A LITERATURE REVIEW ON DEVELOPMENT OF QUEUEING NETWORKS

V. Narmadha[1], P. Rajendran[2,*]

•

[1,*] Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology,
Vellore, Tamil Nadu 632014, India.
[1]narmadha.v@vit.ac.in, [2,*]prajendran@vit.ac.in

Abstract

This study conducts a quantitative research survey on the development of queueing networks over years. Development is a process of gradual change that takes place over many years, during which a theory slowly progress and attain a good state. Queueing theory has been through many developments which made its existence inevitable in every field. Queueing networks can be considered as a collection of nodes, where each node stands for a service facility. It has been proved to be a powerful and versatile tool for modelling facilities in manufacturing units and telecommunication networks. This paper presents the development in Queueing networks and its types over years. This paper's main objective is to give all the analysts and researchers the knowledge about the evolution that happened in Queueing networks over years.

Keywords: Quazi-reversible Queueing networks (QRQN), Stationary distribution, Automated Manufacturing Systems(AMS), Recurrent neural networks(RNN).

1. Introduction

The goal of queuing theory is to create efficient, cost-effective systems that can serve customers promptly and effectively. Agner Krarup Erlang, a Danish mathematician, statistician and engineer conducted an analysis of the Copenhagen telephone exchange in the early 1900s, which is where queuing theory first emerged. His work paved the way for the development of telephone network assessment and the Erlang concept of effective networks. The notion of queues is used to locate and eliminate bottlenecks in a process. Owing to the fact that queueing models only need a little amount of data and are easy to implement, it is a very effective and useful technique. They can be used to instantly examine and compare different service delivery solutions because of their simplicity and speed. Queuing models can be effective in obtaining insights on the degree of specialisation or flexibility for the utilization of resources in an organisation, which goes beyond the most fundamental task of predicting how much resource is required to accomplish a specific service level. Because of this, there are several prospects for its implementation in several industries.

2. Queueing networks (QN) and its types

In simple words QN are nothing but the jobs moving between interconnected queues in a continuous flow. In some cases, it may be simpler to describe a complex service environment as a queueing network in order to more accurately represent how the service is actually delivered. Telecommunication networks, machine shop problems and computer system are some of the instances of QN.

2.1. Types of QN

Open queueing networks	Closed queueing networks	Mixed queueing networks
•The network must be open for each job class if it has multiple job classes.	•The network must be closed for each job class if it has several job classes.	•The network contains a variety of job classes, some of which are open while the others are closed.

Based on the capacity of the Queue's component there are two types of QN

Blocking

- This happens in a network, when there are one or perhaps more queues with finite capacities.

No blocking

- This happens in a network, when there are many queues of infinite capacity.

Some of the blocking models in QN are listed below

- **Rejection blocking**
  The blocked jobs will be forced to leave the system and it is only applicable for open networks.
- **Transfer blocking**
  The blocked job will wait at source line K_i until the job is accepted at destined line $K_{j.}$
- **Repetitive service blocking**
  The blocked job will again dwell in K_i for another service and the process will be repeated until the job can move out of K_i.
- **Blocking before service**
  At Ki, the service starts for the job only when the destined line K_j is free and is ready to accept the jobs from K_i.

3. Literature Review

The queueing theory has undergone numerous developments, which draws scholars to use it in the best way possible. The Erlang formula, which became a cornerstone of contemporary telecommunication network studies, was developed as a result of Erlang's 1920 publication of telephone waiting times [1], which examined the use of local, exchange, and trunk telephone lines in a close knit community to interpret the empirical needs of an efficient network. One model combines and extends a number of distinct results from open, closed, and mixed network of queues with various kinds of clients [2]. To determine the generic model's equilibrium state probability, this study integrates past results from networks of queues with different customer classes across a range of service areas and a large spectrum of estimated time distributions.QN with individuals from various groups [3] are viewed as a generalisation of mechanisms that enable customers in a particular queue to be heterogeneous. The robustness of QN where customers may be of various types is the subject of this study. QN [4] examines how networks behave when they are in an equilibrium, and in some instances it is demonstrated that the status of one queue does not rely on the condition of the overall network. Since the processes in this study are irreversible, the restrictions placed on the potential customer paths by earlier writers are further loosened. The link between regional balance and product form in QN [5] seems to provide insight into why some domains yield product form answers to problems for queues and networks using nonexponential service domains compared to others. A queueing regimen satisfies station balance if the pace at which customers receive service at every position in the queue is proportionate to the possibility that a customer also could appear at that position. The paper's conclusions extrapolate past research on local balance to any stochastic, differentiable service distribution endeavours that would result in regional balance and product form characterization. In steady state, the state’s distribution has the product form, and interconnections of markov chains and QRQN [6] demonstrate that a network created by joining queues, each of which is QR when taken separately, is also QR. An aspect of convergence of the source performance function component in a closed Jackson network(JN) is enabled by the sample's convergence property in closed Jackson queuing networks [7] research. This finding offers some fresh perspectives on QN theory that might not be found in the well-known product form solution. A category of QN with rejection inhibition has a product form steady state flow pattern, and the overall population is insensitive, according to Exact Solutions for Open, Closed, and Mixed QN with Rejection Blocking [8]. The outcomes are startlingly comparable to those for conventional (non-blocking) networks. The stability of open QN [9] only requires the additional assumption that service time ranges have finite first moments in order to prove stability for the open network. It is permitted for the inter - arrival time distribution to have an infinite first instant. The results are expanded to include multi-server nodes, non-Markovian routing, and Markov modulated arrivals. Recent developments in QN: a survey with applications to AMS [10] highlighted the use of QN models for the performance assessment of AMSs, separately addressing the problems of computing larger parts of performance metrics, blocking events, and analysing an open network of AMS model’s multiqueues.

Using the Right-continuous Markov processes(MP) with values theory which provides a single method for finding both optimal and suboptimal feedback control laws in some QN [11]. This method can be used with QN made up of machines and buffers. The results of single-server QN under optimal control [12] provide effective ways to compute the indices. The greatest remaining index approach is presented in its general form in this publication. We can now locate every index for our universal single server QN model. Unique features of the optimum static routing solution in open BCMP QN [13] determine the relay nodes of the underlying optimal policy and demonstrate that they may not be strange, but the overall determination of the usage of each repair facility is unique. We also take into account a policy that is individually optimal and routes jobs. If each job is aware of the average time delay for each path, it can feel as though its own anticipated response time is decreased The iterative process for a class of Batch-movement QN, which is a natural generalisation of the mean-value analysis of JN, was illustrated in [14]. The recurrence relations used in this approach can be easily extended to the generic group of product-form batch-movement QN/petrinets with equilibrium probability distribution. Using the Z-transform, computing the normalisation constants in QN is made easier [15].

The computation required by the proposed strategy in [16] is relatively simpler than that in Gordon's paper in the scenarios of networks with numerous repair queues or with a single server queue and equal traffic levels. Numerous forms of monotone routing strategies for limited capacity QN have successfully used perfect simulation of index-based routing QN. This research could be expanded to batch arrival or batch services as well as monotone network events more broadly. Through the use of current approximate inference methods from graphical models, probabilistic inference in QN [17] provided a new family of tools for studying queueing models. Networks and queueing systems, models, and applications in [18] uses traditional Markovian systems in queueing systems that combines individual service , an exponential service times and a Poisson arrival procedure. Given that academics are interested in using queueing systems and QN to modelling human performance, it provides an architecture known as the Queueing Network-Model Human Processor. In order to represent an industrial system, a four-input, three-stage QN technique was used [19]. This approach computed the best route that results in the shortest reaction times for the delivery of products to the end destination along the three phases of the network. Modeling a supply chain using a QN , a supply chain is shown as a two-input, three-stage QN [20]. The goal of this study is to determine the minimal response time required to deliver products to their destination along the network's three stages. The total number of products that can be distributed with this quickest response period makes up the QN's maximum capacity.

QN and graphical models are combined in the innovative perspective of reasoning and acquiring knowledge in networks of queues [21], which enables the use of Markov chain Monte Carlo. We use actual data from a standard web application to show how successful the sample is. In order to maximise the throughput of single server, generic QN, a multiobjective technique was devised, called throughput maximisation of QN with concurrent reduction of service rates and buffers [22]. It should be investigated further to see whether more optimum conditions in finite QN can be found using this methodology. Stability in constrained network architectures with queueing lags, queue-storage, blocking back, and control [23] has presented numerous techniques for spatial queuing appropriately without using dynamic assignment, hence the strategy is alternative to the methodology used by Bliemer et al (2012). QN with a single shared server: light and high traffic [24] provide a significant two-fold contribution. First, we examine the system under consideration's precise heavy-traffic asymptotics. Second, based on an approximation between the light-traffic and high-traffic limitations, we construct a closedform approximation for the average lag for random loads. The analysis presented in this work can be expanded in a number of ways, for as by considering various server configuring policies or service standards. However, these results are not explored because of compactness. An approximation technique for the assessment of a finite open QN with Transfer blocking and feedback was described in a restricted open QN application to healthcare systems [25]. An unbounded topology network with a focus on a single server finite capacity model based fertility clinic healthcare system is discussed which uses an expansion approach to determine each node's performance measures and throughput. Deadlock in open restricted QNs has been studied in [26]. It has been demonstrated that analysing the corresponding state digraph of a QN is sufficient to identify stalemate. Three deadlocking QN Markov models have been created. The open two-node, multi-server restricted queueing network requires the development of a Markov model with paths between nodes and feedback loops. Modeling urban taxi services with e-hailings: a QN strategy [27] places an emphasis on the macro-interactions between the urban roadway and taxi systems, but it leaves out the intention of changing speed of the individuals and how they react to the taxi prices. Future research will generalise the suggested QN to take into account the complete dynamics of the taxi market and individual behaviour, giving us keen insight into system control. In a single-class open

QN with Markovian routing, infinite waiting space, and the first-come, first-served , A Robust Queueing Network Analyzer Based on Indices of Dispersion(RQNA-IDC) [28] offers practical algorithms to approximate the performance metrics of the steady state. Future research should focus on a number of excellent directions, such as (i) approximations for flows that use multi-dimensional robust models than one dimensional robust models and (ii) expanding RQNA-IDC to additional open QN models, such as models with more than one servers and other service domains.

It has been demonstrated that there is a direct correlation between the architecture of the QN fluid estimation and the typical activation functions and layers of an recurrent neural networks(RNN) in [29]. As far as we are aware, this is the first method that formally unites the vividness of quantitative performance models with the learning potential of machine learning, favourably contributing to the discussion of whether 'AI will be at the centre of performance engineering'. Using stock critical intensities in a QN , handling shared mobility systems [30] investigates a closed JN taking into account nodes for stops and paths. Focusing on the Mean Value Analysis(MVA) approach, a genetic algorithm was created to solve the issue, and an approximation method was offered to determine the crucial parts from the answer. The model can be expanded to take into account static or dynamic equilibrium techniques as part of additional studies. Hospitals can utilise workflow forecasting to manage healthcare systems in practise, as demonstrated by Simulation and betterment of Patients' Workload in cardiac clinics during COVID-19 pandemic using Timed Colored petri nets [31]. This method would be helpful in these trying times because nosocomial transmission puts the health of the personnel and other patients at danger. A new approach to reducing flight delay rates in airports was presented in Reduction of Delay Rate in Open QN[32] in conjunction with deterministic timed petri nets (DTPN) and open QN. The Federal Aviation Administration's performance is evaluated using the flight delay information gathered by the Operations Network (ON). A numerical example is used to demonstrate the significant reduction in the delay rate. A fresh approach to the problem of multistage semi-open queuing networks(SOQN) i.e., A innovative and all-encompassing method for estimating the work departure process parameters from a SOQN is shown with an application in shuttle-based compact storage systems [33], which adds to the body of knowledge in this area. An accurate assessment of the work departure process from the SOQN is very difficult to perform when the work inter-arrival and service times exhibit broad distributions. As a result, it suggest a practical two-moment approximation method in this study.

4. Developmental analyses

It all began with an infinite series to determine whether a call has to be shut out or allowed. Then a number of queues were taken into consideration which paved way for the rise of QN. When same kind of queues are taken into account for research, over time researchers started combining different types of queues which resulted in open, closed and mixed networks with distinct clients. It further paved way for letting customers within a particular queue to be heterogeneous. The next noteworthy development in QN is the product form networks where the state probabilities are given by the products of functions of number of jobs in the queues. This gave rise to Jackson networks which showed that any arbitrary open QN with k servers that follows an exponentially distributed service time has a product form solutuion. Following this BCMP network was described by Baskett, Chandy, Muntz and Palacios which is an extension of Jackson networks. Based on the capacity of queue's components it is further classified into Blocking and no blocking QN. Supply chain has been combined with QN to efficiently manage organisations. QN's development has paved a way for its application in various fields which includes healthcare sector, transport system and banking sector etc.

5. Applications

QN's development has paved a way for its application in various fields which includes healthcare sector, transport system and banking sector etc.

- QN is applied in a variety of different domains, including computer science, civil engineering, and operations research.
- It is also utilised in computer science to optimise communication network performance and to model the behaviour of computer systems.
- QN is used in civil engineering to optimise traffic flow and model the behaviour of traffic networks.
- On the other hand, QN is applied in operations research to enhance the effectiveness of business processes and optimise resource allocation.

Declaration of conflicting interest: The authors declare that there is no conflict of interest.

References

[1] Matematisk Tidsskrift, B. (1920). *Telephone waiting times*, 31(1):25.

[2] Baskett, F., Chandy, K.M., Muntz, R., and Palacios, M.G. (1975). Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, *Journal of the Association for Computing Machinery*, 22(2):248-260.

[3] Kelly, F.P. (1975). Networks of Queues with Customers of Different Types, *Journal of Applied Probability*, 12(3):542-554.

[4] Kelly, F.P. (1976). Networks of queues, *Adv. Appl. Prob*, (8):416-432.

[5] Mani Chandy, K., John H. Howard Jr., And Donald F. Towsley. (1977). Product Form and Local Balance in Queueing Networks, *Journal of the Assoclauon for Computing Machinery*, 24(2):250-263.

[6] Jean Walrand and Pravin Varaiya. (1980). Interconnections Of Markov Chains And Quazi-Reversible Queueing Networks, *Stochastic Processes and their Applications*, 10:209-219.

[7] Xi-Ren Cao. (1989). The Convergence Property Of Sample In Closed Jackson Queuing Networks, *Stochastic Processes and their Applications*, 33:105-122.

[8] Akyildiz, I.F., and Von Brand, H. (1989). Exact Solutions For Open, Closed And Mixed Queueing Networks With Rejection Blocking , *Theoretical Computer Science,* 64:203-219.

[9] Karl Sigman. (1990). The Stability Of Open Queueing Networks, *Stochastic Processes and their Applications,* 35:11-25.

[10] Ram, R. and Viswanadham, N. (1990). Recent Advances In Queueing Networks: A Survey With Applications To Automated Manufacturing.

[11] Yawn, Y. and Frangos, C. (1992). Optimal Control Of Some Queueing Networks, *MathI. Compui. Modelling,* 16(5):3-12.

[12] Svend-Holger Friis, Ulrich Rieder And Jorgen Weishaupt. (1993). Optimal Control of Single-Server Queueing Networks, *ZOR - Methods and Models of Operations Research,* (37):187-205.

[13] Kameda, H. and Zhang, Y. (1995). Uniqueness of the Solution for Optimal Static Routing in Open BCMP Queueing Networks, *Mathl. Comput. Mode*l, 22(12): 119-130.

[14] Coyle, A.J., Henderson, W., Pearce, C.E.M., and Taylor, P.G. (1995). Mean-Value Analysis for a Class of Petri Nets and Batch-Movement Queueing Networks with Product-Form Equilibrium Distributions, *Mathl. Comput. Modelling*, 22(10):27-34.

[15] Yang, H. and Gong, W.B. (1998). Calculating the Normalization Constants in Queueing

Networks , *Appl. Math. Lett.,* 11(6):87-91.

[16] Jean-Marc Vincent and Jerome Vienne. (2005). Perfect simulation of index based routing queueing networks.

[17] Charles Sutton and Michael I. Jordan. (2005). Probabilistic Inference in Queueing Networks.

[18] Filipowicz, B. and Kwiecien, J. (2008). Queueing systems and networks. Models and applications, *Bulletin Of The Polish Academy Of Sciences*, 56(4).

[19] Vidhyacharan Bhaskar and Patrick Lallement. (2009). A four-input three-stage queuing network approach to model an industrial system, *Applied Mathematical Modelling* (33):3465–3487.

[20] Vidhyacharan Bhaskar and Patrick Lallement. (2010). Modeling a supply chain using a network of queues, *Applied Mathematical Modelling*, (34):2074–2088.

[21] Charles Sutton and Michael I. Jordan. (2010). Inference and Learning in Networks of Queues, *Artificial Intelligence and Statistics (AISTATS)*.

[22] Cruz, F.R.B., Kendall, G., While, L., Duarte, A.R. and Brito, N.L.C. (2012). Throughput Maximization of Queueing Networks with Simultaneous Minimization of Service Rates and Buffers, *Mathematical Problems in Engineering.*

[23] Mike Smitha, Wei Huangb and Francesco Vitib. (2013). Equilibrium in capacitated network models with queueing delays, queue-storage, blocking back and control , *Procedia - Social and Behavioral Sciences,* 80:860 – 879.

[24] Boon, M.A.A., Van Der Mei, R.D. and Winands, E.M.M. (2014). Queueing networks with a single shared server: Light and heavy traffic.

[25] Sreekala, M.S. and Manoharan, M. (2016). An Application of Restricted Open Queueing Networks to Healthcare System, *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, 7(1).

[26] Geraint I. Palmer, Paul R. Harper and Vincent A. Knight. (2018). Modelling deadlock in open restricted queueing networks, *European Journal of Operational Research*, 266:609–621.

[27] Wenbo Zhanga, C., Harsha Honnappab and Satish, V. (2019). Modeling Urban Taxi Services with E-Hailings: A Queueing Network Approach, *Transportation Research Procedia*,(38):751–771.

[28] Ward Whitt and Wei You. (2020). A Robust Queueing Network Analyzer Based on Indices of Dispersion, (6).

[29] Giulio Garbi, Emilio Inserto and Micro Tribastone. (2020). Learning Queuing Networks By Recurrent Neural Networks.

[30] Behzad Maleki Vishkaeia, Iraj Mahdavia, Nezam Mahdavi-Amirib and Esmaile Khorramc. (2020). Balancing public bicycle sharing system using inventory critical levels in queuing network, *Computers & Industrial Engineering.*

[31] Masoomeh Zeinalnezhad, Abdoulmohammad Gholamzadeh Chofreh, Feybi Ariani Goni, Jaromir Klemes and Emelia Sari. (2020). Simulation and Improvement of Patients Workflow in Heart Clinics during COVID-19 Pandemic Using Timed Coloured Petri Nets.

[32] Banu Priya, K. and Rajendran, P. (2019). Reduction of Delay Rate in Open Queueing Network, *International Journal Of Scientific & Technology Research,* 8(12).

[33] Govind Lal Kumawat and Debjit Roy. (2021). A new solution approach for multi- stage semi-open queuing networks: An application in shuttle-based compact storage systems, *Computers and Operations Research.*