SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Modelling Of Multi-Node Computer Networks With Task Abandonment And Transient Congestion Analysis

Sachin R. Gurnule[1], V.N. Rama Devi[2], P. Pranay[3], K. Kalyani[4], P. Varaprasada Rao[5], Dr Uttam Mande[6]

[1]Research Scholar, Department of Statistics, Chaitanya Deemed to be University, Warangal, Telangana, sgurnule28@gmail.com
[2]Professor, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, ramadevivn@gmail.com
[3]Associate Professor, Department of Mathematics and Statistics, Chaitanya (Deemed to be university),Telangana, pettempranay@gmail.com
[4]Assistant Professor, Department of Mathematics & Statistics, Vignan's Foundations for Science, Technology & Research (Deemed to be University), Guntur -Tenali Rd, Vadlamudi, kruthiventikalyani@gmail.com
[5]Professor in CSE, Gokaraju Rangaraju institute of engineering and technology, Hyderabad, prasadp.griet @gmail.com
[6]Professor, Depatment of Cse, Raghu engineering college, uttammande@gmail.com
Corresponding Author: Dr. V.N. Rama Devi, Professor, GRIET, Hyderabad (ramadevivn@gmail.com)

Abstract

Modern distributed computing environments, including cloud platforms, data centers, communication networks, web servers and edge computing systems, frequently operate under dynamic workloads, limited processing capacity, transient congestion and delay-sensitive task behavior. This paper presents a transient analysis of a finite-capacity M/M/1 Jackson-type queueing network with impatient tasks for modelling multi-node computer systems. The proposed network consists of two distributed processing nodes connected to a central server. Tasks enter the system from the outside on different nodes, are served, possibly routed to the central server or leave the system because of impatience by balking and reneging processes. The transient state probabilities are developed by Kolmogorov forward differential equations and computed numerically through forth order Runge–Kutta approximation, which is programmed in MATLAB. The performance of the node-wise and central-server is measured in terms of expected buffer occupancy, mean processing delay, congestion level and propagation of workloads. The graphical results show progressive workload accumulation during the transient period, with increasing buffer occupancy and processing delay as the system evolves. The central server has the greatest congestion due to the fact that it receives the external arrivals as well as the tasks that are routed from the distributed nodes, and thus is the main congestion point of the network. Sensitivity analysis suggests that an increase in the arrival rate and the routing probabilities will lead to an increase in congestion and delay, while an increase in service rates will lead to an improvement in processing efficiency and a decrease in waiting. Balking and reneging shorten the length of queues and delay, but are task loss in delay-sensitive applications. The proposed model would be a convenient analytical tool for transient congestion, resource allocation, bottleneck identification and service performance in multi-node computer networks.

Keyword: Distributed computing, Jackson-type queueing network, transient analysis, task abandonment, balking and reneging, Runge–Kutta method, congestion control.

Introduction

Queueing networks are important tools for modelling distributed computing systems, cloud infrastructures, web servers, edge computing systems, Internet of Things (IoT) architectures, and packet-switched communication networks in contemporary applications of computers in science. In such cases, the requests from users may come dynamically and face delays because of congestion, limited processing power, or temporary overload situations. If the time-sensitive requests have to wait too long, they may reject the system, causing task drop, timeout, and decreasing Quality of Service (QoS). Hence, a transient queueing analysis is necessary to analyze the short-term system behaviour and help manage the computational system resources.

The queueing theory has evolved to be a significant approach to the modelling and performance analysis of contemporary computer systems, communication networks, cloud infrastructures and distributed processing systems. Jackson did the foundation work that led to the product-form solution for open queueing networks, allowing for tractable analysis of interconnected service nodes with probabilistic routing [1]. Kelly then extended these ideas to stochastic network reversibility, which greatly deepened the theory of studying large scale distributed systems [2,3].

Queueing networks are widely used in computer science and communication engineering to model various systems, such as packet-switched networks, Web servers, cloud data centres, edge computing systems, and distributed architectures. Congestion control and packet routing were the focus of the application of queueing models to data communication networks discussed by Bertsekas and Gallager [4]. Harchol-Balter presented a comprehensive modeling of computer systems performance using queueing theory, especially for the server scheduling, workload balancing and resource allocation in distributed computing environments [5].

With the dynamic nature of modern computational workloads, transient behaviour in queueing systems has received much attention. Transient analysis can detect congestion, workload variations, and transient overloads that are typical of real time computing applications that are not detected by steady state analysis. Markovian queues were analysed by Abate and Whitt by transient analysis based on Laplace transforms [6] and time-dependent characteristics were analysed by Neuts using matrix-geometric techniques [7]. These techniques are useful for assessing delay sensitive computer networks and communications systems.

In computer-oriented applications, requests might abandon the system when they are delayed too much and this behaviour is known as 'reneging' and the customer impatience is represented by 'balking'. Garnett et al. investigated impatient customers in a call centre system and showed its effect on the system performance [8]. Later, Zeltyn and Mandelbaum analysed the M/M/n+G queue with abandonment and highlighted its importance in service engineering and communication systems [9]. Bu continued his study of transient analysis of impatient queues, and presented mathematical expressions for transient state probabilities [10].

Queueing models in cloud computing and distributed systems have been studied by several researchers. Dey introduced communication related queueing models for networked systems [11] and Rao and Srinivasan studied the queueing behavior of e-governance service platforms with dynamically varying workloads [12]. Ross [13] and Shortle et al. [14] underlined the importance of queueing models in computer networks, distributed processing and performance optimization. The transient queueing systems with catastrophes, vacations, retrial mechanisms and impatience behaviour are also recent topics of study. Sudhesh et al. studied the transient behaviour of queues of M/M/1 type with customer impatience and heterogeneous services [15] and Sharma et al. presented a survey of impatience based queueing models and its applications in modern service systems [16]. But the present papers are mostly concerned with single-node system and steady-state behavior and few papers have been given attention to transient analysis of interconnected Jackson-type computer networks with impatience characteristics.

Although cloud computing and distributed servers and real-time communication systems are gaining in significance, transient queueing analysis of multi-node computer networks with request abandonment is relatively little developed. A transient overload, a congestion of packets, a timeout of a task, or a delay in a routing path have great impact on Quality of Service and the reliability of the system in practical computing environments. Hence, an analytical framework should be developed that is able to model short term workload dynamics in interconnected computing systems

To address these problems, the present work proposes a transient M/M/1 Jackson-type queueing network with impatience for modelling distributed computing and communication system. The behaviour of routing, request abandonment, and transient congestion analysis are integrated into the proposed framework, which is used to analyse performance metrics like expected queue lengths, waiting times and propagation of workloads between interconnected service nodes.

In this study, we set up and study the transient behaviour of an impatient M/M/1 Jackson-type network. There are three contributions to this study. First, the governing Kolmogorov forward equations for transient state probabilities of the system are obtained. Second, numerical approximations of the time-dependent performance measures, such as buffer occupancies and processing delays at each node and at the central server, are obtained using the Runge–Kutta method in MATLAB, and are tractable. Third, sensitivity analysis is conducted to investigate the effect of various arrival rates, service rates, routing probabilities, balking parameters and reneging parameters on queue dynamics.

The proposed model can be applied to many delay-sensitive computing systems, such as cloud computing, packet-switched networks, web servers, distributed processing architectures, and more. The inclusion of transient behaviour and task abandonment leads to more realistic evaluation of the short-term congestion, and to better service efficiency in multi-node computer networks.

Model Description

We considered the Transient Analysis of Jackson type Network Queue consists of two nodes and one Central Server with Task Abandonment as detailed below:

1. The capacity of each node as well as Central Server are assumed as S(finite)
2. The Task Scheduling is allowed from nodes to the Central Server whereas flow from one node to another node is restricted. Also assumed that customers can directly enter in to Central server queue. The mean arrival rates at node-1, node-2 and Central server are assumed to follow Poisson Process with respective mean arrival rates λ_1, λ_2 and λ_3.
3. The mean service rates for first essential services at node-1, node-2 and Central Server are μ_1, μ_2 and μ_3.
4. Customers at nodes who want additional service from Central Server can opt with optional probabilities of p_1 and p_2 respectively.
5. Task Abandonment is characterized by two sets of parameters: the balking parameters
6. (1-b1), (1-b2) and (1-b3) representing the probabilities that a customer declines to join the queues of node-1, node-2 and Central server; and the reneging parameters ξ_1,ξ_2 and ξ_3, representing the rate at which customers abandon the respective queues after waiting.

With the above assumptions and parameter definitions, an infinitesimal generator matrix Q is obtained from derived Kolmogorov forward equations (In order to describe how the state probabilities evolve over time given in appendix.) as follows:

Infinitesimal Generator Matrix Q Formulation

Let

$$X(t) = \{N_1(t), N_2(t), N_3(t)\}$$

denote the state of the system at time t, where N1(t), N2(t), and N3(t) represent the number of tasks at Node-1, Node-2, and the Central Server, respectively. Since the capacity of each service station is finite,

$$0 \leq N_1, N_2, N_3 \leq S.$$

The state space of the system is defined as

$$\Omega = \{(i,j,k): 0 \leq i,j,k \leq S\}$$

with cardinality

$$| \Omega | = (S+1)^3.$$

Define

$$P(t) = [P_{i,j,k}(t)]_{(i,j,k)\in\Omega}.$$

denote the transient probability vector of the system. Then, the transient behaviour of the system is governed by the Kolmogorov forward equation

$$\frac{dP(t)}{dt} = P(t)Q$$

where

$$Q = \left[q_{(i,j,k),(r,s,l)}\right]$$

is an $(S+1)^3 \times (S+1)^3$infinitesimal generator matrix.

Structure of Q

The generator matrix Q has a block tridiagonal structure with respect to the Central Server population level k, and can be written as

$$Q = \begin{bmatrix} D_0 & A_0 & 0 & 0 & \cdots & 0 \\ B_1 & D_1 & A_1 & 0 & \cdots & 0 \\ 0 & B_2 & D_2 & A_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & B_S & D_S & \end{bmatrix}$$

where each block is of dimension

$$(S+1)^2 \times (S+1)^2.$$

Here

- A_k: transitions increasing the Central Server population (k→k+1)
- B_k: transitions decreasing the Central Server population (k→k-1)
- D_k: transitions within the same Central Server level k

Generic Matrix Elements

For any state (i, j, k) ∈ Ω, the non-zero off-diagonal elements of Q are defined as follows.

External Arrivals

$$q_{(i,j,k),(i+1,j,k)} = \lambda_1 b_1, \quad i < S$$
$$q_{(i,j,k),(i,j+1,k)} = \lambda_2 b_2, \quad j < S$$
$$q_{(i,j,k),(i,j,k+1)} = \lambda_3 b_3, \quad k < S$$

Service Completions and Reneging at Node-1

For i ≥1, a task may leave Node-1 after service completion without routing to the Central server. In addition, if i ≥ 2, waiting tasks may abandon the queue due to reneging.

Therefore,

$$q_{(i,j,k),(i-1,j,k)} = (1-p_1)\mu_1 + (\mathrm{i}\ -\ 1)\ \xi_1, \mathrm{i}\ \geq 1$$

When i = 1, the reneging term becomes zero.

Service Completions and Reneging at Node-2

Similarly, for Node-2,

$$q_{(i,j,k),(i,j-1,k)} = (1-p_2)\mu_2 + (\mathrm{j}-1)\ \xi_2, \quad \mathrm{j}\ \geq 1$$

When j = 1, the reneging term becomes zero.

Service Completions and Reneging at the Central Server

For the Central Server,

$$q_{(i,j,k),(i,j,k-1)} = \mu_3 + (\mathrm{k}-1)\ \xi_3, \quad \mathrm{k}\ \geq 1$$

When k = 1, the reneging term becomes zero.

Routing from Nodes to the Central Server

After service completion at Node-1, a task may be routed to the Central Server with probability p1. Therefore,

$$q_{(i,j,k),(i-1,j,k+1)} = p_1\mu_1, i\ \geq\ 1, k\ <\ S.$$

Similarly, after service completion at Node-2, a task may be routed to the Central Server with probability p2. Therefore,

$$q_{(i,j,k),(i,j-1,k+1)} = p_2\mu_2, \quad j\ \geq\ 1, k\ <\ S.$$

Diagonal Elements

The diagonal elements of the infinitesimal generator matrix are defined as the negative sum of all transition rates out of the state (i, j, k). Thus,

$$q_{(i,j,k),(i,j,k)} = -\sum_{(r,s,l)\neq(i,j,k)} q_{(i,j,k),(r,s,l)}.$$

All other elements of Q are zero. This construction ensures that each row of the generator matrix sums to zero, which is required for a valid continuous-time Markov chain representation of the finite-capacity queueing network.

Performance Measures

Some Queueing constants that are calculated to forecast the system through Runge-Kutta method of order 4(RK4) are listed below:

Expected lengths of the Node-1, Node-2 and Central Server (L_N1^((t)),L_N2^((t)) and L_HO^((t))) where,

$L_{N1}^{(t)}$	$L_{N2}^{(t)}$	$L_{HO}^{(t)}$
$L_{N1}^{(t)} = \sum_{i=0}^{s}\sum_{j=0}^{s}\sum_{k=0}^{s} ip_{i,j,k}^{(t)}$	$L_{N2}^{(t)} = \sum_{i=0}^{s}\sum_{j=0}^{s}\sum_{k=0}^{s} jp_{i,j,k}^{(t)}$	$L_{HO}^{(t)} = \sum_{i=0}^{s}\sum_{j=0}^{s}\sum_{k=0}^{s} kp_{i,j,k}^{(t)}$

2. Mean Processing Delays at Node-1, Node-2 and Central server are represented as ($W_{N1}^{(t)}$, $W_{N2}^{(t)}$ and $W_{HO}^{(t)}$)and are given as

$W_{N1}^{(t)}$	$W_{N2}^{(t)}$	$W_{HO}^{(t)}$)
$W_{N1}^{(t)} = \frac{LN_1(t)}{\lambda_1 b_1}$	$W_{N2}^{(t)} = \frac{LN_2(t)}{\lambda_2 b_2}$	$W_{HO}^{(t)} = \frac{LN_2(t)}{\lambda_3 b_3 + p_1\mu_1 + p_2\mu_2}$

Observation and Results and Discussion

Since an analytical solution is not tractable for the considered finite-capacity network with routing and impatience, the resulting system of ordinary differential equations is solved numerically using the fourth-order Runge-Kutta (RK4) method in MATLAB. The computed transient probabilities are subsequently used to evaluate the expected queue lengths and mean waiting times.

The transient state probabilities were updated by using Runge-Kutta method of order 4 with time step size of 0.5(Δt=0.5), truncation error of O(Δt^4) and convergence tolerance of 10^(-6).

The traffic intensities for stability of the model are defined as

Traffic intensities	Formula
$\boldsymbol{\rho_1(i)}$	$\rho_1(i) = \frac{\lambda_1 b_1}{\mu_1 + (i-1) * \xi_1}; 0 \le i \le s$
$\boldsymbol{\rho_2(j)}$	$\rho_2(j) = \frac{\lambda_2 b_2}{\mu_2 + (j-1) * \xi_2};\ 0 \le j \le s$
$\boldsymbol{\rho_3(k)}$	$\rho_3(k) = \frac{\lambda_3 b_3 + p_1\mu_1 + p_2\mu_2}{\mu_3 + (k-1) * \xi_3};\ 0 \le k \le s$

For the numerical illustrations, the values for all the model parameters are chosen which satisfy the traffic intensity conditions are given below:

S=3, λ_1=0.01, λ_2=0.02, λ_3=0.03, μ_1=0.02, μ_2=0.04, μ_3=0.05, ξ_1=0.001, ξ_2=0.002, ξ_3=0.003, b_1=0.5, b_2=0.6, b_3=0.7, p_1=0.001 and p_2=0.002

We have studied the model in four different scenarios [I-IV] to explore various patterns of queue constants, where Scenario-I shows the general behaviour of queue constants over time and Scenarios II-IV deal sensitivity analysis of various parameters in different cases.
They are detailed in the following tables:

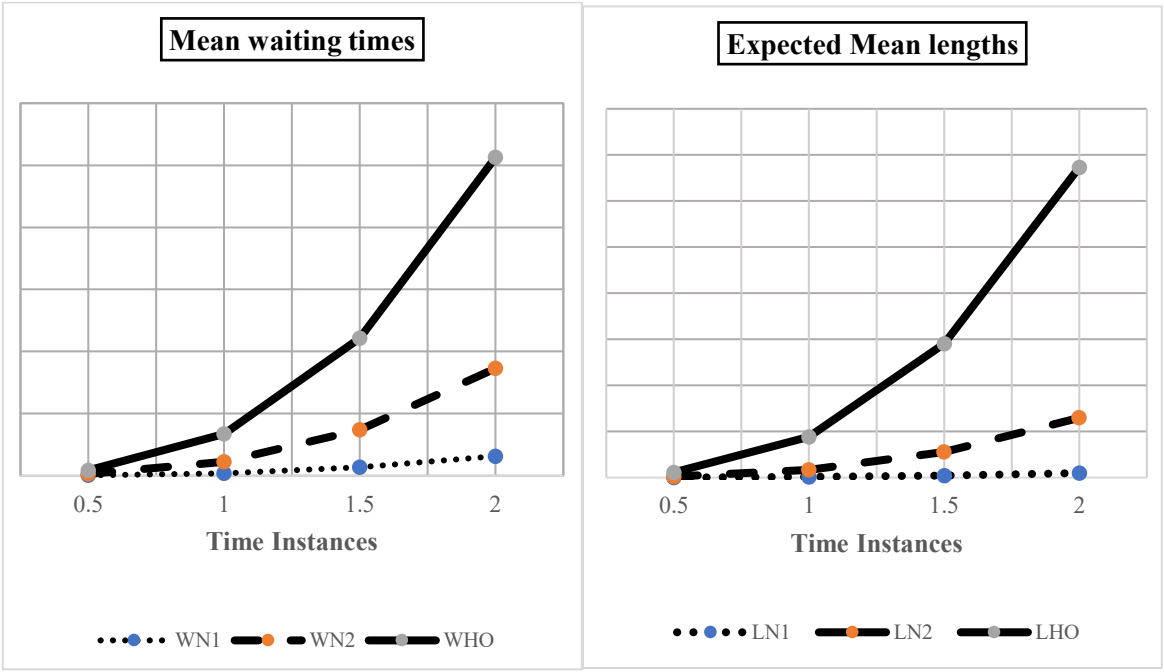
Scenario	Assumption	Objective
I	All the parameter assumes different values	To know the general behaviour of queue dynamics over time
II	Each parameter at two nodes is assumed equal values and are different from that of Central server.	To know the influence of change in parameter at nodes on system dynamics over time i.e., nodes Vs Central Server.

III	All the parameters of each of the nodes are changing and keeping other values constant	To know the influence of change in parameter at each node on system dynamics over time for each of node-1, node-2 and Central server.
IV	Each node's Task Abandonment parameters are changing and keeping other values constant	To know the influence of change in balking and reneging rates at each node on system dynamics over time for each of node-1, node-2 and Central server.

The following graphs explain the system behaviour in each of the above cases in a sequential manner:

- Scenario-1

Graph 1: Pattern of mean lengths and Processing Delays of each node and Central Server over different time instances.



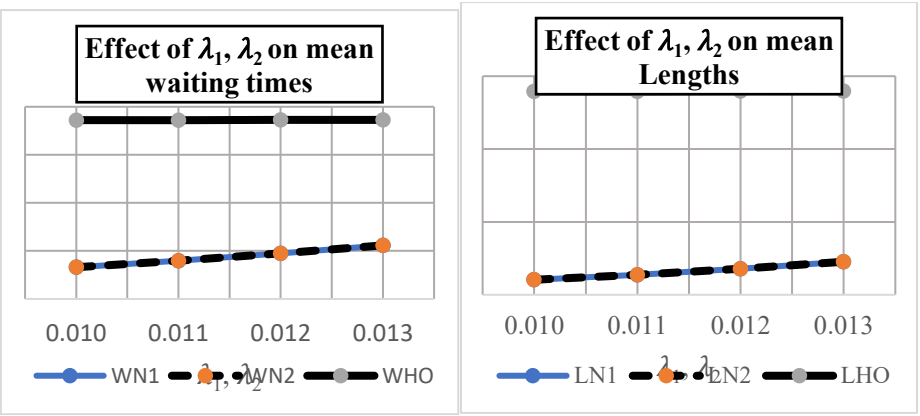
Interpretation: Over a time duration from 0.5-2, it is observed that,

- The mean Buffer Occupancies are increasing at all the 2 nodes and Central Server.
- The mean Processing Delays are increasing at all the 2 nodes and Central Server.

Discussion: Graph 1 shows that the expected mean buffer Occupancies and mean processing delays at Node-1, Node-2, and the Central Server increase with time during the transient period. This indicates gradual accumulation of tasks in the system due to incoming arrivals and internal routing. The Central Server is more congested since it gets both direct arrivals and tasks that are routed from the two nodes. The growth of waiting time is due to the growth of the length of the queue, which indicates the direct effect between congestion and processing delay. So, it can be inferred from Scenario 1 that transient analysis is very important in order to know short term congestion and to increase the capacity of the services, particularly at the Central Server.

- Scenario-2 (Graph 2 to Graph 10)

Graph 2: Effect of arrival rates at nodes $\lambda 1$ or $\lambda 2$ on mean lengths and Processing Delays of each node and Central Server

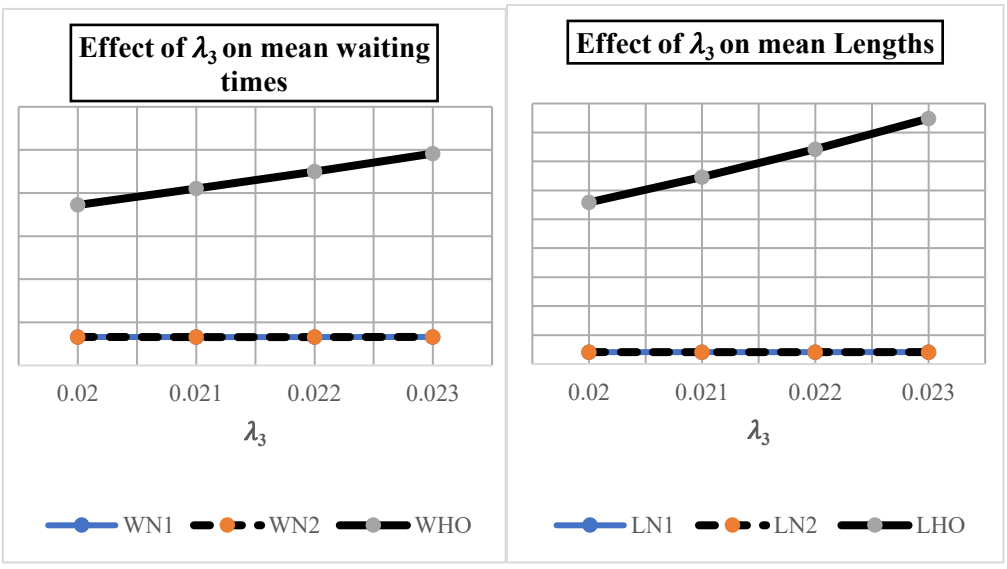


Interpretation: With an increase in $\lambda 1$, $\lambda 2$: it is observed that

- The mean Buffer Occupancies are increasing where they are same at both nodes and different at Central Server.
- The mean Processing Delays are increasing where they are same at both nodes and different at Central Server.

Discussion: Graph 2 shows that an increase in the arrival rates λ1 and λ2 (assume equal) leads to an increase in expected mean Buffer Occupancies and mean Processing Delays times at Node-1, Node-2, and the Central Server. This means that more tasks arrive in the system, which leads to an increase in the workload accumulated in the system. Externally arriving calls to Node-1 and Node-2 are mapped directly, and so the length of the queues increases as the intensity of the arrivals increases. Also, the Central Server is impacted as some of the completed work by both nodes is sent to the Central Server for further services. As seen in Graph 2, then, the higher the arrival rates, the greater the transient congestion and processing delay throughout the network.

Graph 3: Effect of arrival rates at nodes (λ3) on mean lengths and Processing Delays of each node and Central Server



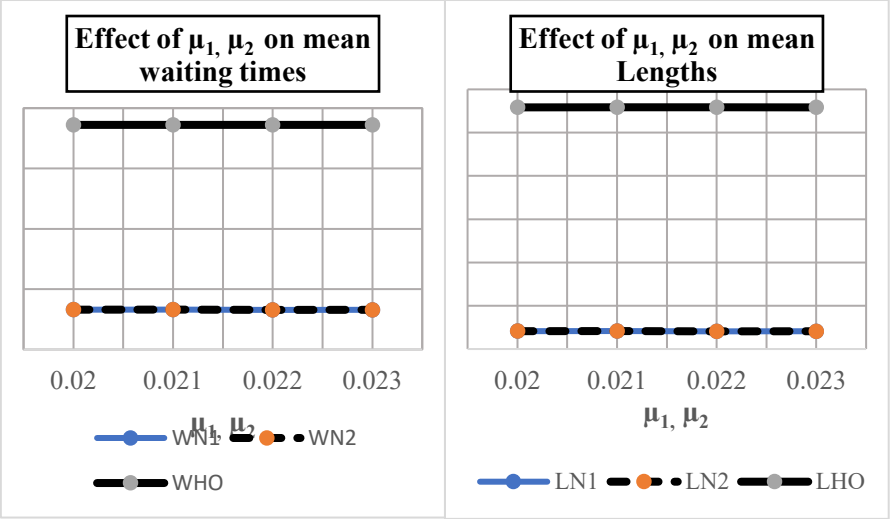
Interpretation: With an increase in $\lambda 3$ it is observed that

- The mean Buffer Occupancies are increasing where they are same at both nodes and different at Central Server and the rate of increase is more at Central Server than at nodes.
- The mean Processing Delays are increasing where they are same at both nodes and different at Central Server and the rate of increase is more at Central Server than at nodes.

Discussion: Graph 3 shows the effect of increasing the external arrival rate $\lambda 3$ at the Central Server on mean Buffer Occupancies and mean Processing Delays. It is observed that both measures increase at Node-1, Node-2, and the Central Server; however, the rate of increase is higher at the Central Server. This is because the Central Server receives direct external arrivals through $\lambda 3$ in addition to tasks routed from both nodes. As a result, an increase in

$\lambda 3$ creates greater workload accumulation and longer processing delays at the Central Server compared with the individual nodes. Thus, Graph 3 indicates that the Central Server is more sensitive to external traffic changes and may become the main congestion point in the network.

Graph 4 : Effect of service rates at nodes (μ1, μ2) on mean lengths and Processing Delays of each node and Central Server

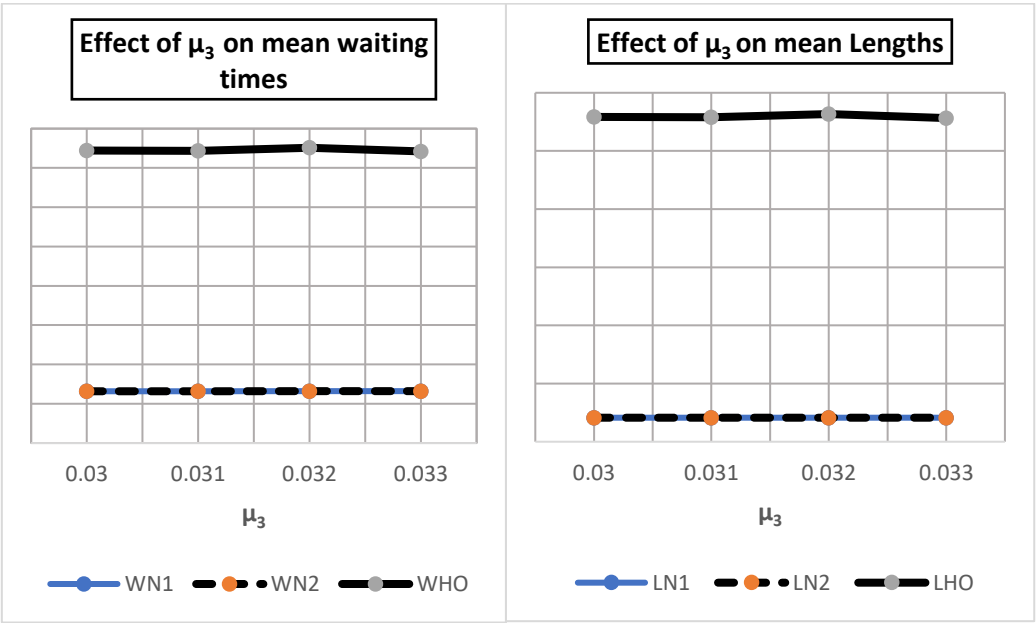


Interpretation: With an increase in μ1, μ2 (assume equal); it is observed that

- The mean Buffer Occupancies are decreasing where they are same at both nodes and different at Central Server.
- The mean Processing Delays are decreasing where they are same at both nodes and different at Central Server.

Discussion: Graph 4 shows the effect of increasing service rates μ1 and μ2 (assume equal) on mean Buffer Occupancies and mean Processing Delays. It is observed that both Buffer Occupancies and Processing Delays decrease at Node-1, Node-2, and the Central Server. This indicates that higher service rates at the nodes improve the processing of incoming tasks and reduce workload accumulation in the system. The trends of both nodes are almost same since the nodes parameter settings are similar; the Central Server has a different trend because of the routed tasks received from nodes. Thus, increasing μ1 and μ2 helps in reducing transient congestion and improving overall system efficiency.

Graph 5: Effect of service rates at node μ3 on mean lengths and Processing Delays of each node and Central Server.

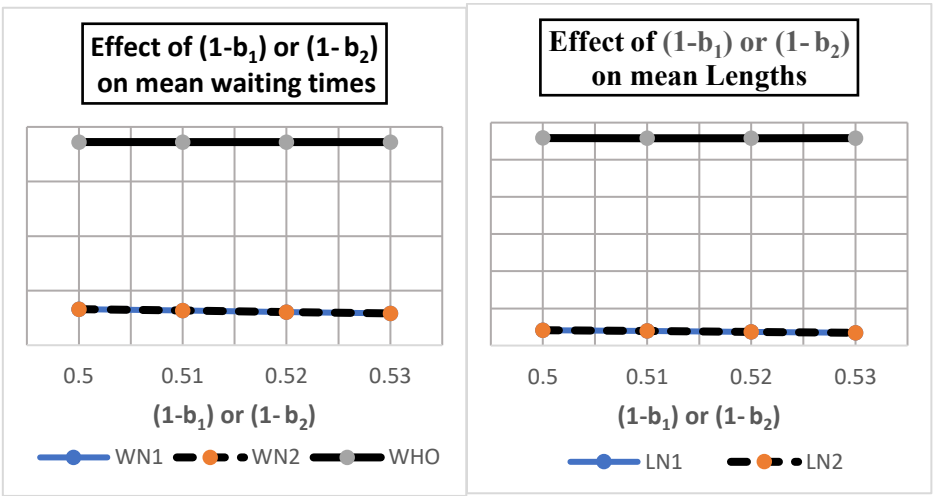


Interpretation: With an increase in μ3 , it is observed that

- The Mean Buffer Occupancies at Node-1 and Node-2 remain almost unchanged, while the Mean Buffer Occupancy at the Central Server shows a slight decrease. This indicates that increasing the service rate of the Central Server mainly reduces workload accumulation at the Central Server, but it has very little effect on both nodes.
- The Mean Processing Delays at Node-1 and Node-2 remain nearly constant. At the Central Server, the Mean Processing Delay shows decrease with an increase in μ3. Therefore, increasing μ3 helps improve service efficiency mainly at the Central Server and reduces congestion and delay at that point.

Discussion: Graph 5 shows that increasing the service rate μ3 mainly affects the Central Server. The Mean Buffer Occupancies and Mean Processing Delays at Node-1 and Node-2 remain almost unchanged, showing that the change in μ3 has very little effect on the two nodes. At the Central Server, both Mean Buffer Occupancy and Mean Processing Delay show a slight decrease, indicating improved service efficiency. Thus, increasing μ3 helps reduce congestion and delay mainly at the Central Server.

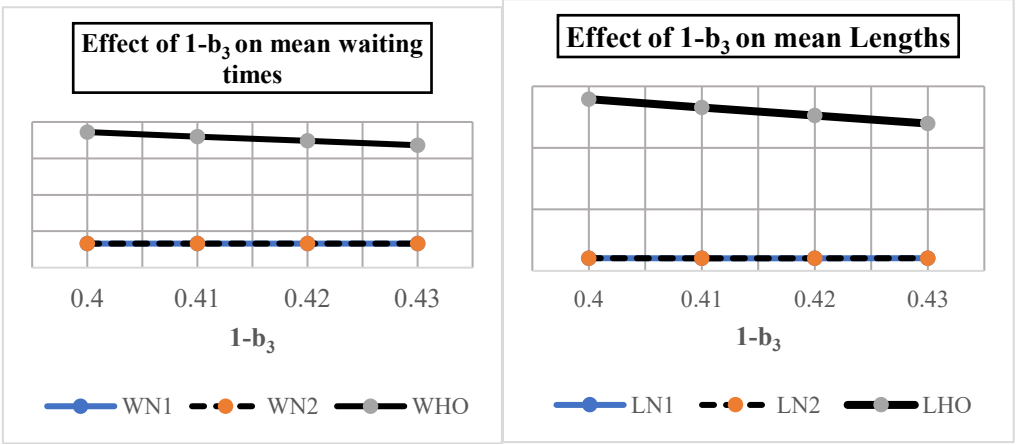
Graph 6: Effect of Balking at nodes (1-b1, 1-b2) on mean lengths and Processing Delays of each node and Central Server



Interpretation: With an increase in (1-b1) or (1- b2) (assume equal) which are assumed same, it is observed that

- The mean Buffer Occupancies are decreasing where they are same at both nodes and different at Central Server.
- The mean Processing Delays are decreasing where they are same at both nodes and different at Central Server.

Discussion: Graph 6 shows the effect of increasing (1-b1) or (1- b2) on Mean Buffer Occupancies and Mean Processing Delays. It is observed that both measures decrease at Node-1, Node-2, and the Central Server. The trends at both nodes are almost the same, while the Central Server shows a different pattern due to the additional routed workload from the nodes. This indicates that changes in the balking-related parameters help reduce task accumulation and processing delay in the system. Thus, Graph 6 suggests that controlling task entry behaviour can improve transient congestion and overall system performance.



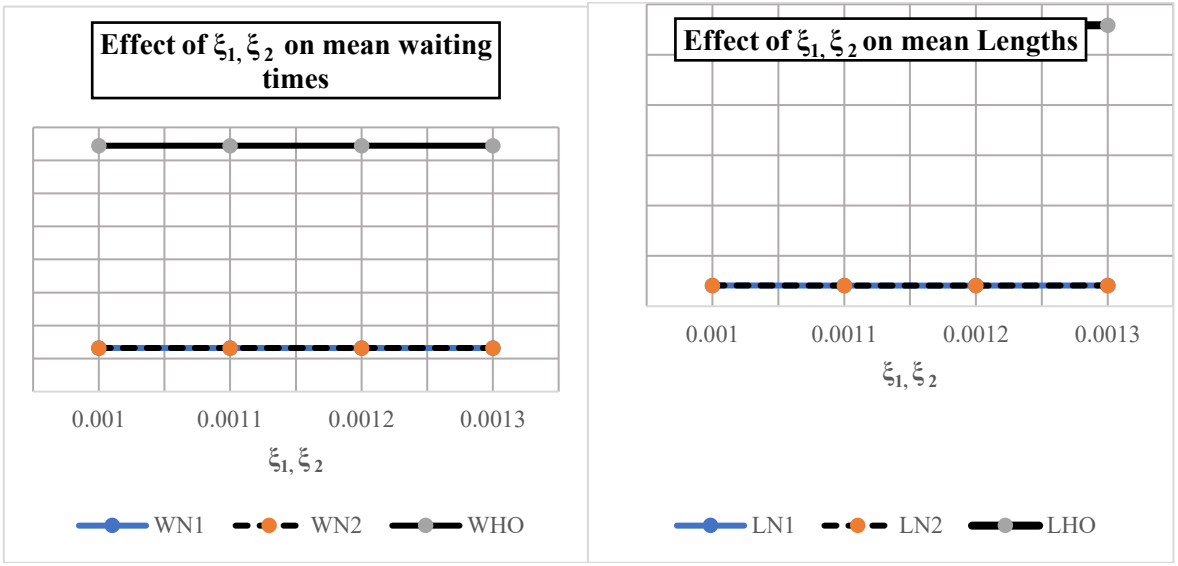
Graph 7: Effect of Balking at central server 1-b3 on mean lengths and Processing Delays of each node and Central Server

Interpretation: With an increase in 1-b3 it is observed that

- The Mean Buffer Occupancy at the Central Server shows a noticeable decrease, while the Mean Buffer Occupancies at Node-1 and Node-2 remain almost the same. This indicates that the change in 1-b3 mainly affects the Central Server and has very little influence on the two nodes.
- The Mean Processing Delay at the Central Server decreases with an increase in 1-b3, whereas the Mean Processing Delays at Node-1 and Node-2 remain nearly unchanged. Therefore, Graph 7 shows that the effect of 1-b3 is mainly observed at the Central Server, where both workload accumulation and processing delay are reduced.

Discussion: Graph 7 shows that increasing 1-b3 mainly affects the Central Server. The Mean Buffer Occupancies and Mean Processing Delays at Node-1 and Node-2 remain almost unchanged, indicating that this parameter has very little effect on the two nodes. At the Central Server, both Mean Buffer Occupancy and Mean Processing Delay decrease, showing reduced workload accumulation and improved processing efficiency. Thus, Graph 7 highlights that changes in 1-b3 are more important for controlling congestion and delay at the Central Server.

Graph 8 : Effect of reneging at nodes ξ1 and ξ2 on mean lengths and Processing Delays of each node and Central Server

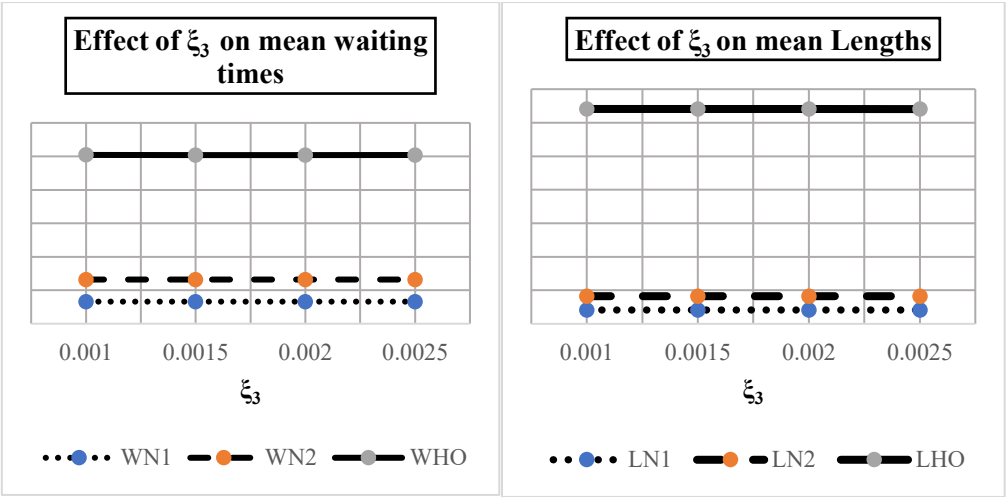


Interpretation: With an increase in ξ1, ξ 2 (assumed same) it is observed that

- The mean Buffer Occupancies are decreasing where they are same at both nodes and different at Central Server.
- The mean Processing Delays are decreasing where they are same at both nodes and different at Central Server.

Discussion: Graph 8 shows the effect of increasing reneging rates ξ1, and ξ 2 on Mean Buffer Occupancies and Mean Processing Delays. It is observed that both measures decrease at Node-1, Node-2, and the Central Server. This means that more waiting tasks exit the system before they are served as reneging increases, thus less workload has to be accumulated, and delay in the processing decreases. The trends at both nodes are same, while the Central Server shows a different pattern due to additional routed tasks from the nodes. As depicted in Graph 8, reneging will make the system less congested, but it can also be interpreted as a loss of tasks in delay sensitive computing systems.

Graph 9: Effect of reneging at Central server ξ3 on mean lengths and Processing Delays of each node and Central Server.

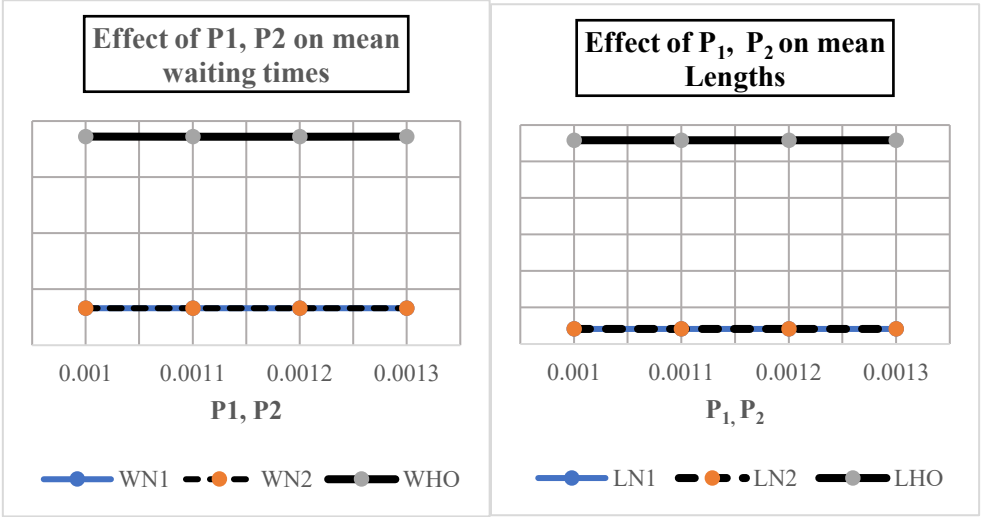


Interpretation: With an increase in ξ3 it is observed that

- The Mean Buffer Occupancy at the Central Server decreases, while the Mean Buffer Occupancies at Node-1 and Node-2 remain almost unchanged. This shows that ξ3 mainly affects the Central Server, because it represents the abandonment rate of waiting tasks at the Central Server.
- The Mean Processing Delay at the Central Server decreases slightly with an increase in ξ3, whereas the Mean Processing Delays at Node-1 and Node-2 remain nearly constant. This indicates that higher reneging at the Central Server reduces workload accumulation and delay at the Central Server, but it has very little effect on the two nodes.

Discussion: Graph 9 shows that increasing the reneging rate ξ3 mainly affects the Central Server. The Mean Buffer Occupancies and Mean Processing Delays at Node-1 and Node-2 remain almost unchanged, while both measures decrease slightly at the Central Server. This indicates that higher reneging at the Central Server reduces workload accumulation and processing delay at that point. Thus, Graph 9 highlights that ξ3, helps control congestion mainly at the Central Server, with very little effect on the two nodes.

Graph 10 : Effect of probabilities at nodes P1 and P2 on mean lengths and Processing Delays of each node and Central Server



Interpretation: With an increase in P1, P2 ; it is observed that

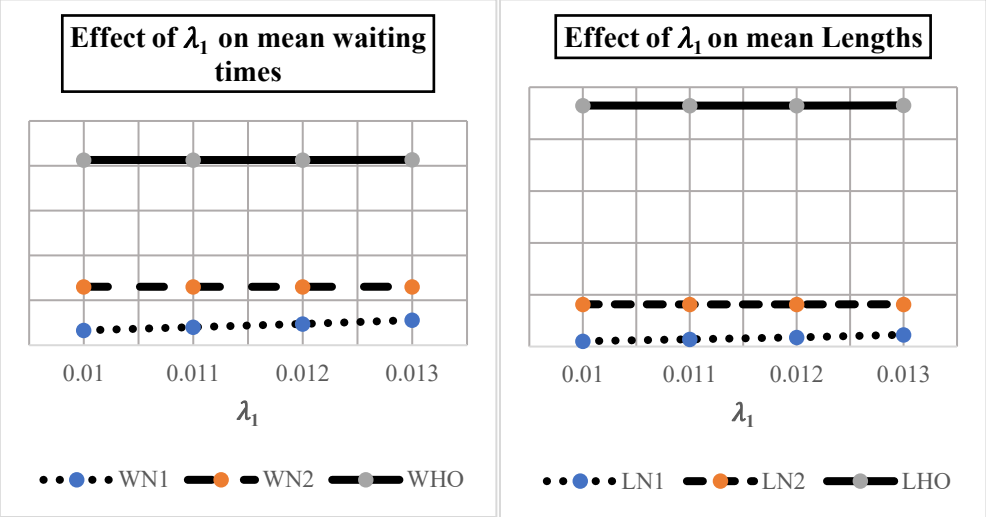
- The mean Buffer Occupancies are increasing at all the three stations, but equal at node 1 and node 2.
- The mean Processing Delays are increasing at all the three stations, but equal at node 1 and node 2

Discussion: Graph 10 shows the effect of increasing routing probabilities (P1) and (P2) on Mean Buffer Occupancies and Mean Processing Delays. It is observed that both measures increase at Node-1, Node-2, and the Central Server. This indicates that higher routing probabilities send more tasks from the nodes to the Central Server, increasing the overall workload in the network. The trend at Node-1 and Node-2 are same, at Central Server

there is a different trend as this server gets more routed tasks. Therefore, as can be seen in Graph 10, the routing probabilities are important for transient congestion and processing delay.

- Scenario 3 (Graph 11 to Graph 14)

Graph 11: Effect of Arrival Rate at Node-1(λ1) on mean lengths and Processing Delays of each node and Central Server.

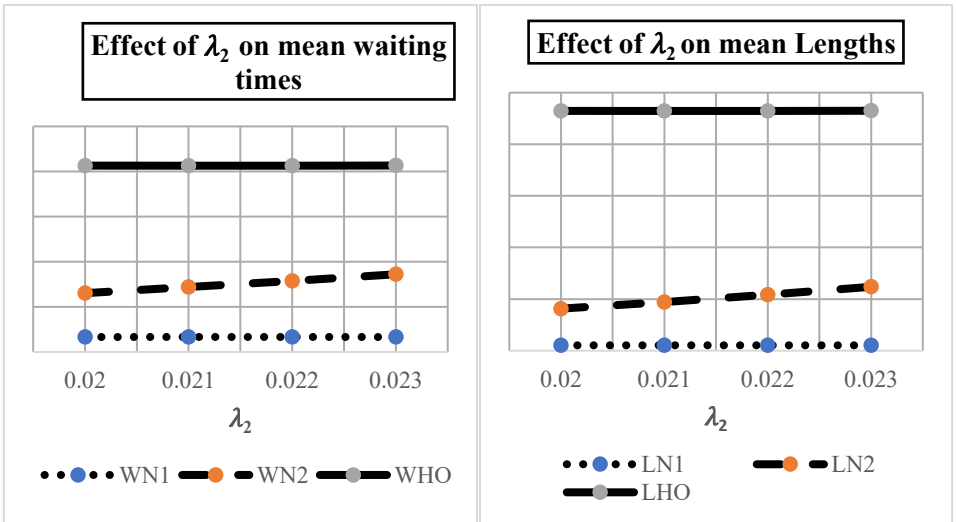


Interpretation: With an increase in λ1 it is observed that

- The Mean Buffer Occupancy increases mainly at Node-1, because λ1 represents the external arrival rate at Node-1. The Central Server also shows some increase because a portion of tasks completed at Node-1 is routed to the Central Server for further processing. The effect on Node-2 is very small because λ1 is not directly related to Node-2 arrivals.
- The Mean Processing Delay increases mainly at Node-1, while the Central Server shows a slight increase due to routed workload from Node-1. Node-2 remains almost unchanged. Thus, Graph 11 indicates that increasing λ1 increases workload accumulation and delay primarily at Node-1 and partially at the Central Server.

Discussion: Graph 11 shows that increasing the arrival rate λ1 mainly increases the Mean Buffer Occupancy and Mean Processing Delay at Node-1. This is because λ1 directly controls the number of tasks entering Node-1. The Central Server is also slightly affected due to tasks routed from Node-1 after service, while Node-2 remains almost unchanged. Thus, Graph 11 indicates that higher arrivals at Node-1 increase workload accumulation and delay mainly at Node-1 and partly at the Central Server.

Graph 12 : Effect of Arrival Rate at Node-1(λ2) on mean lengths and Processing Delays of each node and Central Server

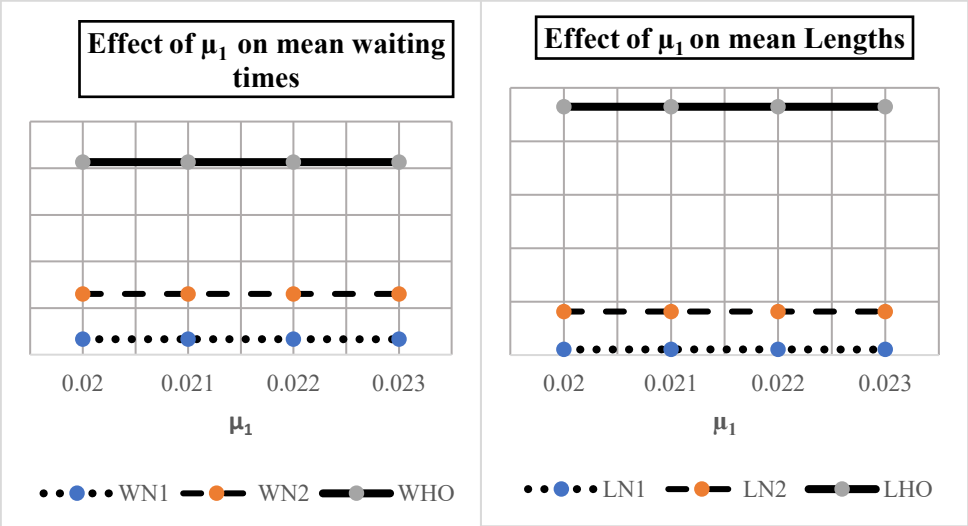


Interpretation: With an increase in λ2, it is observed that

- The Mean Buffer Occupancy increases mainly at Node-2, because λ2 represents the external arrival rate at Node-2. The Central Server also shows a slight increase because some completed tasks from Node-2 are routed to the Central Server for further processing. The effect on Node-1 is very small because λ2 is not directly related to Node-1 arrivals.
- The Mean Processing Delay increases mainly at Node-2, while the Central Server shows a slight increase due to routed workload from Node-2. Node-1 remains almost unchanged. Thus, Graph 12 indicates that increasing λ2 increases workload accumulation and delay primarily at Node-2 and partially at the Central Server.

Discussion: Graph 12 shows that increasing the arrival rate λ2 mainly increases the Mean Buffer Occupancy and Mean Processing Delay at Node-2. This is because λ2 directly represents the task arrival rate at Node-2. The Central Server is also slightly affected due to routed tasks from Node-2, while Node-1 remains almost unchanged. Thus, Graph 12 indicates that higher arrivals at Node-2 increase workload accumulation and delay mainly at Node-2 and partly at the Central Server.

Graph 13: Effect of service rate at node-1(μ1) on mean lengths and Processing Delays of each node and Central Server.

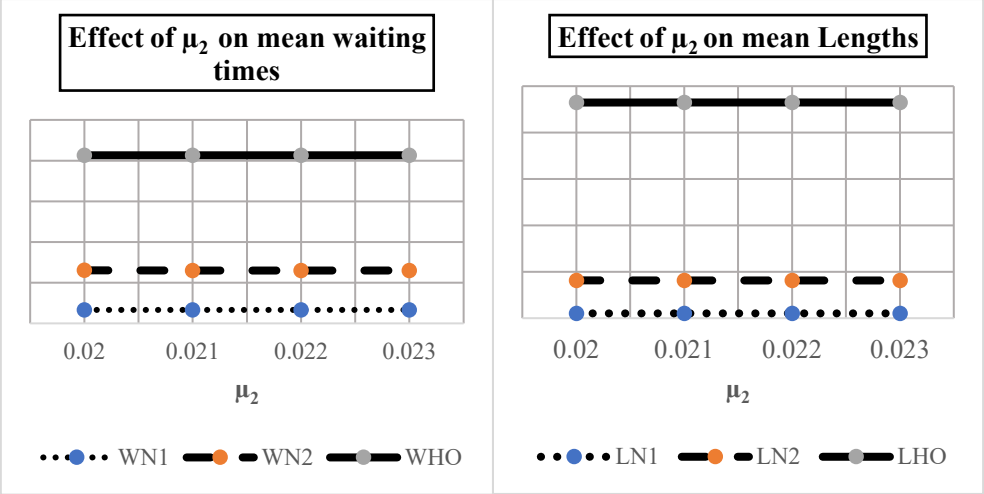


Interpretation: With an increase in μ1 it is observed that

- The Mean Buffer Occupancy decreases mainly at Node-1, because μ1 represents the service rate of Node- When the service rate increases, tasks are processed faster at Node-1, so fewer tasks remain waiting in the buffer.
- The Mean Processing Delay also decreases mainly at Node-1, because faster service reduces the waiting and processing time for tasks at that node. The Central Server may show a slight change because some tasks completed at Node-1 are routed to the Central Server. The effect on Node-2 is very small because μ1 is not directly related to the service mechanism of Node-2.

Discussion: Graph 13 shows that increasing the service rate μ1 mainly reduces the Mean Buffer Occupancy and Mean Processing Delay at Node-1. This happens because higher service rate allows Node-1 to process tasks faster, reducing queue buildup and delay. The Central Server shows only a slight change due to routed tasks from Node-1, while Node-2 remains almost unaffected. Thus, Graph 13 indicates that improving μ1 enhances performance mainly at Node-1.

Graph 14: Effect of service rate at node-2(μ2) or on mean lengths and Processing Delays of each node and Central Server



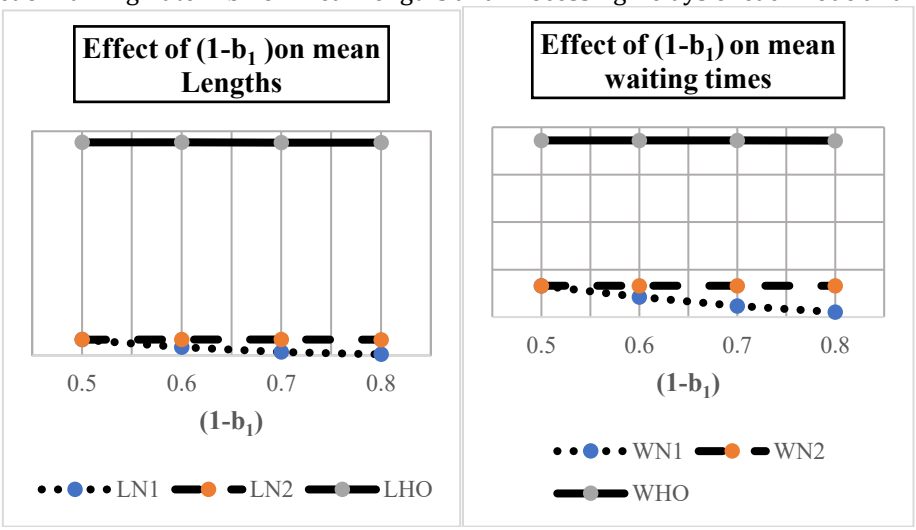
Interpretation: With an increase in μ2 it is observed that

- The Mean Buffer Occupancy decreases mainly at Node-2, because μ2 represents the service rate of Node As the service rate increases, tasks are processed faster at Node-2, so fewer tasks remain waiting in the buffer.
- The Mean Processing Delay also decreases mainly at Node-2, because faster service reduces the waiting and processing time for tasks at that node. The Central Server may show a slight change because some tasks completed at Node-2 are routed to the Central Server. The effect on Node-1 is very small because μ2 is not directly related to the service mechanism of Node-1.

Discussion: Graph 14 shows that increasing the service rate μ2 mainly reduces the Mean Buffer Occupancy and Mean Processing Delay at Node-2. This is because higher μ2 allows Node-2 to process tasks faster, reducing queue buildup and delay. The Central Server shows only a slight change due to routed tasks from Node-2, while Node-1 remains almost unaffected. Thus, Graph 14 indicates that improving μ2 enhances performance mainly at Node-2.

- Scenario 4 (Graph 15 to Graph 18)

Graph 15: Effect of Balking Rate 1-b1 on mean lengths and Processing Delays of each node and Central Server.

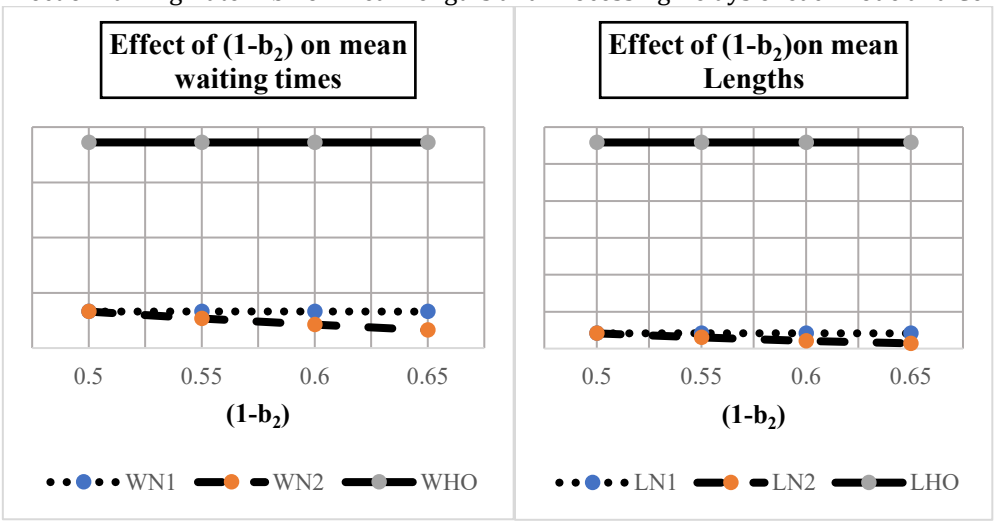


Interpretation: With an increase in (1-b1), it is observed that

- The Mean Buffer Occupancy decreases mainly at Node-1, because b1 is related to the task-entry or balking behaviour at Node-1. The decrease indicates that fewer tasks remain accumulated in the buffer, reducing congestion at Node-1.
- The Mean Processing Delay also decreases primarily at Node-1, indicating that the tasks are subjected to less delay when the workload accumulation is less. Some tasks may be sent to the Central Server after service, which may cause a slight change in the Central Server. Effect on Node-2 is very small as there is no direct relation with (1-b1) on Node-2.

Discussion: Graph 15 shows that variation in (1-b1) mainly affects Node-1. As (1-b1) increases, the Mean Buffer Occupancy and Mean Processing Delay decrease mainly at Node-1, indicating reduced workload accumulation and faster task processing. The Central Server shows only a slight change due to routed tasks from Node-1, while Node-2 remains almost unaffected. Thus, Graph 15 suggests that the task-entry or balking-related behaviour at Node-1 plays an important role in controlling congestion and delay at that node.

Graph 16: Effect of Balking Rate 1-b2 on mean lengths and Processing Delays of each node and Central Server.

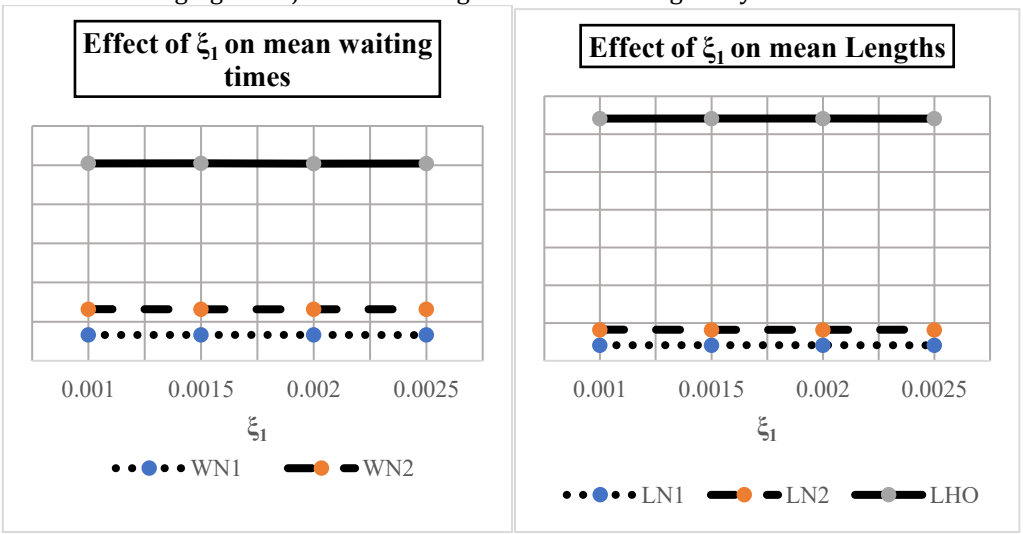


Interpretation: With an increase in (1-b2), it is observed that

- The Mean Buffer Occupancy decreases mainly at Node-2, because (1- b2) is related to the task-entry or balking behaviour at Node-2. This indicates that fewer tasks remain accumulated in the buffer at Node-2, reducing congestion at that node.
- The Mean Processing Delay also decreases mainly at Node-2, showing that tasks experience less delay when workload accumulation is reduced. The Central Server may show a slight change because some tasks from Node-2 are routed to the Central Server after service. The effect on Node-1 is very small because (1-b2) is not directly related to Node-1.

Discussion: Graph 16 shows that variation in (1-b2) mainly affects Node-2. As (1-b2) increases, the Mean Buffer Occupancy and Mean Processing Delay decrease mainly at Node-2, indicating reduced workload accumulation and lower delay. The Central Server shows only a slight change due to routed tasks from Node-2, while Node-1 remains almost unaffected. Thus, Graph 16 suggests that the task-entry or balking-related behaviour at Node-2 plays an important role in controlling congestion and processing delay at that node.

Graph 17: Effect of Reneging Rate ξ1 on mean lengths and Processing Delays of each node and Central Server

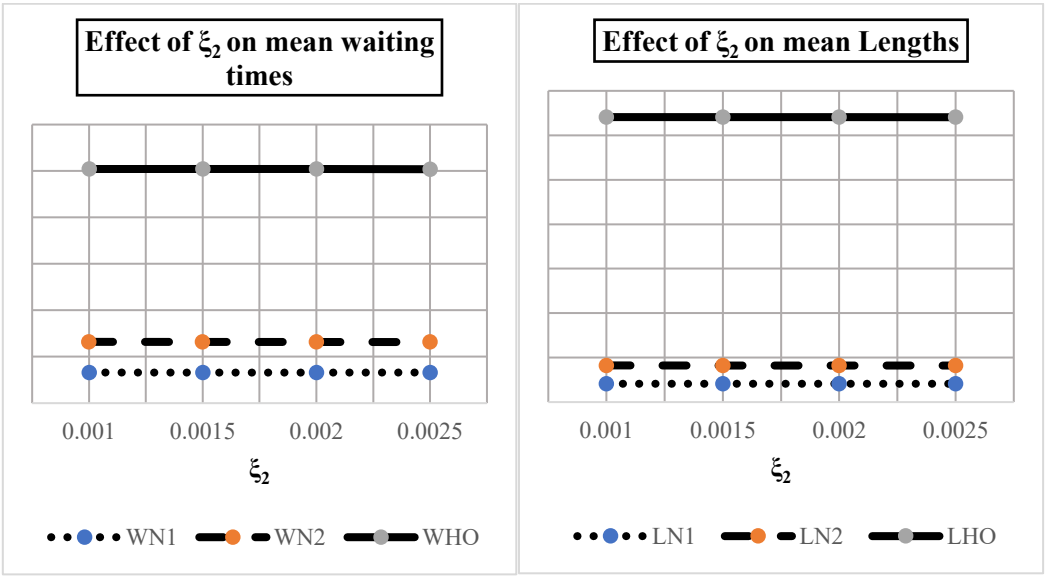


Interpretation: With an increase in ξ1, it is observed that

- The Mean Buffer Occupancies at Node-1, Node-2, and the Central Server remain almost unchanged, with only a very slight decreasing tendency. This shows that increasing ξ1 has very little effect on workload accumulation under the selected parameter values.
- The Mean Processing Delays also remain nearly constant at all three service stations. Since ξ1 represents reneging at Node-1, its effect is mainly expected at Node-1; however, the graph shows that this effect is very small. Thus, it indicates that variation in ξ1 does not significantly reduce transient congestion or processing delay in this scenario.

Discussion: Graph 17 shows that increasing the reneging rate ξ1 has very little effect on the system. The Mean Buffer Occupancies and Mean Processing Delays at Node-1, Node-2, and the Central Server remain almost constant, with only a slight decreasing tendency. This indicates that reneging at Node-1 does not significantly reduce workload accumulation or processing delay under the selected parameter values. Thus, Graph 17 suggests that ξ1 has only a minimal influence on transient congestion in this scenario.

Graph 18 : Effect of Reneging Rate ξ2 on mean lengths and Processing Delays of each node and Central Server



Interpretation: With an increase in ξ2, it is observed that

- The Mean Buffer Occupancy decreases mainly at Node-2, because ξ2 represents the reneging or abandonment rate of waiting tasks at Node-2. As more waiting tasks leave the queue before service, fewer tasks remain accumulated in the buffer.
- The Mean Processing Delay also decreases, primarily at Node-2, as a result of the reduced build up of a queue, thus reducing delay. There might be a slight change on the Central Server due to the tasks that are routed from Node-2 and Node-1 is only slightly affected.

Discussion : As seen in Graph 18, the reneging rate (ξ2) is mainly decreasing the Mean Buffer Occupancy and Mean Processing Delay at Node-2. This happens because there is a lesser number of waiting tasks in the queue which means there is less workload build up and delay. Only a slight change can be seen for the Central Server as the tasks are routed from Node-2 and a minimal change is seen for Node-1. Therefore, Graph 18 shows that (ξ2) is mainly for congestion control at Node-2, but it could also mean task loss in delay sensitive systems.

Conclusion

The present study developed a transient queueing framework for analysing congestion, workload propagation, and task abandonment in a finite-capacity multi-node computer network. The proposed M/M/1 Jackson-type queueing network consists of two processing nodes and one central server, where tasks may arrive externally, receive service, be routed to the central server, or abandon the system due to balking and reneging. The transient state probabilities were formulated using Kolmogorov forward differential equations and solved numerically through the fourth-order Runge–Kutta method.

As observed from the numerical results, it can be seen that the expected buffer occupancies and mean processing delays increase throughout the transient period which shows progressive accumulation of workload in the network. The central server is the most congested of the service stations since it receives direct external arrivals in addition to tasks that are passed on from Node-1 and Node-2. This indicates that the proposed network structure has the central server as the main congestion point.

The sensitivity analysis also reveals that increasing the arrival rate and the probability of routing leads to higher buffer occupancy and higher processing delay while increasing the service rate leads to lower buffer congestion and higher processing efficiency. The results also show that the balking and reneging parameters lead to a decrease in the number of waiting tasks, in the processing delay, but in real computing systems this can also imply a loss of tasks, a timeout occurrence, or a decrease in the number of completed services. Hence, the influence of impatience parameters should be taken from both congestion-control and service-quality points of view.

Conflict of Interest: None to Declare

Funding: None

Authors Contribution: Sachin R. Gurnule contributed to conceptualization, model development, numerical implementation, analysis, and drafting. V.N. Rama Devi contributed to supervision, validation, methodological review, interpretation, and manuscript revision. P. Pranay contributed to study design, model refinement, mathematical guidance, verification, and editing. K. Kalyani contributed to the literature review, numerical support, the organization of the results, and the formatting/editing of the manuscript.

Acknowledgments: The authors sincerely thank Dr. P. Pranay, Associate Professor, Department of Mathematics and Statistics, Chaitanya Deemed to be University, Warnagal, Dr. V. N. Rama Devi, Professor, GRIET, Hyderabad, and all the faculty members of the Department of Statistics and Mathematics, CDU, for their valuable guidance and support.

References

1. Jackson JR (1957). Networks of waiting lines. Operations Research; 5(4):518–521. https://doi.org/10.1287/opre.5.4.518
2. Kelly FP (1979). Reversibility and Stochastic Networks. New York: Wiley. https://www.wiley.com/en-us/Reversibility+and+Stochastic+Networks-p-9780471997885
3. Kelly FP, Yudovina E (2014). Stochastic Networks. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139343265
4. Bertsekas DP, Gallager RG (1992). Data Networks. 2nd ed. Englewood Cliffs (NJ): Prentice-Hall. https://web.mit.edu/dimitrib/www/datanets.html
5. Harchol-Balter M (2013). Performance Modeling and Design of Computer Systems: Queueing Theory in Action. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/performance-modeling-and-design-of-computer-systems/743BEBB137B781EDBAFD807D8F7965DF
6. Abate J, Whitt W (1992). Transient behavior of queues via Laplace transforms. Queueing Systems; 10:5–88. https://doi.org/10.1007/978-1-4615-1803-6_15
7. Neuts MF (1981). Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Baltimore: Johns Hopkins University Press. https://books.google.co.in/books/about/Matrix_geometric_Solutions_in_Stochastic.html?id=rtpQAAAAMAAJ
8. Garnett O, Mandelbaum A, Reiman MI (2002). Designing a call center with impatient customers. Manufacturing & Service Operations Management; 4(3):208–227. https://doi.org/10.1287/msom.4.3.208.7751
9. Zeltyn S, Mandelbaum A (2005). The M/M/n+G queue. Queueing Systems; 51(3–4):361–402. https://doi.org/10.1007/s11134-005-2751-
10. Bu Q (2024). Transient analysis for impatient queues. Mathematics; 12(13):2030. https://doi.org/10.3390/math12132030
11. Dey S (2020). Modelling and Performance Analysis of Some Communication Related Queuing Systems [PhD thesis]. Thiruvananthapuram: Indian Institute of Space Science and Technology. https://events.iist.ac.in/phd/thesis/SC15D020%20FT.pdf
12. Rao S, Srinivasan R (2022). Queueing analysis of e-governance services. Government Information Quarterly; (In press). https://doi.org/10.1016/j.giq.2021.101654

13. Ross SM (2023). Introduction to Probability Models. 13th ed. London: Academic Press. https://shop.elsevier.com/books/introduction-to-probability-models/ross/978-0-443-18761-2
14. Shortle JF, Thompson JM, Gross D, Harris CM (2018). Fundamentals of Queueing Theory. 5th ed. Hoboken (NJ): Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119453765
15. Sudhesh R, Azhagappan A, Dharmaraja S (2017). Transient analysis of M/M/1 queue with working vacation, heterogeneous service and customers' impatience. RAIRO – Operations Research; 51(3):591–606. https://www.numdam.org/item/10.1051/ro/2016046.pdf
16. Sharma S, Kumar R, Soodan BS, Singh P (2023). Queuing models with customers' impatience: a survey. International Journal of Mathematics in Operational Research; 26(4):523–547. https://doi.org/10.1504/IJMOR.2023.135546