IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 22, Issue 2, Ser. 1 (Mar. – Apr. 2026), PP 56-65
www.iosrjournals.org

Analysis Of A Multi-Server Queue With Consultation By Main Server Having Phase Type Distributed Service Time

Resmi T.
Dept. Of Mathematics, KKTM Govt. College, Pullut, Kerala, India

Abstract
This paper analyzes a multi-server queueing system in which a main server provides consultation to regular servers. The main server performs dual roles: serving customers and offering consultation to regular servers on a first-in-first-out (FIFO) basis, with preemptive priority over cus-tomer service. As a result, customer service at the main server may be interrupted whenever it engages in consultation. Both customer arrivals and consultation requests are modeled as mutually independent Poisson processes. The service time of the main server follows a phase-type distri-bution, while each of the k regular servers provides service independently with exponentially distributed service times. The duration of consulta-tions is also assumed to follow an exponential distribution. An explicit stability condition for the system is derived. Furthermore, an optimiza-tion problem is formulated to determine the optimal number of regular servers that maximizes the expected total profit. Numerical analysis is conducted to evaluate key performance measures and to illustrate the impact of system parameters.
Keywords: *multi-server system, main server, regular server, consul-tation, interruption*

Date of Submission: 06-04-2026 Date of Acceptance: 16-04-2026

I. Introduction

Indeed, in today's world, the demand for various services is consistently high, leading service providers to adopt strategies to maintain goodwill and enhance customer satisfaction. Multi-server queueing systems are a common approach to efficiently handle service demands. In multi-server queueing systems, the nature of services offered by different servers can indeed vary significantly. In some systems, all servers may offer the same type of service. This means that each server is capable of handling the same tasks or requests from customers. Alternatively, servers in a multi-server system might provide entirely different types of services. There are some other multi-server systems also in which servers offer some common services and some different services. This means that while the services provided may differ in certain aspects, they share common functionalities or resources.

It is common for some servers, such as trainees or less experienced staff, to require frequent clarifications or assistance. In such cases, experienced servers play a crucial role in providing timely support while simultaneously serving customers. This practice is observed in various industries, including banks, hos-pitals, supermarkets, etc. In a bank, the manager or senior staff members often assist other bank staff in addition to serving customers. They provide guidance, resolve complex issues, and ensure smooth operations within the bank. In hos-pitals, chief physicians or experienced doctors not only treat patients but also clarify doubts and provide guidance to junior doctors or medical staff. This ensures that medical procedures are carried out efficiently and accurately. In supermarkets, senior employees or supervisors may assist other staff members when handling complex customer queries. They ensure that the checkout pro-cess runs smoothly and help resolve any issues that arise.

Consider a multi-server queueing model offering the same type of services. For example, in a call centre, all agents handle customer inquiries or support re-quests related to a particular product or service Similarly, in a fast-food restuar-ant, all counter staff serve customers by taking orders and processing payments. In these cases, the queueing system is designed to effectively distribute incom-ing requests among available servers, optimizing service delivery. By providing timely clarifications and support, experienced servers contribute to a positive customer experience and enhance the reputation of the service provider.

Chakravarthy [1] is credited with introducing a multi-server queuing sys-tem that incorporates a consultation mechanism. This innovative approach to queuing systems integrates the concept of consultations between servers, offer-ing new insights into optimizing service delivery and resource utilization. In this described system with c servers, where one server is designated as the main server and the rest as regular servers, the main server primarily serves customers directly but also provides timely consultations to the regular servers. The con-sultations are given preemptive priority on a first-in-first-out (FIFO) basis to the regular servers. Preemptive priority implies that if a regular server requests a consultation while the main server is serving a customer, the main server will pause the current task, address the consultation request, and then resume the original task. This ensures that the regular servers receive timely assistance from the main server as needed. FIFO means that consultation requests are addressed in the order they are received. The main server does not prioritize consultations based on urgency or any other criteria besides the order of arrival. This fairness principle ensures that all regular servers have an equal opportu-nity to receive consultations from the main server without any bias or preference.

Consequently, whenever a consultation is initiated, the service of the cus-tomer currently being served at the main server is interrupted. The other regular servers can queue up for consultation. Once all consultations are concluded, the main server resumes service for the interrupted customer. When the main server is idle (no customer to serve directly), it can immediately provide consultations to the regular servers without interruption. No restriction falls on the number of consultations to a regular server when he is serving a particular customer. Customers arriving at the main server are served according to the exponential distribution with rate μ1 while those at the regular servers get service according to the same distribution with a different rate μ2.

One notable application of a queueing system with consultations, in the context of an airport is elaborated in this paper [1].

Jeganathan K. et. al. [2] study a multi-server queueing–inventory system consisting of n junior servers, one senior server, and two separate waiting halls designated for customers and junior servers. During service, the junior servers seek guidance from the senior server.

Dudin S. and Dudina O. [3] consider a multi-server queueing system with a finite buffer. The arrival of requests is governed by a Markov arrival process. Service is provided in batches, where both the minimum and maximum group sizes are fixed. The service time of each group follows a phase-type distribution with an irreducible representation that depends on the group size. Additionally, the requests exhibit impatience, with the patience time of each request being exponentially distributed.

Vardanyan A. P. [4] studies a multiprocessor queueing model where tasks demand a random number of processors and are constrained by queue waiting times. In contrast to classical multi-server systems, the model integrates both resource requirements and waiting time limits, enhancing its applicability to real-world computing scenarios.

Krishnamoorthy et al. [5] analyze a single-server queueing system with in-terruptions, in which the occurrence of interruptions is controlled through a super clock mechanism and limited by a predetermined finite upper bound.

Early studies on queueing systems with service interruptions date back to White and Christie [6], who assumed that a customer's service resumes imme-diately after an interruption ends. Later, Gaver [7], Keilson [8], Ibe and Trivedi [9], Avi-Izhak and Naor [10], and Fiems et al. [11] extended the analysis to models with generally distributed service processes and interruption times.

In the present paper a multi-server queueing model with k + 1 servers is examined in which the main server's service time has a phase type distribution.

II. Model Description

Customers arrive according to a Poisson process with rate λ. Upon arrival, a customer immediately enters service if at least one server is available. If the main server is idle, it is given priority, and the arriving customer is served by the main server. The service time at the main server follows a phase-type distribution with representation (α, U) consisting of p phases, whereas the service times at the k regular servers are independent and identically distributed exponential random variables with parameter μ.

If i regular servers are busy ($1 \leq i \leq k$), then the rate at which consultation is required is $1 \leq i \leq k$), then the rate of requirement of consultation is iδ. . Each request for consultation is attended immediately whenever the main server is available. When the main server engages in consultation, its ongoing customer service is said to be interrupted. However, if the main server is currently oc-cupied with a consultation, subsequent requests from regular servers are placed in a queue. These requests are then processed by the main server in a first-in-first-out (FIFO) manner. At any given time, at most k regular servers can require consultation. The service of a customer interrupted at the main server is resumed only after all pending consultation requests in the queue have been completed. The duration of each consultation is assumed to be exponentially distributed with parameter γ. Here, the service of a customer at a regular server that requires consultation during its service is not regarded as interrupted, since such consultations are considered an integral part of the service process.

In this queueing system, the main server can transit between several distinct states depending on its current activities and workload. These states encompass various scenarios involving customer service, consultations, and interruptions. Specifically, the main server can be in one of the following states:

1. In this state, the main server is actively serving customers. During this period none, one, or more regular servers may be in service.

2. Here, the main server is not currently serving any customers, but one or more regular servers are actively engaged in customer service tasks.
3. In this state, the main server is solely focused on providing consultations to other servers. It is not possess any interrupted customer but is instead assisting other servers with their tasks.
4. This state occurs when the main server is interrupted from serving a cus-tomer due to a consultation. While providing consultations, the main server has one interrupted customer whose service remains suspended un-til consultations are completed.

Notations: The following notations are used in this paper.

- $\mathbf{f}_j = \mathbf{e}_j(1)$
- $I^*_j = \begin{bmatrix} I_j & \mathbf{0} \end{bmatrix}_{j\times(j+1)}$
- $\eta = \delta + \mu$

Let the queueing model be $Y = \{Y(\vartheta), \vartheta \geq 0\}$.
Here $Y(\vartheta) = \{N(\vartheta), B(\vartheta), L(\vartheta), M(\vartheta)\}$, where

- $N(\vartheta)$ — the number of customers in the system
- $L(\vartheta)$ — number of regular servers in the queue for consultation
- $M(\vartheta)$ — phase of service of the customer at the main server (the service of that customer may be under interruption)

Here $B(\vartheta)$ denotes the status of the servers at time t such that

$$B(\vartheta) = \begin{cases} \tilde{0}, & \text{if the regular server(s) are busy and main server is idle} \\ 0, & \text{the main along with (or without) regular server is busy} \\ 1, & \text{if the main server is giving consultation only} \\ 2, & \text{if the main server is giving consultation with one interrupted customer at the main server} \end{cases}$$

$\{Y(\vartheta), \vartheta \geq 0\}$ is a Continuous Time Markov Chain with state space

$$\Phi = \{0\} \cup \bigcup_{j=1}^{\infty} \phi(j).$$

The terms $\phi(j)$'s are defined as

$$\begin{aligned} \phi(1) &= \{(1, 0, a) : 1 \leq a \leq p\} \cup \{(1, \tilde{0}) \cup (1, 1)\}, \\ \phi(j) &= \{(j, 0, a) : 1 \leq a \leq p\} \cup \{(j, \tilde{0})\} \cup \{(j, 1, i) : 1 \leq i \leq j\} \\ &\quad \cup \{(j, 2, i, a) : 1 \leq i \leq j-1, 1 \leq a \leq p\}, \text{ for } 2 \leq j \leq k, \ \phi(j) = \\ &\quad \{(j, 0, a) : 1 \leq a \leq p\} \cup \{(j, 1, i) : 1 \leq i \leq k\} \\ &\quad \cup \{(j, 2, i, a) : 1 \leq i \leq k, 1 \leq a \leq p\}, \text{ for } j \geq k+1. \end{aligned}$$

The infinitesimal generator A is

$$A = \begin{bmatrix} -\lambda & E_0 & & & & & & & \\ F_1 & D_1 & E_1 & & & & & & \\ & F_2 & D_2 & E_2 & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & F_c & D_c & E_c & & & \\ & & & & F_{c+1} & P_1 & P_0 & & \\ & & & & & P_2 & P_1 & P_0 & \\ & & & & & & \ddots & \ddots & \ddots \end{bmatrix} \quad (1)$$

where $E_0 = \lambda \begin{bmatrix} \alpha & \mathbf{0} \end{bmatrix}_{1\times p+2}$, $E_1 = \lambda \begin{bmatrix} I_p & O & O \\ \alpha & 0 & O \\ \mathbf{0} & 0 & I^*_1 \end{bmatrix}_{(p+2)\times(2p+3)}$,

$E_j = \lambda \begin{bmatrix} E_{j1} & \mathbf{0} & E_{j2} \end{bmatrix}_{(1+j+jp)\times(j+2+(j+1)p)}$, for $2 \leq j \leq k-1$,

$E_k = \lambda \begin{bmatrix} I_p \\ diag(\alpha, I_k, I^*_{k-1} \otimes I_p) \end{bmatrix}_{(1+k+kp)\times r}$;

$$D_1 = \begin{bmatrix} U & & \\ & -\eta & \delta \\ & \gamma & -\gamma \end{bmatrix}_{p+2\times p+2} - \lambda I;$$

$$D_j = \begin{bmatrix} U - (j-1)\eta I_p & (j-1)\delta \mathbf{f}_{j-1}' \otimes I_p & \\ & -j\eta & j\delta \mathbf{f}_j \\ \xi \mathbf{f}_{j-1} \otimes I_p & \xi \mathbf{f}_j & Q_j \\ & Q_{j-1} \otimes I_p & \end{bmatrix} - \lambda I, \text{ for } 2 \le j \le k;$$

$$F_1 = \begin{bmatrix} U^0 \\ \mu \\ \mathbf{0} \end{bmatrix}_{(p+2)\times 1}, \quad F_2 = \begin{bmatrix} \mu I_p & U^0 & \\ & 2\mu & \\ & & T_1 \\ O & O & O \end{bmatrix}_{(3+2p)\times(p+2)}$$

$$F_j = \begin{bmatrix} (j-1)\mu I_p & U^0 & \\ & j\mu & \\ & & T_{j-1} \\ & & T_{j-2} \otimes I_p \end{bmatrix}_{(j+1+jp)\times(j+(j-1)p)},$$

for $3 \le j \le k$;

$$F_{k+1} = \begin{bmatrix} k\mu I_p & U^0 & \\ & P_{21} & \\ & & T_{k-1} \otimes I_p \end{bmatrix}_{(p+1+kp)\times(k+1+kp)},$$

$$P_0 = \lambda I; \quad P_1 = \begin{bmatrix} U - k\eta I_p & O & k\delta \mathbf{f}_k \otimes I_p \\ \gamma \mathbf{f}_k \otimes \boldsymbol{\alpha} & Q_k & \\ \gamma \mathbf{f}_k \otimes I_p & & Q_k \otimes I_p \end{bmatrix} - \lambda I$$

$$P_2 = \begin{bmatrix} U^0 \otimes \boldsymbol{\alpha} + k\mu I_p & & \\ & P_{21} & \\ & & P_{21} \otimes I_p \end{bmatrix}$$

The sub-matrices are given as follows: ... of order $k \times (k-1)p$.

$$E_{j1} = \begin{bmatrix} I_p \\ \boldsymbol{\alpha} \\ O \end{bmatrix}_{(1+j+jp)\times p}, \quad E_{j2} = diag(I_j^*, \bar{I}_{j-1}^* \otimes I_p)_{(1+j+jp)\times(j+1+jp)};$$

$$Q_j = \begin{bmatrix} -(j-1)\eta & (j-1)\delta & & & \\ \gamma & -(j-2)\eta & (j-2)\delta & & \\ & \gamma & -(j-3)\eta & (j-3)\delta & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ & & & \gamma & -\eta & \delta \\ & & & & \gamma & 0 \end{bmatrix}_{j\times j} -$$

V_{k};

$$T_j = \begin{bmatrix} diag(j\mu, (j-1)\mu, \ldots, \mu) \\ \mathbf{0} \end{bmatrix}_{(j+1)\times j}, \text{ for } 2 \le j \le k;$$

$$P_{21} = diag((k-1)\mu, (k-2)\mu, \ldots, \mu, 0).$$

3 Steady state analysis

This section deals with the steady-state analysis of the model, starting with the derivation of the system's stability condition.

3.1 Stability condition

Let the steady-state probability vector of the generator $P = P_0 + P_1 + P_2$ be denoted by $\boldsymbol{\pi}$. That is,

$$\boldsymbol{\pi}P = 0 \tag{2}$$

$$\boldsymbol{\pi}\mathbf{e} = 1 \tag{3}$$

We now present a theorem that determines the stability of the queueing system under study.

3.2 Theorem

The Markov Chain Y is stable if and only if

$$\lambda < \frac{1}{\zeta}\Big[\mu_1 + \mu \sum_{j=1} \frac{k!}{(j-1)!}\Big(\frac{\delta}{\gamma}\Big)^{k-j}\Big] \tag{4}$$

where $\mu_1 = [\boldsymbol{\alpha}(-U)^{-1}\mathbf{e}]^{-1}$ is the service rate of the main server and

$$\zeta = \sum_{j=0} \frac{k!}{(k-j)!}\Big(\frac{\delta}{\gamma}\Big)^{j} \tag{5}$$

Proof

The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [12]) if and only if

$$\boldsymbol{\pi}P_0\mathbf{e} < \boldsymbol{\pi}P_2\mathbf{e}. \tag{6}$$

Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$, where
$\boldsymbol{\pi}_1 \equiv (\pi_{11}, \ldots, \pi_{1k})$ and $\boldsymbol{\pi}_2 \equiv (\boldsymbol{\pi}_{21}, \ldots, \boldsymbol{\pi}_{2k})$.
Using the structure of P and equation (2), it is easy to verify that

$$\boldsymbol{\pi}_1 = \mathbf{0};$$
$$\gamma\boldsymbol{\pi}_{2j} = (k+1-j)\delta\boldsymbol{\pi}_{2j-1}; \text{ for } 1 \le j \le k. \tag{7}$$

Using equation (7) and the normalizing condition (3), it follows that

$$\zeta\boldsymbol{\pi}_0 = 1 \tag{8}$$

where ζ is given in (5). Also we have

$$\pi P_2 \mathbf{e} = \pi_0 [\mu_1 + \mu \sum_{j=1} \frac{k!}{(j-1)!} (\frac{\delta}{\gamma})^{k-j}]$$

Then the stability condition (6) implies the result (4).

3.3 Steady state probability vector

Let a, partitioned as, $a = (a_0, a_1, a_2, \ldots)$ be the steady state probability vector of the Markov chain $\{Y(\vartheta), \vartheta \geq 0\}$. Note that a_0 is a scalar, $a_1 = (a_{10}, a_{10^-}, a_{11})$, and $a_j = (a_{j0}, a_{j0^-}, a_{j1}, a_{j2})$, for $2 \leq j \leq k$ and $a_j = (a_{j0}, a_{j1}, a_{j2})$, for $j \geq k+1$.

Here a_{10} is an p−dimensional vector whereas a_{1^-0} and a_{11} are scalars.

a_{j0}, a_{i0^-}, a_{j1}, a_{j2}, for $2 \leq j \leq k$ are vectors of dimensions p, 1, j, $(j-1)p$, for $2 \leq j \leq k$, and a_{j0}, a_{j1}, a_{j2}, for $j \geq k+1$ are vectors of dimensions p, k, kp, respectively.

The vector a satisfies the condition $a\,Q = 0$ and $a\,\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub-vectors of a are given by the equation

$$a_i = a_{k+1} R^{i-(k+1)}, i \geq k+1, \tag{9}$$

where R is the minimal non-negative solution of the matrix equation $R^2 P_2 + RP_1 + P_0 = 0$. Knowing the matrix R, the vectors $a_0, a_1, \ldots a_{k+1}$ are obtained by solving the equations

$$-\lambda a_0 + a_1 F_1 = 0$$

$$a_{j-1} E_{j-1} + a_i D_i + a_{j+1} F_{j+1} = 0, \text{ for j=1,…,k-1} \tag{10}$$

$$a_k E_k + a_{k+1}(P_1 + RP_2) = 0$$

subject to the normalizing condition

$$a_0 + a_1 \mathbf{e} + \ldots + a_k \mathbf{e} + a_{k+1}(I - R)^{-1} \mathbf{e} = 1. \tag{11}$$

3.4 Performance measures

Here are some common performance measures we are interested in.

(1) Mean number of customers in the system

$$M_S = \sum_{j=1}^{\infty} j\, a_j \mathbf{e}.$$

(2) Mean number of customers in the queue

$$M_Q = \sum_{j=k+1}^{\infty} (j-k)\, a_{j1} \mathbf{e} + \sum_{j=k+2}^{\infty} (j-k-1)(a_{j0}\mathbf{e} + a_{j2}\mathbf{e}).$$

(3) Mean number of idle regular servers

$$M_{RS} = \sum_{j=1}^{k-1} (k-j)\, a_{j^{*}0}\mathbf{e} + \sum_{j=1}^{k} (k+1-j)\, a_{j0}\mathbf{e}.$$

(4) Effective rate of interruption

$$\epsilon_i = \sum_{j=2}^{k} \Big(j\delta\, a_{j0} + \sum_{i=1}^{j-1} (j-i)\delta\, a_{j2i}\Big)\mathbf{e} + \sum_{j=k+1}^{\infty} \Big(k\delta\, a_{j0} + \sum_{i=1}^{k-1} (k-i)\delta\, a_{j2i}\Big)\mathbf{e}.$$

(5) Effective rate of consultation

$$\epsilon_c = \epsilon_i + \delta\, a_{1^{*}0} + \sum_{j=2}^{k}\sum_{i=1}^{j-1} (j-i)\delta\, a_{j1j} + \sum_{j=k+1}^{\infty}\sum_{i=1}^{k-1} (k-i)\delta\, a_{j1i}.$$

(6) Fraction of time the main server is idle

$$\delta_{mi} = a_0 + \sum_{j=1}^{k} a_{j^{*}0}\mathbf{e}.$$

(7) Fraction of time all the servers are busy serving customers

$$\delta_{ab} = \sum_{j=k+1}^{\infty} a_{j0}\mathbf{e}.$$

3.5 An optimization problem

Now we construct an optimization problem based on the revenue and cost considerations. To define an objective function, we consider revenue generated by the system due to service completions and the costs associated with idle regular servers and waiting spaces. Thus we express per unit time revenue and cost as follows:

1. r be the revenue per customer going out of the system after service completion
2. c_1 be the cost for holding customers in the system
3. c_2 cost for holding regular servers in idle state

To formulate the optimization problem for finding the optimal number of regular servers to be employed to maximize the expected total profit (ETP), define the objective function as

$$ETP = r \times ESR - c_1 \times M_S - c_2 \times M_{RS} \qquad (12)$$

where $ESR = \pi P_2\mathbf{e}$.

4 Numerical examples

To illustrate the impact of the parameter c on the expected total profit ETP we'll consider a numerical example where the service rate of the main server is equal to the service rates of the regular servers. That is, U and $\boldsymbol{\alpha}$ so that $[\boldsymbol{\alpha}(-U)^{-1}\mathbf{e}]^{-1}$ is chosen as μ.

Let $\lambda = 5$, $\delta = 9$, $\mu = 2$, $\gamma = 3$, $U = \begin{pmatrix} -9 & 3 \\ 2 & -8 \end{pmatrix}$, $\boldsymbol{\alpha} = (0.3 \quad 0.7)$

The above data of matrices, vectors and values satisfy the stability condition (4). Fix $r = 25$, $c_1 = 15$ and $c_2 = 100$.

Table 1: Effect of number of k on cost function

k	3	4	5	6	7	8
ESR	9.8130	15.2409	20.3885	25.1093	29.5694	33.8892
M_S	11.7206	3.9603	2.6137	2.1678	1.9604	1.8370
M_{RS}	0.6429	2.1892	3.6003	4.8710	6.0460	7.1640
ETP	5.226	102.698	**110.477**	108.116	105.229	103.275

Table 5.1 shows that an increment in k leads to an increases in the effective service rate ESR. Consequently, customers experience shorter wait times and are served more quickly, reducing the accumulation of customers in the system. So M_S decreases. As the number of regular servers increases, for a particular λ, number of idle regular servers M_{RS} also increases. This aligns with our expectations.

Thus ETP has an optimum value **110.477** when the number of regular servers $k = 5$.

4.1 More numerical examples

In this section, we present some numerical examples that describe the performance characteristics of the queueing model under study.

Let $k = 3$ and λ, δ, μ, γ, U and $\boldsymbol{\alpha}$ are as given in the above example.

Table 2: Effect of δ on various performance measures

δ	3	4	5	6	7
M_S	0.8597	1.0946	1.4578	2.0588	3.1573
M_Q	0.0531	0.1264	0.2809	0.6087	1.3436
δ_{mi}	0.6444	0.5966	0.5401	0.4727	0.3915
δ_{ab}	0.0081	0.0106	0.0137	0.0177	0.0227

Table 5.2 shows that Increasing δ means that each customer spends less time being served before an interruption occurs. This results in a higher rate of interruptions during service. With interruptions occurring at a faster rate, customers spend more time in the system and in the queue, waiting for their services to be

Table 3: Effect of λ on various performance measures

= 3

λ	3	3.5	4	4.5	5
M_s	1.9332	2.8234	4.1764	6.4072	10.266
M_q	0.4030	0.8879	1.8212	3.6097	6.9135
δ_{mi}	0.4658	0.3805	0.2985	0.2197	0.1440
δ_{ab}	0.0386	0.0667	0.1045	0.1516	0.2038

Table 4: Effect of μ on various performance measures

$\delta = 3$

μ	2	2.5	3	3.5	4
M_s	4.1764	2.7497	2.0696	1.6839	1.4400
M_q	1.8212	0.8359	0.4485	0.2660	0.1693
δ_{mi}	0.2985	0.3683	0.4130	0.4434	0.4651
δ_{ab}	0.1045	0.0625	0.0400	0.0271	0.0192

Table 5: Effect of γ on various performance measures

$\delta = 6, \mu = 2.5$

γ	3	3.5	4	4.5	5
M_s	11.7064	11.0572	6.1958	4.2548	3.3207
M_q	8.6051	7.6826	3.4795	1.8598	1.1410
δ_{mi}	0.0283	0.1303	0.2095	0.2662	0.3087
δ_{ab}	0.0556	**0.1018**	0.0909	0.0813	0.0743

completed. This leads to an increased number of customers in the system MS and MQ. As the system becomes more congested due to increased customer ac-cumulation, the fraction of time all servers are busy serving customers ab tends to increase. This indicates that the servers are more fully utilized. Conversely, with the increased congestion and higher interruption rates, the fraction of time servers are idle δmi tends to decrease. This is because servers are more fre-quently engaged in serving customers or experiencing interruptions, leaving less idle time.

Table 5.3 shows that as the arrival rate increases, customers arrive at the system more frequently, leading to higher customer accumulation in the sys-tem and in the queue. This results in an increase in both the total number of customers in the system MS and the number of customers waiting in the queue MQ. With more frequent customer arrivals, the servers are kept busy for a larger portion of time, leading to an increase in the fraction of time all servers are busy serving customers δab. Since the servers are busier due to higher customer arrivals, the idle time of the main server δmi decreases. This means that the main server spends less time idle and more time engaged in serving customers or experiencing interruptions.

From table 5.4, we can see that as the service rate μ of the regular servers increases, customers are served at a faster rate, resulting in a slower accumu-lation of customers in the system and in the queue MS and MQ. With faster service rates, customers spend less time waiting in the system and in the queue before being served, leading to a decrease in the total number of customers in the system and in the queue. Since more customers are being served by the regular servers, fewer customers approach the main server. Consequently, the main server experiences more idle time δmi, as it has fewer customers to serve. With the main server experiencing more idle time and the regular servers serv-ing customers at a faster rate, the overall fraction of time all servers are busy δab decreases. This indicates that the system is less congested and that servers are not fully utilized all the time.

We see from table 5.5 that as γ increases, the rate of consultation completion increases as well. This allows servers to have more uninterrupted time to serve customers. With interruptions occurring for lesser duration, servers can serve customers fasty, leading to a decrease in the accumulation of customers in the system and in the queue MS and MQ. Since a larger number of customers are served by the regular servers due to duration of consultation is less, the main server experiences more idle time δmi. Now consider δab. We see that δab increases till γ=3.5. If again γ increases, since the value of λ is fixed, the servers need less amount of time for service completion and therefore the fraction of time all servers are busy serving customers δab will decrease.

References

[1] Chakravarthy S. R. : A Multi-Server Queueing Model With Server Consulta-Tions, Eur.J. Of Oper. Res., 233(3), 625-639, 2014.
[2] Jeganathan K, Koffer VA, Lakshmi KP, Anbazhagan N, Joshi GP, Cho W.: Analysis Of Junior Servers Approaching A Senior Server In The Multi-Server Queueing-Inventory System. Sci Rep. ;15(1):15860, 2025 May 6.
[3] Dudin S, Dudina O.:Analysis Of A Multi-Server Queue With Group Service And Service Time Dependent On The Size Of A Group As A Model Of A Delivery System. Mathematics, 11(22):4587, 2023.
[4] Vardanyan, A. P.: Advanced Queueing Model Of A Multiprocessor Com-Puting System. Mathematical Problems Of Computer Science, 62, 43–51, 2024.
[5] Krishnamoorthy A., Pramod P. K., Chakravarthy S. R. : A Note On Char-Acterizing Service Interruptions With Phase Type Distribution, Stochastic Analysis And Applications , 31(4), 671-683, 2013.
[6] White H., Christie L. S. : Queuing With Preemptive Priorities Or With Breakdown, Operations Research, 6, 79–95, 1958.
[7] Gaver D. P. : A Waiting Line With Interrupted Service Including Priorities, Journal Of Royal Statistical Society, B 24(1), 73-90, 1962.
[8] Keilson J. : Queues Subject To Service Interruptions, The Annals Of Math-Ematical Statistics 33(4), 1314-1322, 1962.
[9] Ibe O. C., Trivedi K. S.: Two Queues With Alternating Service And Server Breakdown, Queueing Systems 7(3), 253-268, 1960.
[10] Avi-Itzhak B, Naor P : Some Queueing Problems With The Service Station Subject To Breakdowns, Oper. Res. 11(3), 303-320, 1963.
[11] Fiems D., Maertens T., Bruneel H. : Queueing Systems With Different Types Of Interruptions, Eur J Oper Res 188(3), 838-845, 2008.
[12] Neuts M.F. : Matrix-Geometric Solutions In Stochastic Models, An Algo-Rithmic Approach, The Johns Hopkins University Press, Baltimore, 1981.
[13] Neuts, M.F. And Lucantoni, D.M. : A Markovian Queue With N Servers Subject To Breakdowns And Repairs, Management Science, 25(9), 849-861, 1979.