Queuing Theory and its Applications

A PROJECT REPORT ON THE TOPIC OF ADVANCE OPERATION RESEARCH

Prepared by
DEBASIS MOHANTY (UEBA19001)

05 Aug 2020

Advance Operation Research
Term IV, EMBA – BA

Contents

Queueing Theory: An Introduction

Queueing theory is the mathematical study of waiting in lines, or queues. Queueing theory, along with simulation, are the most widely used operations-research and management-science techniques. Its main objective is to build a model to predict queue lengths and waiting times to make effective business decisions related to resources' management and allocation to provide a given service.

In this project, I will discuss about Queueing Theory and some of its practical applications. We all have experienced the annoyance of having to wait in a queue. We wait in line at supermarkets to check out, we wait in line in banks and post offices and we wait in line at fast food restaurants. But we as customers do not like waiting. And the managers of these establishments also don't like their customers to wait as it may cost them business.

So, the first question that arises is that "Why is there waiting?"

To which the answer is that there is more demand for service than there is an available facility for that service.

And "Why is this so?"

For which there could be a number of reasons such as, shortage of available servers, limitation of space, economic limitations etc.

These limitations can be removed with the expenditure of capital. And to know how much service should then be made available, one needs to know:

1. How many people will form the queue?
2. How long must a customer wait?

Queuing theory attempts to answer these questions through detailed mathematical analysis.

GOAL: The ultimate goal is to achieve an economic balance between cost of service and the cost associated with the waiting for that service.

Basic Structure of a Queuing System

Please see **Exhibit 1**.

Input Source (Calling population): The size is the total number of customers that might require service from time to time, i.e., the total number of distinct potential customers. This population from which arrivals come is referred to as the calling population.

Arrival process: It describes how the customers arrive to the system, and the distribution of the customers' arrival.

Service mechanism: It is articulated by the number of servers, and whether each server has its own queue or there is one queue feeding all servers, and the distribution of customer's service times

Queue discipline: It refers to the rule that a server uses to choose the next customer from the queue when the server completes the service of the current customer (e.g. *FIFO*: first-in, first-out; *LIFO*: last-in, first-out; priority-based; random selection).

Service Mechanism: It refers at a given facility, the customer enters one of the parallel service channels and is completely serviced by that server. A queueing model must specify the arrangement of the facilities

and the number of servers (parallel channels) at each one. Most elementary models assume one service facility with either one server or a finite number of servers.

The time elapsed from the commencement of service to its completion for a customer at a service facility is referred to as the **service time** (or holding time).

An Elementary Queuing Process

A basic business Queue model have two approaches:

- Analytical Model: The one which is built on differential equations which has arrivals and services.
- Simulation: Complex system with multiple stations (For e.g. CAT Exam, Air Traffic controller)

The basic goal as stated earlier is to minimize the service and waiting costs, to reduce the waiting time of customers and optimize the service time because the resources are limited. Please see **Exhibit 2**.

As we have already suggested, queueing theory has been applied to many different types of waiting-line situations. However, the most prevalent type of situation is a single waiting line (which may be empty at times) forms in the front of a single service facility, within which are stationed one or more servers. Each customer generated by an input source is serviced by one of the servers, perhaps after some waiting in the queue (waiting line). We can see **Exhibit 3** for a depiction.

Kendall's Notations

The notations is given by M/M/s/K where, 1st two M stands for Markovian probability distributions, follow by s which stands for number of servers. The K stands for Maximum Queue length or the Queue capacity.

We have other queue models like M/D/s, M/G/s and M/E_k/s were D is Deterministic, G is General (normal) and E_k is Erlang Distribution.

M = exponential distribution (Markovian)

D = degenerate distribution (constant times)

G = general distribution (any arbitrary distribution allowed)

For example, the M/M/s model, assumes that both interarrival times and service times have an exponential distribution and that the number of servers is s (any positive integer). The M/G/1 model discussed assumes that interarrival times have an exponential distribution, but it places no restriction on what the distribution of service times must be, whereas the number of servers is restricted to be exactly 1.

Terminology, Notations and Formulas:

Unless otherwise noted, the following standard terminology and notation will be used:

State of system = number of customers in queueing system.

Queue length = number of customers waiting for service to begin. OR (state of system minus number of customers being served.)

$\boldsymbol{N(t)}$ = number of customers in queueing system at time t ($t \geq 0$).

$\boldsymbol{P_n(t)}$ = probability of exactly n customers in queueing system at time t.

s = number of servers (parallel service channels) in queueing system.

$\boldsymbol{\lambda_n}$= mean arrival rate (expected number of arrivals per unit time) of new customers when n customers are in system.

$\boldsymbol{\mu_n}$ = mean service rate for overall system (expected number of customers completing service per unit time) when n customers are in system. Note: $\mu_n$represents combined rate at which all busy servers (those serving customers) achieve service completions.

$\boldsymbol{\rho}$ = overall system utilization. $\rho = \frac{\lambda}{s\mu}$ (utilization factor)

$\boldsymbol{L_q}$ = expected number of customers in queue

$\boldsymbol{L_s}$ = expected number of customers in the system

$\boldsymbol{W_q}$ = expected time spent in the queue

$\boldsymbol{W_s}$ = expected time spent in the system

Relationships between L,W, $\boldsymbol{L_q}$, and $\boldsymbol{W_q}$

Assume that λ_n is a constant λ for all n. It has been proved that in a steady-state queueing process,

$L = \lambda W$ (Because John D. C. Little provided the first rigorous proof, this equation sometimes is referred to as Little's formula.) Furthermore, the same proof also shows that $\boldsymbol{L_q = \lambda W_q}$. Under these circumstances, $1/\lambda$ and $1/\mu$ are the *expected interarrival time* and the *expected service time*, respectively. The arrival process is assumed to follow Poisson Distribution and the service rate is assumed to follow a and Exponential distribution. Please see **Exhibit 4**.

If the λ_n are not equal, then λ can be replaced in these equations by $\bar{\lambda}$, the average arrival rate over the long run. (We shall show later how $\bar{\lambda}$ can be determined for some basic cases.)

Now assume that the mean service time is a constant, $1/\mu$ for all $n \geq 1$. It then follows that

$W_s = W_q + (1/\mu)$

These relationships are extremely important because they enable all four of the fundamental quantities L,W, L_q, and W_q to be immediately determined as soon as one is found analytically. This situation is fortunate because some of these quantities often are much easier to find than others when a queueing model is solved from basic principles.

Queue Formulas for M/M/1 (Performance Measures):

Average Arrival rate : λ

Average Service rate : μ

Average number of customers in service: $\rho = \frac{\lambda}{\mu}$

Probability of exactly "n" customers in the system: $P_n = \rho^n(1-\rho)$

Probability of "k" or more customers in the system: $P(n \geq k) = \rho^k$

Average number of customers in the system: $L_s = [\lambda/(\mu - \lambda)]$

Average number of customers in the queue: $L_q = L_s - (\lambda/\mu)$ OR $[\rho\lambda/(\mu - \lambda)]$

Average time in the system: $W_s = [1/(\mu - \lambda)]$

Average time in the queue: $W_q = W_s - (1/\mu)$ OR $[\rho/(\mu - \lambda)]$

Queue Formulas for M/M/s (Performance Measures):

Average server utilization: $\rho = \frac{\lambda}{s\mu}$

Average number of customers in queue: $L_q = \frac{\left(\left(\frac{\lambda}{\mu}\right)^s\right)\lambda\mu}{(s-1)!\,(s\mu-\lambda)}$

Average number of customers in system: $L = L_q + \frac{\lambda}{\mu}$

Average number of customers spends waiting in the queue: $W_q = \frac{L_q}{\lambda}$

Average number of customers spends waiting in the system: $W = W_q + \frac{1}{\mu}$

Probability of zero customers: $P_0 = \frac{1}{\sum_{k=0}^{s-1}\left(\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s + \frac{s\mu}{s\mu-\lambda}\right.}$

Probability of n customers in system: $P_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0 \; for \; n \leq s$

$$P_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{s!\,s^{(n-s)}} \; P_0 \; for \; n > s$$

Problem Description

PROBLEM STATEMENT I

The County Hospital Example with the M/M/s Model. For the County Hospital emergency room problem, the management engineer has concluded that the emergency cases arrive pretty much at random (a Poisson input process), so that interarrival times have an exponential distribution. She also has concluded that the time spent by a doctor treating the cases approximately follows an exponential distribution. Therefore, she has chosen the M/M/s model for a preliminary study of this queueing system. By projecting the available data for the early evening shift into next year, she estimates that patients will arrive at an average rate of 1 every 1/2 hour. A doctor requires an average of 20 minutes to treat each patient. What is the queue performance at steady state? Do we require one more doctor?

To answer this the management engineer did his primary analysis and using the above queuing formula and found out the performance measure. As per the server utilization is concern he/she was pretty much convinced that the utilization percentage came down from 67% to 33% (Please see **Exhibit 5**), however he/she wanted to further analyse the emergency situation to give an all round report to the management as everyone might question him because it is obvious that whenever you increase resource then there utilization gets divided.

PROBLEM STATEMENT II

The County Hospital Example with Priorities. For the County Hospital emergency room problem, the management engineer has noticed that the patients are not treated on a first-come-first-served basis. Rather, the admitting nurse seems to divide the patients into roughly three categories: (1) critical cases, where prompt treatment is vital for survival; (2) serious cases, where early treatment is important to prevent further deterioration; and (3) stable cases, where treatment can be delayed without adverse medical consequences. Patients are then treated in this order of priority, where those in the same category are normally taken on a

first-come-first-served basis. A doctor will interrupt treatment of a patient if a new case in a higher-priority category arrives. Approximately 10 percent of the patients fall into the first category, 30 percent into the second, and 60 percent into the third. Because the more serious cases will be sent to the hospital for further care after receiving emergency treatment, the average treatment time by a doctor in the emergency room actually does not differ greatly among these categories. The management engineer has decided to use a priority-discipline queueing model as a reasonable representation of this queueing system, where the three categories of patients constitute the three priority classes in the model. Because treatment is interrupted by the arrival of a higher-priority case, the preemptive priorities model is the appropriate one. Given the previously available data (μ=3 and λ=2), the preceding percentages yield λ_1=0.2, λ_2=0.6 *and* λ_3=1.2. Calculate the expected waiting times in the queue (so excluding treatment time) for the respective priority classes when there is one (s=1) or two (s=2) doctors on duty. Also find out the corresponding results for the preemptive and non preemptive priorities model to show the effect of preempting.

Modelling approach

Priority-Discipline Queues

In priority-discipline queueing models, the queue discipline is based on two priority system.

- Static : Doesn't change throughout the execution
- Dynamic: Changes after some time

We have two models based on priority system. They are:

- Non-Preemptive Priorities model
- Preemptive Priorities model

With non-preemptive priorities, a customer being served cannot be ejected back into the queue (preempted) if a higher-priority customer enters the queueing system. Therefore, once a server has begun serving a customer, the service must be completed without interruption. With preemptive priorities, the lowest-priority customer being served is preempted (ejected back into the queue) whenever a higher-priority customer enters the queueing system. A server is thereby freed to begin serving the new arrival immediately.

When a server does succeed in finishing a service for a higher priority customer, so a preempted customer normally will get back into service again and will eventually finish its treatment after waiting again in the queue. Both the queue model will follow queue discipline of FIFO according to its priority.

Non-Preemptive Priorities model

To calculate the waiting time let W_k be the steady state expected waiting time in the system (including service time) for a member of priority class k. Then

$W_k = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu}$ for k = 1,2, . . . , N

Where $A = s! \frac{s\mu - \lambda}{\rho^s} \sum_{j=0}^{s-1} \frac{\rho^j}{j!} + s\mu$

$B_0 = 1$,

$B_k = 1 - \frac{\sum_{i=1}^{k} \lambda_i}{s\mu}$

Using the above equations, we can create a excel model or any functions which can solve the purpose. However, the underlying algorithm in shown in **Exhibit 6.**

Preemptive Priorities model

Similarly, for preemptive model waiting time let W_k be the steady state expected waiting time in the system and it is expressed as per below formula:

$$W_k = \frac{1/\mu}{B_{k-1}B_k}$$

Using the above equations, we can create a excel model, however, the underlying algorithm in shown in **Exhibit 7.**

Problem Analysis

Analysis

Steady-state results from the priority-discipline models for the County Hospital problem can be solved by using the excel templates available in the CD resources provided along with our book.

Using the sheet, we can compute all the four fundamental quantities L,W, L_q, W_q and the resource utilization. The details are present in the **Exhibit 8** and **Exhibit 9** tables.

However, for the preemptive queue model specific the probabilities of arrivals are provided which is not considered when run in the excel model. As the excel formulas treat the arrivals as a Poisson distribution. So, manually subtracting the expected service time $1/\mu$ from each waiting time of the queue calculating the cumulative arrival rate using the model used in the previous problem will do the needful. The calculation is done below for each priority class separately.

Priority Class 1 (10% probability):

Arrival rate (λ)	0.2
Service rate (μ)	3
Number of servers (s)	2
Average time in the system (W_1)	**0.33370**

Waiting time = $W_1 - \frac{1}{\mu}$ = 0.33370 - 0.33333 = 0.00037

Priority Class 2 (30% probability):

Arrival rate (λ)	0.8
Service rate (μ)	3
Number of servers (s)	2
Average time in the system (W_1)	**0.33937**

$W_{1-2} = \frac{1}{4}W_1 + \frac{3}{4}W_2$ so $W_2 = \frac{4}{3}\left[0.33937 - \frac{1}{4}(0.33370)\right] = 0.34126$

Waiting time = $W_2 - \frac{1}{\mu}$ = 0.34126 - 0.33333 = 0.00793

Priority Class 3 (60% probability):

Arrival rate (λ)	2
Service rate (μ)	3
Number of servers (s)	2
Average time in the system (W_1)	**0.375**

$W_{1-3} = 0.1W_1 + 0.3W_2 + 0.6W_3$ so, $W_3 = \frac{1}{0.6}[0.375 - 0.1(0.33370) - 0.3(0.34126)] = 0.39875$

Waiting time = $W_3 - \frac{1}{\mu}$ = 0.39875 - 0.33333 = 0.06542

Please **Exhibit 10** for the updated results for the preemptive priorities case model.

Conclusion

When s=1, the $W_k - 1/\mu$ values for the preemptive priorities case indicate that providing just a single doctor would cause critical cases to wait about 1 1/2 minutes (0.024 hour) on the average, serious cases to wait more than 9 minutes, and stable cases to wait more than 1 hour. (Contrast these results with the average wait of W_q = 2/3 hour for all patients that was obtained in table for the first model under the first-come-first-served queue discipline.) However, these values represent statistical expectations, so some patients have to wait considerably longer than the average for their priority class. This wait would not be tolerable for the critical and serious cases, where a few minutes can be vital. By contrast, the s=2 results in above table (preemptive priorities case) indicate that adding a second doctor would virtually eliminate waiting for all but the stable cases. Therefore, the management engineer recommended that there be two doctors on duty in the emergency room during the early evening hours next year. The board of directors for County Hospital adopted this recommendation and simultaneously raised the charge for using the emergency room!

Cost Benefit Analysis

The management engineer also calculated total cost to strengthen its above conclusion. So, we could see that when we have two doctor it does costs less considering the waiting cost to be $100 per hour. Please see **Exhibit 11** for the details.

Some other analysis

A simulation using R simmer package is also conducted for better understanding of the above problem.

The details of the simulation result in provided in the **Exhibit 12** and **Exhibit 13**.

The kink in the usage curve shows that the time consumed by the server in emptying the customers for the sake of priority ones.

Exhibits

Exhibit 1

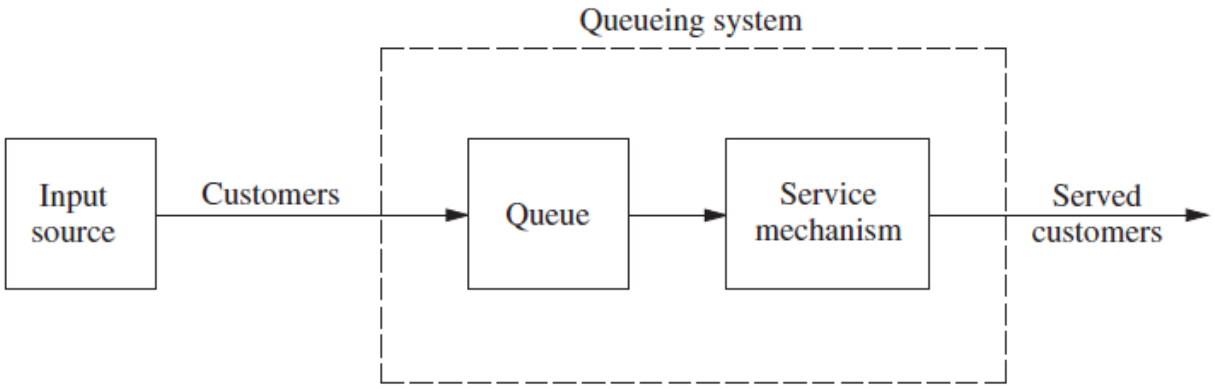
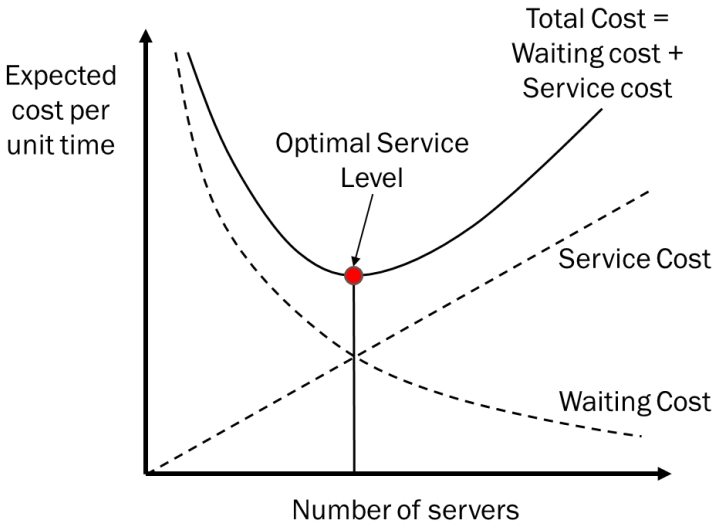


Exhibit 2



Total Cost = Waiting Cost + Service Cost

$$C_w * L \quad + \quad C_s * s$$

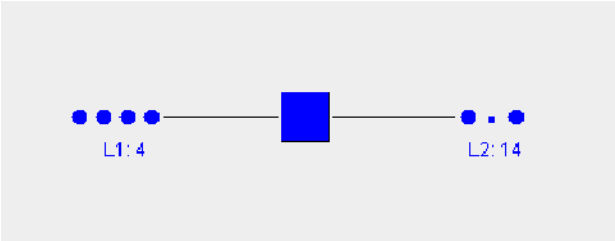
$C_w = Waiting\ cost\ per\ unit\ time$

$C_s = Cost\ per\ server\ per\ unit\ time$

$L = Average\ number\ of\ customers\ in\ system$

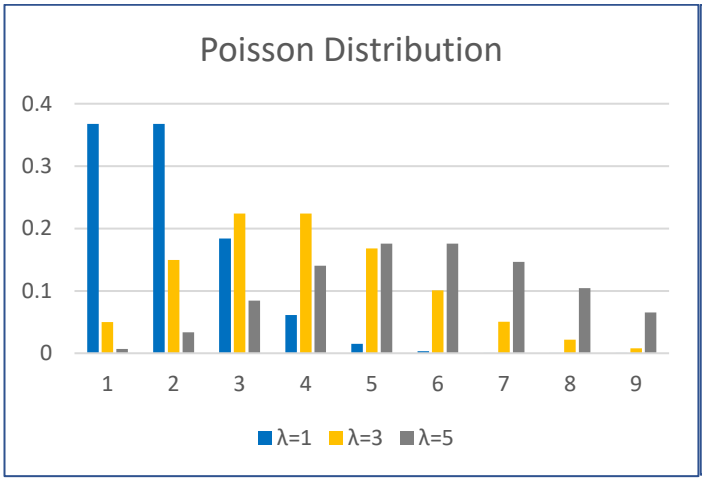
$s = Number\ of\ servers$

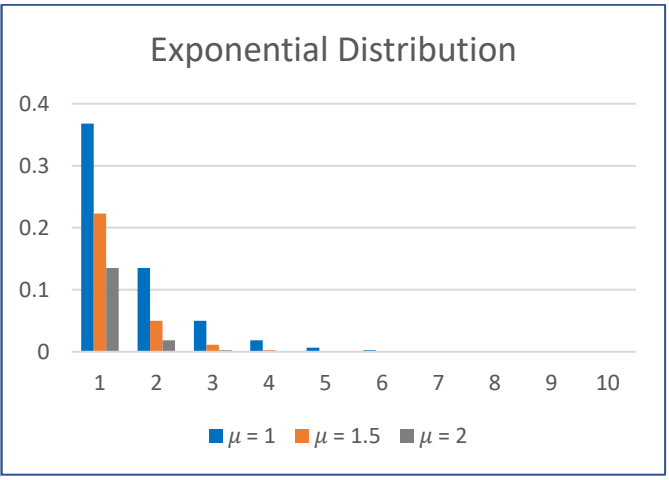
Exhibit 3



A single server queue model (M/M/1)

Exhibit 4





Given an average arrival rate λ, the Poisson distribution gives the probability a certain number of customer x arrive in a given time. P(x) = $\frac{e^{-\lambda}\lambda^{x}}{x!}$

Given an average service time μ, the Exponential distribution gives the probability of service provided by a server at time t. P(x) =$e^{-\mu t}$

Exhibit 5

Queue Performance/Operating Characteristics at steady state	s = 1	s = 2	measure
Average server utilization (ρ)	67%	33%	busy
Average number of customers in the queue (Lq)	1.3333	0.0833	in queue
Average number of customers in the system (L)	2.0000	0.7500	in system
Average waiting time in the queue (Wq)	0.6667	0.0417	hour
Average time in the system (W)	1.0000	0.3750	hour
Probability (% of time) system is empty (Po)	0.3333	0.5000	empty

Exhibit 6

Non-Preemptive Priority Queue Algorithm						
Patients	Priority	Arrival Time	Burst Time	System Time (Completion Time)	Turnaround Time (ST - AT)	Waiting Time (TT - BT)
P1	3	0	8	8	8	0
P2	4	1	2	17	16	14
P3	4	3	4	21	18	14
P4	5	4	1	22	18	17
P5	2	5	6	14	9	3
P6	6	6	5	27	21	16
P7	1	10	1	15	5	4
Total				124	95	68

Queue	P1	P5	P7	P2	P3	P4	P6
0	8	14	15	17	21	22	27

Exhibit 7

Preemptive Priority Queue Algorithm							
Patients	Priority	Arrival Time	Burst Time	New BT	System Time (Completion Time)	Turnaround Time (ST - AT)	Waiting Time (TT - BT)
P1	3	0	8	3	15	15	7
P2	4	1	2		17	16	14
P3	4	3	4		21	18	14
P4	5	4	1		22	18	17
P5	2	5	6	1	12	7	1
P6	6	6	5		27	21	16
P7	1	10	1		11	1	0
Total					125	96	69

Queue	P1	P5	P7	P5	P1	P2	P3	P4	P6
0	5	10	11	12	15	17	21	22	27

Exhibit 8

Queue Performance/Operating Characteristics at a steady state	Preemptive Priorities					
	s = 1			s = 2		
Priority Classes (λ1 = 0.2, λ2 = 0.6, λ3 = 1.2)	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Average server utilization (ρ)	67%	67%	67%	33%	33%	33%
Average number of customers in the queue (Lq)	0.00476	0.09221	1.23636	0.00230	0.03873	0.29231
Average number of customers in the system (L)	0.07143	0.29221	1.63636	0.06897	0.23873	0.69231
Average waiting time in the queue (Wq)	0.02381	0.15368	1.03030	0.01149	0.06454	0.24359
Average time in the system (W)	0.35714	0.48701	1.36364	0.34483	0.39788	0.57692

Exhibit 9

Queue Performance/Operating Characteristics at a steady state	Non-Preemptive Priorities					
	s = 1			s = 2		
Priority Classes (λ1 = 0.2, λ2 = 0.6, λ3 = 1.2)	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Average server utilization (ρ)	67%	67%	67%	33%	33%	33%
Average number of customers in the queue (Lq)	0.04762	0.19481	1.09091	0.00575	0.01989	0.05769
Average number of customers in the system (L)	0.11429	0.39481	1.49091	0.07241	0.21989	0.45769
Average waiting time in the queue (Wq)	0.23810	0.32468	0.90909	0.02874	0.03316	0.04808
Average time in the system (W)	0.57143	0.65801	1.24242	0.36207	0.36649	0.38141

Exhibit 10

Queue Performance/Operating Characteristics at a steady state	Preemptive Priorities		
	s = 2		
Priority Classes (λ1 = 0.2, λ2 = 0.6, λ3 = 1.2)	Class 1	Class 2	Class 3
Average server utilization (ρ)	33%	33%	33%
Average number of customers in the queue (Lq)	0.00230	0.03873	0.29231
Average number of customers in the system (L)	0.06897	0.23873	0.69231
Average waiting time in the queue (Wq)	0.00037	0.00793	0.06542
Average time in the system (W)	0.34483	0.39788	0.57692

Exhibit 11

Input Data		
Arrival rate (λ)	2	hour
Service rate (μ)	3	hour
Cost of Service	100	hour
Cost of Waiting	100	hour

Total Cost based on Queue			
Servers	Service Cost	Wait Cost	Total Cost
1	100.00	200.00	300.00
2	200.00	75.00	275.00
3	300.00	67.60	367.60
4	400.00	66.77	466.77
5	500.00	66.68	566.68

Exhibit 12

Preemptive QUEUE OUTPUT

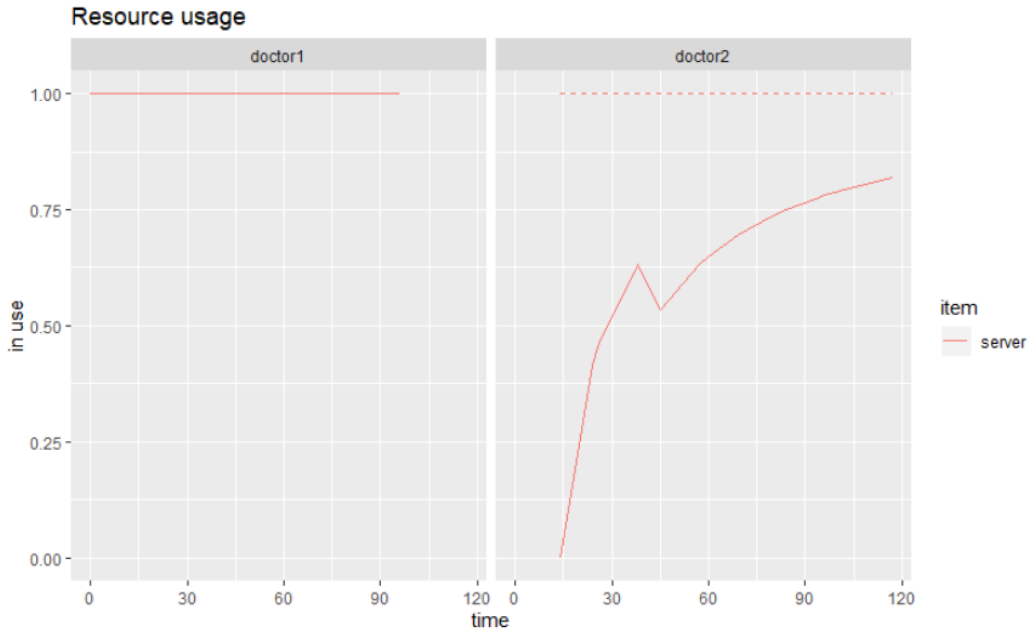
SlNo.	name	start_time	end_time	activity_time	finished	replication	waiting_time
1	John0	2.00	26	24	TRUE	1	0.00
2	Patient0	0.00	38	24	TRUE	1	14.00
3	Jimmy0	33.00	57	24	TRUE	1	0.00
4	Maria0	43.00	69	24	TRUE	1	2.00
5	Patient1	0.18	81	24	TRUE	1	56.82
6	Patient2	8.64	93	24	TRUE	1	60.36
7	Patient3	21.13	105	24	TRUE	1	59.87
8	Patient4	28.09	117	24	TRUE	1	64.91

Non - Preemptive QUEUE OUTPUT

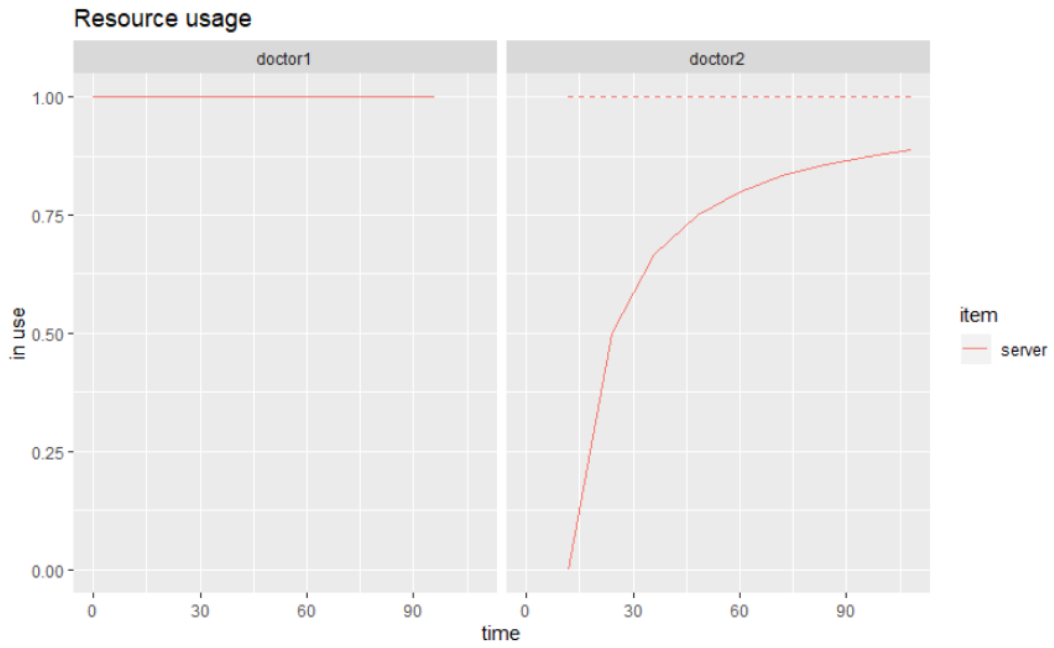
SlNo.	name	start_time	end_time	activity_time	finished	replication	waiting_time
1	Patient0	0.00	24	24	TRUE	1	0.00
2	John0	2.00	36	24	TRUE	1	10.00
3	Patient1	0.18	48	24	TRUE	1	23.82
4	Jimmy0	33.00	60	24	TRUE	1	3.00
5	Maria0	43.00	72	24	TRUE	1	5.00
6	Patient2	8.64	84	24	TRUE	1	51.36
7	Patient3	21.13	96	24	TRUE	1	50.87
8	Patient4	28.09	108	24	TRUE	1	55.91

Exhibit 13

Preemptive QUEUE RESOURCE USAGE



Non - Preemptive QUEUE RESOURCE USAGE



Bibliography

1. Problem Statement taken from Introduction to Operation Research (10th Edition) by Frederick S. Hillier, Gerald J. Lieberman, Bodhibrata Nag, Preetam Basu.
2. The materials for analysis is taken from the CD resource provided along with the above book.
3. R studio to run simulation using simmer package.