Chapter 5

A multi-server queue with consultations

5.1 Introduction

We analysed queueing models with two and three servers in the previous chapters. The interruptions to the main server and consultations to the regular servers are controlled by certain parameters such as upper bounds on number of interruptions, number of consultations and total duration of service interruption of a customer at the main server. Service times at all the servers follow phase type distributions. In this chapter, we analyse a multi-server queueing model with $c + 1$ servers, namely a main server and $c$ identical regular servers. There is a common queue of customers. Service time at the main server follows phase type distribution and that at the regular servers follow i.i.d. exponential distribution. There are no
upper bounds on the number of interruptions to the main server and of consultations to the regular servers. We derive an explicit expression for the system stability. An expression for expected number of interruptions to a customer at the main server is derived. We discuss an optimization problem to determine the number of regular servers to be employed to maximize the expected total profit $ETP$. Some important performance measures are studied numerically.

5.1.1 Model description

Customers arrive according to a Poisson process with rate $\lambda$. An arriving customer enters into service immediately if at least one of the servers is free. Whenever the main server is free, the arriving customer will be served by the main server. The service time of the customers at the main server has phase type distribution with representation $(\alpha, T)$. The service times of the $c$ regular servers are i.i.d. exponential random variables with parameter $\mu$. If there are $i$ regular servers busy ($1 \leq i \leq c$), then the rate of requirement of consultation is $i\theta$. The request for consultation will be attended immediately. Then the main server (and hence the customer’s service at the main server) is said to be interrupted. However, if the main server is busy offering consultation to a regular server, any other regular server requiring consultation will be queued up. Thus, the regular servers are offered consultation on a FIFO basis by the main server. Note that at any given time there can be a maximum of $c$ servers requiring consulting work. The interrupted customer’s service will be resumed by the main server only after all consultations in the queue are completed. We assume that the duration of consultation is exponentially distributed
5.1. Introduction

with parameter $\xi$. Here the service of customers at the regular servers whose servers need consultation during their services is not considered to be interrupted since such consultations are considered to be a part of their services.

The main server in the system can be in any one of the following states:

1. Main server together with none, one or more regular servers are busy serving customers
2. Main server is idle and none, one or more regular servers are busy
3. Main server is giving consultation only
4. Main server is giving consultation with one interrupted customer

**Notations:** We use the following notations in this chapter.

- $f_i = e_i(1)$
- $\tilde{I}_i = \begin{bmatrix} I_i & 0 \end{bmatrix}_{i \times (i+1)}$
- $\omega = \theta + \mu$

Consider the queueing model $X = \{X(t), t \geq 0\}$, where $X(t) = \{N(t), S(t), J(t), K(t)\}$, where

- $N(t)$ – the number of customers in the system
- $J(t)$ – number of regular servers in the queue for consultation
126 Chapter 5. A multi-server queue with consultations

- $K(t)$—phase of service of the customer at the main server
  (the service of that customer may be under interruption)

Here $S(t)$ denotes the status of the servers at time $t$ such that
\[
S(t) = \begin{cases} 
    0, & \text{if the regular server(s) are busy and main server is idle} \\
    0, & \text{the main along with (or without) regular server is busy} \\
    1, & \text{if the main server is giving consultation only} \\
    2, & \text{if the main server is giving consultation} \\
    & \quad \text{with one interrupted customer at the main server}
\end{cases}
\]

Here $S(t)$ denotes the status of the servers at time $t$ such that
\[
S(t) = \begin{cases} 
    0, & \text{if the regular server(s) are busy and main server is idle} \\
    0, & \text{the main along with (or without) regular server is busy} \\
    1, & \text{if the main server is giving consultation only} \\
    2, & \text{if the main server is giving consultation} \\
    & \quad \text{with one interrupted customer at the main server}
\end{cases}
\]

$\{X(t), t \geq 0\}$ is a Continuous Time Markov Chain with state space
\[
\Psi = \{0\} \cup \bigcup_{i=1}^{\infty} \psi(i).
\]

The terms $\psi(i)$’s are defined as
\[
\psi(1) = \{(1, 0, t_1) : 1 \leq t_1 \leq a\} \cup \{(1, \tilde{0}) \cup (1, 1)\},
\]
\[
\psi(i) = \{(i, 0, t_1) : 1 \leq t_1 \leq a\} \cup \{(i, \tilde{0})\} \cup \{(i, 1, j) : 1 \leq j \leq i\}
\]
\[
\quad \cup \{(i, 2, j, t_1) : 1 \leq j \leq i - 1, 1 \leq t_1 \leq a\}, \text{ for } 2 \leq i \leq c,
\]
\[
\psi(i) = \{(i, 0, t_1) : 1 \leq t_1 \leq a\} \cup \{(i, 1, j) : 1 \leq j \leq c\}
\]
\[
\quad \cup \{(i, 2, j, t_1) : 1 \leq j \leq c, 1 \leq t_1 \leq a\}, \text{ for } i \geq c + 1.
\]
The infinitesimal generator $Q$ is given by

$$Q = \begin{bmatrix}
-\lambda & C_0 \\
D_1 & B_1 & C_1 \\
D_2 & B_2 & C_2 \\
& \ddots & \ddots & \ddots \\
D_c & B_c & C_c \\
D_{c+1} & A_1 & A_0 \\
& A_2 & A_1 & A_0 \\
& & \ddots & \ddots & \ddots
\end{bmatrix}$$

(5.1)

where $C_0 = \lambda \left[ \begin{array}{cc} \alpha & 0 \end{array} \right]_{1 \times (a+2)}$;

$C_1 = \lambda \left[ \begin{array}{ccc} I_a & 0 & 0 \\
\alpha & 0 & 0 \\
0 & 0 & \tilde{I}_1 \end{array} \right]_{(a+2) \times (2a+3)}$; where

$C_i = \lambda \left[ \begin{array}{cc} C_{i1} & 0 \\
0 & C_{i2} \end{array} \right]_{(1+i+ia) \times (i+2+(i+1)a)}$; for $2 \leq i \leq c - 1$, with

$C_{i1} = \left[ \begin{array}{c} I_a \\
\alpha \\
O \end{array} \right]_{(1+i+ia) \times a}$ and $C_{i2} = \left[ \begin{array}{c} O \\
diag(\tilde{I}_i, \tilde{I}_{i-1} \otimes I_a) \end{array} \right]_{(1+i+ia) \times (i+1+ia)}$;

$C_c = \lambda \left[ \begin{array}{c} I_a \\
diag(\alpha, I_c, \tilde{I}_{c-1} \otimes I_a) \end{array} \right]_{(1+c+ca) \times r}$;

$A_0 = \lambda I$;
Chapter 5. A multi-server queue with consultations

\[
B_1 = \begin{bmatrix}
T & -\omega & \theta \\
\xi & -\xi & \alpha
\end{bmatrix}_{a+2 \times a+2} - \lambda I;
\]

\[
B_i = \begin{bmatrix}
T - (i-1)\omega I_a & -\omega & (i-1)\theta I_a \\
\xi & -\xi & F_i \\
\xi F_{i-1} \otimes I_a & F_i \otimes I_a & F_{i-1} \otimes I_a
\end{bmatrix} - \lambda I,
\]

for \(2 \leq i \leq c;\)

\[
A_1 = \begin{bmatrix}
T - c \omega I_a & O & c\theta \xi \alpha I_a \\
\xi \theta f_c \otimes I_a & F_c \\
\xi \theta f_c \otimes I_a & F_c \otimes I_a
\end{bmatrix} - \lambda I, \text{ where}
\]

\[
F_i = \begin{bmatrix}
-(i-1)\omega & (i-1)\theta \\
\xi & -(i-2)\omega & (i-2)\theta \\
\xi & -(i-3)\omega & (i-3)\theta \\
& & \ddots & \ddots \\
& & & \xi & -\omega & \theta \\
& & & \xi & 0 & \theta \\
& & & & \ddots & \ddots
\end{bmatrix}_{i \times i} - \xi I_i;
\]

\[
D_1 = \begin{bmatrix}
T^0 \\
\mu \\
0
\end{bmatrix}_{(a+2) \times 1}; \quad D_2 = \begin{bmatrix}
\mu I_a & T^0 \\
2\mu & E_1 \\
O & O & O
\end{bmatrix}_{(3+2a) \times (a+2)};
\]

\[
D_i = \begin{bmatrix}
(i-1)\mu I_a & T^0 \\
i\mu & E_{i-1}
\end{bmatrix}_{(i+1+i) \times (i+(i-1)a)}.
\]
5.2 Steady state analysis

In this section we discuss the steady-state analysis of the model under study. We first establish the stability condition of the queueing system.

5.2.1 Stability condition

Let the steady-state probability vector of the generator \( A = A_0 + A_1 + A_2 \) be denoted by \( \pi \). That is,

\[
\pi A = 0
\]  

(5.2)
\[ \pi e = 1 \] (5.3)

The following theorem gives the stability of the queueing system under study.

**Theorem 5.3.1** : The Markov Chain X is stable if and only if

\[ \lambda < \frac{1}{\zeta} \left[ \mu_1 + \mu \sum_{i=1}^{c} \frac{c!}{(i-1)!} \left( \frac{\theta}{\xi} \right)^{c-i} \right] \] (5.4)

where \( \mu_1 \) and \( \mu \) are the service rates of the main and regular servers respectively and

\[ \zeta = c \sum_{i=0}^{c} \frac{c!}{(c-i)!} \left( \frac{\theta}{\xi} \right)^i. \] (5.5)

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

\[ \pi A_0 e < \pi A_2 e. \] (5.6)

Let \( \pi = (\pi_0, \pi_1, \pi_2) \), where
\( \pi_1 = (\pi_{11}, ..., \pi_{1c}) \) and \( \pi_2 = (\pi_{21}, ..., \pi_{2c}). \)

Using the structure of \( A \) and equation (5.2), it is easy to verify that

\[ \pi_1 = 0; \]

\[ \zeta \pi_{2i} = (c+1-i)\theta \pi_{2i-1}; \text{ for } 1 \leq i \leq c. \] (5.7)

Using equation (5.7) and the normalizing condition (5.3), it follows that

\[ \zeta \pi_0 = 1 \] (5.8)
5.2. Steady state analysis

where $\zeta$ is given in (5.5). Also we have

$$\pi A_2 e = \pi_0 [\mu_1 + \mu \sum_{i=1}^{c} \frac{c^i}{(i-1)!} \left( \frac{\theta}{\xi} \right)^{c-i}]$$

Then the stability condition (5.6) implies the result (5.4).

### 5.2.2 Expected number of interruptions to a customer at the main server

Since we are not imposing any upper bounds to the number of interruptions to a customer at the main server, we intend to find the expected number of interruptions before the service completion of a customer at the main server. For this, we consider the Markov process

$$Y(t) = \{(N_1(t), \hat{S}(t), J(t), K(t)) : t \geq 0\},$$

where $N_1(t)$ is the number of interruptions already befell to a customer at the main server.

$\hat{S}(t) = S(t) - \{0, 1\}$ and all other variables are as defined earlier.

The state space is

$$\{(i, 0, t_1) \cup (i, 2, j, t_1) : 1 \leq j \leq c, 1 \leq t_1 \leq a\} \cup \{\Delta\}.$$

The absorbing state $\Delta$ denote the customer at the main server leaves the system after service. Thus the infinitesimal generator $\tilde{V}$ of the process
Chapter 5. A multi-server queue with consultations

\[ Y(t) \text{ takes the form } \tilde{V} = \begin{bmatrix} V & V^0 \\ 0 & 0 \end{bmatrix}. \]

Here

\[ V = \begin{bmatrix} G_1 & G_0 \\ G_1 & G_0 \\ \vdots & \vdots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots \\ G_1 & G_0 \end{bmatrix}, \quad V^0 = \begin{bmatrix} G_2 \\ G_2 \\ \vdots \end{bmatrix}, \]

where \( G_2 = f_{c+1} \otimes T^0, \ G_1 = \text{diag}(T, O)_{(c+1)a \times (c+1)a} + \tilde{G}_1 \otimes I_a \) and \( G_0 = \xi \begin{bmatrix} 0 & 0 \\ I_c & 0 \end{bmatrix} \otimes I_a \), with

\[ \tilde{G}_1 = \begin{bmatrix} -c\theta & c\theta & (c-1)\theta & \ldots & \theta \\ -c\theta & \ldots & \ldots & \ldots & \theta \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -c\theta & \ldots & \ldots & \ldots & 0 \end{bmatrix}_{c+1 \times c+1} - \xi \text{diag}(0, I_c). \]

If \( z_j \) is the probability that there are exactly \( j \) interruptions during the service of a customer at the main server, then

\[ z_j = \eta(-G_1^{-1}G_0)^j(-G_1^{-1}G_2), \quad j = 0, 1, \ldots \]

where \( \eta = (\alpha, 0) \).

Expected number of interruptions during the service of a customer at the main server is given by,

\[ E(NI) = \sum_{j=0}^{\infty} j z_j. \]
5.2. Steady state analysis

5.2.3 Steady state probability vector

Let $x$, partitioned as, $x = (x_0, x_1, x_2, x_3, \ldots)$ be the steady state probability vector of the Markov chain $\{X(t), t \geq 0\}$. Note that $x_0$ is a scalar, $x_1 = (x_{10}, x_{1\tilde{0}}, x_{11})$, and $x_i = (x_{i0}, x_{i\tilde{0}}, x_{i1}, x_{i2})$, for $2 \leq i \leq c$ and $x_i = (x_{i0}, x_{i1}, x_{i2})$, for $i \geq c + 1$.

Here $x_{10}$ is an $a$-dimensional vector whereas $x_{1\tilde{0}}$ and $x_{11}$ are scalars.

$x_{i0}$, $x_{i\tilde{0}}$, $x_{i1}$, $x_{i2}$, for $2 \leq i \leq c$ are vectors of dimensions $a, 1, i, (i - 1)a$, for $2 \leq i \leq c$, and $x_{i0}$, $x_{i1}$, $x_{i2}$, for $i \geq c + 1$ are vectors of dimensions $a, c, ca$, respectively.

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

$$x_j = x_{c+1}R^{j-(c+1)}, \ j \geq c + 1,$$

(5.9)

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 $x_0, x_1, \ldots, x_{c+1}$ are obtained by solving the equations

$$-\lambda x_0 + x_1 D_1 = 0$$

$$x_{i-1}C_{i-1} + x_iB_i + x_{i+1}D_{i+1} = 0, \text{ for } i=1,\ldots,c-1$$

(5.10)

$$x_cC_c + x_{c+1}(A_1 + RA_2) = 0$$

subject to the normalizing condition

$$x_0 + x_1 e + \ldots + x_c e + x_{c+1}(I - R)^{-1}e = 1.$$  

(5.11)
5.2.4 Performance measures

Now we compute some performance measures.

(1) Expected number of customers in the system

\[ ES = \sum_{i=1}^{\infty} i \; x_i e. \]

(2) Expected number of customers in the queue

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

(3) Expected number of idle regular servers

\[ E(IS) = \sum_{i=1}^{c-1} (c - i) \; x_{i0} e + \sum_{i=1}^{c} (c + 1 - i) x_{i0} e. \]

(4) Effective rate of interruption

\[ EI = \sum_{i=2}^{c} (i \theta \; x_{i0} + \sum_{j=1}^{i-1} (i - j) \theta \; x_{i2j}) e + \sum_{i=c+1}^{\infty} (c \theta \; x_{i0} + \sum_{j=1}^{c-1} (c - j) \theta \; x_{i2j}) e. \]

(5) Effective rate of consultation

\[ ECo = EI + \theta \; x_{10} + \sum_{i=2}^{c} \sum_{j=1}^{i-1} (i - j) \theta \; x_{i1j} + \sum_{i=c+1}^{\infty} \sum_{j=1}^{c-1} (c - j) \theta \; x_{i1j}. \]
(6) Fraction of time the main server is idle

\[ F_{mi} = x_0 + \sum_{i=1}^{c} x_i \tilde{e}. \]

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

\[ F_{ab} = \sum_{i=c+1}^{\infty} x_i \tilde{e}. \]

### 5.2.5 An optimization problem

In this section we propose an optimization problem and discuss it through an illustrative example. To construct an objective function, we assume that the service completion produces a revenue to the system whereas idle regular servers and waiting spaces involve expenditure to the system. Thus we produce per unit time revenue and cost as follows:

1. \( r \) be the revenue per customer leaving the system after service completion
2. \( c_1 \) be the holding cost monetary of customers in the system
3. \( c_2 \) holding cost of idle regular servers

The problem of interest is to find an optimum value of the number of regular servers to be employed so that the expected total profit \( ETP \) is maximum. The objective function is given below:

\[ ETP = r \times ESR - c_1 \times ES - c_2 \times E(IS) \quad (5.12) \]
where $ESR = \pi A_2 e$.

## 5.3 Numerical examples

Now we present numerical results for implementing the qualitative nature of the model under study. The purpose of this example is to see the impact of parameter $c$. Here we consider that the service rate of the main server and that of the regular servers are equal. That is, we choose $T$ and $\alpha$ so that $[\alpha(-T)^{-1}e]^{-1} = \mu$.

Let $\lambda = 5$, $\theta = 9$, $\mu = 2$, $\xi = 3$, $T = \begin{bmatrix} -9 & 3 \\ 2 & -8 \end{bmatrix}$, $\alpha = \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix}$.

The above data of matrices, vectors and values satisfy the stability condition (5.4). Fix $r = 25$, $c_1 = 15$ and $c_2 = 100$.

<table>
<thead>
<tr>
<th>$c$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ESR$</td>
<td>9.8130</td>
<td>15.2409</td>
<td>20.3885</td>
<td>25.1093</td>
<td>29.5694</td>
<td>33.8892</td>
</tr>
<tr>
<td>$ES$</td>
<td>11.7206</td>
<td>3.9603</td>
<td>2.6137</td>
<td>2.1678</td>
<td>1.9604</td>
<td>1.8370</td>
</tr>
<tr>
<td>$E(IS)$</td>
<td>0.6429</td>
<td>2.1892</td>
<td>3.6003</td>
<td>4.8710</td>
<td>6.0460</td>
<td>7.1640</td>
</tr>
<tr>
<td>$ETP$</td>
<td>5.226</td>
<td>102.698</td>
<td>110.477</td>
<td>108.116</td>
<td>105.229</td>
<td>103.275</td>
</tr>
</tbody>
</table>

From the table 5.1 we can see that as $c$ increases the effective service rate $ESR$ increases. Since the services are done in a faster rate, accumulation of customers becomes less. Thus $ES$ decreases. If the number of regular servers increases, for a fixed $\lambda$, number of idle regular servers $E(IS)$ also increases as is to be expected.
Thus $ETP$ has an optimum value $110.477$ when the number of regular servers $c = 5$.

### 5.3.1 More numerical examples

In this section, we present some numerical examples that describe the performance characteristics of the queueing model under study. Let $c = 3$ and $\lambda, \theta, \mu, \xi, T$ and $\alpha$ are as given in the above example.

#### Table 5.2: Effect of $\theta$ on various performance measures

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ES$</td>
<td>0.8597</td>
<td>1.0946</td>
<td>1.4578</td>
<td>2.0588</td>
<td>3.1573</td>
</tr>
<tr>
<td>$EQ$</td>
<td>0.0531</td>
<td>0.1264</td>
<td>0.2809</td>
<td>0.6087</td>
<td>1.3436</td>
</tr>
<tr>
<td>$F_{mi}$</td>
<td>0.6444</td>
<td>0.5966</td>
<td>0.5401</td>
<td>0.4727</td>
<td>0.3915</td>
</tr>
<tr>
<td>$F_{ab}$</td>
<td>0.0081</td>
<td>0.0106</td>
<td>0.0137</td>
<td>0.0177</td>
<td>0.0227</td>
</tr>
</tbody>
</table>

#### Table 5.3: Effect of $\lambda$ on various performance measures

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ES$</td>
<td>1.9332</td>
<td>2.8234</td>
<td>4.1764</td>
<td>6.4072</td>
<td>10.266</td>
</tr>
<tr>
<td>$EQ$</td>
<td>0.4030</td>
<td>0.8879</td>
<td>1.8212</td>
<td>3.6097</td>
<td>6.9135</td>
</tr>
<tr>
<td>$F_{mi}$</td>
<td>0.4658</td>
<td>0.3805</td>
<td>0.2985</td>
<td>0.2197</td>
<td>0.1440</td>
</tr>
<tr>
<td>$F_{ab}$</td>
<td>0.0386</td>
<td>0.0667</td>
<td>0.1045</td>
<td>0.1516</td>
<td>0.2038</td>
</tr>
</tbody>
</table>

From the table 5.2, we see that as $\theta$ increases rate of consultation (and hence interruption) increases. So the customers have to spend longer time to get their services completed. Thus accumulation of customers in system
Table 5.4: Effect of $\mu$ on various performance measures

\[ \theta = 3 \]

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ES$</td>
<td>4.1764</td>
<td>2.7497</td>
<td>2.0696</td>
<td>1.6839</td>
<td>1.4400</td>
</tr>
<tr>
<td>$EQ$</td>
<td>1.8212</td>
<td>0.8359</td>
<td>0.4485</td>
<td>0.2660</td>
<td>0.1693</td>
</tr>
<tr>
<td>$F_{mi}$</td>
<td>0.2985</td>
<td>0.3683</td>
<td>0.4130</td>
<td>0.4434</td>
<td>0.4651</td>
</tr>
<tr>
<td>$F_{ab}$</td>
<td>0.1045</td>
<td>0.0625</td>
<td>0.0400</td>
<td>0.0271</td>
<td>0.0192</td>
</tr>
</tbody>
</table>

Table 5.5: Effect of $\xi$ on various performance measures

\[ \theta = 6, \mu = 2.5 \]

<table>
<thead>
<tr>
<th>$\xi$</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$EQ$</td>
<td>8.6051</td>
<td>7.6826</td>
<td>3.4795</td>
<td>1.8598</td>
<td>1.1410</td>
</tr>
<tr>
<td>$F_{mi}$</td>
<td>0.0283</td>
<td>0.1303</td>
<td>0.2095</td>
<td>0.2662</td>
<td>0.3087</td>
</tr>
<tr>
<td>$F_{ab}$</td>
<td>0.0556</td>
<td>0.1018</td>
<td>0.0909</td>
<td>0.0813</td>
<td>0.0743</td>
</tr>
</tbody>
</table>

and queue happens in a faster rate, which results in increased number of $ES$ and $EQ$. So the fraction of time all servers are busy serving customers $F_{ab}$ increases. Naturally, $F_{mi}$ will decrease.

Table 5.3 shows that as $\lambda$ increases, the server is fed with customers more frequently and so $ES$ and $EQ$ increase. Busy time of each server increases, therefore $F_{ab}$ increases and thus idle time of main server $F_{mi}$ decreases.

From table 5.4, we see that as $\mu$ increases, regular servers serve cus-
tomers in a faster rate. Thus there is a slow accumulation of customers in system and in queue which results in decrease in $ES$ and in $EQ$. As the customers get served in a faster rate at the regular servers, less number of customers approach the main server. So the main server gets more idle time, i.e, $F_{mi}$ increases. So as a whole, the fraction of time all servers are busy $F_{ab}$ decreases.

We see from table 5.5 that an increase in $\xi$ results in a faster rate of consultation completion. So the servers get more time to serve customers and so the accumulation of customers in the system and in the queue decrease. Thus $ES$ and $EQ$ decrease. As larger number of customers are served by the regular servers, main server gets more idle time. So $F_{mi}$ increases. Now we consider $F_{ab}$. We can see that $F_{ab}$ increases until $\xi=3.5$. If again $\xi$ increases, since the value of $\lambda$ is fixed, the servers need less amount of time for service completion and therefore the fraction of time all servers are busy serving customers $F_{ab}$ will decrease.