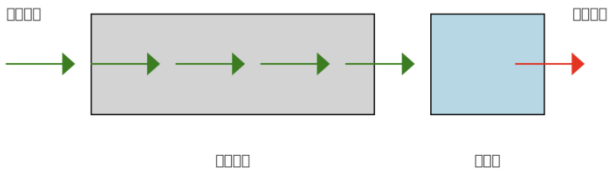

Fig. 1 Schematic diagram of the queuing theory model

2.3. Common Queuing Models

Queuing models are of central importance in the field of queuing theory, and are frequently employed to describe and analyze mathematical models of queuing systems. The most commonly used queuing models are M/M/1, M/G/1, M/M/c and G/G/1.

The M/M/1 model represents the fundamental queuing model. In the event that the arrival process is exponentially distributed with the service process, the system is characterized by a single server. The primary parameters are based on the arrival rate λ and the service rate μ, which are more straightforward.

The M/G/1 model is analogous to the M/M/1 model, but based on the property that service times can be arbitrarily distributed, this model is more general and equally more complex.

The M/M/c model is a multi-server model. In this model, there are c servers, and the arrival times of clients are exponentially distributed concerning the service times. This model is appropriate for the analysis of multi-server systems.

The G/G/1 server model is the most commonly used and allows for an arbitrary distribution of client arrivals and service times. This flexibility, however, is often accompanied by a greater difficulty in parsing the algorithm.

2.4. Common Scheduling Algorithms

The scheduling algorithm in a queuing system determines the sequence in which customers receive services. The most commonly employed scheduling systems are FCFS, LCFS and SJF.

FCFS is the simplest and fairest algorithm for customers to receive services in the order of their arrival. It is the simplest and the fairest algorithm, but it usually consumes a longer waiting time.

The last customer to arrive receives the highest priority service. There are many special scenarios where this algorithm may be used.

Priority service is given to the shortest customers. This is an algorithm based on the need to estimate service time, which can somewhat effectively reduce the overall average waiting time.

The scheduling algorithm incorporates a preemptive scheduling component. Thus, the operating system can interrupt the currently running task while it is in progress and transfer control to the client with the highest priority. Preemptive scheduling is a widely utilized approach in multi-tasking environments and is also applicable to emergency task scheduling. To illustrate, the Windows operating system employs preemptive scheduling to oversee system processes and threads, guaranteeing that mission-critical tasks are accorded priority and that high-priority tasks are responded to promptly.

The selection of the most appropriate scheduling algorithm in different scenarios is of paramount importance. Furthermore, the queuing system can be optimized in different scenarios in order to meet the varying needs of different scenarios. The application of comprehensive queuing theory modelling and algorithms can facilitate the effective improvement of system performance.

3. Queuing Theory Applications

Queuing theory is a significant mathematical theory that has a wide range of applications in various resource service systems. With the advancement of modern technology, queuing theory has become increasingly pivotal in numerous contemporary fields, offering the potential to enhance efficiency and resource utilization. This section will provide a comprehensive overview of the real-world contexts where queuing theory is employed, illustrating its far-reaching influence in modern technology.

3.1. Applications in Information Technology

In the field of information technology, the server represents the core of the system, responsible for transmitting and processing information. As such, it plays an essential role in ensuring the efficient operation of the system as a whole. In the context of queuing theory, the arrival and service rates of servers are subjected to analysis, to develop more efficient models that enhance the operational efficiency of servers and reduce the waiting time for tasks in the queue. For example, by analyzing the arrival and service rates of servers, one can optimize the service delivery of ATM machines, thereby reducing waiting times and associated costs. This optimization not only improves the operational efficiency of the ATM but also has a positive impact on the server's overall performance [4]. In this process, the values of waiting time, customer satisfaction and extraction problems were subjected to analysis. It was determined that an increase in the number of servers would serve to reduce both waiting time and cost. This is a typical example of calculating load through a queuing model with the objective of improving system responsiveness and stability. Queuing theory is also a significant field of study in the context of data transportation. The scheduling process entails analyzing network traffic to optimize the selection of transmission paths, accelerating the transfer of data and reducing latency.

3.2. Applications in Network Communications

In the context of network communication, call centers play a pivotal role in the delivery of customer service, with a significant impact on customer satisfaction levels. Queuing theory is similarly pivotal in call centers, with the analysis of arrival rates and service times enabling the optimization of resource allocation, thereby reducing customer waiting times and enhancing operational efficiency. In the case of emergency calls, such as those from hospitals and the police, determining the optimal number of operators over a given period can lead to improved service quality and operational cost savings [2]. To illustrate, queuing systems may employ dynamic seat allocation during periods of peak access in order to accommodate peak call periods. In network communication, broadband allocation exerts a significant influence on network quality. In response to the evolving needs of users, relative optimization models are devised to achieve optimal resource allocation in the

network while enhancing energy efficiency [5]. The optimization of broadband allocation through queuing theory entails dynamic adjustment, including priority for meeting the data transmission requirements of the emergency call category, to improve resource utilization while enhancing the user experience.

3.3. Applications in Transport

The field of traffic flow represents a significant domain where the utilization of queuing systems is of paramount importance within the transport sector. The analysis of traffic flow queuing is employed to optimize signal allocation, aiming to reduce vehicle waiting time and congestion. Mathematical queuing theory models can also be used to determine appropriate traffic light durations based on vehicle arrivals, aiming to reduce queuing waiting times to below the driver time pressure threshold [6]. The optimization of traffic signals represents a crucial aspect of traffic flow management. However, traditional time-sequence signals are often ill-equipped to effectively manage and optimize the flow of dynamic traffic conditions. By analyzing the arrival and waiting time of vehicles, a strategy for signal regulation can be designed to achieve an adaptive adjustment that reduces the waiting time of vehicles. To illustrate, in a stationary state, the statistical variables of the traffic system, namely the number of vehicles and the waiting time through the road section, can provide effective information for drivers. The efficacy of the model in simulating the theoretical distribution of traffic and vehicle numbers has been effectively demonstrated [7]. Furthermore, an increase in vehicle waiting time at toll booths on highways can also result in traffic congestion. Implementing an appropriate arrangement of electronic toll collection system (ETC) lanes and manual lane distribution, establishing high-occupancy vehicle lanes (HOV), and determining the location and number of HOV lanes can contribute to reducing congestion at toll stations. For instance, a queuing theory-based PSO-LSTM model can be employed to analyze and model the toll station lanes, thereby facilitating the attainment of an optimal toll station duration [8]. The optimization of resource allocation and regulation is a key strategy for reducing waiting times and improving efficiency in airport gates, security channels and baggage systems. It is recommended that the transport system undergo an intelligent and digital transformation.

3.4. Applications in Medical Services

It is not uncommon for lengthy queues and a considerable number of individuals to be observed in healthcare facilities. The issue of patient flow allocation is of significant importance. Furthermore, the provision of superior service can be achieved by optimizing the allocation of resources through the application of queuing theory. To illustrate, during the epidemic, Nanjing Children's Hospital employed the sophisticated M/G/2 queuing model to assess the admission status of patients, while concurrently utilizing the M/M/1 model to examine the characteristics of mild and severe diseases [9]. Significant improvements have been made with regard to the quality of medical services and the rational allocation of medical resources. Furthermore, the analysis of the various stages of the healthcare process enables the reduction of waiting times and the identification and removal of bottlenecks within the process. The analysis of the workload of medical practitioners and the requirements of patients can be employed to enhance scheduling and guarantee adequate services during periods of peak demand. Implementing improvements to healthcare services based on queuing theory has been demonstrated to effectively reduce waiting times for patients while simultaneously rationalizing the allocation of healthcare resources.

3.5. Applications in Manufacturing

Effective production line scheduling represents a significant application in the field of manufacturing, with the potential to reduce the production cycle time significantly. The conventional approach to production line scheduling is based on established rules and is inadequate for addressing the complexities inherent in dynamic production demands. The application of queuing theory enables the design of more reasonable scheduling schemes through the modelling of queuing and service in

a range of dynamic production line scenarios. For instance, queuing theory modelling is employed as a representative of the development matrix of experimental algorithms and implicit numerical methods to guarantee the production department's optimal size, production costs, and supply. This approach was successfully applied (and subsequently validated in practice) in the design of the production component of a tangible industrial engineering unit [10]. Furthermore, queuing theory is also applied to multi-link distribution processes subsequent to the production line. The optimization of distribution routes and modes across multiple links in a distribution network has the potential to reduce the economic and time costs of transport. Emergency planning can be employed to guarantee the dependability of distribution in the event of exceptional, unpredictable demand. The mathematical model of queuing theory can be employed to achieve cost savings and optimization at various stages of the manufacturing and distribution process.

In conclusion, queuing theory has been demonstrated to be an invaluable mathematical modelling tool, with a wide range of applications in diverse fields including information technology, network communication, transportation, medical services and manufacturing. The analysis of customer queuing time, arrival rate and satisfaction level data allow for optimizing resource allocation and improving efficiency, which brings significant benefits to the modern economy. The advancement of contemporary technology has created new avenues for queuing systems to assume a more pivotal role in intricate scenarios. In the context of the ongoing digital transformation across a range of industries, queuing theory offers a valuable set of tools. Furthermore, the future holds additional challenges for queuing theory.

4. Challenges and Prospects

The field of queuing theory plays a pivotal role in optimizing the utilization of resources and enhancing system efficiency. Nevertheless, in practice, queuing theory encounters numerous challenges. The optimization of modelling, the management of uncertainty and the acquisition of dynamic data are all areas that require further development. The following section presents a discussion of the suggestions as mentioned earlier and challenges.

4.1. Modelling Optimization Recommendations

The efficacy of modelling is contingent upon the quality of the data input, and the accuracy of the data serves as the foundation for modelling optimization. In the event that the data is not sufficiently accurate, the modelling process may yield inaccurate results and predictions. It is therefore evident that data screening represents a crucial stage in the data collection process. The accuracy and adaptability of the selected model are the primary determinants of model optimization. The selection and adjustment of models according to the specific requirements of a given system can lead to significant improvements in the accuracy and adaptability of the models in question. In contrast to the conventional approach of optimizing model accuracy, the averaging and weighting of multiple models, as described in reference [11], can enhance the precision and resilience of the resulting model. The application of simulation and actual data to the dynamic adjustment of parameters in a queuing system can facilitate the improvement of the accuracy of the ideal model, thereby enabling the identification of the optimal parameter adjustments.

4.2. Uncertainty Prevention

In order to accommodate fluctuations in demand, unforeseen occurrences, and data inconsistencies, it is necessary to enhance the adaptability of the model and align it with the evolving demands of the system. Analyses based on historical data can effectively predict and identify fluctuations in demand. The unforeseen time will result in an inaccurate model prediction; thus, the establishment of contingency plans and redundancy mechanisms can effectively address this issue. Furthermore, the presence of data noise in the actual data will also affect the accuracy of the data. The implementation of data smoothing and filtering techniques can effectively enhance the accuracy of the data. The

application of techniques such as moving average and exponential smoothing can facilitate the extraction of genuine data trends.

4.3. Dynamic Data Acquisition

The queuing model will be adjusted in real time in accordance with the actual situation, and the dynamic data will reflect the actual demand and changes. Consequently, the implementation of real-time monitoring facilitates the acquisition of dynamic data, which is of paramount importance for the optimization of the queuing model. Timely updating is conducive to facilitating parameter changes and enabling timely feedback and resolution of any issues that may arise. The application of big data analysis and training techniques, such as machine learning (ML), enables the identification of regular patterns and the enhancement of the model's applicability. It is important to regularly update the system model to maintain its validity and ensure its continued applicability in new environments. This process of updating and adapting is a fundamental aspect of model iteration and optimization.

The emergence of numerous challenges also accompanies the extensive application of queuing theory in contemporary technology. The application of modelling, the prevention of uncertainty and the utilization of dynamic data facilitate the enhancement of the model's precision and adaptability. The optimization of resource allocation and queuing capabilities has been significantly enhanced. As technology continues to advance, it is evident that queuing theory will assume a more prominent role in the advancement of intelligence and digitalization.

Since queuing theory inception, it has played an important role in addressing practical issues in a range of fields. As a result of its integration with future predictions and technological advancement, the practical utility of this field is set to grow.

First and foremost, the conjunction of queuing theory and artificial intelligence represents a significant research avenue in the forthcoming era of machine learning, with the objective of attaining an intelligent process for the system. The application of machine learning enables the prediction of customer behavior, the optimization of resource allocation and a reduction in waiting times. Artificial intelligence can analyze and learn from historical data, thereby assisting the system in identifying potential trends. Such applications are beneficial in both the prediction of proposals and the adjustment of dynamic strategies.

Secondly, the current application of queuing theory in practice is more conventional in the domains of transportation and communication. The application of queuing models is evolving towards a more interdisciplinary and synergistic approach. To illustrate, in the domain of e-commerce and logistics, the optimization of order processing and distribution can effectively enhance customer satisfaction, thereby fostering economic growth. In the intelligent construction of cities, queuing theory can be employed to optimize the utilization of public resources. Furthermore, the multidisciplinary collaboration of signal lights and public transport can facilitate the efficiency of urban transport and enhance accessibility.

Concurrently, the accessibility of real-time data processing and dynamic optimization is increasing in tandem with technological advancement. It is anticipated that future queuing systems will be capable of collecting and analyzing data in real time through the utilization of bespoke methodologies, thereby enabling the dynamic adjustment of strategies in real time. Utilizing specialized methodologies, such as sensors, the Internet of Things and other real-time technologies, facilitates data collection with greater accuracy. To illustrate, the monitoring of production processes, the implementation of dynamic customer service staffing and the provision of other services can collectively enhance customer satisfaction.

The combination of artificial intelligence, machine learning, multi-disciplinary use and dynamic optimization conditions will enhance the future of queuing theory. This will further improve the real-world application of queuing systems, with greater efficiency and customer satisfaction. As technology advances, the role of the updated queuing system will become more extensive and powerful.

5. Conclusion

This paper introduces the fundamental concepts and definitions of queuing theory and provides an analysis of its basic components. The paper goes on to discuss common queuing models and scheduling algorithms in detail. The deployment of queuing theory in a number of contemporary industries, including server optimization for enhanced efficiency, reduction of waiting times in call centers, congestion reduction on transportation networks, the rational allocation of medical resources and the scheduling of production lines, is illustrated through the presentation of specific case studies.

The study concludes that queuing theory has the potential to enhance system efficiency and resource utilization, although it is not without its challenges. The modelling of optimization, the prevention of uncertainty and the acquisition of dynamic data present significant difficulties. In the process of mathematical modelling, it is of the utmost importance to ensure the quality and accuracy of the data being used. In an environment characterized by uncertainty, ensuring that the model in question exhibits a high degree of flexibility is paramount. The analysis of historical data, the formulation of contingency plans and the implementation of measures to ensure the system's smooth functioning can enhance the model's adaptability and precision. The acquisition of dynamic data through real-time monitoring and big data analysis enables the system to be more flexible and responsive, facilitating timely adjustments to the plan.

It is anticipated that in the future, queuing theory will be combined with artificial intelligence and machine learning in order to further exploit its value in reality. To illustrate, the application of queuing theory to e-commerce, logistics and urban intelligent construction will facilitate the optimisation of public resources and enhance the system's efficiency. The processing of real-time data and the optimisation of dynamic systems will represent a significant area of development within the field of queuing theory. This will be achieved through the utilisation of sensors, the Internet of Things and other technologies, facilitating the collection of data, the adjustment of data in real-time and enhancing the quality and efficiency of services.

In essence, the field of queuing theory in modern technology will continue to expand the use of technology as its utility in addressing real-world problems improves, leading to significant gains for a range of industries.

References

[1] Mas, L., Vilaplana, J., Mateo, J., & Solsona, F. (2022). A queuing theory model for fog computing. The Journal of Supercomputing, 78(8), 11138-11155.

[2] Afolalu, S. A., Ikumapayi, O. M., Abdulkareem, A., Emetere, M. E., & Adejumo, O. (2021). A short review on queuing theory as a deterministic tool in sustainable telecommunication system. Materials Today: Proceedings, 44, 2884-2888.

[3] Proudlove, N. C. (2020). The 85% bed occupancy fallacy: The use, misuse and insights of queuing theory. Health services management research, 33(3), 110-121.

[4] Olu, O. T. (2019). Application of queuing theory to a bank's automated teller machine (ATM) service optimization. Mathematics Letters, 5(1), 8-12.

[5] Yu, P., Zhou, F., Zhang, X., Qiu, X., Kadoch, M., & Cheriet, M. (2020). Deep learning-based resource allocation for 5G broadband TV service. IEEE Transactions on Broadcasting, 66(4), 800-813.

[6] Harahap, E., Darmawan, D., Fajar, Y., Ceha, R., & Rachmiatie, A. (2019, March). Modeling and simulation of queue waiting time at traffic light intersection. In Journal of Physics: Conference Series (Vol. 1188, No. 1, p. 012001). IOP Publishing.

[7] Li, Y., Chen, H., & Feng, M. (2020, December). A novel model for the traffic of urban roads based on queuing theory. In 2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS) (pp. 190-194). IEEE.

[8] Wang, P., Zhao, J., Gao, Y., Sotelo, M. A., & Li, Z. (2020). Lane work-schedule of toll station based on queuing theory and PSO-LSTM model. Ieee Access, 8, 84434-84443.

[9] Hu, J., Hu, G., Cai, J., Xu, L., & Wang, Q. (2020). Hospital Bed Allocation Strategy Based on Queuing Theory during the COVID-19 Epidemic. *Computers, Materials & Continua*.

[10] Rece, L., Vlase, S., Ciuiu, D., Neculoiu, G., Mocanu, S., & Modrea, A. (2022). Queueing Theory-Based Mathematical Models Applied to Enterprise Organization and Industrial Production Optimization. Mathematics, 10(14), 2520.

[11] Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., ... & Schmidt, L. (2022, June). Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In International conference on machine learning (pp. 23965-23998). PMLR.