A Fuzzy-Based Server Incremental Technique for N-Policy M/G/n Queue Network

.

ABSTRACT

The study presents an N-Policy M/G/n queue model having multiple servers with possible increment. In this model, the server is turned off as soon as the queue is empty. However, as customers arrive the network according to a homogeneous Poisson process with rate *λ*, the server is not immediately turned on untilwhen the number of customers reaches a pre-determined threshold value, *N*. Service of customers is on first-come, first-served basis. The model has two categories of servers viz: active and reserved servers. In a situation where all active servers are busy and a customer arrives, two options are available. It is either the arriving customer wait until one of the active servers becomes idle or request for an additional server from among reserved servers. The decision of the customer to wait for one of the active servers to become idle or request an additional server in order to reduce wait time is taken by an Expert System rather than making such on the basis of network performance metrics.

Keywords: [Queue size, Expert System, Network performance metrics,Threshold value, Throughput]

1. INTRODUCTION

Queues are part of everyday's life. This is so because people wait in cars, banks, hotels, supermarkets, box offices, airports, hospitals and so on. These are examples of visible queues. In fact, queues of voice calls or data packets in communication channels are common but invisible. The large size of traffic in communication lines or computer networks is also a reason why queues cannot be easily avoided [1]. Queues are often undesirable because they cost time, money and resources. They exist because the service resources are not sufficient to satisfy demand. This is because of a number of reasons. Servers may be unavailable because of space or cost limitations, or it may not always pay to provide the level of service necessary to prevent waiting.

Queuing theory is the branch of mathematics which deals with the study of waiting lines. A queue is formed when customers arrive to a service location expecting to be served with limited resources. If the server is not immediately available, the customers need to join a waiting line. The use of queuing theory allows the study of different processes associated with queues including arrivals, waiting and service[2]. The applications of queuing theory in traffic flow, telecommunications and facility design, provides a clear usage of the method in solving a wide range of industrial and domestic problems.

Queuing theory uses mathematical tools to predict the behaviour of queuing systems. Predictions deal with the probability to have *n* customers in the system, mean length of queues, mean waiting time, throughput and so on. A queuing system consists of a stream of arriving customers, a queue and a service process as well as the number of servers. Generally, a queue has the following components:

a. A stochastic process describing the arrivals of customers;
b. A stochastic process describing the service system of customers;

c. The system capacity;
d. The size of customer population; and
e. The queue discipline such as First In, First Out (FIFO); Last In, First Out and so on.

Traditionally, queuing theory considers models with a fixed number of servers. In most of these cases, the main performance metrics considered are queue length and waiting time [3], consequently posing great restrictions on the performance of such system. Advancements in service requirement as well as flexibility in service's delivery had changed this pattern. Consequently, it is more optimal to consider queuing systems with a changing number of servers depending on the queue length.

Queueing models have wider range of applications in service organizations as well as in manufacturing firms, where customers receive service by different kinds of servers in accordance with the queue discipline. In particular, the inter-arrival times and service times are restricted to follow specific probability distributions [4].In some queuing systems, it is required that a certain level of queuing performance, such as the mean queuing delay or the blocking (queuing) probability, be guaranteed for its customers. In classical queuing systems, meeting stringent performance requirements usually results in inefficient server utilization. In some cases, such as in traditional telephone networks, frequently adjusting the number of servers may not be economically justifiable. In order to improve servers' utilization in this situation, the number of servers can be adjusted over a relatively large time scale such as on a daily or weekly basis, according to the forecast of future demand.

An N-policy queue refers to a queuing system in which the server does not start its service until there are *N* customers waiting in the queue. This policy is often used to avoid excessively frequent setups and to minimize servers' cost. The need to adequately determine the number of servers to provide required services in a queue network is paramount as it ensures that expected services are not only offered, but that such are offered within the shortest possible time. In addition, it is not only important to ensure adequate availability of required server(s), it is equally important to ensure that available ones are put to optimal use. In static queue networks, consideration is given to models with fixed number of servers. This poses great restrictions on the performance of such system [3].

One of the important methods to resolving conflict between meeting stringent performance requirements and achieving optimal server utilization is to adjust the number of servers dynamically over time rather than keeping a fixed number of servers all the time. This problem was formulated by [5]. In this study, the authors associated the number of servers S_t in a queue network as a function of time. This is minimized subject to the constraint that the probability of a delay never exceeds a target probability, given the characteristics of the time-dependent arrival process as a function of time. When the change in the number of servers in a queuing system is economically feasible, server utilization could be improved by adjusting the number of servers according to the number of customer(s) in the system at time *t* [6].

In an N-Policy system, the turning on of server depends on the number of customers in the system. When the number of customers in the system reaches a threshold of $N (N \geq 1)$, the server is turned on but not immediately accessible to waiting customers until start-up is completed. After this, the server immediately begins serving waiting customers [7]. A common type of N-Policy is called (v, N)-policy, with $0 \leq \mathrm{v} \leq \mathrm{N} < +\infty$, according to which the server is turned on when *N* customers are present and the server is turned off when it terminates a service with *v* customers left in the system [8]. This duration of the 'start-up' are independent and identically distributed random variables of the general distribution function *U(t),* where $t \geq 0$ with a mean startup time μ_U and a finite $\frac{2}{\partial U}$. Similarly, the service times for a customer are independent and identically distributed random variables for arbitrary distribution function S_t, where $t \geq 0$, a mean service time μ_S and a finite variance ∂_S^2.

Although efficient methods have been developed for analyzing queueing system when arrival rate and service rate of customers are known, however in practical applications when the arrival rate and service rate are described using linguistic terms rather than numerical values, it becomes complicated to evaluate the performance measures of such a queueing system using statistical theory [4]. In this situation, [9] introduced the concept of “fuzziness”. Consequently, fuzzy queueing model was first introduced by [10]. As an extension of this, [11], [12] and [13] improved the concept of “fuzziness” as introduced by [9].

Research findings by [14], the author analyzed fuzzy queueing models using Day/Stout/Warren (DSW) algorithm while [15] analyzed fuzzy N-Policy queues with infinite capacity. Similar to this, [16] implemented DSW algorithm for the brief description of his fuzzy queueing model while [17] studied fuzzy queueing model with multiple servers using the same algorithm and also executed its performance measures.

Automated systems based on fuzzy logic have been used widely in control systems,household appliances, decision-making systems, the medical and automobile industries [9]. While Boolean algebra set values only include “1” and “10” or “True” and “False”, it is believed through fuzzy logic that there are other values between “1” and “0” or “True” and “False”, which are sometimes referred to as in-between values. In other words, Boolean logic engages the principles of totally inclusive and exclusive rules on its set of “1” and “0” while the principles of totally inclusive, exclusive and ‘in between values’ rules is engaged in fuzzy logic [18].

Related works to this study can be broadly grouped into two as follows:

a. A queue network with servers having different service rates and researchers aim at allocating incoming customers to optimize network performance. This gives opportunity to customers to move from long queues to shorter ones or even leave the queue. In this case, [19] proposes that the optimal number of servers is of the form $\lambda + \gamma\sqrt{\lambda}$ depending on the total arrival rate λ for a given grade of service γ. In order to ensure performance optimality, multi-threshold strategies could be adopted as it gives customers opportunity to take decisions when the queue to a given server exceeds a certain threshold [20]; and

b. A queue network with identical servers and researchers aim to distribute customers among available servers which can become active or inactive [5].

This study is aimed at the use of a fuzzy-based Expert System in servers' management in a queue network. In essence, flexibility in the management of servers becomes highly inevitable. The study becomes necessary taking cognizance of the level of flexibility needed in today business environment involving the application of queues in service delivery. Unlike previous studies in which dynamic servers' management is premised on increasing the number of servers correspondingly as the number of customers arriving the network increases in order to save time, this study aims at the use of dynamic management of servers using a fuzzy - based Expert System.

The Expert system manages servers in the queue network by the application of fuzzy rules on input variables to produce an output and consequently applying other fuzzy procedures to arrive at decisions as far as servers' management is concerned in the queue network. A practical application of the proposed model is a customers' service unit of a telecommunication firm. Customers make calls and also use various unstructured supplementary service data codes on their phones to make inquiries on services, request for service upgrade, migrate from one service plan to another, buy airtime and data, among others. In most cases, these requests and services are managed using automated systems which are limited in number. The queuing system considered in this application is illustrated in figure 1.

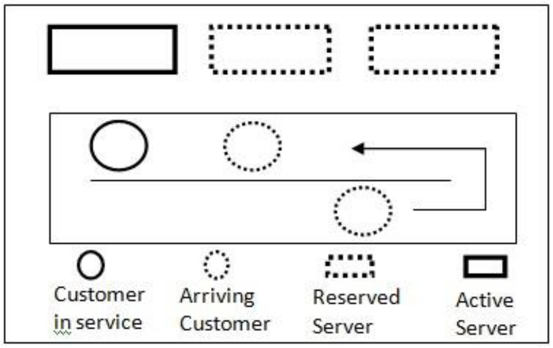

Figure 1: Queuing system in a typical customer care unit of a telecommunication firm

Active and reserved servers which are the customers' care service providers as in our case, share the same queue and are shown as rectangles. If the active server is busy attending to a subscriber and a request is made by another subscriber, the system decides whether or not to take anadditional server from among reserved servers to attend to the arriving request or the new request wait in queue until an active server is idle. In figure 1, there is one active server and two reserved servers, a request in service and an arriving request. The proposed model is flexible such that the fuzzy-based Expert System manages the number of servers in the network taking cognizance of input variables and fuzzy rules which are applied in order to arrive at a decision.

In a multi-server queue system, customers arrive at rate λ. Each customer is served by one server and an arriving customer waits in queue when all servers are busy. There are s servers so that the maximum service rate of the queue is μ, where $\mu$is the service rate of individual servers. If the number of customers in the queue, n, is less than the number of servers, s, the service rate equals n_{μ}. Similarly, in order to ensure queue stability, it is required that the amount of work that arrives per unit time ρ is less than the maximum service rate, i.e. $\rho = \lambda \, \mathrm{E}[S] < s$ [21]. In this case, the equilibrium distribution is obtained using (1) as follows:

$$\lambda P_0 = \mu P_1$$

$$(\lambda + n\mu)\, P_n = \lambda\, P_{n-1} + (n+1)\mu P_{n+1} \quad for \quad n < s,$$

$$(\lambda + s\mu)\, P_n = \lambda\, P_{n-1} + (s+1)\mu P_{n+1} \quad for \quad n \geq s.$$

Consequently, $$P_n = \frac{\rho n}{m(n)} P_0, \qquad (1)$$

where $m(n) = \begin{cases} n! & 0 \leq n < s \\ s^{n-s}\, s! & n \geq s \end{cases}$

for $0 \leq n$

2. MATERIAL AND METHODS

This is discussed under the following sub-headings: schematic structure of the proposed system and fuzzy approach to servers' increment in the proposed model.

2.1. Schematic structure of the proposed system

The proposed model considers a case of *N* customers traveling through and contending for service in a queue network as depicted in figure 2.

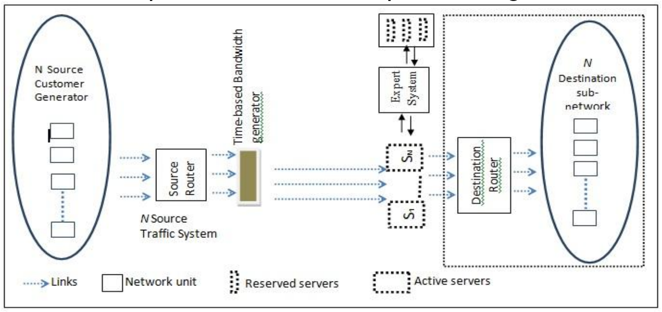

Figure 2: Proposed queue and servers' system

The proposed model in figure 2 has the following components:

a. *N* source customers' generator: This generates sequence of customers and transmits them into the network over available links;
b. *N* source traffic system: This is a source router used to transmit customers to the idle servers;
c. Time-based bandwidth generator: This generates random numbers typical of bandwidth sizes or capacity over a known range of time $0 \leq t \leq T$;
d. Active servers: These serve available customers as they arrive the network;
e. Expert System: When an arriving customer gets into the system and found no idle server, the Expert System chooses whether the numbers of active servers be increased by a unit from among reserved servers in order to serve the arriving customer or to allow it wait in the system until one of the busy servers becomes idle;
f. Reserved servers: These are reserved servers from among which the system chooses to increase the number of active servers whenever the need arises; and
g. *N* destination sub-network: This route served customers to their respective destinations.

2.2. Fuzzy approach to servers' increment in the proposed model

A fuzzy control system is a rule-based system in which a set of rules, called fuzzy rules, define a control mechanism to adjust the system [22]. Generally, a fuzzy logic controller for queues comprises of four principal components: a fuzzification interface, a knowledge base, an inference engine as well as a de-fuzzification interface [23].The output of the fuzzy logic controller is used to tune the system parameters according to some predefined program which is based on the state of the system and it is adaptive in nature.

2.2. Simulation

Matlab trial version was used to simulate the model. As customers arrive, the system computes corresponding values of the average wait time (AW_t) and average service time (AS_t). At a point, the values of AW_t and AS_t were 59.5 and 44.9 respectively giving the current throughputbased on the set of rules using the proposed fuzzy controller depicted in figure 3.

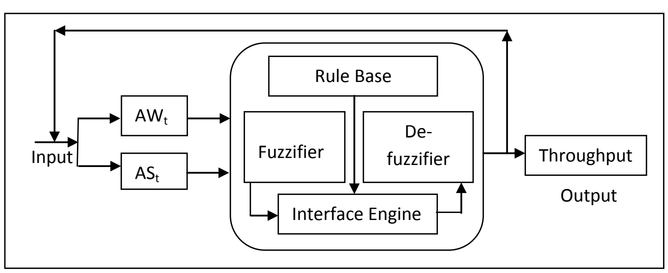

Figure 3: Fuzzy controller used

Each of the two crisp inputs, i.e. AW_t and AW_s were classified into five linguistic variables of “Extremely Low”, “Low”, “Normal”, “High” and “Extremely High” represented as “EL”, “L”, “N”, “H” and “EH” respectively. The throughput which is the output was also classified into five linguistic variables as applicable to the input variables. When a customer arrives, the current values of AW_t and AW_s were obtained and the corresponding throughput is calculated based on the two inputs and the set of rules. There are rules on the basis of which the system operates. The rule function, f is defined as follows:

$$f = \{F, G, V, E\}$$

where “F” is "Fair", “G” is "Good", “V” is “Very Good” and “E” is "Excellent". The corresponding fuzzy rules table is given in table 1.

Average Service Time (AS_t)	**Average Wait Time (AW_t)**				
	EL	L	N	H	EH
EL	F	F	G	G	V
L	F	G	G	V	E
N	G	G	V	V	E
H	G	V	V	V	E
EH	G	V	V	E	E

Table 1: Fuzzy rules table adopted

Consequently, the membership of AW_t is given in figure 4.

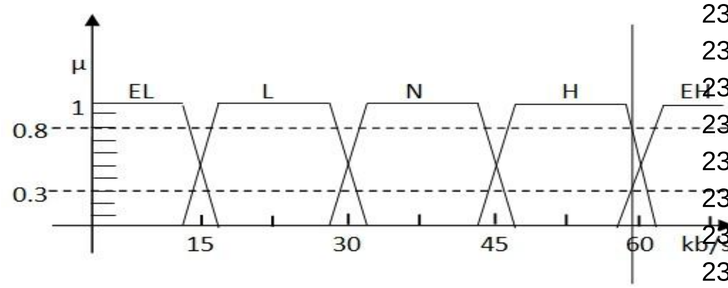

Figure 4: Membership of AW_t

In a similar way, the membership of AS_t is given in figure 5.

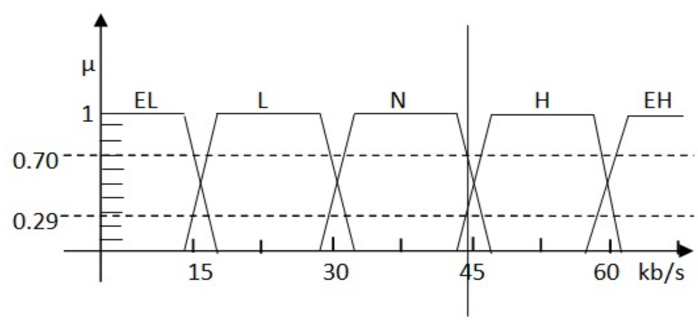

Figure 5: Membership of AS_t

The final decision (FD) is generated based on the minimum operations as indicated in figures 4 and 5 after the minimum value operation. The values are 0.3, 0.8, 0.29 and 0.70. The computation of the FD was made using centroid method as indicated in (2) below:

$$\text{FD} = \frac{\mu_1 D_1 + \mu_2 D_2 + \ldots + \mu_n D_n}{\mu_1 + \mu_2 + \mu_3} \qquad (2)$$

Substituting the minimum values in this equation gives:

$$\text{FD} = \frac{(0.29x0.2) + (0.70x0.4) + (0.29x0.6) + (0.3x0.8)}{0.29 + 0.70 + 0.29 + 0.3}$$

$$\text{FD} = 0.5$$

The FD value of 0.5 is plotted to derive the decision index as indicated in figure 6.

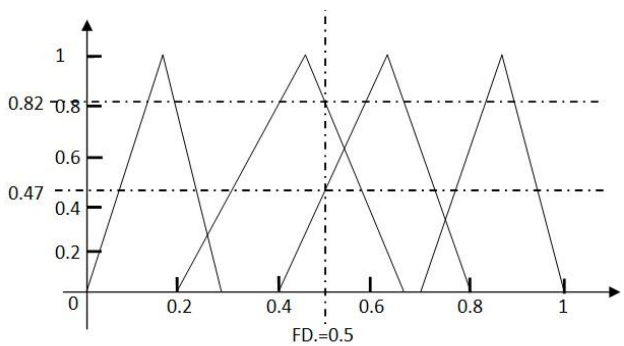

Figure 6: Fuzzified decision index

2.3. Discussion

From figure 6, it is obvious that at 0.82, the ‘wait time’ of customers is high. This implies that thereis a significant number of customers waiting for service turns. Consequently the Expert System request for additional one server from among the reserved servers to complement the services of the active ones in order to reduce the wait time of the customers in the buffer. Similarly, at 0.47, the‘wait time’ of customers is reasonable. Thisimplies that the rate of customers’ arrival has not exceeded the capacity of the active servers.Consequently, if a customer arrives and does not find any idle server to service it, it waits until oneof the busy servers become idle. This decision is taken by the Expert System within the system and not taken on the basis of any network performance metric. This implies that every time a customer arrives the system and found no idle server, the current value of AW_t and AS_t are obtained and the throughput is calculated based on the two inputs and the set of rules and decision is taken on the basis of the output.

3. CONCLUSION

The study describes an N-Policy M/G/n queue model with possible server increment in which customers are served on first-come, first-served basis by active servers. As customers arrive the system and found no idle server, the Expert System determines the action totake using fuzzy logic approach, consequently making the management of servers dynamic. The contribution of the study is that the model is able todynamically manage the number of servers in queue network using fuzzy logic approach as against the usual idea of correspondingly increasing the number of servers in a queue system as the number of customers increases or by considering certain network performance metrics. The proposed model does not necessarily increment servers correspondingly to service requests of arriving customers but also ensures that available servers are put to optimal use.

ACKNOWLEDGEMENTS

Nil

COMPETING INTERESTS

Authors have declared that no competing interests exist.

AUTHORS' CONTRIBUTIONS

All the authors prepared the manuscript, carried out the simulation, read the manuscript and approved it for submission.

4. REFERENCES

[1] Thamotharan, S. "AStudy on Multi-Server Fuzzy Queueing Model in Triangular and Trapezoidal Fuzzy Numbers Using α-Cuts." *International Journal of Science and Research*.5(1).2016. 226-230.

[2] Mohammed, S. A. "Fuzzy Queue with Erlang Service Model using DSW Algorithm." *International Journal of Engineering Sciences & Research Technology*.5(1).2016. 50-54.

[3] Alenany, E. and El-Baz, M. A. "Modelling a Hospital as a Queuing Network: Analysis for improving Performance". *Journal of Industrial Manufacturing Engineering*.11(5). 2017. 1181-1187

[4] Sujatha, N., Murthy-Akella, V. S. N. and Deekshitulu, G. V. S. R. "Analysis of Multiple Server Fuzzy Queueing Model using α – Cuts."*International Journal of Mechanical Engineering and Technology*.8(10). 2017. 35–41.

[5] Jenings, O. B., Mandelbaum, A., Massey, W. A. and Whit, W. Server Staffing to Meet Time-Varying Demand: A presentation at the 2nd INFORMS Telecommunication Conference, Florida. 1996. 24-26

[6] Yu, Z., Liu, M. and Ma, Y. "Steady State Queue Length Analysis of a Batch Arrival Queue under N-Policy with Single Vacation and Set-Up Times". *Intelligent Information Management*. 2. 2020. 365-374.

[7] Li, H. and Yang, T. "Queues with a Variable Number of Servers". *European Journal of Operations Research.* 124. 2000. 613-628

[8] Ghimire, S., Ghimire, R. P., Thapa, G. B. and Fernandes, S. "Multi-Server Batch Service Queuing Model with Variable Service Rates". *International Journal of Applied Mathematics and Statistical Sciences* 6(4). 2017. 43-54

[9] Zadeh L. A."Fuzzy Sets as Basis for a Theory of Possibility." Fuzzy Sets andSystems.1(1). 1978. 3-28.

[10] Li, R.J. and Lee, E. S.“Analysis of Fuzzy Queues.” *Computers and Mathematics withApplications*.17(7). 1989. 1143-1147.

[11] Buckley, J.J.“Elementary Queueing Theory based on Possibility Theory”.*Fuzzy Sets and Systems*.37. 1990. 43-52.

[12] Nege, D. S. and Lee, E.S.“Analysis and Simulation of Fuzzy Queue.”*Fuzzy Sets andSystems*. 46. 1992. 321-330.

[13] Chen, S. P.“A Mathematics Programming Approach to the Machine Interference Problem with Fuzzy Parameters.”*Applied Mathematics and Computation*. 174. 2006. 374-387.

[14] Ritha, W. and Sreelekha-Menon, B. “Fuzzy N policy Queues with Infinite Capacity.”*Journal of Physical Sciences*. 15. 2011. 73-82.

[15] Shanmuga-Sundaram, S. and Venkatesh, B. B. “Fuzzy Multi-Server Queueing Model Through DSW Algorithm.” International Journal of Latest Trends in Engineering and Technology. 2015. 452-457.

[16] Manish, R. and Sedamkar, R. R. “Design of Expert System for Medical Diagnosis Using Fuzzy Logic.” *International Journal of Scientific and Engineering Research.*4(6). 2013.2914-2921

[17] Faran, B., Saleem-Khan, M., Yasir, N. and Imran, N.“Design Model of Fuzzy Logic MedicalDiagnosis Control System”.*International Journal onComputer Science and Engineering* (IJCSE).3(5). 2011. 2093-2108.

[18] Hu, B. and Banjaafar, S. “Partitioning of Servers in Queuing System During Rush Hour”. *Manufacturing and Service Operations Management* 11. 2009. 416-428.

[19] Yang, D. Y. and Wu, Y. Y. Analysis of a Finite-Capacity System with working Breakdowns and Retention of Impatient Customers. Journal of Manufacturing Systems. 44. 2017. 207-216

[20] Rajadurai, P., Saravanarajan, M. C. and Chandrasekeran, V. M. “Analysis of MX/G/1 Retrial Queue with Two Phases Service under Bernoulli Vacation Schedule and Random Breakdown”. *Mathematics in Operations Research*.7(1). 2015. 19-31.

[21] Munoz, E. and Ruspini, E. H. “Simulation of Fuzzy Queuing Systems with a Variable Number of Servers, Arrival and Service Rates”. IEEE. 2014. Available at http://dx.doi.org/10.1109/TFUZZ.2013.2278407

[22] Ayyappan, G. and Nirmala, M. “An MX/G(a,b)/1 Queue with Breakdown and Delay Time to Two-Phase Repair Under Multiple Vacation”. *Applications and Applied Mathematics* 13(2). 2018. 639-663.