CARITAS UNIVERSITY AMORJI-NIKE, EMENE, ENUGU STATE

Caritas Journal of Engineering Technology

CJET, Volume 3, Issue 2 (2024)

Article History: Received: 15th August, 2024 Revised: 20th September, 2024 Accepted: 10th October, 2024

Application of Utilization Rate Approach for the Evaluation of Effective Numbers of Servers for A Parallel System of Single-Server Queues

[1]Owu F. U.
[2]Oboba C. U.
[1]Department of Mechanical/Mechatronics Engineering, Federal University Otuoke, Bayelsa State Nigeria
[2]Department of Production Engineering, Faculty of Engineering, University of Benin, Edo State Nigeria.
Corresponding authors' email: festek1028@gmail.com

Abstract

Prospective service delivery and manufacturing firms as the case may be, are desirous of knowing the number of servers and service delivery operating equipment to be installed in their proposed facilities, to ensure adequate service delivery. Perhaps, this underlining interest may not be unconnected with the fact that finite facility space and limited income constraints probably will not allow for infinite severs and manpower. In this paper, the M/M/K (an exponentially distributed interarrival, service time and multiple servers) parallel system of single-sever queues model is modified and deployed to analyse the services of a petroleum product loading depot. The adequacy or otherwise of the loading process is measured by a stability equation developed from the utilization rate equation. The constraints of the stability equation required to be satisfied to ensure effective service delivery (bounded queues and decongested facility) establish a range of new queue parameters. The least mean service rate value within this range can be evaluated with a server equation to produce the minimum number of effective servers, required by any firm that will yield a utilization factor of less than one. A numerical evaluation of the stability equation and the server equation is carried out with primary data sourced from the Warri refinery depot. It is found that the current operating system in the depot with nine parallel servers is ineffective with a utilization factor of 1.17, hence the resultant congestion in the depot. Upon satisfying the stability equations constraints, a new mean service rate value is estimated that gives a utilization factor of 0.97 less than one. The new service time value corresponds to a minimum effective number of eleven servers by the server equation, to give a utilization factor of 0.97. Better services with lower utilization rate values can be achieved by varying the mean service time along the established range of the stability equation's constraint.

Keywords: Paralley system, Petroleum, Servers, Service delivery, Stability equation.

1 Introduction

Queues or waiting lines are common in service delivery and manufacturing facilities. The literature on the use of queue theory in modelling various queueing problems is limitless. The variants of these queue models are a result of the different approaches in handling queue features such as; arrival and service processes, service discipline, queue capacity, numbers and configuration of servers and network structure. One of the variants of the queue models that takes into cognizance the numbers and configuration of the servers, is the parallel system of single-sever queues, as opposed to the multiple-severs single-queue system. The parallel system of single-server queues is common to supermarkets, where customers queue behind any server of their choice at the check out point [7]. The choice of a queue is informed by the queue size, which by extension speaks of a faster service rate.

Again, the literature on the parallel system of single-server queues is in-exhaustive, as they come in different shades, characterized by constant or variable service rate, [8], attitude of customers in the queue as to whether they can jockey their position [1], finite or infinite queue capacity [1, 8, 10].

Increasing the number of servers in a parallel system of single-server queues will certainly improve the utilization rate [7], as the expected number of customers in service will be increased, which invariably means, the proportion of time the server is busy over the service time will be reduced. The challenge of this method of improving utilization rate as a decongestion strategy is that of facility space constraint and limited income sources, which negates infinite servers, limitless manpower and service time.

The above scenario requires that a feasible number of servers alongside a practicable service time be determined concerning an effective utilization factor to offer quality and efficient service delivery.

In [1], two parallel queues of M/M/I type were considered. Here, each queue's capacity is restricted and customers are allowed to Jockey from one queue to the other, but the arrival is controlled such that on arrival, customers are made to join the shortest queue. An optimal assigning policy was developed that assigns customers to a faster queue when that server has a shorter queue.

The work reported in [2] analysed a system of parallel queues and a single server. A combination of simulation and neural network techniques was used to obtain the optimal scheduling policy. For [3], a parallel server queuing system consisting of buffers for holding incoming jobs and flexible servers for processing these jobs was considered. A dynamic scheduling system was designed for assigning jobs to available servers.

[4] considered a parallel system of multiple servers that operate two periods of uptime and downtime. The amount of service a customer receives during the uptime before the downtime is conserved for the next uptime, as only arrivals are allowed during downtime without service. In all the reported literature and many more not presented, the emphasis is on the distributions of the arrivals and service times, the assignment of customers to queues, service discipline and better approximation of queue parameters. In this study, a deliberate effort is made to evaluate the minimum number of feasible servers and service time that will support adequate service delivery in a parallel system of single-server queues through the utilization rate consideration.

2 Materials and Method

2.1 Data Collection

The data used for the numerical evaluation in this paper are primary data sourced from the record of the loading operation of the Warri refinery depot.

2.2 Model Development

Consider a petroleum product loading facility with a service system of k parallel servers and permissible $(k + c)$ servers, where c is the allowable additional server due to facility space constraint. It is contemplated that there is a waiting zone (calling source) within the facility for arriving trucks with an arrival of ψ for a given time with a mean value of $\bar{\psi}$. The trucks are assigned equally to each server and the k servers are expected to complete service in a service time of α with a mean value of $\bar{\alpha}$. Let λ and μ be the mean arrival rate to the facility and the mean service rate of the k servers respectively. Assuming the k servers are serving at the same rate, then the utilization rate of the k servers for a parallel system of single server queues can be expressed as;

$$\rho = \frac{\lambda}{\mu} \qquad (1)$$

See [12&13],

where ρ is the utilization rate of the k servers.

The condition for the k servers to be effective concerning bounded queues and a decongested facility is,

$$\rho = \frac{\lambda}{\mu} < 1 \quad (2)$$

Supposing this is not the case, i.e. the k servers are inadequate, then a constraint has to be imposed to satisfy eqn. (2) for the selection of effective servers k_o. The constraint or stable utilization rate equation can be expressed as follows;

$$\rho = \frac{\lambda}{\mu_o} : \ 0 < \lambda < \mu_o` \quad (3)$$

where $\mu_o$is the mean service rate of the k_o servers to serve the same mean arrival rate of λ. In this paper, equation (3) above is the stability equation, since the satisfaction of the given constraint will result in a utilization rate of less than one. Since the same mean arrival of λ served by k servers at a rate of μ is sought to be served by k_o servers at a service rate of μ_o for an effective result (utilization rate of less than one), it is possible by inverse proportion to draw a relation between the parameters of k, k_o, μ and μ_o, provided, $k_o > k$ and $\mu > \mu_o$ as it is with the case under study.

Simply put;

$$\frac{\mu}{\mu_o} = \frac{k}{k_o} \quad (4)$$

See [14, 15 & 16]

$$\Longrightarrow k_o = \frac{k\mu_o}{\mu} \quad (5)$$

Equation (5) is the effective servers equation, based on the satisfaction of the constraint of equation (3). Note that the desire to satisfy the constraint of equation (3) by the selection of a new service time is because the service time is endogenous to the facility, compared to the arrival that is exogenous and is practically impossible to influence from within the facility.

After satisfying all required logistics for improved services, a Facility manager or operator must insist on competence measures estimated on a mean service rate evaluated over the range of $\mu_o = \lambda + n$ where $n = 0.1, 0.2, 0.3\ldots..\infty$

Suppose a facility operates a system of k servers with a resultant ineffective service delivery and wishes to improve services by upgrading to a system of $(k + c)$ servers, it is possible to determine if $(k + c)$ will produce an effective service, base on a utilization rate of less than one, where c is the permissible additional servers due to facility space constraint.

This is achieved by substituting $(k + c)$ for k_o and $\mu_{(k+c)}$ for μ_o In eqn. (5) and comparing the value of $\mu_{(k+c)}$ with the stable values of μ_o In the range established by the constraints of eqn. (3). The procedure is as follows;

$$\mu_{(k+c)} = \frac{(k+c)\mu}{k} \quad (6)$$

The condition of sever effectiveness or otherwise for $(k + c)$ servers is as follows.

$\mu_{(k+c)} > \lambda$ stable system

$\mu_{(k+c)} < \lambda$ unstable system

3 Result and Analysis

3.1 Result

3.1.1 Computational Application of the Developed Model.

In the collection of data for the numerical evaluation of this study, twenty observations of six (6) hours of interarrival time were carried out. The result is presented in the table below.

Table 3.1: Arrivals per Six hours

Observation	Arrivals to the Facility per Six Hours	Service time (α)
1	21	7
2	19	7
3	22	7
4	23	8
5	17	7
6	15	8
7	19	7
8	20	7
9	24	8
10	20	7
11	18	7
12	20	7
13	21	7
14	22	7
15	23	7
16	24	8
17	23	7
18	21	7
19	24	7
20	20	7
	$\sum \psi = 416$	$\sum \alpha = 146$

From Table 3.1, the mean arrival per six hours

$\bar{\psi} = \frac{416}{20} = 20.8 \approx 21$ trucks per six hours. where $\bar{\psi}$ is the mean arrival per six hours

Similarly, the mean service time $\bar{\alpha} = \frac{146}{20} = 7.3 \approx 7$hrs per 21 trucks

The server's mean arrival rate (λ) is given as follows

$\lambda = \frac{\bar{\psi}}{6} = \frac{21}{6} = 3.5$ trucks per hour

The server's mean service rate (μ) is,

$\mu = \frac{\bar{\alpha}}{7} = \frac{21}{7} = 3$trucks per hour.

The Warri refinery currently has nine (9) operational servers (k) and a space for an additional two ($a = 2$) if the need arises.

Given that $k = 9$ and $c = 2$

From eqn. (3), the constraint for stable system or effective servers requires that,

$0 < \lambda < 3$

As this is not the case for the above scenario, the utilization rate (ρ) is likely going to be above one, indicating high congestion or high traffic within the facility.

Recall eqn. (1) $\rho = \frac{\lambda}{\mu} = \frac{3.5}{3} = 1.17 \approx 1.2$

Expectedly, the value of ρ is above one indicating high congestion or high traffic within the facility

Recall that for a stable or low traffic intensity service system according to the constraint of eqn. (3);

$0 < 3.5 < \mu_o$

Assuming that the new values of the mean service rate (μ_0) will increase in the following order;

$\mu_0 = \lambda + n$

Where $n = 0.1, 0.2, 0.3 \ldots . \infty$

then the first value of $\mu_o = 3.6$, Hence,

$$\rho_0 = \frac{\lambda}{\mu_o} = \frac{3.5}{3.6} = 0.972 \approx 0.97$$

where ρ_0 and μ_0 are the utilization rate and the mean service rate of the k_o Effective servers.

Recall eqn. 5;

$$k_o = \frac{k\mu_o}{\mu} = \frac{9 \times 3.6}{3} = \frac{32.4}{3} = 10.8 \approx 11$$

From the above evaluation of the servers' value, eleven servers will generate a utilization value that is below one. Though slightly high, this value will decrease gradually as the value of the mean service rate increases within the specified range.

Given that $k = 9$ and $c = 2$,

then from eqn. 6, we have that,

$$\mu_{11} = \frac{k_{(11)} \times \mu}{k} = \frac{11 \times 3}{9} = 3.67 \approx 3.7$$

Since $\mu_{11} > \lambda$ i.e $\mu_{11} > 3.5$, then the corresponding number of servers i.e eleven is an effective number of servers with a utilization rate of less than one as shown below;

From eqn. (1);

$$\rho = \frac{\lambda}{\mu} = \frac{\lambda}{\mu_{11}} = \frac{3.5}{3.7} = 0.95.$$

Consider the service delivery system for $n = 0.5$

$\Rightarrow \mu_o = 3.5 + 0.5 = 4$ trucks per hour.

Given that $\mu_o = 4$, then the utilization rate ρ_o for the new service rate is evaluated below as;

$$\rho_o = \frac{\lambda}{\mu_o} = \frac{3.5}{4} = 0.875 \approx 0.88.$$

Hence the value of the corresponding k_o is given as follows;

$$k_o = \frac{k\mu_o}{\mu} = \frac{9 \times 4}{3} = 12 \text{ servers}$$

3.2 Analysis

It was estimated, using the developed model that the current operating system in the Warri refinery with nine operating servers has an individual traffic intensity of 1.17. This value accounts for the high traffic congestion experienced at the deport. Again, when the value of $\mu_o(3.6)$ was selected from the range of the stable values on account of eqn. (3), eleven (11) effective servers were obtained, this being the least number of effective servers since the selected μ_o is the least stable value in the established range of $0 < 3.5 < \mu_o$. The selected value of μ_o produced a utilization factor of 0.97, which provided a better service than that of nine servers. While investigating whether the allowable number of servers in the Warri refinery deport for a value of $c = 2$ will yield an effective service, eqn. (6) was employed and a mean service rate ($\mu_o = 3.7$) was obtained. This falls within the range of stable values of $0 < 3.5 < \mu_o$. This indicates that management must as a matter of urgency install the unutilized servers to remedy the traffic congestion in the deport, as eleven servers will produce a utilization rate value of 0.97. This new value of the utilization rate showed an

improvement of 17.1% efficient service over the previous value of 1.17 for the nine (9) servers. Selecting $\mu_o = 4$ at random from the range of stable values of $0 < 3.5 < \mu_o$, an effective server value of twelve (12) servers was obtained. As it concerns the Warri refinery deport, even though twelve (12) servers will be effective with a utilization rate of 0.88, it is not permissible due to space constraints. Though this number is beyond the acceptable number of servers due to space constraints, the facility operator must strike a balance in the trade-off between a reduction in traffic by over 25% over the 9 servers or 9.3% over the 11 servers and the extra cost of expanding the facility space in accommodating two or three servers.

In the application of the developed model to general service delivery operations, one simply needs to have a fair knowledge of arrivals per some periods and an affordable service rate from a pre-selected stable value to regulate the number of servers that will provide adequate services concerning traffic congestion.

Since the service rate is endogenous to any facility, it is more amenable to adjustment given better services compared to arrivals that are much more subject to uncertainty influenced by issues exogenous to the facility. The developed model therefore allows for the flexibility of this adjustment.

4 Conclusion

A model to evaluate the effective number of servers in a parallel system of single-server queues, that will produce adequate services in terms of facilities congestion concerning utilization rate has been developed. Although the effectiveness of service as defined for a utilization rate of less than one may not reflect an optimal server selection procedure, arbitrary selection of servers has been curtailed with the development of this model, as a stable range of service time for any given number of servers has been introduced for which a utilization rate of below one which is required for adequate services can be selected. The range of the mean service values exposes management to an option of a server value that is feasible based on improvements on basic requirements. The study findings will not only help operators of already functioning service delivery facilities to improve their services but will also allow prospective facility owners to project for adequate services and choose between the options of better server configurations as it relates to single queues parallel servers or parallel systems of single server queues.

Declarations

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

Not applicable

Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors contribution

F.U.O and O.C.U – Conceptualization, Drafting of the article, Data sharing, technical support, Funding acquisition, , methodology, interpretation of data and materials support and final approval.

Funding

The authors received no funding for this study.

Acknowledgement

The authors wish to thank the management and technical staff of the Department of Chemical Engineering at Federal University Otuoke Nigeria and the Department of Production Engineering, Faculty of Engineering, University of Benin, Edo State Nigeria.

REFERENCES

[1] Ahmed M. K. Tarabia, (2008). Analysis of two queues in parallel with Jockeying and Restricted Capacities. Applied Mathematical Modelling.
DOI : 10.1016/j.amp.2007.02.0110.1016/j.amp.2007.02.014.

[2] Efrosinin, D., Vishnersky, V., Stepanova, W. (2023). Optimal Scheduling in General Multi-Queue System by Combining Simulation and Neural Network Techniques. Sensor 2023, 23, 5479. https.//doi.org/10.3390/523/25479.

[3] Bell, S. L. And Williams R. J. (2005). Dynamic Scheduling of a parallel server system in Heavy Traffic with Complete Resource Pooling: Asymptotic Optimality of a Threshold Policy. Electronic Journal of Probability. Vol 10 (2005). Paper no.33, pp.1044 – 1115

[4] Gnodong, P. And Yuhung Z. (2016). G/G/∞ queues with alternatry interruptions. Advances in Applied Probability. Vol. 48, issue 3, pp 812 – 831

[5] Bo Li and Gnodong Pang (2023). Heavy – Traffic Limits for Parallel single–server queues with randomly split Hawkes arrival process. Journal of Applied Probability, first view, pp. 1=25. Doi: https://doi.org/10.1017/Jpr.2023.50.

[6] Charles Carlos Roberto. (2016). Approximation for Single – Channel muti-server queues and queuing Networks with generally distributed inter-arrival and service times. Doctoral Dissertation. 2644. https.//scholarsoniare.rrist.edu/doctoral-dissertations/2644

[7] Steven Nabmias (2009). Production and Operations Analysis. Sixth Editions. Published by McGraw-Hill/hrwin business unit of McGraw-Hill companies inc., 1221 Avenue of the Americas, New York, 4y100020.

[8] Arie hordisk and Ger Koole, (2009). On the assignment of customers to parallel queues. Probability in the engineering information Science vol. 6, issue 4.

[9] Crane, M., and Ighebart, O. (1974). Stimulating Stable Stochastic Systems, 1: General multi-server queues. Journal of association for computing machinery, Vol. 21, No.1. pp.103-113.

[10] Kebarighotbi, A. And Cossandras, C. (2011). Optimal Scheduling of Parallel queues using stochastic flow models. Discrete Event Dynamic Systems. Vol.21. pp. 547 – 576

[11] Rowland, J. O. Ekeocha, and Victor, I. Ihebom (2018). The use of Queuing Theory in the Management of Traffic Intensity. International Journal of Science Vol. 7. no.3

[12] Anyin, P. B. And Etika, A. A. (2022). Analytical Determination of Queuing System Performance for Sustainable Economic Development. Arid Zone Journal of Engineering, Technology and Environment. Vol. 8(3) pp. 517 – 526

[13] The Ramesh Babu, K. V., Sravani, D., Bhargara, D. K., Venkateswarlu, W. And lingarao, B. (2019). The effect of traffic intensity in social networking with a special reference to WhatsApp. A Case Study. Turkish Journal of Computer and Mathematics Education. Vol. 10. No. 03 pp. 1361 – 1366

[14] Parmjit Singh (2015). Understanding the Concept of Proportion and Ratio Constructed by Two Grade Six Students. Educational Studies in Mathematics, Vol. 43, No. 3 (2000), pp. 271-292.

[15] Nisa Azzahra. Tatang Herman and Dadan Dasari (2022). Analysis of Inverse Proportion in Mathematics Textbook Based on Praxelogical Theory. Journal Analisa, 8(2), pp. 152-167.

[16] Muhhammad Irfan, Toto Nusantera, Subanji, Suwool (2019). Direct Proportion of Inverse Proportion? The Occurrence of Student Thinking Interference Vol. 8, Issue 07.