Analysis of a two-server queue with consultation by main server having protected phases of service

Abstract

In a queueing system where service is carried out in multiple phases and is subjected to interruptions, the idea of protecting certain critical phases from interruption becomes highly significant. This paper investigates a two-server queueing system in which a primary (main) server provides consultation to a secondary (regular) server. The main server performs dual functions: servicing customers and offering consultation to the regular server, with consultations being assigned preemptive priority over customer service. Customer service at the main server is subjected to stochastic interruptions. But interruptions are prohibited when the service process is in designated protected phases. Some constraints are imposed through upper bounds on the number of allowable interruptions experienced by a customer at the main server and the number of consultations provided to the regular server during an individual service period. There is a supervisory timing mechanism super clock which governs the admissibility of additional interruptions to an ongoing service at the main server. A threshold clock regulates the restart or resumption of service at both servers following each consultation event. Customer arrivals and consultation requests are modelled as mutually independent Poisson processes. Service times at

both the main and regular servers are assumed to follow independent phase-type distributions. The system's stability condition is derived, and key performance measures are evaluated through numerical analysis.

Key words : main server, regular server, consultation, interruption, protected phases

1 Introduction

Multi-server queueing systems incorporating consultation mechanisms, wherein a designated primary server provides assistance to subordinate servers, are widely encountered in practical settings. For instance, in healthcare systems, a chief physician offers expert guidance to attending doctors regarding patient diagnosis and treatment decisions. The inclusion of consultation mechanisms plays a significant role in enhancing the overall quality and reliability of service.

Chakravarthy [1] introduced a c-server queueing system with consultations, in which one of the servers is designated as the main server and the remaining as regular servers. The main server gives preemptive priority to consultation requests from the regular servers over its own customers. Whenever a regular server requests consultation, the main server immediately attends to it, even if it is currently serving a customer. Consequently, the service of the customer at the main server is interrupted and remains suspended until the consultation is completed. In contrast, the service at the regular server is not considered interrupted, since consultation is treated as an integral part of its service process. If multiple regular servers require consultation simultaneously, they form a queue, and consultations are provided on a first-come, first-served (FIFO) basis. The interrupted service at the main server resumes only after all pending consultations are completed. Notably, there is no restriction on the number of interruptions a customer at the main server may experience; similarly, a regular server may request any number of consultations during the service of a customer.

Klimenock et al. [2] analyze a multi-server queueing model with heterogeneous customers, where two types of customers with different priority levels are considered. The service times at the servers are assumed to follow independent phase-type distributions.

Samouylov et. al. [3] considers a multi-server queueing model that serves two correlated streams of requests. A non-preemptive priority mechanism is implemented by introducing a preliminary delay for one of the streams through intervening buffers, from which requests are retrieved at different rates.

Krishnamoorthy et. al. [4] investigate a two-server queueing model where both servers have service times represented by phase-type distributions of identical order, with one server having a comparatively lower service rate.

Ayyappan G. and Archana G. [5] elaborate a classical queueing system by considering two types of heterogeneous servers. Optional services are provided to customers who are not satisfied with the initial service. The service at server 2 is subject to interruptions due to breakdowns during any stage of service; however, despite such breakdowns, server 2 continues and completes the ongoing service at a reduced rate.

T. Resmi and K Ravikumar [6] analyze a three-server queueing system in which a main server provides consultations to two regular servers. A finite buffer is maintained at the main server, and preemptive priority is given to consultation requests over customer service.

Resmi et. al. [7] investigate a two-server queueing model with mutual consultation between servers, where each server provides timely assistance to the other, and the service times are governed by independent phase-type distributions.

Krishnamoorthy et.al [8] investigate a single-server queueing model experiencing interruptions, with the number of such interruptions being controlled through a super clock mechanism and a prescribed finite upper bound.

Early work on queueing systems with service interruptions can be traced back to White and Christie in their paper [9]. The customer's service is resumed immediately upon completion of the interruption.
Gaver [10] , Keilson [11], Ibe and Trivedi [12], Avi-Izhak and Naor [13] and Fiems et. al [14] include works that analyze queueing systems with generally distributed service processes and interruption durations.

Some phases of service may involve costly or delicate operations, where any interruption could lead to substantial loss of time, resources, or service quality. By designating such phases as "protected," interruptions are either delayed or temporarily disallowed during their execution. This ensures smoother service completion for sensitive stages, reduces inefficiencies caused by repeated restarts, and improves overall system performance. Consequently, incorporating protected phases into interrupted service models provides a more realistic and practical framework for analyzing complex service systems.

Klimenok et. al. [15] consider a multi-server queueing model with finite buffer and negative customers, where arrivals follow a BMAP and services are PH-type. Negative customers eliminate customers in service only during unprotected phases, while protected phases remain immune.

Klimenok and Dudin [16] consider admission control policies of complete acceptance and complete rejection, and assumes an infinite buffer capacity.

Krishnamoorthy et. al. [17] considers the concept of protection in a queueing system with service interruptions. If the service process is in any of the first n phases of an Erlang service process with m phases, it is subject to interruption, whereas the remaining mn phases are protected. Following an interruption, the service is either resumed or restarted. Interruptions do not result in any loss of customers from the system.

In the present study, we analyze a two-server queueing model in which a primary (main) server provides consultation to a secondary (regular) server. The service times at both servers are assumed to follow mutually independent phase-type distributions. The service process at the main server is subjected to interruption upon the occurrence of consultation requests from the regular server.

Although consultation improves the overall quality of service, excessive interruptions to customers at the main server may lead to undesirable delays. To address this, restrictions are imposed on the provision of consultation while the main server is actively serving a customer. Specifically, interruptions are regulated through upper bounds on the number of allowable interruptions and consultations. An additional layer of control can be introduced through a supervisory timing mechanism, often referred to as a super clock. If the super clock has not expired, then an interruption is allowed to the customer at the main server, otherwise, the regular server has to

wait until the completion of the service of the customer at the main server.

We assume that some phases of the service at the main server are too costly, so these pahases are protected from interruption. These phases are called proteceted phases while the other pahses where interruptions are allowed, are said to be non-protected pahses. Consultation to the regular server will be denied if the main server is at the protected phase. In this situation, the regular server must wait until the completion of the ongoing service at the main server before receiving assistance.

This threefold mechanism—phase-based protection, an upper bound on the number of interruptions, and time-based supervision—significantly enhances the flexibility and realism of the model. It enables the system to achieve an effective balance between efficiency and quality of service by preventing disruptions during critical phases, while still permitting necessary interruptions in a controlled and limited manner. Such a framework is particularly valuable in practical applications such as maintenance systems, communication networks, and healthcare services, where both precise timing and continuity of service play a crucial role in overall system performance.

A threshold clock governs the restart or resumption of services at both the main and regular servers following each consultation. This clock is initiated whenever the regular server temporarily suspends its service to seek consultation. If the regular server receives consultation immediately, the consultation process and the threshold clock commence simultaneously. However, if the regular server must wait for consultation, the threshold clock starts ticking during the waiting period and continues throughout the consultation phase once the main server becomes available. If the threshold clock expires before the completion of the consultation process, the services at both servers are restarted afresh. Otherwise, the services are resumed from the exact phases at which they were interrupted.

2 Description of the model

In this model, we consider a queueing system consisting of one main server and one regular server. The customers arrive to the system according to a Poisson process with rate λ. An arriving customer gets service immediately if at least one server is free, else joins the queue of waiting customers. The service times at the main and regular servers are independently distributed and follow phase-type distributions, characterized by$(\boldsymbol{\alpha}, U)$ and $(\boldsymbol{\beta}, V)$ with

p and r, phases, respectively. Write $\mathbf{U}^0 = -U\mathbf{e}$ and $\mathbf{V}^0 = -V\mathbf{e}$. Here, $\mathbf{e}$ denotes a column vector of ones of appropriate dimension. It is assumed that, among the p phases of service at the main server, $m \leq p$ phases are designated as protected.

The requirement for consultation follows a Poisson process with rate θ. Let L and M denote the maximum allowable numbers of interruptions to a customer at the main server and consultations for a customer at the regular server, respectively. These bounds are imposed to prevent excessive delays and to ensure that customers receiving service at the regular server do not become too impatient and leave the system.

The durations of the super clock, threshold clock, and consultation clock are assumed to be mutually independent and follow phase-type distributions, with representations given by $(\boldsymbol{\gamma}, G)$, $(\boldsymbol{\eta}, E)$, $(\boldsymbol{\delta}, D)$ with number of phases c, d and f, respectively. We have $\mathbf{G}^0 = -G\mathbf{e}$, $\mathbf{E}^0 = -E\mathbf{e}$, and $\mathbf{D}^0 = -D\mathbf{e}$, respectively.

Notations :- We use the following notations in this model.

- $L_0 = L(c+1), L_1 = L_0 + 1$
- $\tilde{\boldsymbol{\alpha}} = \mathbf{e}'_{L_1}(1) \otimes \boldsymbol{\alpha}$, $\tilde{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}, 0)$, $\tilde{\boldsymbol{\eta}} = (\boldsymbol{\eta}, 0)$
- $\tilde{G} = \begin{bmatrix} G & G^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$, $\tilde{E} = \begin{bmatrix} E & E^0 \\ \underline{\mathbf{0}} & 0 \end{bmatrix}$, $D^* = D \oplus \tilde{E}$, $G^* = \tilde{G} \oplus D^*$
- $\boldsymbol{\delta}^* = \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}}$ and $\boldsymbol{\gamma}^* = \tilde{\boldsymbol{\gamma}} \otimes (\boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}})$
- $\dot{I} = \begin{bmatrix} \mathbf{0} & I_{L_0} \end{bmatrix}_{L_0 \times L_1}$, $\bar{I}_m = \tilde{\boldsymbol{\eta}} \otimes \begin{bmatrix} O & O \\ O & I_{p-m} \end{bmatrix}_{p \times p}$
- $\tilde{\mathbf{e}}_c = \begin{bmatrix} \mathbf{e}_c \otimes \bar{I}_m \\ \tilde{\boldsymbol{\eta}} \otimes I_p \end{bmatrix}$, $I_m^* = \begin{bmatrix} I_m \\ O \end{bmatrix}_{p \times m}$

Consider the queueing model $P = \{P(\tau), \tau \geq 0\}$, where
$P(\tau) = \{N(\tau), S(\tau), K_1(\tau), K_2(\tau), \sigma_1(\tau), \sigma_2(\tau), \sigma_3(\tau), J_1(\tau), J_2(\tau)\}$.
Here $N(\tau)$ is the number of customers in the system, $K_1(\tau)$ is the number of consultations already enjoyed by the regular server during the service of a particular customer, $K_2(\tau)$ is the number of interruptions already befell to a customer at the main server. $\sigma_1(\tau)$, $\sigma_2(\tau)$ and $\sigma_3(\tau)$ represent the phases of the super clock, the consultation process and the threshold clock, respectively

and $J_1(\tau)$ and $J_2(\tau)$ represent the phases of the main and the regular servers, respectively.

Here $S(\tau)$ denotes the status of the servers at time τ such that

$$S(\tau) = \begin{cases} \tilde{0}, & \text{if only the regular server is busy} \\ 0, & \text{if the main together with or without} \\ & \text{the regular server is busy} \\ 1, & \text{if the main server is giving consultation only} \\ 2, & \text{if the main server is giving consultation} \\ & \text{with one interrupted customer at the main server} \\ 3, & \text{if the regular server is waiting for getting consultation} \\ & \text{after the present service at the main server} \end{cases}$$

Note that $K_2(\tau)$ is '0' means the customer at the main server has not interrupted yet and so super clock has not started. In this case the super clock has no role to play. So we do not consider the super clock variable $\sigma_1(\tau)$ when $K_2(\tau) = 0$. Also, since super clock is associated with the interruption to a customer at the main server and no customer is present at the main server during the 'consultation only' mode, super clock is not 'present' at this mode.

Since there are no interruption from the $(m+1)^{th}$ phase on wards, these phases are not present when $S(\tau) = 2$ or 3.

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

$$\Omega = \{0\} \cup \bigcup_{n=1}^{\infty} \omega(n).$$

The terms $\omega(i)$'s are defined as
$\omega(1) = \omega(1,0) \cup \omega(1,\tilde{0}) \cup \omega(1,1)$,
$\omega(n) = \omega(n,0) \cup \omega(n,1) \cup \omega(n,2) \cup \omega(n,3), n \geq 2$,

where

$$
\begin{aligned}
\omega(1,0) &= \{(1,0,0,\tau_1)\} \cup \{(1,0,k,j_1,\tau_1) : 1 \le k \le L, 1 \le \tau_1 \le p\} \\
\omega(1,\tilde{0}) &= \{(1,\tilde{0},i,\tau_2) : 0 \le i \le M\} \\
\omega(1,1) &= \{(1,1,i,j_2,j_3,\tau_2) : 0 \le i \le M-1\} \\
\omega(n,0) &= \{(n,0,i,0,\tau_1,\tau_2) \cup (n,0,i,k,j_1,\tau_1,\tau_2) : 0 \le i \le M, \\
&\quad 1 \le k \le L, 1 \le \tau_1 \le p\}, \\
\omega(n,1) &= \{(n,1,i,j_2,j_3,\tau_2) : 0 \le i \le M-1\} \\
\omega(n,2) &= \{(n,2,i,k,j_1,j_2,j_3,\tau_1,\tau_2) : 0 \le i \le M-1, 0 \le k \le L-1, \\
&\quad 1 \le \tau_1 \le m\} \\
\omega(n,3) &= \{(n,3,i,j_3,\tau_1,\tau_2) : 0 \le i \le M-1, 1 \le \tau_1 \le p\}, \\
&\quad \text{for } 0 \le j_1 \le c, 1 \le j_2 \le d, 0 \le j_3 \le f, 1 \le \tau_2 \le r.
\end{aligned}
$$

The infinitesimal generator Z is given by

$$
Z = \begin{bmatrix}
-\lambda & A_1 & & & & \\
A_2 & A_3 & A_4 & & & \\
 & A_5 & B_1 & B_0 & & \\
 & & B_2 & B_1 & B_0 & \\
 & & & \ddots & \ddots & \ddots \\
 & & & & & \\
\end{bmatrix} \tag{1}
$$

Here B_0, B_1 and B_2 are square matrices of order C_0; A_3 is a square matrix of order C_1 and A_1, A_2, A_4, A_5 are matrices of orders $1 \times C_1$, $C_1 \times 1$, $C_1 \times C_0$ and $C_0 \times C_1$, respectively, where
$C_0 = L_1(M+1)pr + Md(f+1)r + L_0Md(f+1)mr + M(f+1)mr$ and
$C_1 = L_1p + (M+1)r + Mrd(f+1)$ and these matrices are defined as follows:

$$
A_1 = \begin{bmatrix} \boldsymbol{\alpha} & \underline{\mathbf{0}} \end{bmatrix}, A_2 = \begin{bmatrix} \mathbf{e}_{L_1} \otimes U^0 \\ \mathbf{e}_{M+1} \otimes V^0 \\ \mathbf{0} \end{bmatrix}, A_3 = \begin{bmatrix} I_{L_1} \otimes U & O & O \\ O & A_{31} & A_{32} \\ O & A_{33} & I_M \otimes D^* \otimes I_r \end{bmatrix},
$$

$$
A_4 = \begin{bmatrix} A_{41} & A_{42} & O \end{bmatrix}, A_5 = \begin{bmatrix} A_{51} & A_{52} & A_{53} \end{bmatrix},
$$

$$
B_0 = \lambda I, B_1 = \begin{bmatrix} B_{11} & O & B_{12} & B_{13} \\ B_{14} & B_{15} & O & O \\ B_{16} & O & B_{17} & O \\ O & O & O & B_{18} \end{bmatrix}, B_2 = \begin{bmatrix} B_{21} & A_{53} & O \end{bmatrix}.
$$

Here the block matrices are

$$A_{31} = I_{M+1} \otimes V - \theta \left[\begin{array}{cc} I_M & \mathbf{0} \\ \underline{\mathbf{0}} & 0 \end{array} \right]_{M+1} \otimes I_r,\ A_{32} = \theta \left[\begin{array}{c} I_M \\ \underline{\mathbf{0}} \end{array} \right]_{(M+1)\times M} \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}} \otimes I_r,$$

$$A_{33} = \left[\begin{array}{cc} O & I_M \otimes D^0 \otimes \tilde{\Delta}_r \end{array} \right]_{Md(f+1)r\times(M+1)r},$$

$$A_{41} = \left[\begin{array}{c} \mathbf{e}'_{M+1}(1) \otimes I_{L_1} \otimes I_p \otimes \boldsymbol{\beta} \\ I_{M+1} \otimes \tilde{\boldsymbol{\alpha}} \otimes I_r \\ O \end{array} \right]_{C_1\times(M+1)L_1pr},\ A_{42} = \left[\begin{array}{c} O \\ I_{Md(f+1)r} \end{array} \right]_{C_1\times Md(f+1)r},$$

$$A_{51} = \left[\begin{array}{c} \mathbf{e}_{M+1} \otimes I_{L_1} \otimes I_p \otimes V^0 \\ O \end{array} \right]_{C_0\times L_0p},\ A_{52} = \left[\begin{array}{c} I_{M+1} \otimes \mathbf{e}_{L_1} \otimes U^0 \otimes I_r \\ O \end{array} \right]_{C_0\times(M+1)r},$$

$$A_{53} = \left[\begin{array}{c} O \\ I_M \otimes \boldsymbol{\delta} \otimes I_{f+1} \otimes U^0 \otimes I_r \end{array} \right]_{C_0\times Md(f+1)r},$$

$$B_{11} = I_{M+1} \otimes I_{L_1} \otimes (U \oplus V) - \theta \left[\begin{array}{cc} I_M & \mathbf{0} \\ \underline{\mathbf{0}} & 0 \end{array} \right] \otimes I_{L_1} \otimes I_{pr},$$

$$B_{12} = \theta \left[\begin{array}{c} I_M \\ \underline{\mathbf{0}} \end{array} \right]_{(M+1)\times M} \otimes Q \otimes I_m^* \otimes I_r,\ B_{13} = \theta \left[\begin{array}{c} I_M \\ \underline{\mathbf{0}} \end{array} \right]_{(M+1)\times M} \otimes Q^* \otimes I_r,$$

$$B_{14} = \left[\begin{array}{cc} O & I_M \otimes D^0 \otimes \Delta^0 \end{array} \right]_{Md(f+1)r\times L_1(M+1)pr},\ B_{15} = I_M \otimes D^* \otimes I_r,$$

$$B_{16} = \left[\begin{array}{cc} O & I_M \otimes \dot{I} \otimes D^0 \otimes \tilde{\Delta} \end{array} \right]_{L_0Md(f+1)mr\times L_1(M+1)pr},$$

$$B_{17} = I_M \otimes I_L \otimes G^* \otimes I_{mr},\ B_{18} = I_{Md} \otimes (\tilde{E} \oplus U) \otimes I_r,$$

$$B_{21} = \left[\begin{array}{c} \tilde{U}^0 + \tilde{V}^0 \\ O \end{array} \right]_{C_0\times L_1(M+1)pr}.$$

Here

$$Q = \left[\begin{array}{c} diag(\tilde{\boldsymbol{\gamma}}, I_{L-1} \otimes \hat{I}_c) \\ O \end{array} \right]_{L_1\times L_0} \otimes \boldsymbol{\delta} \otimes \tilde{\boldsymbol{\eta}},\ Q^* = \left[\begin{array}{c} \bar{I}_m \\ \mathbf{e}_{L-1} \otimes \tilde{\mathbf{e}}_c \\ \mathbf{e}_{c+1} \otimes \tilde{\boldsymbol{\eta}} \otimes I_p \end{array} \right]_{L_1p\times(f+1)p},$$

$$\tilde{U}^0 = I_{M+1} \otimes \mathbf{e}_{L_1} \otimes U^0 \otimes \boldsymbol{\alpha} \otimes I_r\ ,\ \tilde{V}^0 = \mathbf{e}_{M+1} \otimes I_{L_1} \otimes I_p \otimes V^0 \otimes \boldsymbol{\beta}\ ,$$

$$\tilde{\Delta}_q = \left[\begin{array}{c} \mathbf{e}_f \otimes I_r \\ \mathbf{e}_r \otimes \boldsymbol{\beta} \end{array} \right],\ \Delta^0 = \left[\begin{array}{c} \mathbf{e}_f \otimes \tilde{\boldsymbol{\alpha}} \otimes I_r \\ \mathbf{e}_r \otimes \tilde{\boldsymbol{\alpha}} \otimes \boldsymbol{\beta} \end{array} \right],\ \tilde{\Delta} = \left[\begin{array}{c} \mathbf{e}_f \otimes (I_m^*)' \otimes I_r \\ \mathbf{e}_{mr} \otimes \boldsymbol{\alpha} \otimes \boldsymbol{\beta} \end{array} \right].$$

3 Steady state analysis

This section presents the steady-state analysis of the queueing model under consideration. We first derive the stability condition required for the system to operate in equilibrium.

3.1 Stability condition

Consider the generator matrix $B_0 + B_1 + B_2$. Let $\boldsymbol{\nu}$ be its the steady-state probability vector. Then ν and the generator satisfy the equations $\boldsymbol{\nu}(B_0 + B_1 + B_2) = 0$ and $\boldsymbol{\nu}\mathbf{e} = 1$.

Theorem 3.1: The Markov chain $\{P(\tau), \tau \geq 0\}$ is stable if and only if

$$\boldsymbol{\nu}B_0\mathbf{e} < \boldsymbol{\nu}B_2\mathbf{e}. \tag{2}$$

Proof: The system is stable if and only if the downward drift from any given level is greater than the upward drift to the next higher level. See Neuts [18].

The stability condition in equation (2) is equivalent to $\rho < 1$, where the traffic intensity ρ is given by

$$\rho = \frac{\boldsymbol{\nu}B_0\mathbf{e}}{\boldsymbol{\nu}B_2\mathbf{e}}. \tag{3}$$

The above condition is validated through numerical analysis in Section 4.

3.2 Steady state probability vector

Let $\boldsymbol{q} = (\boldsymbol{q}_0, \boldsymbol{q}_1, \boldsymbol{q}_2, ...)$ denotes the steady state probability vector of the Markov chain $\{P(\tau), \tau \geq 0\}$, where $\boldsymbol{q}_n$ represents the probability vector corresponding to level n.

Note that $\boldsymbol{q}_0$ is a scalar, $\boldsymbol{q}_1 = (\boldsymbol{q}_{10}, \boldsymbol{q}_{1\tilde{0}}, \boldsymbol{q}_{11})$ and $\boldsymbol{q}_n = (\boldsymbol{q}_{n0}, \boldsymbol{q}_{n1}, \boldsymbol{q}_{n2}, \boldsymbol{q}_{n3})$, for $n \geq 2$. The condition $\boldsymbol{q}Z = 0$ with $\boldsymbol{q}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector of appropriate dimension is satisfied by the vector $\boldsymbol{q}$. When the stability condition is satisfied, the sub vectors of $\boldsymbol{q}$ can be obtained by solving the corresponding system of balance equations

$$\boldsymbol{q}_n = \boldsymbol{q}_2 R^{n-2}, n \geq 3 \tag{4}$$

where R is the minimal non-negative solution of the matrix equation

$$R^2B_2 + RB_1 + B_0 = 0. \tag{5}$$

Once the matrix R is determined, the sub vectors $\boldsymbol{q}_0, \boldsymbol{q}_1$ and $\boldsymbol{q}_2$ can be obtained using the matrix-geometric equation

$$\begin{bmatrix} \boldsymbol{q}_0 & \boldsymbol{q}_1 & \boldsymbol{q}_2 \end{bmatrix} \begin{bmatrix} -\lambda & A_1 & \\ A_2 & A_3 & A_4 \\ & A_5 & B_1 + RB_2 \end{bmatrix} = 0, \tag{6}$$

together with the normalizing condition

$$\boldsymbol{q}_0 + \boldsymbol{q}_1\mathbf{e} + \boldsymbol{q}_2(I - R)^{-1}\mathbf{e} = 1. \tag{7}$$

3.3 Performance measures

Some important system performance measures are listed below, together with their respective computational formulae, in order to elucidate the qualitative behavior of the model. To facilitate this, the vectors $\boldsymbol{q}_n$, $n \geq 1$ are partitioned as

$$\boldsymbol{q}_1 = (\boldsymbol{q}_{10}, \boldsymbol{q}_{1\tilde{0}}, \boldsymbol{q}_{11})$$

and

$$\boldsymbol{q}_n = (\boldsymbol{q}_{n0}, \boldsymbol{q}_{n1}, \boldsymbol{q}_{n2}, \boldsymbol{q}_{n3}), n \geq 2.$$

Observe that $\boldsymbol{q}_0$ is a scalar, $\boldsymbol{q}_{10}$, $\boldsymbol{q}_{1\tilde{0}}$, $\boldsymbol{q}_{11}$, $\boldsymbol{q}_{n0}$, $\boldsymbol{q}_{n1}$, $\boldsymbol{q}_{n2}$ and $\boldsymbol{q}_{n3}$ are vectors of dimensions L_1p, $(M+1)r$, $Mrd(f+1)$, $(M+1)L_1pr$, $Md(f+1)r$, $L_0Md(f+1)mr$ and $M(f+1)mr$, respectively.

(1) Mean number of customers in the system

$$N_S = \sum_{n=1}^{\infty} n\boldsymbol{q}_n\mathbf{e}. \tag{8}$$

(2) Mean number of customers in the queue

$$N_Q = \sum_{n=2}^{\infty}(n-1)\boldsymbol{q}_{n1}\mathbf{e} + \sum_{n=3}^{\infty}(n-2)(\boldsymbol{q}_{n0}\mathbf{e} + \boldsymbol{q}_{n2}\mathbf{e} + \boldsymbol{q}_{n3}\mathbf{e}). \tag{9}$$

(3) Interruption's effective rate

$$\rho_i = \theta \sum_{n=2}^{\infty} \sum_{j=0}^{M-1} \sum_{\tau_1=1}^{m} \boldsymbol{q}_{n0j0\tau_1}\mathbf{e} + \theta \sum_{n=2}^{\infty} \sum_{j=0}^{M-1} \sum_{k=1}^{L-1} \sum_{j_1=1}^{c} \sum_{\tau_1=1}^{m} \boldsymbol{q}_{n0jkj_1\tau_1}\mathbf{e}. \quad (10)$$

(4) Consultation's effective rate

$$\rho_c = \theta \sum_{j=0}^{M-1} \boldsymbol{q}_{1\tilde{0}j}\mathbf{e} + \theta \sum_{n=2}^{\infty} \sum_{j=0}^{M-1} \boldsymbol{q}_{n0j}\mathbf{e}. \quad (11)$$

(5) Probability the main server is in idle state

$$\psi_m = \boldsymbol{q}_0\mathbf{e} + \boldsymbol{q}_{1\tilde{0}}\mathbf{e}. \quad (12)$$

(6) Probability that the regular server is in idle state

$$\psi_r = \boldsymbol{q}_0\mathbf{e} + \boldsymbol{q}_{10}\mathbf{e}. \quad (13)$$

(7) Probability that the main server is serving a customer

$$\zeta_m = \boldsymbol{q}_{10}\mathbf{e} + \sum_{n=2}^{\infty} \boldsymbol{q}_{n0}\mathbf{e} + \sum_{n=2}^{\infty} \boldsymbol{q}_{n3}\mathbf{e}. \quad (14)$$

(8) Probability that the regular server is serving a customer

$$\zeta_r = \boldsymbol{q}_{1\tilde{0}}\mathbf{e} + \sum_{n=2}^{\infty} \boldsymbol{q}_{n0}\mathbf{e}. \quad (15)$$

(9) Probability that the main server continues to be interrupted

$$\mu = \sum_{n=2}^{\infty} \boldsymbol{q}_{n2}\mathbf{e}. \quad (16)$$

(10) Probability that the regular server is receiving consultation

$$\nu_c = \sum_{n=1}^{\infty} \boldsymbol{q}_{n1}\mathbf{e} + \sum_{n=2}^{\infty} \boldsymbol{q}_{n2}\mathbf{e}. \quad (17)$$

(11) Probability that the regular server is waiting to receive consultation

$$\nu_w = \sum_{n=2}^{\infty} \boldsymbol{q}_{n3}\mathbf{e}. \quad (18)$$

4 Numerical results

Let us assume
$M = 3$, $L = 3$

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix}, U = \begin{bmatrix} -12 & 3 & 1 & 2 \\ 3 & -15 & 1 & 2 \\ 0 & 0 & -5 & 1 \\ 0 & 0 & 2 & -7 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, V = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}, \boldsymbol{\delta} = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, D = \begin{bmatrix} -6 & 4 \\ 3 & -4 \end{bmatrix},$$

$$\boldsymbol{\eta} = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix}, \boldsymbol{\gamma} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}, G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}.$$

These matrices, vectors, and values are specified so that ρ remains less than 1.

Table 1: Effect of θ on various performance measures

$\lambda = 4$

θ	1	1.5	2	2.5	3
ρ	0.5911	0.6910	0.7835	0.8694	0.9492
N_S	2.4103	3.2431	3.9898	4.6096	5.1017
N_Q	1.4926	2.2582	2.9576	3.5459	4.0178
ρ_i	0.0759	0.1139	0.1472	0.1748	0.1973
ρ_c	0.2585	0.3685	0.4600	0.5339	0.5929
ψ_m	0.4021	0.3503	0.3102	0.2800	0.2572
ψ_r	0.5487	0.4798	0.4265	0.3861	0.3558
ζ_m	0.4355	0.4193	0.4021	0.3853	0.3697
ζ_r	0.2591	0.2471	0.2327	0.2175	0.2030
μ	0.0320	0.0486	0.0636	0.0765	0.0872
ν_c	0.1613	0.2271	0.2806	0.3228	0.3559
ν_w	0.0298	0.0427	0.0533	0.0616	0.0682

Table 2: Effect of λ on various performance measures

$\theta = 3$

λ	3	3.5	4	4.5	5
ρ	0.5877	0.6856	0.7835	0.8815	0.9794
N_S	2.0385	2.9399	3.9898	5.0653	6.0266
N_Q	1.2498	2.0195	2.9576	3.9487	4.8572
ρ_i	0.0927	0.1206	0.1472	0.1699	0.1871
ρ_c	0.3242	0.3964	0.4600	0.5105	0.5454
ψ_m	0.4510	0.3757	0.3102	0.2553	0.2103
ψ_r	0.5977	0.508	0.4265	0.3558	0.2965
ζ_m	0.3462	0.3777	0.4021	0.4187	0.4268
ζ_r	0.1641	0.2006	0.2327	0.2581	0.2756
μ	0.0405	0.0525	0.0636	0.0727	0.0791
ν_c	0.2024	0.2448	0.2806	0.3070	0.3228
ν_w	0.0355	0.0448	0.0533	0.0602	0.0651

Referring to Table 1, as the rate of consultation θ increases, the traffic intensity ρ increases and hence there is an increase in ρ_i and ρ_c. So μ and ν_c will increase. As θ increases, consultation is more frequent, the upper bound of number of interruptions is reached rapidly or super clock may realise frequently and thus the main server is compelled to complete the service of the customer at him before further consultations. Thus the regular server has to wait more time to get consultation. So ν_w also increases. Since μ, ν_c and ν_w increase, the customers stay in the system and in the queue for a longer time and thus there is an increase in N_S and N_Q. This makes a decrease in ψ_m and ψ_r. Since main server is forced to spend more time in consultation, it gets less time to serve customers. So ζ_m and ζ_r decreases.

From Table 2, we can see that as the arrival rate λ increases, there is an increase in the traffic intensity ρ. So the system is fed with more and more customers and this results in an accumulation of customers. So N_S and N_Q will increase. Thus ρ_i and ρ_c also increase. Therefore there is a hike in μ and ν_c. Thus ν_w also increases. As λ increases, there are more customers in the queue and thus the servers have to spend longer time in service. So ζ_m and ζ_r increase. This in turn make a decrease in ψ_m and ψ_r.

Concluding remarks and suggestions for further study

In this paper we analyse a two-server queueing model with consultation by main server to the regular server. The service of the customer at the main server is interrupted when a request for consultation by the regular server arises. Some phases of the service at the main server are kept protected from interruption. Even though quality of service is enhanced by consultations, it is not fair to spend a lot of time of the customer at the main server in interrupted state. So we introduce some restrictions in permitting consultation to the regular server if the main server is serving a customer. The interruptions to a customer at the main server are controlled by upper bounds of interruptions and consultations and a super clock. We establish stability condition and provide numerical illustrations. It will be interesting to study models without protected phases of service at the main server. Then interruption may be permitted at any phase.

References

[1] *Chakravarthy, S. R.* . (2014). A multi-server queueing model with server consultations, European Journal of Operational Research, 233(3), 625-639.

[2] Klimenok, V., Dudin, A., Vishnevsky, V. (2020). Priority multi-server queueing system with heterogeneous customers. Mathematics , 8, 1501.

[3] Samouylov, K., Dudina, O., Dudin, A. (2023). Analysis of Multi-Server Queueing System with Flexible Priorities. Mathematics, 11, 1040.

[4] Krishnamoorthy, A., Divya, V. (2020). A Two-Server Queueing System with Processing of Service Items by a Server. In Applied Probability and Stochastic Processes; Springer Nature: Singapore, 307–333.

[5] Ayyappan, G., Archana, G. (2023). Analysis of MAP/PH1, PH2/2 Queueing Model with Working Breakdown, Repairs, Optional Service, and Balking. Appl. Appl. Math. Int. J. (AAM), 18, 1.

[6] Thekkiniyedath Resmi., K. Ravikumar. (2021). Three-Server Queue with Consultations by Main Server with a Buffer at the Main Server, Information Technologies and Mathematical Modelling, Queueing Theory and Applications. ITMM 2020. Communications in Computer and Information Science, 1391, 131–142.

[7] Resmi, T., Lakshmy, B., Krishnamoorthy, A. (2018). A Two-Server Queue with Mutual Consultations. J Indian Soc Probab Stat 19, 201–215.

[8] *Krishnamoorthy, A., Pramod, P.K. and Chakravarthy S. R.* . (2013). A note on characterizing service interruptions with phase type distribution, Stochastic Analysis and Applications, 31(4), 671-683.

[9] White H., Christie L. S. (1958). Queuing with Preemptive Priorities or with Breakdown, Operations Research, 6, 79–95.

[10] Gaver D. P., Jacobs P. A., Latouche G. (1984). Finite birth and death models in randomly changing environments, Advances in Applied Probability, 16(4), 715-731.

[11] Keilson J. (1962) . Queues subject to service interruptions, The Annals of Mathematical Statistics 33(4), 1314-1322.

[12] Ibe O. C., Trivedi K. S. (1960). Two queues with alternating service and server breakdown, Queueing Systems 7(3), 253-268.

[13] Avi-Itzhak B, Naor P. (1963). Some queueing problems with the service station subject to breakdowns, Oper. Res. 11(3), 303-320.

[14] Fiems D., Maertens T., Bruneel H. (2008). Queueing systems with different types of interruptions, Eur J Oper Res 188(3), 838-845.

[15] *Klimenok, V., Kim, C. S. and Kuznetsov,V..* (2006). A multi-server queue with negative customers and partial protection of service, Proceedings of 13^{th} International Conference on analytical and stochastic Modelling Techniques and applications (ASMTA 06), Bonn, Germany; Eds. K Al-Begain; 143-148, 28-31 May.

[16] *Klimenok, V. and Dudin, A. N..* (2012). A BMAP/PH/N queue with negative customers and partial protection of service, Communications in Statistics- Simulation and Computation, 41(7), 1062-1082.

[17] *Krishnamoorthy, A, Gopakumar,B. and Viswanath,C. N..* (2010). An M/Em/1 queue with protected and unprotected phases from interruptions, 5^{th} International conference on Queueing Theory and Network Applications, Beijing, China, July 24-26.

[18] *Neuts, M.F..* (1981). Matrix-geometric solutions in stochastic models, An Algorithmic Approach, The Johns Hopkins University Press, Baltimore.