Research and Analysis on the Interaction between Queuing Theory and Artificial Intelligence

Yuanhe Liu [1, *,†], Ruiming Quan [2, †]

[1, 2] Tsinghua Experimental School, Beijing, 100084, China
*3553385182@qq.com

†These authors contributed equally.

Abstract. This paper provides a theoretical exploration of queueing theory and its intersection with the rapidly expanding field of artificial intelligence (AI). In an effort to employ AI methodologies in addressing queueing theory problems, an innovative approach is proposed that blends simulation techniques with artificial neural networks (ANNs). This marriage not only exhibits high effectiveness but also holds promise for considerable advancements in applying machine learning methods to queueing theory dilemmas. Beyond theoretical underpinnings, the paper also illuminates practical applications of queueing theory within AI's domain. It showcases examples from real-life scenarios, such as online bookstores and e-commerce platforms, to demonstrate the strategic deployment of various queueing models to enhance user experiences and system efficiencies. Discussions delve into the advantages of resource allocation, load balancing, and service time optimization, among others. Furthermore, the paper speculates about the evolving relationship between queueing theory and AI, anticipating a future where this connection becomes even more profound and impactful. As AI research continues to advance, novel insights into complex queueing issues and innovative solutions for managing queues in an array of real-world scenarios are expected. This analysis emphasizes the potential richness of integrating queueing theory with AI, paving the way for exciting prospects for future research and applications.

Keywords: Queuing theory, Artificial intelligence, Simulation, Artificial neural networks, Virtual bookstore, E-Commerce

1 Introduction

To shape modern telecommunication networks, mathematical portrayals based on queuing theory are utilized to represent networks and structures. These models scrutinize the traits of existing and emerging network protocols, network topologies, and routing algorithm optimization. Beginning with the pioneering inquiries of luminaries like A. Erlang, A. Hinchin, and L. Kleinrock, and progressing to contemporary times, numerous studies have been conducted on different

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_40

infrastructures and queueing systems, all in relation to their application in telecommunication networks [1].

Current computer networks often exhibit these flows, leading to an increased interest from queuing theory scholars in the analysis of queueing networks with associated input flows, such as MAP and MMAP. The substantial contributions of researchers like N. Newts, A.N. Dudin, and V.I. Klimenok have played a critical role in shaping and progressing this field of study [2].The application of queuing theory in the realm of artificial intelligence has been thoroughly investigated by academia. As a vital component of operations research, queuing theory aims to achieve optimal design and control of a system by studying the probability laws inherent to queueing systems, striving to maximize system benefits at the lowest cost [3].

Several international conferences are dedicated to Agent theory and Agent systems, such as the International Conference on Multi-Agents, and the International Conference on Multi-Agents in the Asia Pacific. Furthermore, numerous renowned artificial intelligence conferences have recognized Agent theory as a separate topic [4]. Nonetheless, exploration into Agent theory, both in China and globally, remains in its infancy, with many significant concerns necessitating further investigation.

Current Agent research is broadly divided into three interrelated facets: Agent, Multi-Agent Systems, and Agent-Oriented Programming. Within this trinity, Agent forms the foundation of multi-agent system research and can be viewed as the microstructure of the multi-agent system. Meanwhile, the study of inter-agent relationships forms the macro level of the multi-agent system. The successful application of Agents and multi-agents is bolstered by Agent-oriented programming..

2 Theoretical Overview of Queuing Theory

2.1 Definition of Queuing Theory

Queuing theory is a theory focusing on understanding how queues work and how to increase efficiency 5. Within a computer system with only one job, the job enters, uses certain resources, and then exits. Furthermore, because there are no queues, no delay is anticipated.

When queues appear, queuing theory is relevant. In fact, the core of every computer system is a queue 6. A time-sharing scheduler is used by the CPU to serve. Memory banks provide memory chunks to queues of threads. See Fig. 1. Servers are shown in Fig. 1 as circles, queues as groups of rectangles, and the routing network as arrows. Although the study of the transient state is essential in many applications, one often attempts to acquire the network's equilibrium distribution when studying queue networks.

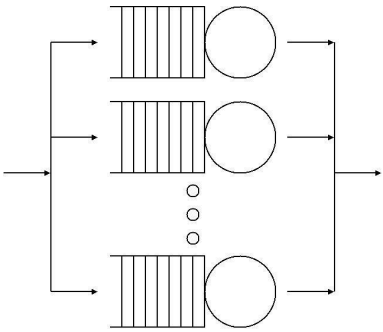

Fig. 1. Schematic Representation of the Queueing System - Servers, Queues, and Routing Network

2.2 Function and Purpose of Queuing Theory

A queuing theorist has two objectives. Predicting system performance is the first. Typically, this entails estimating the mean delay, the variability of the delay, or the likelihood that the delay would exceed a certain Service Level Agreement (SLA). It can also refer to forecasting the number of jobs that will be in a queue, the average number of servers being used (for example, total power requirements), or any other measure of this nature 7 . Prediction is vital, but creating a better system design to boost performance is even more crucial. Usually, this takes the form of capacity planning, when one decides which new resources to buy in order to meet delay targets (for instance, is it better to buy a faster disk or a faster CPU, or is it preferable to add a second sluggish disk). But frequently, one can boost performance without investing in any new resources at all by only implementing a more intelligent scheduling policy or a different routing scheme to cut down on delays.

2.3 Scheduling Policies of the Queuing Theory

At queuing nodes, different scheduling policies can be applied 8.

The policies are as follows.

First-Come, First-Served (FCFS)

Also called *First in, First out*. In accordance with this guideline, customers receive service one at a time, beginning with the one who has waited the longest.

Non-Preemptive Last-Come, First-Served (LCFS)

The customer who has the shortest wait time will be served first according to this approach, which also serves consumers one at a time.

Preemptive Last-Come, First-Served (PLCFS)

The work currently in service is promptly preempted if a new arrival enters the system. The preempted job is only permitted to resume operations once that arrival is finished.

RANDOM

When the server frees up, it chooses a random job to run next.

Processor Sharing (PS)

Service capacity is shared equally between customers.

Priority

Priority customers receive first-class treatment. Preemptive priority queues allow a higher-priority job to interrupt a lower-priority job in service while non-preemptive priority queues do not allow this to happen. In either paradigm, no labor is sacrificed.

Shortest Job First

The task with the smallest size is the one that will be completed next.

2.4 Overview of the Three Main Queuing Models: M/M/1, M/G/1, M/M/k

In a system with a single server, where arrivals are governed by a Poisson process and task service timing follow an exponential distribution, an M/M/1 queue operates. See Fig. 2.

Fig. 2. Illustration of an M/M/1 Queue in a Single Server System [2]

An M/G/1 queue (See Fig. 3) is a queue model with a single server, a General distribution for service times, and Poisson-modulated arrivals. The M/G/1 queue is often considered an extension of the M/M/1 queue.

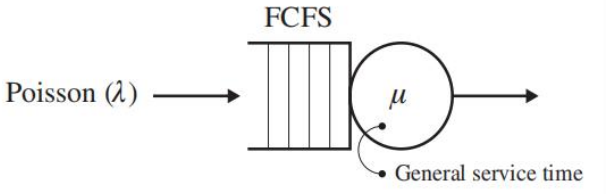

Fig. 3. Poisson-Modulated General Queue [3]

However, nearly no websites, computing facilities, or call centers in today's high-volume world possess just a single server, which is when a “server farm” is introduced. A group of servers that collaborate to manage incoming requests is known as a server farm. Each request may go to a random server, allowing the servers to jointly handle the incoming workflow 9. The M/M/k queue is one Queuing theory functioning in the multi-server system. Fig. 4 demonstrates the M/M/k queue.

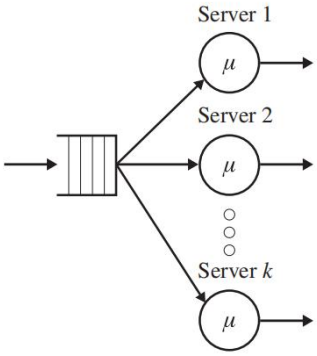

Fig. 4. Representation of an M/M/k Queue in a Multi-Server System [4]

3 Interactions between the Queuing Theory and the Artificial Intelligence Field

3.1 Applications of Artificial Intelligence in the Queuing Theory Research

Artificial Neutral Networks and Queuing Theory. One of the main types of problems that Machine Learning algorithms solve is forecasting--to predict system behaviors according to its previous reactions. It is possible to simplify this forecasting to the problem of estimating the function of numerous variables. In the past research, scientists have widely agreed that one of the top machine learning techniques for estimating functions is Artificial Neutral Networks [5].

The two authors, Vishnevsky and Gorbunova, provide an in-depth exploration of the approximation function in their paper [6], arguing that machine learning approaches, particularly the algorithm for generating a decision tree, are methods that can be used to solve Queuing Theory difficulties. Vishnevsky and Gorbunova further address the promise of machine learning techniques in general, as well as some novel ways of mimicking complicated system behavior and the benefits of simulating a single queuing system. By combining traditional methods with machine learning techniques such as simulation and artificial neural networks, this paper offers a more streamlined process for obtaining estimates of performance measures. This can be especially helpful when dealing with complex queuing models where simulation time can be prohibitive. Rather than taking the effort to simulate all of the necessary input values for the parameters, this unique approach allows for training an intelligent model that can provide estimates for any intermediate values without restrictions on their number. This can ultimately save time, increase accuracy, and provide valuable insights into complex queuing systems 10.

A Novel Approach to Dealing With Queuing Theory Problems Using Machine Learning Methods. Currently, researchers are looking into how to merge various data mining techniques with conventional Queuing Theory methodologies, particularly simulation and Artificial Neural Networks. Current publications on the application of Machine Learning in the queuing Theory field are quite scattered, making it difficult to extract a general idea from them, let alone develop a distinct and novel approach to addressing complex Queuing Theory issues that is comparable to traditional methods. In their paper, Vishnevsky and Gorbunova propose a new approach of the combination that traditional Queuing Theory methods be combined with simulation and ANN. In queuing models, simulation may offer accurate estimations of performance metrics. However, the intricacy of the queuing model, the simulation software environment, and the computing system hardware, all affect how long it takes to achieve a specific value. The two authors further the argument that it is feasible to train neural networks using simulations to generate estimates of intermediary input values for parameters.. This approach reduces the need for simulation modeling of all required input parameter values, but explicit neural network or another intelligent model training is still necessary. The forecast process is time-efficient. Simulation models can be created using a variety of software programs, from specialized tools like AnyLogic, and Arena to custom models developed in the Python programming language, which has a broad range of features and archives, including those for training artificial neural networks. The model run length is a challenging job that refers to the number of requests needed to produce a single output value or output value set. Data obtained from a single run are correlated, which means that numerous realizations are necessary to calculate the average value of any investigated variable. Run length varies depending on the input parameters, increasing the simulation time. However, if an algorithm is used to evaluate a network or queuing system's probabilistic-temporal characteristics, and it has high computational costs, estimates can be obtained for a limited input parameter set before building a neural network and resolving the forecasting issue.

Review on the Application of Machine Learning Methods in the Queuing Theory Field. The employment of machine learning techniques and algorithms in queuing theory is not well represented in the global literature, despite the fact that they are used in an extensive variety of applications in science and technology, particularly the study on cutting-edge broadband wireless networks.

The study of physical queuing is important in areas such as sales and service industries to reduce waiting time and increase service efficiency. Queuing theory is a traditional method used for assessing waiting time, but more recent research has focused on using machine learning methods to predict queuing times. In one study, Sundaria and Palaniammalb uses artificial neural networks to represent the conventional queuing mechanism M/M/1, with one network utilizing a backpropagation technique with a single hidden layer and the other using input and output layers [7]. The results of this study show that neural network models are remarkably compatible with analytical frameworks and are capable of accurately forecasting the parameters of the target provided input data. Another study conduct by

the same authors Sundaria and Palaniammalb uses a neural network to simulate the classical queuing system to organize and enhance lines on an airport runway [8].

The application of neural networks to analyze non-Markov Queueing Systems (QSs) is a promising research topic in the Queuing Theory studies. There are few publications dedicated to this topic, but it is gaining attention as non-Markov systems model the majority of actual physical processes and structures. One of the first works on using neural networks to analyze non-Markov QS models was based on the non-Markov QS with a "warm-up," which can duplicate the activation procedure of a vacant system as soon as it receives a request for the first time after a break. The system was successfully "markovized" by approximating it via a QS with the phase-type distribution of the incoming flow or service time. Due to the lengthy and resource-intensive nature of the mathematical procedures used to determine the static frequencies of states in this kind of QS, neural networks have been developed to simplify the problem without sacrificing accuracy. The article's neural network was built using a double-layer perceptron, and its input parameters were the intensity of the serving and receiving flows, "warm-up," and coefficient of variation. The average waiting and sojourn periods in the system, along with the fixed distribution of the total amount of clients, served as the output parameters. Additionally, research discover that the Bayesian regularization technique is the most accurate of the numerous strategies employed to train the ANN.

3.2 Applications of Queuing Theory in the Artificial Intelligence Field

Virtual Bookstores. To sufficiently understand the multi-agent system asynchronous leave M/MIC queuing model, this section of the paper presents a basic application based on the multi-agent system asynchronous leave M/M/C queuing model, which has a clear and wide application in practice as can be seen by this specific example.

In this application example, the development environment chosen for this section is the TuCsoN Coordinated Architecture Development Environment, in which a virtual bookstore was developed. This virtual bookstore is an example of a virtual organization, and this section will explore two specific processes that may be present in the bookstore scenario.

The two specific processes that occur in the bookstore scenario are: A Single Book Purchase Workflow (See Fig. 5) and Multiple Books Purchase Workflow (See Fig.6).

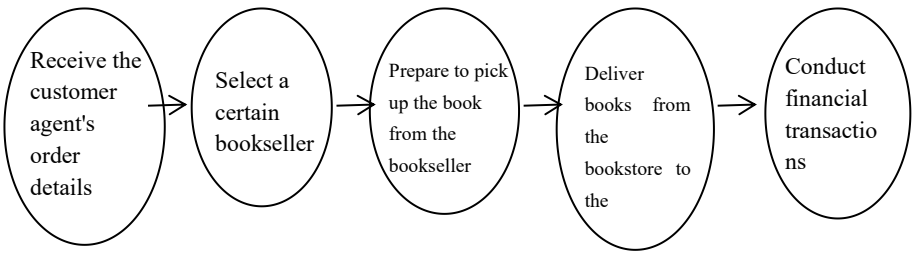

Fig. 5. A single book purchase workflow

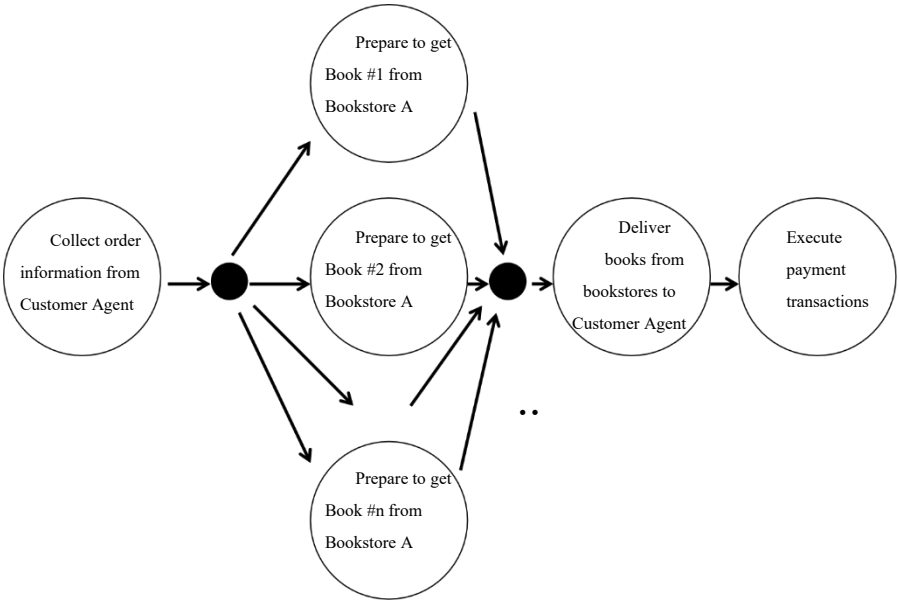

Fig. 6. Multiple books purchase workflow

Fig. 5 depicts the sequence of activities included in a single book buying activity. First, the network collects information about book orders from the user (any supplier of the network participating in the virtual bookstore). The book is then ready to be purchased at the selected bookseller. When the book arrives, the delivery activity is carried out, the delivery person gives the customer the book that has been ordered from the bookseller, and then the payment process is carried out, including the use of an interactive bank to give the customer's money to the bookseller. [9]

Fig. 6 depicts the process of purchasing multiple books from various booksellers. Once the order details have been obtained, a series of required books performs activities in parallel to get the required books from each specific bookseller. When all included booksellers have their books ready, the book delivery activity starts as in the first case.

E-Commerce. The rapid growth of information technology has profoundly changed the conventional business operation mode. Therefore, modern business operation mode should be based on the Internet to realize all business activities such as information release, marketing, purchase, payment and settlement of funds on the Internet. According to this idea, this section selects Electronic Commerce (abbreviated as EC) as an application example of multi-agent help desk leave queuing system for analysis.

The operations office of an E-Commerce software company is primarily responsible for both e-mail response and electronic money order settlement. As soon

as two employees become available (no E-Mail queue), they work together to perform the Electronic Funds Transfer (EFT) clearing operations. Due to financial submission and supervision requirements, any EFT clearing must be performed by both. At this point, the other employees are no longer engaged in auxiliary work if further vacancies occur and are ready to handle new E-Mai1 arrivals. Whenever they finish processing an E-Mai1, they have to check the storage to see if there is a queued E-Mail, and if so, both employees return to their posts at the same time to answer technical requests for help. Otherwise they will process the next E-Mai1 at the same time according to the analysis.

After analyzing, this paper intends to use the M/M/C queuing model to solve the problem in this multi-agent system [10]. Assume that the processing time of an electronic money order obeys the exponential distribution of parameters, according to the partial service desk synchronous leave M/M/C queuing model in multi-agent systems.

Based on these results, the average captain and average wait time in the system can be found by combining the mean value formula of the synchronous leave M/MIC queuing model for some service desks in a multi-agent system. It is clear that the model can be used to study the relationship between operational indicators and parameters, for example: how many additional Agent employees are needed to keep the average waiting time for help E-Mai1 from exceeding 0.3 minutes. As shown in Table 1.

Table 1. Average queue length and average wait time for different number of Agent employees

C_1	6	7	8	9	10	11
$E(W_d)$	1.192124	1.754527	1.945341	1.846647	1.598043	1.298599
$E(L_d)$	3.576373	6.140846	7.781366	1.846647	7.990213	7.142295

C1	12	13	14	15	16	17
E(Wd)	1.003998	0.382798	0.016877	0.001146	0.000046	0
E(Ld)	6.023991	2.488185	0.118141	0.008598	0.000365	0

4 Conclusion

This comprehensive study investigates the intriguing intersection between Queuing Theory and Artificial Intelligence, casting light on the reciprocal advantages these fields can offer each other. Within this examination, a novel and promising approach is proposed that leverages AI's powerful machine learning capabilities, combined with simulation techniques, as an innovative way to analyze and optimize queuing systems. This method has shown its effectiveness by significantly reducing complexity and processing time in determining the numerical characteristics of intricate queuing systems, as demonstrated in a thorough review of globally published literature. In addition, the study delves into the practical application of queuing theory in the realm

of Artificial Intelligence. Notably, the focus is drawn towards its potential impact in emerging digital domains like Virtual Bookstores and E-Commerce. These areas have been increasingly intertwined with AI to optimize user experience, manage resources, and streamline operations, and queuing theory can play a pivotal role in these optimizations. The study hypothesizes that incorporating queuing theory can lead to substantial improvements in system efficiency, customer satisfaction, and overall performance in these digital platforms. Furthermore, this investigation offers a stepping stone for further research and suggests a promising future for the marriage of Queuing Theory and Artificial Intelligence in various practical domains..

References

1. Li, J., Zhang, D., & Xu, Y. (2021). Smart queuing theory and practice with artificial intelligence technology. Applied Soft Computing, 107, 107641.
2. Yan, Y., Li, Y., & Du, H. (2021). Queuing theory and artificial intelligence: A literature review. Journal of Intelligent & Fuzzy Systems, 1-13.
3. Wei, J., Miao, C., Li, J., & Xiang, Q. (2021). Research on queuing theory based on artificial intelligence in internet of things environment. IEEE Access, 9, 73918-73927.
4. Luo, J., Zhang, W., Ji, Y., & Luo, Y. (2020). Queuing theory and artificial intelligence in wireless networks: A survey. IEEE Access, 8, 45859-45872.
5. Li, Y., Guan, W., & Li, W. (2020). Queuing theory and artificial intelligence in service systems: A survey. Mathematics, 8(5), 692.
6. Li, L., Li, J., Zhang, Y., & Zheng, J. (2020). Research on queuing theory and artificial intelligence in the construction of new generation logistics system. IEEE Access, 8, 95661-95670.
7. Chen, Y., Li, Y., & Wang, J. (2020). Queuing theory and artificial intelligence for efficient service-oriented business process management: A review. Sustainability, 12(23), 10009.
8. Li, X., Li, K., & Yu, H. (2019). Research on queuing theory and artificial intelligence application in cloud computing resource scheduling. Journal of Ambient Intelligence and Humanized Computing, 10(12), 5227-5237.
9. Lv, P., Zhang, Z., Xu, Y., & Zhou, X. (2019). Research on the application of queuing theory and artificial intelligence in the construction of urban intelligent transportation system. IEEE Access, 7, 99008-99017.
10. Tang, J., Li, Q., & Wang, J. (2018). Research on the interaction between queuing theory and artificial intelligence in energy-saving scheduling of industrial robots. Energy Procedia, 152, 285-290.