Chapter 2

Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

In this chapter we study three two-server queueing models with consultations given by the main server to the regular server. The service of the customer at the main server is interrupted when he is being served by the main server at the time of request of the regular server for consultation. It is not fair to interrupt a customer at the main server infinitely many times or to receive infinitely many consultations, if he is at the
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

regular server, during his service. So we impose some upper bounds to control consultations and interruptions. In this aspect our model differs from that of Chakravarthy [6] in which a multi-server queueing system with consultations is discussed. There is no boundary on the number of interruptions to a customer at the main server and the regular server is free to get any number of consultations during the service of a customer. The main server gives immediate consultations to the regular servers. The request for consultation of the regular server is attended by the main server, even if there is a customer being served at the main server. Then that customer at the main server has to wait until the consultation is completed. At this stage the service of the customer at the main server is said to be interrupted. (So the word ‘interruption’ is associated with the customer at the main server when the main server is providing consultation to the regular server.) The service times at these servers follow independent phase type distributions.

We introduce upper bounds for interruptions, consultations; a super clock to get an ‘approximate measure’ of the total duration of interruption. We say it is an ‘approximate measure’ because the total interruption time can be greater than the duration of super clock if super clock expires during interruption and interruption continues to be completed. If super clock does not expire, duration of super clock is the total interruption time during the service of a customer at the main server. (But in model 2, duration of super clock is strictly equal to the total duration of interruption since the interruption will be removed and the service of the customer at the main server will be continued as soon as the super clock expires.) A maximum of $M$ interruptions are allowed to a customer at the main
Chapter 2: Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption server. No further interruptions are allowed to that customer after $M$ interruptions. If the regular server needs further consultation at this time, he/she has to wait until the service at the main server is completed. After the service completion of the interrupted customer, the main server will immediately attend the consultation before taking a new customer from the queue for service. The maximum number of consultations possible to the regular server during the service of a particular customer is $K$.

If the super clock expires during consultation with one interrupted customer at the main server, then the present consultation is permitted to complete and no more interruption is allowed to befall to that particular customer at the main server. At this stage, if the regular server again needs a consultation, he has to wait until the completion of the service at the main server. After finishing the service, the main server will immediately attend the consultation. At this time, no customer is interrupted at the main server and so no super clock is present here.

So the main server offers consultation in the following manner:

(i) If the main server is idle, then the request for consultation will be attended immediately.

(ii) If the number of interruptions already befell to the customer at the main server is less than $M$ and the super clock has not expired, then also the consultation will be provided immediately.

(iii) If either the customer at the main server has interrupted $M$ times or the super clock has expired, then the regular server has to wait until the completion of the service of the present customer at the main server.
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

(iv) The regular server needs a further consultation only when the number of consultations already taken by him for the same customer is strictly less than $K$.

In model 3, we assume that interruption is not allowed to a customer at the main server if the service is in any one of the protected phases of service (which may be so costly to afford an interruption at these phases). So the consultation to the regular server will be denied if the main server is at the protected phases. All other assumptions are same as those in model 1.

A comparison of the three models is provided towards the end of this chapter.

2.1 Description of model 1

Here we consider a service system equipped with one main server and one regular server to which customers arrive according to a MAP with representation $(L_0, L_1)$, where $L_0$ and $L_1$ are matrices of order $r$. An arriving customer enters into 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 follow independent phase type distributions with representations $(\alpha, T)$ and $(\beta, U)$ with number of phases $a$ and $b$, respectively. Write $T^0 = -Te$ and $U^0 = -Ue$ where $e$ is a column vector of 1's of appropriate order. The main server offers consultation to the regular server whenever it is needed. Requirement of consultation is a Poisson process with rate $\theta$. The request for consultation by the regular
server is attended by the main server. If there is a customer being served at the main server, that customer at the main server has to wait until the consultation is completed. At this stage the service of the customer at the main server is said to be interrupted. (So the word ‘interruption’ is associated with the customer at the main server when the main server is providing consultation to the regular server.) At most $M$ interruptions are allowed to a customer at the main server. No further interruption is permitted to that customer after $M$ interruptions. If the regular server needs consultation at this time, he/she has to wait until the service of the customer at the main server is completed. Once his service is completed, the main server will attend the consultation before taking a new customer from the queue for service. The maximum number of consultations possible to the regular server during the service of a particular customer is set as $K$. This is to ensure that customers in service at the regular server do not get too impatient to leave the system.

The duration of super clock, threshold clock and consultation clock follow independent phase type distributions with representations $(\gamma, G)$, $(\eta, E)$, $(\delta, D)$ with number of phases $c, d$ and $f$, respectively. We have $G^0 = -Ge$, $E^0 = -Ee$, and $D^0 = -De$, respectively.

The threshold clock determines the restart or resumption of services at both the servers. Every time this clock starts anew when the regular server temporarily stops his service for consultation. If the regular server is waiting to get consultation, this clock starts ticking and continue during the time of consultation after the service at the main server. On the other hand, if regular server gets consultation immediately, the consultation process and threshold clock start together. If the threshold clock expires before the consultation process, then the services at both the servers are to
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption be restarted. Otherwise the services will be resumed at the phases where they are interrupted.

A super clock is set to determine whether further interruption to a customer at the main server is to be allowed or not. This clock starts at the epoch of the first interruption of a particular customer at the main server and is freezed at the moment the consultation is over. When the next interruption to the same customer strikes, the super clock starts from the earlier position where it stopped ticking and so on. If the super clock expires during consultation with one interrupted customer at the main server, then the present consultation is permitted to continue until completion and no more interruption is allowed to befall to that particular customer at the main server. At this stage, if the regular server again needs a consultation, he has to wait until the completion of the service at the main server. After completing the service, the main server will immediately attend the consultation. Since there is no interrupted customer at the main server, super clock is in the ’off’ mode (indicating that service is not interrupted at the main server.)

So the main server offers consultation to the regular server in the following manner:

(i) If the main server is idle, then the request for consultation will be attended immediately.

(ii) If the number of interruptions already befell to the customer at the main server is less than $M$ and the super clock has not expired, then also the consultation will be provided immediately.
(iii) If either the customer at the main server has interrupted $M$ times or the super clock has expired, then the regular server has to wait until the completion of the service of the present customer at the main server.

(iv) The regular server needs further consultation only when the number of consultations already taken by him for a particular customer is strictly less than $K$.

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

- $M_0 = M(c + 1)$ and $M_1 = M_0 + 1$
- $\tilde{\alpha} = e_{M_1}(1) \otimes \alpha$
- $\tilde{\gamma} = (\gamma, 0), \tilde{\eta} = (\eta, 0)$
- $\tilde{G} = \begin{bmatrix} G & G^0 \\ 0 & 0 \end{bmatrix}$ and $\tilde{E} = \begin{bmatrix} E & E^0 \\ 0 & 0 \end{bmatrix}$
- $D^* = D \oplus \tilde{E}$ and $G^* = \tilde{G} \oplus D^*$
- $\hat{I} = \begin{bmatrix} 0 & I_{M(c+1)} \end{bmatrix}$

Consider the queueing model

$X = \{X(t), t \geq 0\}$,

where $X(t) = \{N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t), U(t)\}$.
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

The variables are defined as follows:

\[ N(t) \] the number of customers in the system

\[ B_1(t) \] number of consultations already enjoyed by the regular server during the service of a particular customer

\[ B_2(t) \] number of interruptions already befell to a customer at the main server

\[ S_1(t) \] phase of the super clock

\[ S_2(t) \] phase of the consultation process

\[ S_3(t) \] phase of the threshold clock

\[ J_1(t) \] phase of the main server

\[ J_2(t) \] phase of the regular server

\[ U(t) \] phase of the arrival process

Here \( S(t) \) denotes the status of the servers at time \( t \) such that

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

Note that \( B_2(t) \) 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 $S_1(t)$ when $B_2(t) = 0$. 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.

\[
\{X(t), t \geq 0\} \text{ is a Continuous Time Markov Chain with state space}
\]

\[
\Psi = \bigcup_{i=0}^{\infty} \psi(i).
\]

The terms $\psi(i)$’s are defined as

\[
\psi(0) = \{(0, u)\},
\]

\[
\psi(1) = \psi(1, 0) \cup \psi(1, \bar{0}) \cup \psi(1, 1) \text{ and}
\]

\[
\psi(i) = \psi(i, 0) \cup \psi(i, 1) \cup \psi(i, 2) \cup \psi(i, 3), \text{ for } i \geq 2,
\]

where

\[
\psi(1, 0) = \{(1, 0, 0, t_1, u)\} \cup \{(1, 0, k, l_1, t_1, u) : 1 \leq k \leq M\}
\]

\[
\psi(1, \bar{0}) = \psi\{(1, \bar{0}, j, t_2, u) : 0 \leq j \leq K\}
\]

\[
\psi(1, 1) = \{(1, 1, j, l_2, l_3, t_2, u) : 0 \leq j \leq K - 1\}
\]

\[
\psi(i, 0) = \{(i, 0, j, 0, t_1, t_2, u) \cup (i, 0, j, k, l_1, t_1, t_2, u) : 0 \leq j \leq K, 1 \leq
\]
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption $\{k \leq M\}$

\[
\psi(i, 1) = \{(i, 1, j, l_2, l_3, t_2, u) : 0 \leq j \leq K - 1\}
\]

\[
\psi(i, 2) = \{(i, 2, j, k, l_1, l_2, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1, 0 \leq k \leq M - 1\}
\]

\[
\psi(i, 3) = \{(i, 3, j, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1\}
\]

with $0 \leq l_1 \leq c, 1 \leq l_2 \leq d, 0 \leq l_3 \leq f, 1 \leq t_1 \leq a, 1 \leq t_2 \leq b, 1 \leq u \leq r$ and for $i \geq 2$.

The infinitesimal generator $Q$ is given by

\[
Q = \begin{bmatrix}
L_0 & B_1 & & & \\
B_2 & B_3 & B_4 & & \\
& B_5 & A_1 & A_0 & \\
& & A_2 & A_1 & A_0 \\
& & & \ddots & \ddots & \ddots
\end{bmatrix}
\]

(2.1)

where $B_1 = \begin{bmatrix} \alpha & 0 \end{bmatrix} \otimes L_1, B_2 = \begin{bmatrix} e_{M_i} \otimes T^0 \\ e_{K+1} \otimes U^0 \\ 0 \end{bmatrix} \otimes I_r$, $B_3 = \begin{bmatrix}
I_{M_i} \otimes T & O & O \\
O & B_{31} & B_{32} \\
O & B_{33} & I_K \otimes D^* \otimes I_b
\end{bmatrix} \oplus L_0.$
2.1. Description of model 1

\[ B_4 = \begin{bmatrix} B_{41} & B_{42} & O \end{bmatrix} \otimes L_1, \quad B_5 = \begin{bmatrix} B_{51} & B_{52} & B_{53} \end{bmatrix} \otimes I_r, \]

\[ A_0 = I \otimes L_1, \quad A_1 = \begin{bmatrix} A_{11} & O & A_{12} & A_{13} \\ A_{14} & A_{15} & O & O \\ A_{16} & O & A_{17} & O \\ O & O & O & A_{18} \end{bmatrix} \oplus L_0, \]

\[ A_2 = \begin{bmatrix} A_{21} & B_{53} & O \end{bmatrix} \otimes I_r. \]

Here \( A_0, A_1 \) and \( A_2 \) are square matrices of order \( C_0 \), \( B_3 \) is a square matrix of order \( C_1 \) and \( B_1, B_2, B_4, B_5 \) are matrices of orders \( r \times C_1, C_1 \times r, C_1 \times C_0 \) and \( C_0 \times C_1 \), respectively,

where

\[ C_0 = [M_1(K + 1)ab + Kd(f + 1)b + M_bKd(f + 1)ab + K(f + 1)ab]r, \]

and

\[ C_1 = [M_1a + (K + 1)b + Kbd(f + 1)]r. \]

We have

\[ B_{31} = \begin{bmatrix} I_K \otimes (U - \theta I) & O \\ O & U \end{bmatrix}_{(K+1)b \times (K+1)b}, \]

\[ B_{32} = \theta \begin{bmatrix} I_K \otimes \delta \otimes \tilde{\eta} \\ O \end{bmatrix}_{(K+1) \times Kd(f+1)} \otimes I_b, \]

\[ B_{33} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \tilde{\Delta}_b \end{bmatrix}_{Kd(f+1)b \times (K+1)b}. \]
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

\[ B_{41} = \begin{bmatrix} e_{K+1} \otimes I_{M_1} \otimes I_a \otimes \beta \\ I_{K+1} \otimes \alpha \otimes I_b \\ 0 \end{bmatrix}, \quad B_{42} = \begin{bmatrix} O \\ I \\ C_1 \times (K+1)M_1ab \end{bmatrix}, \]

\[ B_{51} = \begin{bmatrix} e_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \\ 0 \end{bmatrix}, \quad C_0 \times M_0a, \]

\[ B_{52} = \begin{bmatrix} I_{K+1} \otimes e_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}, \quad C_0 \times (K+1)b, \quad B_{53} = \begin{bmatrix} O \\ F \end{bmatrix}, \quad C_0 \times Kbd(f+1), \]

\[ A_{11} = \begin{bmatrix} I_K \otimes I_{M_1} \otimes (T \otimes U - \theta I) \\ O \\ I_{M_1} \otimes (T \otimes U) \end{bmatrix}, \quad M_1(K+1)ab \times M_1(K+1)ab, \]

\[ A_{12} = \theta \begin{bmatrix} I_K \otimes P \\ O \end{bmatrix}, \quad \otimes I_{ab}, \quad M_1(K+1) \times M_0(d(f+1)), \]

\[ A_{13} = \theta \begin{bmatrix} I_K \otimes P^* \\ O \end{bmatrix}, \quad \otimes \tilde{\eta} \otimes I_{ab}, \quad M_1(K+1) \times K, \]

\[ A_{14} = \begin{bmatrix} O \\ I_K \otimes D^0 \otimes \Delta^0 \end{bmatrix}, \quad Kd(f+1)b \times M_1(K+1)ab, \]

\[ A_{15} = I_K \otimes D^* \otimes I_b, \]

\[ A_{16} = \begin{bmatrix} O \\ I_K \otimes \hat{I} \otimes D^0 \otimes \hat{\Delta} \end{bmatrix}, \quad M_0Kd(f+1)ab \times M_1(K+1)ab, \]

\[ A_{17} = I_K \otimes I_M \otimes G^* \otimes I_{ab}, \quad A_{18} = I_K \otimes (\hat{E} \otimes T) \otimes I_b, \]

\[ A_{21} = \begin{bmatrix} \hat{F} \\ O \end{bmatrix}, \quad C_0 \times M_1(K+1)ab. \]
2.2 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study by first establishing the stability condition of the queueing system.

2.2.1 Stability condition

Let \( \pi \) denote the steady-state probability vector of the generator \( A_0 + A_1 + A_2 \). That is, \( \pi(A_0 + A_1 + A_2) = 0; \pi e = 1. \)
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption.

The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

\[ \pi A_0 e < \pi A_2 e. \]  \hspace{1cm} (2.2)

That is, the rate of drift to the left has to be higher than that to the right. The vector \( \pi \) cannot be obtained explicitly in terms of the parameters of the model, and hence the stability condition is known only implicitly. If we partition the vector \( \pi \) as

\[ \pi = (\pi_0, \pi_1, \pi_2, \pi_3) \]

and then using the structure of the matrices \( A_0 \) and \( A_2 \), equation (2.2) is given by

\[ \lambda < \pi_0 \bar{F} e + \pi_3 F e. \]  \hspace{1cm} (2.3)

For future reference, we define the traffic intensity \( \rho_1 \) as

\[ \rho_1 = \frac{\pi A_0 e}{\pi A_2 e}. \]  \hspace{1cm} (2.4)

Note that the stability condition in (2.2) is equivalent to \( \rho_1 < 1 \). We will discuss the impact of the input parameters of the model on the traffic intensity in Section 2.3.

2.2.2 Steady state probability vector

Since the model studied as a QBD process, its steady-state distribution has a matrix-geometric solution under the stability condition. Assume that
the stability condition holds. Let $\mathbf{x}$ denote the steady-state probability vector of the generator $Q$ given in (2.1). That is,

$$\mathbf{x} Q = 0; \mathbf{x} e = 1.$$  \hspace{1cm} (2.5)

Partitioning $\mathbf{x}$ as

$$\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots),$$ \hspace{1cm} (2.6)

we see that, under the assumption that the stability condition (2.2) holds, the sub-vectors $\mathbf{x}_i$, $i \geq 3$ are obtained as (see, Neuts [44])

$$\mathbf{x}_j = \mathbf{x}_2 R^{j-2}, j \geq 3,$$ \hspace{1cm} (2.7)

where $R$ is the minimal non-negative solution to the matrix quadratic equation:

$$R^2 A_2 + RA_1 + A_0 = 0.$$ \hspace{1cm} (2.8)

$\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$ are obtained using the boundary equations

$$\mathbf{x}_0 L_0 + \mathbf{x}_1 B_2 = 0$$

$$\mathbf{x}_0 B_1 + \mathbf{x}_1 B_3 + \mathbf{x}_2 B_5 = 0$$ \hspace{1cm} (2.9)

$$\mathbf{x}_1 B_4 + \mathbf{x}_2 (A_1 + RA_2) = 0$$

The normalizing condition of (2.5) results in

$$\mathbf{x}_0 e + \mathbf{x}_1 e + \mathbf{x}_2 (I - R)^{-1} e = 1.$$ \hspace{1cm} (2.10)
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption.

Once the rate matrix $R$ is obtained, the vector $x$ can be computed by exploiting the special structure of the coefficient matrices. We can use the iterative formulas (see Neuts [44])

$$R_n = -A_0(A_1 + R_{n-1}A_2)^{-1}, \text{ for } n \geq 1,$$

with an initial value $R_0$, which converges to $R$ if $sp(R) < 1$.

2.2.3 Expected waiting time in queue

For computing expected waiting time in queue of a particular customer who joins as the $m^{th}$ customer, where $m > 0$, in the queue, we consider the Markov process

$$Z(t) = \{(\tilde{N}(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t)) : t \geq 0\}$$

where

$\tilde{N}(t)$ is the rank of the customer and all other variables defined as earlier. The rank $\tilde{N}(t)$ of the customer is assumed to be $i$ if he is the $i^{th}$ customer in the queue at time $t$. His rank may decrease to 1 as the customers ahead of him leave the system either after completing their services (if $S(t) = 0$) or completing the consultation (if $S(t) = 1$). Since the customers who arrive after the tagged customer cannot change his rank, level-changing transitions in $Z(t)$ can only take place to one side of the diagonal. The absorbing state $\Delta_2$ denote the tagged customer is selected for service. Thus the infinitesimal generator $\tilde{V}$ of the process $Z(t)$ takes the form

$$\tilde{V} = \begin{bmatrix} V & V^0 \\ 0 & 0 \end{bmatrix},$$
2.2. Steady state analysis

where

\[
V = \begin{bmatrix}
\tilde{A}_1 & \tilde{A}_2 \\
\tilde{A}_1 & \tilde{A}_2 \\
\vdots & \vdots \\
\tilde{A}_1 & \tilde{A}_2 \\
\tilde{A}_1 & \tilde{A}_1
\end{bmatrix}, \quad \quad V^0 = \begin{bmatrix}
0 \\
\mathbf{e}_{M_1} \otimes (T^0 \oplus U^0) \\
0 \\
T^0 \otimes \mathbf{e}_b
\end{bmatrix},
\]

with \( \tilde{A}_1 = A_1^* - U_2 \) and \( \tilde{A}_2 = A_2^* + U_2 \), where \( A_1^* \) and \( A_2^* \) are obtained from \( A_1 \) and \( A_2 \) if they are written as \( A_1 = A_1^* \oplus L_0 \) and \( A_2 = A_2^* \otimes I_r \). Here \( U_2 = \begin{bmatrix}
O & O \\
A_{14} & O \\
O & O
\end{bmatrix} \).

Now, the waiting time \( V \) of a customer, who joins the queue as the \( j^{th} \) customer is the time until absorption of the Markov chain \( V(t) \). Thus the expected waiting time of this particular customer is given by the column vector,

\[
E^{(j)}_V = [-\tilde{A}_1^{-1}(I + \sum_{i=1}^{j-1}(-\tilde{A}_2\tilde{A}_1^{-1})^i)]\mathbf{e}.
\]

The second moment of waiting time of the tagged customer is given by the column vector \( E^2_V \) which is the first block of the matrix \( 2(-\tilde{V})^{-2}\mathbf{e} \). Hence the expected waiting time of a general customer in the queue is,

\[
V_L = \sum_{j=1}^{\infty} x(j) E^{(j)}_V.
\]
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption.

The second moment of $V$ is

$$V_L^{(2)} = \sum_{j=1}^{\infty} x(j)E_{V^2}^j.$$ 

### 2.2.4 Performance measures

In this section we list a number of key system performance measures to bring out the qualitative aspects of the model under study. These are listed below along with their formulae for computation. Towards this end, we further partition the vectors $x_i, i \geq 1$ as

$$x_1 = (x_{10}, x_{10}, x_{11})$$

and

$$x_i = (x_{i0}, x_{i1}, x_{i2}, x_{i3}), i \geq 2.$$ 

Note that $x_0, x_{10}, x_{10}, x_{11}, x_{i0}, x_{i1}, x_{i2}$ and $x_{i3}$ are vectors of dimensions $r, M_1 ar, (K + 1)br, Kbd(f + 1)r, M_1(K + 1)abr, Kd(f + 1)br$, $M_0Kd(f + 1)abr, K(f + 1)abr$ respectively.

1. Expected number of customers in the system

$$ES = \sum_{i=1}^{\infty} ix_i e. \quad (2.11)$$

2. Expected number of customers in the queue

$$EQ = \sum_{i=2}^{\infty} (i - 1)x_{i1}e + \sum_{i=3}^{\infty} (i - 2)(x_{i0}e + x_{i2}e + x_{i3}e). \quad (2.12)$$
(3) Effective rate of consultation

\[
EC_0 = \theta \sum_{j=0}^{K-1} x_{10j} e + \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} x_{i0j} e. \tag{2.13}
\]

(4) Effective rate of interruption

\[
EI = \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} x_{i0j0} e + \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=1}^{M-1} \sum_{l_1=1}^{c} x_{i0jkl_1} e \tag{2.14}
\]

(5) Fraction of time the main server is idle

\[
F_{mi} = x_{0} e + x_{10} e. \tag{2.15}
\]

(6) Fraction of time the regular server is idle

\[
F_{ri} = x_{0} e + x_{10} e. \tag{2.16}
\]

(7) Fraction of time the main server is busy serving a customer

\[
F_{mb} = x_{10} e + \sum_{i=2}^{\infty} x_{i0} e + \sum_{i=2}^{\infty} x_{i3} e. \tag{2.17}
\]

(8) Fraction of time the regular server is busy serving a customer

\[
F_{rb} = x_{10} e + \sum_{i=2}^{\infty} x_{i0} e. \tag{2.18}
\]
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption.

(9) Fraction of time regular server is getting consultation

\[ F_{rc} = \sum_{i=1}^{\infty} x_{i1} e + \sum_{i=2}^{\infty} x_{i2} e. \] (2.19)

(10) Fraction of time regular server is waiting to get consultation

\[ F_{wc} = \sum_{i=2}^{\infty} x_{i3} e. \] (2.20)

(11) Fraction of time main server remains interrupted

\[ F_{\min} = \sum_{i=2}^{\infty} x_{i3} e. \] (2.21)

(12) Rate at which interruption completion takes place before threshold is realised

\[ R_{cb}^{i} = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \sum_{l_2=1}^{d} \sum_{l_3=1}^{f} D_{l_2}^{0} x_{i2jkl_1l_2l_3} e. \] (2.22)

(13) Rate at which interruption completion takes place after threshold is realised

\[ R_{cb}^{ia} = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{c} \sum_{l_2=1}^{d} \sum_{l_3=1}^{f} D_{l_2}^{0} x_{i2jkl_1l_2} e. \] (2.23)

(14) Rate at which consultation completion takes place before threshold
2.2. Steady state analysis

is realised

$$R^c_C b = \sum_{i=1}^{\infty} \sum_{j=0}^{K-1} \sum_{l_2=1}^{d} \sum_{l_3=1}^{f} D_{l_2}^0 x_{i1jvl_3} e + R^c_I b. \quad (2.24)$$

(15) Rate at which consultation completion takes place after the threshold is realised

$$R^c_C a = \sum_{i=1}^{\infty} \sum_{j=0}^{K-1} \sum_{l_2=1}^{d} D_{l_2}^0 x_{i1jvl_0} e + R^c_I a. \quad (2.25)$$

(16) Rate at which service completion at the main server takes place without any interruption

$$R^c_S wi = \sum_{t_1=1}^{a} T_{t_1}^0 x_{100t_1} e + \sum_{i=2}^{\infty} \sum_{j=0}^{K} \sum_{t_1=1}^{a} T_{t_1}^0 x_{i0jt_1} e. \quad (2.26)$$

(17) Rate at which service completion (with at least one interruption) at the main server takes place before super clock is realised

$$R^c_S b = \sum_{i=2}^{\infty} \sum_{j=0}^{K} \sum_{k=1}^{M} \sum_{c=1}^{c} \sum_{l_1=1}^{a} T_{l_1}^0 x_{i0jkl_1t_1} e + \sum_{j=0}^{\infty} \sum_{k=1}^{K} \sum_{l_1=1}^{M} \sum_{c=1}^{c} \sum_{l_1=1}^{a} T_{l_1}^0 x_{0jkl_1t_1} e. \quad (2.27)$$

(18) Rate at which service completion (with at least one interruption) at the main server takes place after super clock is realised

$$R^c_S a = \sum_{i=2}^{\infty} \sum_{j=0}^{K} \sum_{k=1}^{M} \sum_{c=1}^{a} T_{t_1}^0 x_{i0jkl_0t_1} e + \sum_{j=0}^{\infty} \sum_{k=1}^{K} \sum_{l_1=1}^{M} \sum_{c=1}^{a} T_{t_1}^0 x_{0jkl_0t_1} e. \quad (2.28)$$
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption.

\[(19)\] Rate at which service completion at the regular server takes place without any consultation

\[
R_{Sc}^{wc} = \sum_{t_2=1}^{b} U_{t_2}^0 x_{100t_2} e + \sum_{i=2}^{\infty} \sum_{t_1=1}^{a} \sum_{t_2=1}^{b} U_{t_2}^0 x_{i00t_1t_2} e
\]

\[
+ \sum_{i=2}^{\infty} \sum_{k=1}^{M} \sum_{c=0}^{a} \sum_{l_1=1}^{b} \sum_{l_2=1}^{b} U_{t_2}^0 x_{i00kl_1t_2} e. \quad (2.29)
\]

\[(20)\] Rate at which service completion (with at least one consultation) at the regular server takes place

\[
R_{Sc}^{c} = \sum_{j=1}^{K} \sum_{t_2=1}^{b} U_{t_2}^0 x_{10jt_2} e + \sum_{i=2}^{\infty} \sum_{j=1}^{K} \sum_{t_1=1}^{a} \sum_{t_2=1}^{b} U_{t_2}^0 x_{i0jt_1t_2} e
\]

\[
+ \sum_{i=2}^{\infty} \sum_{j=1}^{K} \sum_{k=1}^{M} \sum_{c=0}^{a} \sum_{l_1=1}^{b} \sum_{l_2=1}^{b} U_{t_2}^0 x_{i0jklt_1t_2} e. \quad (2.30)
\]

2.3 Numerical results

For the arrival process we consider the following five sets of matrices for $L_0$ and $L_1$.

(i) Erlang (ERA) $L_0 = \begin{bmatrix} -5 & 5 \\ -5 & 5 \\ -5 & \end{bmatrix}$, $L_1 = \begin{bmatrix} 5 \\ \end{bmatrix}$.

(ii) Exponential (EXA)
2.3. Numerical results

\[ L_0 = [-1], \quad L_1 = [1] \]

(iii) Hyper Exponential (HEA)

\[ L_0 = \begin{bmatrix} -10 & 0 \\ 0 & -1 \end{bmatrix}, \quad L_1 = \begin{bmatrix} 9 & 1 \\ 0.9 & 0.1 \end{bmatrix}. \]

(iv) MAP with negative correlation (MNA)

\[ L_0 = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{bmatrix}, \quad L_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0.02 & 1.98 & 445.995 \\ 0.38 & 4.505 & 445.995 \end{bmatrix}. \]

(v) MAP with positive correlation (MPA)

\[ L_0 = \begin{bmatrix} -2 & 2 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -450.5 \end{bmatrix}, \quad L_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1.98 & 0.02 & 4.505 \\ 0 & 0 & 445.995 \end{bmatrix}. \]

All these five MAP processes are normalized so as to have an arrival rate of 4. However, these are qualitatively different in that they have different variances and correlation structures. The first three arrival processes, namely ERA, EXA, and HEA, correspond to renewal processes and so the correlation is 0. The arrival process labelled MNA has correlated arrivals with correlation between two successive inter-arrival times given by \(-0.4889\) and the arrival process corresponding to the one labelled MPA has a positive correlation with value 0.4889. The ratio of the standard deviations of the inter-arrival times of these five arrival processes with respect to ERA are, respectively, 1, 2.2361, 5.0194, 3.1518, and 3.1518.

The purpose of this example to see how various performance measures behave under different scenario.
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

Let

\[ T = \begin{bmatrix} -9 & 2 \\ 3 & -8 \end{bmatrix}, \quad U = \begin{bmatrix} -12 & 5 \\ 6 & -10 \end{bmatrix}, \quad D = \begin{bmatrix} -6 & 3 \\ 4 & -4 \end{bmatrix}, \]

\[ E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix}, \quad G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}, \]

\[ \alpha = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}, \quad \beta = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \quad \delta = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \quad \eta = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, \]

\[ \gamma = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}, \quad K = 3, \quad M = 3. \]

We choose the above matrices, vectors and values so that the stability condition \( \rho_1 < 1 \) is not violated.

We look at the effect of varying \( \theta \) on the performance measures \( \rho_1, ES, EQ, EI \) and \( ECo \). From the table 2.1 we can see that as \( \theta \) increases the traffic intensity also increases. This results in a rapid accumulation of customers in system and in queue. Thus \( ES \) and \( EQ \) increase. The effective rates for interruption \( EI \) and for consultation \( ECo \) also increase as \( \theta \) increases.

### 2.4 Description of model 2

In model 1, the interruption is allowed to continue even when the super clock is saturated. In model 2, the interruption will be stopped at the moment the super clock realises and the service at the main server will be restarted or resumed according to the threshold clock. The regular server has to wait until the service completion at the main server to get the remaining consultation. After consultation, the regular server resumes or restarts the service in accordance with the threshold clock. Thus the total
2.4. Description of model 2

Table 2.1: Effect of $\theta$ on various performance measures

<table>
<thead>
<tr>
<th></th>
<th>$\theta$</th>
<th>ERA</th>
<th>EXA</th>
<th>HEA</th>
<th>MNA</th>
<th>MPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_1$</td>
<td>1.0</td>
<td>0.3697</td>
<td>0.3808</td>
<td>0.3623</td>
<td>0.399</td>
<td>0.399</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.3829</td>
<td>0.3988</td>
<td>0.3724</td>
<td>0.4257</td>
<td>0.4257</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.3963</td>
<td>0.4169</td>
<td>0.3826</td>
<td>0.4526</td>
<td>0.4526</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>0.4098</td>
<td>0.4349</td>
<td>0.3924</td>
<td>0.4796</td>
<td>0.4796</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>0.4234</td>
<td>0.4529</td>
<td>0.4029</td>
<td>0.5068</td>
<td>0.5068</td>
</tr>
<tr>
<td>ES</td>
<td>1.0</td>
<td>2.4097</td>
<td>2.9833</td>
<td>4.6033</td>
<td>3.4646</td>
<td>0.9471</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.2763</td>
<td>5.1493</td>
<td>7.2913</td>
<td>5.822</td>
<td>1.1145</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.2931</td>
<td>8.2445</td>
<td>9.9817</td>
<td>9.0074</td>
<td>1.3083</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>10.962</td>
<td>11.5465</td>
<td>12.013</td>
<td>12.2457</td>
<td>1.5302</td>
</tr>
<tr>
<td>EQ</td>
<td>1.0</td>
<td>1.4299</td>
<td>1.9467</td>
<td>3.5202</td>
<td>2.3434</td>
<td>0.1685</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>3.1185</td>
<td>3.9301</td>
<td>6.0511</td>
<td>4.5111</td>
<td>0.2788</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>5.9646</td>
<td>6.8711</td>
<td>8.6464</td>
<td>7.5533</td>
<td>0.4129</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>9.5381</td>
<td>10.1048</td>
<td>10.6604</td>
<td>10.7442</td>
<td>0.5733</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>11.8909</td>
<td>12.0959</td>
<td>11.7247</td>
<td>12.6007</td>
<td>0.7622</td>
</tr>
<tr>
<td>EI</td>
<td>1.0</td>
<td>0.1505</td>
<td>0.201</td>
<td>0.2245</td>
<td>0.2193</td>
<td>0.0489</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.2389</td>
<td>0.3095</td>
<td>0.3128</td>
<td>0.3161</td>
<td>0.0724</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.3269</td>
<td>0.4076</td>
<td>0.3729</td>
<td>0.3938</td>
<td>0.0954</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>0.3906</td>
<td>0.4717</td>
<td>0.3999</td>
<td>0.4362</td>
<td>0.1180</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>0.3964</td>
<td>0.4765</td>
<td>0.3964</td>
<td>0.4271</td>
<td>0.1403</td>
</tr>
<tr>
<td>ECo</td>
<td>1.0</td>
<td>0.2356</td>
<td>0.2755</td>
<td>0.3079</td>
<td>0.3456</td>
<td>0.0898</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.3760</td>
<td>0.4221</td>
<td>0.4477</td>
<td>0.5002</td>
<td>0.1379</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.5182</td>
<td>0.5571</td>
<td>0.5564</td>
<td>0.6282</td>
<td>0.1874</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>0.6271</td>
<td>0.6502</td>
<td>0.6210</td>
<td>0.7058</td>
<td>0.2378</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>0.6492</td>
<td>0.6660</td>
<td>0.6392</td>
<td>0.7059</td>
<td>0.2888</td>
</tr>
</tbody>
</table>

time duration of interruption to the main server is equal to the duration of the super clock whereas in model 1, the duration of interruption can be greater than the duration of super clock. The assumption in model 2 reduces the duration of interruption at the main server and so reduces the
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption effective service time of a customer at the main server. (Effective service time at the main server includes the actual time necessary to get served and the sum of the durations of intermediate interruptions.)

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

- \( M_1 = Mc + 1 \)
- \( \tilde{\alpha} = e'_{M_1}(1) \otimes \alpha \)
- \( \tilde{\eta} = (\eta, 0) \)
- \( \tilde{E} = \begin{bmatrix} E & E^0 \\ 0 & 0 \end{bmatrix} \)
- \( D^* = D \oplus \tilde{E} \)

Consider the queueing model \( X = \{ X(t), t \geq 0 \} \), where \( X(t) = \{ N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t), U(t) \} \). \( N(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t) \) and \( U(t) \) have the same meaning as described in model 1.

\( S(t) \) takes one more value ‘4’ in addition to the values taken by that in model 1 of this chapter.

\( S(t) = 4 \) if the regular server is waiting to get the remaining part of consultation after the completion of the present service at the main server. This happens when the super clock expires in the midst of a consultation. As soon as the super clock expires during the process of an interruption, the main server restarts or resumes the service of the customer with him according to the status of the threshold. At this stage, the regular server has to wait until the service completion at the main server to get his consultation completed. Since the interruption will be removed at the time
the super clock saturates, the super clock saturation point ‘0’ is not in the interruption state; but it is present in the busy state.

\( \{X(t), t \geq 0\} \) is a Continuous Time Markov Chain with state space

\[ \Phi = \bigcup_{i=0}^{\infty} \phi(i). \]

The terms \( \phi(i) \)'s are defined as

\( \phi(0) = \{(0, u)\} \),

\( \phi(1) = \phi(1, 0) \cup \phi(1, \tilde{0}) \cup \phi(1, 1) \) and

\( \phi(i) = \phi(i, 0) \cup \phi(i, 1) \cup \phi(i, 2) \cup \phi(i, 3) \cup \phi(i, 4) \), for \( i \geq 2 \), where

\( \phi(1, 0) = \{(1, 0, 0, t_1, u)\} \cup \{(1, 0, k, l_1, t_1, u) : 1 \leq k \leq M\}, \)

\( \phi(1, \tilde{0}) = \{(1, \tilde{0}, j, t_2, u) : 0 \leq j \leq K\}, \)

\( \phi(1, 1) = \{(1, 1, j, l_2, l_3, t_2, u) : 0 \leq j \leq K - 1\} \)

\( \phi(i, 0) = \{(i, 0, j, 0, t_1, t_2, u) \cup (i, 0, j, k, l_1, t_1, t_2, u) : 1 \leq k \leq M\}, \)

for \( 0 \leq j \leq K \); and

\( \phi(i, 1) = \{(i, 1, j, l_2, l_3, t_2, u) : 0 \leq j \leq K - 1\}, \)

\( \phi(i, 2) = \{(i, 2, j, k, l_1, l_2, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1, 0 \leq k \leq M - 1\}, \)

\( \phi(i, 3) = \{(i, 3, j, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1\}, \)

\( \phi(i, 4) = \{(i, 4, j, l_2, l_3, t_1, t_2, u) : 0 \leq j \leq K - 1\}, \)

for \( i \geq 2 \) with \( 1 \leq l_1 \leq c, 1 \leq l_2 \leq d, 0 \leq l_3 \leq f, 1 \leq t_1 \leq a, \)

\( 1 \leq t_2 \leq b, \) and \( 1 \leq u \leq r. \)
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

The infinitesimal generator $Q$ is given by

$$Q = \begin{bmatrix} L_0 & B_1 & B_2 & B_3 & B_4 \\ B_5 & A_1 & A_0 & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots \end{bmatrix}$$

(2.31)

where

$$B_1 = \begin{bmatrix} \alpha & 0 \end{bmatrix} \otimes L_1 , \quad B_2 = \begin{bmatrix} e_{M_1} \otimes T^0 \\ e_{K+1} \otimes U^0 \\ 0 \end{bmatrix} \otimes I_r ,$$

$$B_3 = \begin{bmatrix} I_{M_1} \otimes T \\ B_{31} & B_{32} \\ B_{33} & I_K \otimes D^* \otimes I_b \end{bmatrix} \oplus L_0 ,$$

$$B_4 = \begin{bmatrix} B_{41} & B_{42} & 0 \end{bmatrix} \otimes L_1 , \quad B_5 = \begin{bmatrix} B_{51} & B_{52} & B_{53} \end{bmatrix} \otimes I_r ,$$

$$A_0 = I \otimes L_1 , \quad A_1 = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{14} & A_{15} & A_{16} & A_{17} & A_{18} \\ & & & A_{19} & A_{110} \end{bmatrix} \oplus L_0 ,$$

$$A_2 = \begin{bmatrix} A_{21} & A_{22} & 0 \end{bmatrix} \otimes I_r .$$

Here $A_0, A_1$ and $A_2$ are square matrices of order $C_0$, $B_3$ is a square matrix of order $C_1$ and $B_1, B_2, B_4, B_5$ are matrices of orders $r \times C_1, C_1 \times r, C_1 \otimes C_0$ and $C_0 \times C_1$ respectively, where

$$C_0 = [M_1(K+1)ab + Kd(f+1)b + MKcd(f+1)ab + Kab + K(f+1)ab]r,$$

and

$$C_1 = [M_1a + (K+1)b + Kbd(f+1)]r.$$
The blocks are defined as follows:

\[ B_{31} = \begin{bmatrix} I_K \otimes (U - \theta I) & O \\ O & U \end{bmatrix}_{(K+1)b}, \]

\[ B_{32} = \theta \begin{bmatrix} I_K \otimes \delta \otimes \tilde{\eta} \\ O \end{bmatrix}_{(K+1)b \times Kd(f+1)}, \]

\[ B_{33} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \tilde{\Delta}_{\theta} \end{bmatrix}_{Kd(f+1) \times (K+1)b}, \]

\[ B_{41} = \begin{bmatrix} e_{K+1}^{0} & I_{M_1} \otimes I_a \otimes \beta \\ I_{K+1} \otimes \bar{\alpha} \otimes I_b \\ O \end{bmatrix}_{C_1 \times (K+1)M_1 ab}, \]

\[ B_{42} = \begin{bmatrix} O \\ I \end{bmatrix}_{C_1 \times Kbd(f+1)}, \]

\[ B_{51} = \begin{bmatrix} e_{K+1}^{0} & I_{M_1} \otimes I_a \otimes U^0 \\ O \end{bmatrix}_{M_0 \times M_0 a}, \]

\[ B_{52} = \begin{bmatrix} I_{K+1} \otimes \theta_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}_{C_0 \times (K+1)b}, \]

\[ B_{53} = \begin{bmatrix} O \\ F \end{bmatrix}_{C_0 \times Kbd(f+1)}, \]

\[ A_{11} = \begin{bmatrix} I_K \otimes I_{M_1} \otimes (T \oplus U - \theta I) & O \\ O & I_{M_1} \otimes (T \oplus U) \end{bmatrix}_{M_1(K+1)ab}, \]

\[ A_{12} = \theta \begin{bmatrix} I_K \otimes P \\ O \end{bmatrix}_{M_1(K+1) \times McKd(f+1)}, \]

\[ A_{13} = \theta \begin{bmatrix} I_K \otimes P^* \\ O \end{bmatrix}_{M_1(K+1) \times K}, \]

\[ A_{14} = \begin{bmatrix} O & I_K \otimes D^0 \otimes \Delta^0 \end{bmatrix}_{Kd(f+1) \times M_1(K+1)ab}, \]

\[ A_{15} = I_K \otimes D^* \otimes I_b, \]

\[ A_{16} = \begin{bmatrix} O & I_K \otimes \Delta^* \end{bmatrix}_{Med(f+1)ab \times M_1(K+1)ab}, \]

\[ A_{17} = I_K \otimes I_M \otimes (G \oplus D \oplus \bar{E}) \otimes I_{ab}, \]

\[ A_{18} = I_K \otimes e_M \otimes G^0 \otimes e_d \otimes \bar{\Delta}, \]

\[ A_{19} = I_K \otimes T \otimes I_b, \]

\[ A_{110} = I_K(f+1) \otimes T \otimes I_b. \]
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

\[
A_{21} = \begin{bmatrix} F \\ O \end{bmatrix}_{C_0 \times Kd(f+1)b}, \quad A_{22} = \begin{bmatrix} O \\ \tilde{F} \end{bmatrix}.
\]

Here

\[
F = I_K \otimes \delta \otimes \tilde{\eta} \otimes T^0 \otimes I_b,
\]

\[
P = diag(\gamma, I_{M-1} \otimes I_c) \otimes \delta \otimes \tilde{\eta}, \quad P^* = \begin{bmatrix} 0 \\ e_M \otimes e_c \end{bmatrix},
\]

\[
\Delta^* = \begin{bmatrix} O & I_M \otimes I_c \otimes D^0 \end{bmatrix} \otimes \tilde{\Delta},
\]

\[
\tilde{\Delta} = \begin{bmatrix} e_f \otimes I_{ab} & 0 \\ O & e_a \otimes \alpha \otimes I_b \end{bmatrix},
\]

\[
\tilde{F} = I_{K+1} \otimes e_{M_1} \otimes T^0 \otimes \alpha \otimes I_b + e_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \otimes \beta,
\]

\[
\hat{F} = I_K \otimes diag(I_f \otimes T^0 \otimes I_b, T^0 \otimes e_b \otimes \beta),
\]

\[
\tilde{\Delta}_b = \begin{bmatrix} e_f \otimes I_b \\ e_b \otimes \beta \end{bmatrix}, \quad \Delta^0 = \begin{bmatrix} e_f \otimes \tilde{\alpha} \otimes I_b \\ e_b \otimes \tilde{\alpha} \otimes \beta \end{bmatrix}, \quad \tilde{\Delta} = \begin{bmatrix} e_f \otimes I_{ab} \\ e_{ab} \otimes \alpha \otimes \beta \end{bmatrix}.
\]

\(P\) and \(P^*\) are matrices of orders \(M_1 \times Mcd(f+1)\) and \(M_1 \times 1\) respectively.

### 2.5 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study by first establishing the stability condition of the queueing system.
2.5. Steady state analysis

2.5.1 Stability condition

Let $\pi$ denote the steady-state probability vector of the generator $A_0 + A_1 + A_2$. That is, $\pi(A_0 + A_1 + A_2) = 0; \pi e = 1$.

The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

$$\pi A_0 e < \pi A_2 e. \quad (2.32)$$

is satisfied. That is, the rate of drift to the left has to be higher than that to the right. The vector $\pi$ cannot be obtained explicitly in terms of the parameters of the model, and hence the stability condition is known only implicitly. If we partition the vector $\pi$ as

$$\pi = (\pi_0, \pi_1, \pi_2, \pi_3, \pi_4)$$

and then using the structure of the matrices $A_0$ and $A_2$, (2.32) is given by

$$\lambda < \pi_0 \tilde{F} e + \pi_3 F e + \pi_4 \hat{F} e. \quad (2.33)$$

For future reference, we define the traffic intensity $\rho_2$ as

$$\rho_2 = \frac{\pi A_0 e}{\pi A_2 e}. \quad (2.34)$$

Note that the stability condition in (2.32) is equivalent to $\rho_2 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 2.6.
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

### 2.5.2 Steady state probability vector

Let $\mathbf{x}$, partitioned as, $\mathbf{x} = (x_0, x_1, x_2, x_3, \ldots)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$. Note that $x_1 = (x_{10}, x_{11}, x_{11})$ and $x_i = (x_{i0}, x_{i1}, x_{i2}, x_{i3}, x_{i4})$, for $i \geq 2$. The vector $x$ satisfies the condition $xQ = 0$ and $xe = 1$, where $e$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub-vectors of $x$ are given by the equation

$$x_j = x_2R_{j-2}, j \geq 3 \quad (2.35)$$

where $R$ is the minimal non-negative solution of the matrix equation $R^2A_2 + RA_1 + A_0 = 0$. Knowing the matrix $R$, the vectors $x_0, x_1$ and $x_2$ are obtained by solving the boundary equations

$$x_0L_0 + x_1B_2 = 0$$
$$x_0B_1 + x_1B_3 + x_2B_5 = 0$$
$$x_1B_4 + x_2(A_1 + RA_2) = 0$$

subject to the normalizing condition

$$x_0e + x_1e + x_2(I - R)^{-1}e = 1.$$

### 2.5.3 Performance measures

In this section we list a number of key system performance measures to bring out the qualitative aspects of the model under study. These are
2.5. Steady state analysis

listed below along with their formulae for computation. Towards this end, we further partition the vectors \( x_i, i \geq 1 \) as

\[
\begin{align*}
  \mathbf{x}_1 &= (x_{10}, x_{1\tilde{0}}, x_{11}) \\
  \mathbf{x}_i &= (x_{i0}, x_{i1}, x_{i2}, x_{i3}, x_{i4}), i \geq 2.
\end{align*}
\]

Note that \( x_0, x_{10}, x_{1\tilde{0}}, x_{11}, x_{i0}, x_{i1}, x_{i2}, x_{i3} \) and \( x_{i4} \) are vectors of dimensions \( r, M_1r, (K+1)br, Kbd(f+1)r, M_1(K+1)abr, Kd(f+1)br, MKcd(f+1)abr, K(f+1)abr \) and \( Kabr \) respectively.

Even though the vectors \( x_0, x_1, x_2, \) etc. in model 2 are different from those vectors in model 1, the expressions for \( ES, ECo, EI, F_{mi}, F_{ri}, F_{rb}, F_{rc}, F_{min}, R_{c}\tilde{c}b, R_{s}\tilde{c}a, R_{s\tilde{w}i}, R_{scw}c \) and \( R_{c}\tilde{c}c \) are similar to those in model 1. These values are obtained by using the equations (2.11), (2.13), (2.14), (2.15), (2.16), (2.18), (2.19), (2.21), (2.24), (2.25), (2.26), (2.27), (2.29) and (2.30). We get the following performance measures also.

1. Expected number of customers in the queue

\[
EQ = \sum_{i=2}^{\infty} (i-1)x_{i1}e + \sum_{i=2}^{\infty} (i-2)[x_{i0}e + x_{i2}e + x_{i3}e + x_{i4}e].
\]  (2.36)

2. Fraction of time the main server is busy serving a customer

\[
F_{mb} = x_{1e} + \sum_{i=2}^{\infty} x_{i0}e + \sum_{i=2}^{\infty} x_{i3}e + \sum_{i=2}^{\infty} x_{i4}e.
\]  (2.37)
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption.

(3) Fraction of time regular server is waiting to get consultation

\[ F_{wc} = \sum_{i=2}^{\infty} x_{i3}e + \sum_{i=2}^{\infty} x_{i4}e. \] (2.38)

(4) Rate at which interruption completion takes place before threshold is realised

\[ R^b_{ic} = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=1}^{c} \sum_{l_2=1}^{d} \sum_{l_3=1}^{f} D_{l_2}^0 x_{i2jkl_1l_2l_3}e. \] (2.39)

(5) Rate at which interruption completion takes place after threshold is realised

\[ R^a_{ic} = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=1}^{c} \sum_{l_2=1}^{d} \sum_{l_3=1}^{f} D_{l_2}^0 x_{i2jkl_1l_2l_3}e. \] (2.40)

(6) Rate at which service completion (with at least one interruption) at the main server takes place after super clock is realised

\[ R^c_{Sc} = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{l_2=1}^{d} \sum_{l_3=1}^{f} \sum_{l_1=1}^{a} T_{t_1}^0 x_{i4jl_2l_3l_1}e. \] (2.41)

2.6 Numerical results

We consider the arrival processes ERA, EXA, HEA, MNA, MPA defined in the example of model 1. The purpose of this example to see how various performance measures behave under different scenario. Choose the matrices, vectors and values so that the stability condition \( \rho_2 < 1 \) is
satisfied. We fix

\[ T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}, \quad U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}, \quad D = \begin{bmatrix} -6 & 4 \\ 3 & -4 \end{bmatrix}, \]

\[ E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix}, \quad G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 0.3 & 0.7 \end{bmatrix}, \]

\[ \beta = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \quad \delta = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \]

\[ \eta = \begin{bmatrix} 0.5 & 0.05 \end{bmatrix}, \quad \gamma = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}, \quad K = 3, M = 3. \]

We look at the effect of varying \( \theta \) on the performance measures \( \rho_2 \), \( ES \), \( EQ \), \( EI \) and \( ECo \).

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>ERA</th>
<th>EXA</th>
<th>HEA</th>
<th>MNA</th>
<th>MPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_2 )</td>
<td>1.0</td>
<td>0.6043</td>
<td>0.6043</td>
<td>0.6043</td>
<td>0.6043</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.7180</td>
<td>0.7180</td>
<td>0.7180</td>
<td>0.7180</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.8227</td>
<td>0.8227</td>
<td>0.8227</td>
<td>0.8227</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>0.9186</td>
<td>0.9186</td>
<td>0.9186</td>
<td>0.9186</td>
</tr>
<tr>
<td>( ES )</td>
<td>1.0</td>
<td>2.3627</td>
<td>2.9628</td>
<td>4.8839</td>
<td>3.4395</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.2166</td>
<td>5.2298</td>
<td>8.1508</td>
<td>5.9142</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>7.6748</td>
<td>8.9864</td>
<td>11.4539</td>
<td>7.5644</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>12.4492</td>
<td>13.0625</td>
<td>12.9123</td>
<td>13.692</td>
</tr>
<tr>
<td>( EQ )</td>
<td>1.0</td>
<td>3.1734</td>
<td>4.1368</td>
<td>7.0363</td>
<td>4.7486</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>6.4915</td>
<td>7.7686</td>
<td>10.2969</td>
<td>6.5255</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>11.2291</td>
<td>11.8550</td>
<td>11.8539</td>
<td>12.4519</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>12.4492</td>
<td>13.0625</td>
<td>12.9123</td>
<td>13.692</td>
</tr>
<tr>
<td>( EI )</td>
<td>1.0</td>
<td>0.3491</td>
<td>0.4253</td>
<td>0.5211</td>
<td>0.5054</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.5842</td>
<td>0.6772</td>
<td>0.7718</td>
<td>0.7702</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.8502</td>
<td>0.9295</td>
<td>0.9447</td>
<td>0.8044</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>1.0411</td>
<td>1.0612</td>
<td>0.9605</td>
<td>1.1117</td>
</tr>
<tr>
<td>( ECo )</td>
<td>1.0</td>
<td>0.2375</td>
<td>0.1784</td>
<td>0.0994</td>
<td>0.0700</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.3032</td>
<td>0.2025</td>
<td>0.0963</td>
<td>0.0757</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.2819</td>
<td>0.1821</td>
<td>0.0850</td>
<td>0.1336</td>
</tr>
<tr>
<td></td>
<td>2.5</td>
<td>0.1979</td>
<td>0.1273</td>
<td>0.06559</td>
<td>0.0433</td>
</tr>
</tbody>
</table>

From the table 2.2 we can see that as \( \theta \) increases the traffic intensity also increases. This results in rapid accumulation of customers in system and in queue. Thus \( ES \) and \( EQ \) increase. The effective rates for interruption \( EI \) and for consultation \( ECo \) also increase as \( \theta \) increases.
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

2.7 Description of model 3

In a queueing system where the service process consists of certain number of phases, with service subject to interruptions, the concept of protecting a few phases of service (which may be so costly to afford an interruption) from interruption could be an important idea. Klimenok et. al. [29] studied a multi-server queueing system with finite buffer and negative customers where the arrival is BMAP and service is PH-type. They assumed that a negative customer can delete an ordinary customer in service if the service of a customer goes on in any of the unprotected phases; whereas if the service of the customer is protected from the effect of the negative customers. Klimenok and Dudin [28] extended the above paper by considering disciplines of complete admission and complete rejection. Further, Klimenok and Dudin [28] assumed an infinite buffer. Krishnamoorthy et. al. [32] introduced the idea of protection in a queueing system where the service process is subject to interruptions. They assume that the final \( m - n \) phases of the Erlang service process with \( m \) phases are protected from interruption. Whereas if the service process belongs to the first \( n \) phases, it is subject to interruption and an interrupted service is resumed/repeated after some random time. There is no reduction (removal) in the number of customers due to interruption and no bound was assumed on the number of interruptions that can possibly occur in the course of a service.

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

- \( M_0 = M(c + 1) \), \( M_1 = M_0 + 1 \)
- \( \bar{\alpha} = e'_{M_1}(1) \otimes \alpha \)
• \( \tilde{\gamma} = (\gamma, 0) \) and \( \tilde{\eta} = (\eta, 0) \)

• \( \tilde{G} = \begin{bmatrix} G & G^0 \\ 0 & 0 \end{bmatrix} \) and \( \tilde{E} = \begin{bmatrix} E & E^0 \\ 0 & 0 \end{bmatrix} \)

• \( \delta^* = \delta \otimes \tilde{\eta} \) and \( \gamma^* = \tilde{\gamma} \otimes (\delta \otimes \tilde{\eta}) \)

• \( D^* = D \oplus \tilde{E} \) and \( G^* = \tilde{G} \oplus D^* \)

• \( \tilde{I} = \begin{bmatrix} 0 & I_{M_0} \\ I_{M_0} \end{bmatrix}_{M_0 \times M_1} \)

• \( \tilde{I}_m = \tilde{\eta} \otimes \begin{bmatrix} O & O \\ O & I_{a-m} \end{bmatrix}_{a \times a} \)

• \( \tilde{e}_c = \begin{bmatrix} e_c \otimes \tilde{I}_m \\ \tilde{\eta} \otimes I_a \end{bmatrix} \)

• \( I_m^* = \begin{bmatrix} I_m \\ O \end{bmatrix}_{a \times m} \)

In this model, we assume that out of the 'a' phases at the main server, \( m \leq a \) phases have the property that no interruptions are allowed to the main server (and therefore to the customer being served at the main server) if the service is at any one of these phases. If the regular server needs a consultation at this time, he/she has to wait until the service at the main server is completed. All other assumptions are same as those in model 1 of this chapter. Thus if either the customer at the main server has already interrupted \( M \) times or the super clock has expired or the service at the main server is at any one of the last \( a - m \) protected phases, then the regular server has to wait until the completion of the service of the present customer at the main server.
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

Consider the queueing model \( X = \{X(t), t \geq 0\} \), where

\[
X(t) = \{N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t), U(t)\}.
\]

\( N(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t) \) and \( U(t) \) have the same meaning as those in model 1. Since there are no interruption from the \((m+1)^{th}\) phase onwards, these phases are not present when \( S(t) = 2 \) or 3.

\[ \{X(t), t \geq 0\} \] is a Continuous Time Markov Chain with state space

\[ \Psi = \bigcup_{i=0}^{\infty} \psi(i). \]

The terms \( \psi(i)'s \) are defined as

\( \psi(0) = \{(0, u)\}, \)

\( \psi(1) = \psi(1, 0) \cup \psi(1, \bar{0}) \cup \psi(1, 1), \)

\( \psi(i) = \psi(i, 0) \cup \psi(i, 1) \cup \psi(i, 2) \cup \psi(i, 3), i \geq 2, \)
where

\[
\begin{align*}
\psi(1,0) &= \{(1,0,0,t,u)\} \cup \{(1,0,k,l,t,u) : 1 \leq k \leq M, 1 \leq t \leq a\} \\
\psi(1,\bar{0}) &= \{(1,\bar{0},j,t,u) : 0 \leq j \leq K\} \\
\psi(1,1) &= \{(1,1,j,l,t,u) : 0 \leq j \leq K-1\} \\
\psi(i,0) &= \{(i,0,j,0,t,u) \cup (i,0,j,k,l,t,u) : 0 \leq j \leq K, \\
& \hspace{1cm} 1 \leq k \leq M, 1 \leq t \leq a\}, \\
\psi(i,1) &= \{(i,1,j,l,t,u) : 0 \leq j \leq K-1\} \\
\psi(i,2) &= \{(i,2,j,k,l,t,u) : 0 \leq j \leq K-1, 0 \leq k \leq M-1, \\
& \hspace{1cm} 1 \leq t \leq m\} \\
\psi(i,3) &= \{(i,3,j,l,t,u) : 0 \leq j \leq K-1, 1 \leq t \leq a\}, \\
& \text{for } 0 \leq l \leq c, 1 \leq l_2 \leq d, 0 \leq l_3 \leq f, 1 \leq t_2 \leq b, 1 \leq u \leq r.
\end{align*}
\]

The infinitesimal generator \(Q\) is given by

\[
Q = \begin{bmatrix}
L_0 & B_1 \\
B_2 & B_3 & B_4 \\
B_5 & A_1 & A_0 \\
A_2 & A_1 & A_0 \\
\vdots & \ddots & \ddots
\end{bmatrix}
\] \hspace{1cm} (2.42)

Here \(A_0, A_1\) and \(A_2\) are square matrices of order \(C_0\); \(B_3\) is a square matrix of order \(C_1\) and \(B_1, B_2, B_4, B_5\) are matrices of orders \(r \times C_1, C_1 \times r, C_1 \times C_0\) and \(C_0 \times C_1\), respectively, where \(C_0 = [M_1(K + 1)ab + Kd(f + 1)b + M_0Kd(f + 1)mb + K(f + 1)mb]r\) and
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption $C \varepsilon = [M_1 a + (K + 1)b + Kbd(f + 1)]r$ and these matrices are defined as follows:

$$B_1 = \left[ \begin{array}{c} \alpha \\ 0 \end{array} \right] \otimes L_1, B_2 = \left[ \begin{array}{c} e_{M_1} \otimes T^0 \\ e_{K+1} \otimes U^0 \\ 0 \end{array} \right] \otimes I_r,$$

$$B_3 = \left[ \begin{array}{ccc} I_{M_1} \otimes T & O & O \\ O & B_{31} & B_{32} \\ O & B_{33} & I_K \otimes D^* \otimes I_b \end{array} \right] \oplus L_0,$$

$$B_4 = \left[ \begin{array}{cc} B_{41} & B_{42} \\ B_{43} & O \end{array} \right] \otimes L_1, B_5 = \left[ \begin{array}{ccc} B_{51} & B_{52} & B_{53} \end{array} \right] \otimes I_r,$$

$$A_0 = I \otimes L_1, A_1 = \left[ \begin{array}{ccc} A_{11} & O & A_{12} \\ A_{14} & A_{15} & O \\ A_{16} & O & A_{17} \end{array} \right] \oplus L_0,$$

$$A_2 = \left[ \begin{array}{cc} A_{21} & B_{53} \\ B_{53} & O \end{array} \right] \otimes I_r.$$

Here the block matrices are

$$B_{31} = I_{K+1} \otimes U - \theta \left[ \begin{array}{c} I_K \\ 0 \\ 0 \end{array} \right]_{(K+1) \times (K+1)} \otimes I_b,$$

$$B_{32} = \theta \left[ \begin{array}{c} I_K \\ 0 \end{array} \right]_{(K+1) \times K} \otimes \delta \otimes \eta \otimes I_b,$$

$$B_{33} = \left[ \begin{array}{cc} O & I_K \otimes D^0 \otimes \tilde{\Delta}_b \end{array} \right]_{Kd(f+1)b \times (K+1)b}^b,$$

$$B_{41} = \left[ \begin{array}{c} e_{K+1}^{(1)} \otimes I_{M_1} \otimes I_a \otimes \beta \\ I_{K+1} \otimes \tilde{\chi} \otimes I_b \\ O \end{array} \right]_{C_1 \times (K+1)M_1ab}.$$


2.7. Description of model 3

\[B_{42} = \begin{bmatrix} O \\ I_{Kd(f+1)b} \end{bmatrix}_{C_1 \times Kd(f+1)b},\]

\[B_{51} = \begin{bmatrix} e_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \\ O \end{bmatrix}_{C_0 \times M_0 a},\]

\[B_{52} = \begin{bmatrix} I_{K+1} \otimes e_{M_1} \otimes T^0 \otimes I_b \\ O \end{bmatrix}_{C_0 \times (K+1)b},\]

\[B_{53} = \begin{bmatrix} O \\ I_{K} \otimes \delta \otimes I_{f+1} \otimes T^0 \otimes I_b \end{bmatrix}_{C_0 \times Kd(f+1)b},\]

\[A_{11} = I_{K+1} \otimes I_{M_1} \otimes (T \oplus U) - \theta \begin{bmatrix} I_{K} & 0 & 0 \end{bmatrix} \otimes I_{M_1} \otimes I_{ab},\]

\[A_{12} = \theta \begin{bmatrix} I_{K} \\ 0 \end{bmatrix}_{(K+1) \times K} \otimes P \otimes I_m^* \otimes I_b,\]

\[A_{13} = \theta \begin{bmatrix} I_{K} \\ 0 \end{bmatrix}_{(K+1) \times K} \otimes P^* \otimes I_b,\]

\[A_{14} = \begin{bmatrix} O & I_{K} \otimes D^0 \otimes \Delta^0 \end{bmatrix}_{Kd(f+1)b \times M_1(K+1)ab},\]

\[A_{15} = I_{K} \otimes D^* \otimes I_b,\]

\[A_{16} = \begin{bmatrix} O & I_{K} \otimes \hat{I} \otimes D^0 \otimes \hat{\Delta} \end{bmatrix}_{M_0Kd(f+1)b \times M_1(K+1)ab},\]

\[A_{17} = I_{K} \otimes I_M \otimes G^* \otimes I_{mb},\]

\[A_{18} = I_{Kd} \otimes (\tilde{E} \oplus T) \otimes I_b,\]

\[A_{21} = \begin{bmatrix} \tilde{T}^0 + \tilde{U}^0 \\ O \end{bmatrix}_{C_0 \times M_1(K+1)ab}.\]

Here

\[P = \begin{bmatrix} \text{diag}(\hat{\gamma}, I_{M-1} \otimes \hat{I}_c) \\ O \end{bmatrix}_{M_1 \times M_0} \otimes \delta \otimes \tilde{\eta}.\]
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

\[
P^* = \begin{bmatrix} I_m \\ e_{M-1} \otimes \tilde{e}_c \\ e_{c+1} \otimes \bar{\eta} \otimes I_a \end{bmatrix}_{M_1a \times (f+1)a},
\]

\[
\bar{T}^0 = I_{K+1} \otimes e_{M_1} \otimes T^0 \otimes \alpha \otimes I_b, \quad \bar{U}^0 = e_{K+1} \otimes I_{M_1} \otimes I_a \otimes U^0 \otimes \beta,
\]

\[
\bar{\Delta}_b = \begin{bmatrix} e_f \otimes I_b \\ e_b \otimes \beta \end{bmatrix}, \quad \Delta^0 = \begin{bmatrix} e_f \otimes \tilde{\alpha} \otimes I_b \\ e_b \otimes \tilde{\beta} \otimes \beta \end{bmatrix}, \quad \bar{\Delta} = \begin{bmatrix} e_f \otimes (I_{M_1}^*)' \otimes I_b \\ e_{mb} \otimes \alpha \otimes \beta \end{bmatrix}.
\]

### 2.8 Steady state analysis

In this section we perform the steady-state analysis of the queueing model under study by first establishing the stability condition of the queueing system.

#### 2.8.1 Stability condition

Let \( \pi \) denote the steady-state probability vector of the generator \( A_0 + A_1 + A_2 \). That is, \( \pi(A_0 + A_1 + A_2) = 0; \pi e = 1 \).

The LIQBD description of the model indicates that the queueing system is stable (see, Neuts [44]) if and only if

\[
\pi A_0 e < \pi A_2 e. \quad (2.43)
\]

That is, the rate of drift to the left has to be higher than that to the right. The vector \( \pi \) cannot be obtained explicitly in terms of the parameters of the model.
For future reference, we define the traffic intensity $\rho_3$ as

$$\rho_3 = \frac{\pi A_0 e}{\pi A_2 e}. \quad (2.44)$$

Note that the stability condition in (2.43) is equivalent to $\rho_3 < 1$. We will discuss the impact of the input parameters of the model on the traffic intensity in Section 2.9.

### 2.8.2 Steady state probability vector

Let $x$, partitioned as, $x = (x_0, x_1, x_2, x_3, \ldots, \ldots, \ldots)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$.

Note that $x_1 = (x_{10}, x_{10}, x_{11})$ and $x_i = (x_{i0}, x_{i1}, x_{i2}, x_{i3})$, for $i \geq 2$. The vector $x$ satisfies the condition $xQ = 0$ and $xe = 1$, where $e$ is a column vector of appropriate dimension. When the stability condition is satisfied, the sub vectors of $x$ are given by the equation

$$x_j = x_2R^{j-2}, j \geq 3 \quad (2.45)$$

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

$$R^2 A_2 + RA_1 + A_0 = 0. \quad (2.46)$$
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of 

Knowing the matrix $R$, the vectors $x_0$, $x_1$ and $x_2$ are obtained by solving the equation

$$
\begin{bmatrix}
x_0 & x_1 & x_2
\end{bmatrix}
\begin{bmatrix}
L_0 & B_1 & & \\
B_2 & B_3 & B_4 & \\
B_5 & A_1 + RA_2 & & \\
\end{bmatrix} = 0,
$$

subject to the normalizing condition

$$
x_0e + x_1e + x_2(I - R)^{-1}e = 1.
$$

2.8.3 Expected waiting time in queue

For computing expected waiting time in queue of a particular customer who joins as the $k^{th}$ customer, where $k > 0$, in the queue, we consider the Markov process

$$
Z(t) = \{(\tilde{N}(t), S(t), B_1(t), B_2(t), S_1(t), S_2(t), S_3(t), J_1(t), J_2(t)) : t \geq 0\},
$$

where $\tilde{N}(t)$ is the rank of the customer and all other variables defined as earlier. The rank $\tilde{N}(t)$ of the customer is assumed to be $i$ if he is the $i^{th}$ customer in the queue at time $t$. His rank may decrease to 1 as the customers ahead of him leave the system either after completing their services (if $S(t) = 0$) or completing the consultation (if $S(t) = 1$). Since the customers who arrive after the tagged customer cannot change his rank, level-changing transitions in $Z(t)$ can only take place to one side of the diagonal. The absorbing state $\Delta_2$ denote the tagged customer is
selected for service. Thus the infinitesimal generator $\tilde{V}$ of the process $Z(t)$
takes the form

$$\tilde{V} = \begin{bmatrix} V & V^0 \\ 0 & 0 \end{bmatrix},$$

where $V = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 \\ \tilde{A}_1 & \tilde{A}_2 \\ & \ddots & \ddots \\ & & \tilde{A}_1 & \tilde{A}_2 \\ & & & \tilde{A}_1 \end{bmatrix}$ and $V^0 = \begin{bmatrix} 0 \\ e_{M_1} \otimes (T^0 \oplus U^0) \\ 0 \\ T^0 \otimes e_b \end{bmatrix},$

with $\tilde{A}_1 = A_1^* - U_2$ and $\tilde{A}_2 = A_2^* + U_2$, where $A_1^*$ and $A_2^*$ are obtained from $A_1$ and $A_2$ if they are written as $A_1 = A_1^* \oplus L_0$ and $A_2 = A_2^* \otimes I_r$. Here

$$U_2 = \begin{bmatrix} O & O \\ A_{14} & O \\ O & O \end{bmatrix}.$$

Now, the waiting time $V$ of a customer, who joins the queue as the $j^{th}$
customer is the time until absorption of the Markov chain $V(t)$. Thus the expected waiting time of this particular customer is given by the column vector,

$$E^{(j)}_V = \{ -\tilde{A}_1^{-1} [I + \sum_{i=1}^{j-1} (-\tilde{A}_2 \tilde{A}_1^{-1})^i] \} e.$$

The second moment of waiting time of the tagged customer is given by the column vector $E^{(j)}_{V^2}$ which is the first block of the matrix $2(-\tilde{V})^{-2} e$.

Hence the expected waiting time of a general customer in the queue is,

$$V_L = \sum_{j=1}^{\infty} x(j) E^{(j)}_V.$$
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption.

The second moment of $V$ is

\[ V^{(2)}_L = \sum_{j=1}^{\infty} x(j)E_{V_2}^j. \]

2.8.4 Performance measures

The vectors $x_0$, $x_1$, $x_2$, etc. in model 3 are different from those vectors in model 1. The expressions for $ES$, $EQ$, $ECo$, $F_{mi}$, $F_{ri}$, $F_{mb}$, $F_{rb}$, $F_{rc}$, $F_{wc}$, $F_{min}$, $R_{c}^{c}b$, $R_{c}^{a}a$, $R_{c}^{a}a$, $R_{c}^{a}wc$ and $R_{c}^{c}c$ etc. are similar to those in model 1. These values are obtained by using the equations (2.11), (2.12), (2.13), (2.15), (2.16), (2.17), (2.18), (2.19), (2.20), (2.21), (2.24), (2.25), (2.26), (2.27), (2.28), (2.29) and (2.30).

Note that $x_0$, $x_{10}$, $x_{11}$, $x_{i0}$, $x_{i1}$, $x_{i2}$, $x_{i3}$, for $i \geq 2$ are vectors of dimensions $r$, $M_1 ar$, $(K + 1)br$, $Kbd(f + 1)r$, $(K + 1)M_1 abr$, $Kd(f + 1)br$, $M_0 Kd(f + 1)mbr$ and $K(f + 1)mbr$, respectively.

We get the following performance measures also.

(1) Effective rate of interruption

\[ EI = \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{t_1=1}^{m} x_{i0j0t_1} e + \theta \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=1}^{M-1} \sum_{l_1=1}^{c} \sum_{t_1=1}^{m} x_{i0jkl_1t_1} e. \]  

(2.49)

(2) Rate at which interruption completion takes place before threshold
2.9. Numerical results

is realised

\[ R_c^b = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{d} \sum_{l_2=1}^{m} D_{l_2}^0 x_{i2jkl_1l_2t_1t_1} \]  \hspace{1cm} (2.50)

(3) Rate at which interruption completion takes place after threshold is realised

\[ R_c^a = \sum_{i=2}^{\infty} \sum_{j=0}^{K-1} \sum_{k=0}^{M-1} \sum_{l_1=0}^{d} \sum_{l_2=1}^{m} D_{l_2}^0 x_{i2jkl_1l_2t_1t_1} \]  \hspace{1cm} (2.51)

2.9 Numerical results

Let us assume

\[ T = \begin{bmatrix} -12 & 3 & 1 & 2 \\ 3 & -15 & 1 & 2 \\ 0 & 0 & -5 & 1 \\ 0 & 0 & 2 & -7 \end{bmatrix}, \quad U = \begin{bmatrix} -12 & 6 \\ 5 & -10 \end{bmatrix}, \quad D = \begin{bmatrix} -6 & 4 \\ 3 & -4 \end{bmatrix}, \]

\[ E = \begin{bmatrix} -12 & 3 \\ 3 & -12 \end{bmatrix}, \quad G = \begin{bmatrix} -12 & 8 \\ 8 & -12 \end{bmatrix}, \quad \alpha = \begin{bmatrix} 0.4 & 0.3 & 0.1 & 0.2 \end{bmatrix}, \]

\[ \beta = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \quad \delta = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}, \quad \eta = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, \quad \gamma = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}, \]

\( K = 3, \ M = 3. \)

We choose these matrices, vectors and values such that \( \rho_3 \) is less than 1.

Referring to Table 2.3, as the rate of consultation \( \theta \) increases, the traffic intensity \( \rho_3 \) increases and hence \( EI \) and \( ECo \) will increase. This results in an increase in \( F_{min} \) and \( F_{rc} \). As \( \theta \) increases, consultation is more frequent, so the main server will reach the upper bounds of number of
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

Table 2.3: Effect of $\theta$ on various performance measures

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\theta$</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_3$</td>
<td></td>
<td>0.5913</td>
<td>0.6910</td>
<td>0.7835</td>
<td>0.8694</td>
<td>0.9492</td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>2.4103</td>
<td>3.2431</td>
<td>3.9898</td>
<td>4.6096</td>
<td>5.1017</td>
</tr>
<tr>
<td>EQ</td>
<td></td>
<td>1.5026</td>
<td>2.3582</td>
<td>3.0976</td>
<td>3.5359</td>
<td>4.0178</td>
</tr>
<tr>
<td>EI</td>
<td></td>
<td>0.0759</td>
<td>0.1139</td>
<td>0.1472</td>
<td>0.1748</td>
<td>0.1974</td>
</tr>
<tr>
<td>Eco</td>
<td></td>
<td>0.2585</td>
<td>0.3685</td>
<td>0.4600</td>
<td>0.5339</td>
<td>0.5929</td>
</tr>
<tr>
<td>$F_{mi}$</td>
<td></td>
<td>0.4021</td>
<td>0.5503</td>
<td>0.6102</td>
<td>0.6200</td>
<td>0.6374</td>
</tr>
<tr>
<td>$F_{ri}$</td>
<td></td>
<td>0.5487</td>
<td>0.4798</td>
<td>0.4265</td>
<td>0.3864</td>
<td>0.3558</td>
</tr>
<tr>
<td>$F_{mb}$</td>
<td></td>
<td>0.4555</td>
<td>0.4193</td>
<td>0.4021</td>
<td>0.3853</td>
<td>0.3697</td>
</tr>
<tr>
<td>$F_{rb}$</td>
<td></td>
<td>0.2551</td>
<td>0.2471</td>
<td>0.2327</td>
<td>0.2175</td>
<td>0.2036</td>
</tr>
<tr>
<td>$F_{min}$</td>
<td></td>
<td>0.0326</td>
<td>0.0346</td>
<td>0.0366</td>
<td>0.0376</td>
<td>0.0384</td>
</tr>
<tr>
<td>$F_{rc}$</td>
<td></td>
<td>0.1613</td>
<td>0.2271</td>
<td>0.2806</td>
<td>0.3228</td>
<td>0.3559</td>
</tr>
<tr>
<td>$F_{rw}$</td>
<td></td>
<td>0.0298</td>
<td>0.0442</td>
<td>0.0533</td>
<td>0.0618</td>
<td>0.0682</td>
</tr>
</tbody>
</table>

Table 2.4: Effect of $\lambda$ on various performance measures

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\theta$</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_3$</td>
<td></td>
<td>0.5877</td>
<td>0.6836</td>
<td>0.7835</td>
<td>0.8845</td>
<td>0.9794</td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td>2.0385</td>
<td>2.9399</td>
<td>3.9898</td>
<td>5.0653</td>
<td>6.0260</td>
</tr>
<tr>
<td>EQ</td>
<td></td>
<td>1.8498</td>
<td>2.7159</td>
<td>3.7276</td>
<td>4.4197</td>
<td>4.5872</td>
</tr>
<tr>
<td>EI</td>
<td></td>
<td>0.0927</td>
<td>0.1472</td>
<td>0.1699</td>
<td>0.1874</td>
<td>0.1874</td>
</tr>
<tr>
<td>Eco</td>
<td></td>
<td>0.2816</td>
<td>0.3964</td>
<td>0.4600</td>
<td>0.5105</td>
<td>0.5454</td>
</tr>
<tr>
<td>$F_{mi}$</td>
<td></td>
<td>0.5350</td>
<td>0.3798</td>
<td>0.3102</td>
<td>0.2553</td>
<td>0.2105</td>
</tr>
<tr>
<td>$F_{ri}$</td>
<td></td>
<td>0.5977</td>
<td>0.5088</td>
<td>0.4265</td>
<td>0.3558</td>
<td>0.2965</td>
</tr>
<tr>
<td>$F_{mb}$</td>
<td></td>
<td>0.3462</td>
<td>0.3777</td>
<td>0.4021</td>
<td>0.4187</td>
<td>0.4268</td>
</tr>
<tr>
<td>$F_{rb}$</td>
<td></td>
<td>0.1841</td>
<td>0.2006</td>
<td>0.2327</td>
<td>0.2581</td>
<td>0.2750</td>
</tr>
<tr>
<td>$F_{min}$</td>
<td></td>
<td>0.0435</td>
<td>0.0525</td>
<td>0.0636</td>
<td>0.0727</td>
<td>0.0791</td>
</tr>
<tr>
<td>$F_{rc}$</td>
<td></td>
<td>0.2624</td>
<td>0.2448</td>
<td>0.2806</td>
<td>0.3079</td>
<td>0.3228</td>
</tr>
<tr>
<td>$F_{rw}$</td>
<td></td>
<td>0.0355</td>
<td>0.0448</td>
<td>0.0533</td>
<td>0.0602</td>
<td>0.0651</td>
</tr>
</tbody>
</table>

interruptions rapidly or super clock may realise frequently and thus the main server compels to complete the service of the customer at him before further consultations and this results in more waiting time of the regular server to get consultation. So $F_{rw}$ also increases. Since $F_{min}$, $F_{rc}$ and $F_{rw}$ increase, the customers have to stay in the system and in queue for a longer time and this results in an increase in $ES$ and $EQ$. This in turn make a decrease in $F_{mi}$ and $F_{ri}$. Since main server has to spend more time in consultation, it gets less time to serve customers. So $F_{mb}$ and $F_{rb}$ decreases.
2.10. Comparison of the three models

Referring to Table 2.4, as the arrival rate $\lambda$ increases, the traffic intensity $\rho_3$ increases. The system is fed with more and more customers and therefore accumulation of customers increases. So $ES$ and $EQ$ increase. Thus $EI$ and $EC_0$ will also increase. This results in a hike in $F_{\text{min}}$ and $F_{rc}$. Thus $F_{rw}$ also increases. As the arrival rate increases, there are more customers in the queue and therefore the servers have to spend longer time in service. Thus $F_{mb}$ and $F_{rb}$ increase. This in turn make a decrease in $F_{mi}$ and $F_{ri}$.

2.10 Comparison of the three models

Now we present a comparison of the three models analysed in this chapter. Recall that in model 1, the interruption is allowed to continue until its completion even when the super clock realises, whereas in model 2, the interruptions instantly terminated when the super clock realises. In model 3, some phases of service at the main server are protected from interruption. We compare the performance measures, namely, the traffic intensities $\rho_1$, $\rho_2$, $\rho_3$, expected number of customers in the system $ES$ and expected number of customers in the queue $EQ$. Let expected number of customers in the system $ES$ and expected number of customers in the queue $EQ$ be denoted by $ES_i$ and $EQ_i$ for the respective models $i = 1, 2$ and 3. Here a model with least number of customers waiting in the queue $EQ$ is considered as the most efficient one. We check the traffic intensities $\rho_i$’s in each case. Let $G_r$ denote the rate of the super clock. For convenience, we denote the models 1, 2 and 3 as $M_1$, $M_2$ and $M_3$, respectively. Let us assume the following matrices and parameters:
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption

\[
T = \begin{bmatrix}
-12 & 3 & 1 & 2 \\
3 & -15 & 1 & 2 \\
0 & 0 & -5 & 1 \\
0 & 0 & 2 & -7 \\
\end{bmatrix}, 
U = \begin{bmatrix}
-12 & 6 \\
5 & -10 \\
-6 & 4 \\
\end{bmatrix}, 
D = \begin{bmatrix}
-6 & 4 \\
3 & -4 \\
\end{bmatrix},
\]

\[
E = \begin{bmatrix}
-12 & 3 \\
3 & -12 \\
\end{bmatrix}, 
\alpha = \begin{bmatrix}
0.4 & 0.3 & 0.1 & 0.2 \\
\end{bmatrix},
\beta = \begin{bmatrix}
0.4 & 0.6 \\
\end{bmatrix}, 
\delta = \begin{bmatrix}
0.4 & 0.6 \\
\end{bmatrix}, 
\eta = \begin{bmatrix}
0.5 & 0.5 \\
\end{bmatrix}, 
\lambda = 4, \theta = 1, M = 3 \text{ and } K = 3.
\]

Table 2.5: Effect of $G_r$ on various performance measures

<table>
<thead>
<tr>
<th>$G_r$</th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>$\rho_3$</th>
<th>$ES_1$</th>
<th>$ES_2$</th>
<th>$ES_3$</th>
<th>$EQ_1$</th>
<th>$EQ_2$</th>
<th>$EQ_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.6200</td>
<td>0.5891</td>
<td>0.6438</td>
<td>3.2826</td>
<td>2.8073</td>
<td>3.1793</td>
<td>2.1836</td>
<td>1.7059</td>
<td>2.1690</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6210</td>
<td>0.5560</td>
<td>0.6438</td>
<td>3.2850</td>
<td>2.3273</td>
<td>3.1792</td>
<td>2.1935</td>
<td>1.2732</td>
<td>2.1693</td>
</tr>
<tr>
<td>1</td>
<td>0.6213</td>
<td>0.5209</td>
<td>0.6438</td>
<td>3.2869</td>
<td>1.9438</td>
<td>3.1789</td>
<td>2.1935</td>
<td>0.9380</td>
<td>2.1695</td>
</tr>
<tr>
<td>2</td>
<td>0.6217</td>
<td>0.4838</td>
<td>0.6438</td>
<td>3.2883</td>
<td>1.6469</td>
<td>3.1785</td>
<td>2.1981</td>
<td>0.6900</td>
<td>2.1696</td>
</tr>
<tr>
<td>4</td>
<td>0.6220</td>
<td>0.4528</td>
<td>0.6438</td>
<td>3.2891</td>
<td>1.4057</td>
<td>3.1780</td>
<td>2.2015</td>
<td>0.5487</td>
<td>2.1697</td>
</tr>
</tbody>
</table>

Referring to table 2.5 we can see that as $G_r$ increases, there is a slight increase in $\rho_1$, a considerable decrease in $\rho_2$ whereas $\rho_3$ remains a constant. We see that if $G_r$ is high, the super clock realises at a faster rate and this will result in a faster completion of interruption and hence the service completion in case of model 2. Remember that service completion happens only at the main server and the regular server waits to get the remaining consultation after the main server’s service completion. Thus in $M_2$, even though the regular server waits to get the remaining consultation, by this time the service at main server will be completed. This decreases $\rho_2$ of the system.

In $M_2$, the service at the main server will be restarted or resumed as soon as the super clock expires during an interruption, while in $M_1$ and $M_3$ the interruption will be continued even after the expiry of the super
2.10. Comparison of the three models

In $M_3$ there are no interruption at all from some protected phases of service at the main server. Thus we get a comparison of expected number of customers in the system and in the queue as $ES_2 < ES_3 < ES_1$ and $EQ_2 < EQ_3 < EQ_1$.

As $G_r$ has a rapid increase, $ES_2$ and $EQ_2$ decrease rapidly. But $ES_1$ and $EQ_1$ increase slightly as $G_r$ increases. $ES_3$ has a slight decrease and $EQ_3$ has a negligible increase. This shows that the rate of the super clock $G_r$ has a considerable effect in model $M_2$ when compared with the other two models. This is exactly what we are expected because in $M_2$, the interruption will be stopped as soon as the super clock expires, but the interruption will be continued until its completion in the other two models. So the rate of super clock has no direct effect on the values of $ES$ and $EQ$ in models $M_1$ and $M_3$.

Thus we can see that $M_2$ is the most efficient model and $M_1$ is the least one for the data in hand.
Chapter 2. Two-server queues with consultations controlled by upper bounds on number of interruptions, consultations and duration of interruption