MANAGING SUPPLY IN THE INDIAN CONTEXT: FLEXIBLE WORKERS, FULL-TIME EMPLOYEES AND FREELANCERS

A Project Report Submitted

for the Course

MA498 Project I

by

Manne Hema Priya

(Roll No. 170123032)

Tanvi Ohri

(Roll No. 170123051)

to the

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

November 2020

CERTIFICATE

This is to certify that the work contained in this report entitled **"Managing Supply in the Indian Context: Flexible Workers, Full-Time Employees And Freelancers"** submitted by **Manne Hema Priya** (**Roll No: 170123032**) and **Tanvi Ohri** (**Roll No: 170123051**) to Department of Mathematics, Indian Institute of Technology Guwahati towards partial requirement of Bachelor of Technology in Mathematics and Computing has been carried out by them under my supervision.

It is also certified that this report is a survey work based on the references in the bibliography.

Turnitin Similarity: 18%

Guwahati - 781 039   (Dr. N.Selvaraju)

November 2020   Project Supervisor

ABSTRACT

The main aim of the project is to adapt the model proposed by Dong and Ibrahim (2020) [2] for the Indian market. There is an important extension we propose for this. With the increasing trend of startups in India, we feel we need to include another type of employee- the freelancer. The flexible employee is hired for a short interval, say when we typically expect higher traffic or a certain kind of work. The freelancers are hired for a particular task, that is, the profession's or startup's area of expertise. This is delegation of a task, it need not have a certain minimum duration. Freelancers have the freedom to decide their work schedules. There is a rise in the number of firms that are staffing a blended workforce. A blended workforce is a clever business strategy of employing some flexible workers with full-time employees, and now, freelancers as well. An optimal staffing strategy is to be devised to combat supply-side uncertainty and staffs freelancers, flexible workers and full-time employees. The strategy must take into account both quality and efficiency factors. In this phase of the project, we survey some of the preliminary works in this area. We study staffing under parameter uncertainty [1] and staffing with self-scheduling servers [6].

Contents

List of Figures

List of Tables

Chapter 1

Key Concepts

This section has been added to cater to a reader who does not have prior knowledge about queuing theory .It includes some basic queuing models, approximations, laws and problems that have been used throughout this project.

1.1   Basics of Queuing Theory

The main reason for the usage of queues in modelling service systems is because the resources are limited. Hence it leads to the formation of queues.In these queuing systems, we often see a trade-off between the cost of maintaining the system and customer service, while modelling there is an attempt to optimize on both ends .To analyze any queuing system, we need to know the following parameters.

1.1.1   Arrival Process

We look at this in terms of

- Inter arrival times of different customers

- How the customers arrive (individually or in batches)

- Whether the number of potential customers (also known as calling population) is finite or infinite.

The arrival times could be random or scheduled, in random arrival times one of the important models is the exponential arrival process where the difference between two arrivals is distributed exponentially in a Poisson distribution

1.1.2   Service Mechanism

We look at this in terms of

- The distribution of service times

- No. of servers in the system and how that changes

- Relation between servers in different service centers and preemption in processes.

In general, we assume that the service times are independent of other factors and that they are exponential. We also assume that the number of servers is fixed.

1.1.3   Queue Characteristics

- **Queue Capacity**

  The queue capacity can be finite or infinite.

- **Queue Behaviour**

  The actions of customers when they are waiting in the queue for the service to begin .There are various types of behaviours like

  - **Balk**: A behaviour in which the customers won't join the system when the queue is too long

  - **Renege or Abandon**: A behaviour in which the customers are impatient and leave before getting served.

  - **Jockey**: A behaviour in which the customers move from one queue to a shorter queue

  Customer abandonment or impatient customers is a important factor which we look at when modelling queuing systems.The distribution of customer abandonment times is also taken into consideration when optimizing a system.

- **Queue Discipline**

  The algorithm which determines which customer gets served when a server gets free, there are various algorithms, with their own merits and demerits . For example,

  - Service according to priority (PR).

  - Shortest processing time first (SPT)

- Last in first out (LIFO)

- First in first out or First come first serve (FIFO)

- Customers are served in a random orders (SIRO)

1.1.4 Kendall's Notation

In this notation , any queuing system can be represented as $A/B/c/N/K/D$ where

D is the queue discipline

K is the size of potential number of customers in the system

N is the capacity of the queue

c represents the number of servers in the system

B represents the distribution of service process

A represents the distribution of arrival process

In general , when not specified the last three parameters N, K, D are taken as ∞, ∞, FIFO respectively

For exponential distribution, we use M and for the general distribution we use G, for the arrival and service time distribution.

1.2 Performance Measures

For analysing the performance of a queuing system , we need to know the following values

- Average waiting time

- Average occupancy of queue

- Server utilization

- Probability of waiting time exceeding a given time

- Probability of queue occupancy exceeding a given occupancy

- Service time of nth customer

- Arrival time between n^{th} customer and $(n-1)^{th}$ customer

- Parameters of the system in the long run

The answers to the above questions can help us in designing a optimal system to minimize some cost in the system.

1.3   Some results for M/M/1 system

Consider a model with poisson distributed customer arrival (arrival rate: λ) and exponential service times (rate: μ).

The inter-arrival time distribution is $p_\lambda(t) = \lambda e^{-\lambda t}$

The service time distribution is $p_\mu(t < t_0) = 1 - e^{-\mu t_0}$

For a stable system , the rate of customer arrival should be less than or equal to the rate at which they are being processed $\lambda < \mu$

We define system utilization as $\rho = \frac{\lambda}{\mu} = P[system\,is\,busy]$

Let's define S_n as the state with n customers in the system ($n-1$ in queue , the nth being served) and let $p_n = P[n\,customers\,in\,the\,system]$, for steady state. Then,

$rate_{in} = rate_{out} \to p_n \lambda = p_{n+1} \mu \to p_{n+1} = p_0 \rho^{n+1}$

Since $p_0 = 1 - \rho$

$$P[n\,customers\,in\,the\,system] = p_n = \rho_n(1 - \rho)$$

To find the average number of customers in the system,

$$\bar{N} = \sum k.(\rho_k(1 - \rho)) = (1 - \rho)\sum k\rho_k = \frac{\rho}{1 - \rho} \quad [7] \tag{1.1}$$

Average system time is given by,

$$T = \frac{\bar{N}}{\lambda} \; (Little's\,Result) = \frac{1}{\mu - \lambda} \tag{1.2}$$

For more results and further explanation , refer *Fundamentals of Queuing Systems* [7].

1.4   The Fluid Approximation

In most practical scenarios, the value of number of arrivals is quite large with respect to the time period under consideration. For such systems, customer flow may be approximated as the flow of a continuous fluid. This means we don't consider customers as discrete entities.

Let's imagine the fluid queue's physical analogue to gain better understanding.

Arrivals to the queue (Arrival Rate: λ) $\leftrightarrow$ Water flowing out of a tap

Server $\leftrightarrow$ Drain

Queue $\leftrightarrow$ Water in the sink

Processing Capacity of one server (Service Rate: μ) $\leftrightarrow$ Maximum flow rate through one drain

What does this mean w.r.t. overcrowding in the queue ($\leftrightarrow$ accumulation of water in the sink)?

We are answering this based on single-server model. If we have multiple servers($\leftrightarrow$ drains) then the processing($\leftrightarrow$ drainage) capacity becomes $b * \mu$ where b is the number of servers/drains.

- If $\lambda < \mu$, water gets drained out faster than it flows out of the tap, no accumulation of water $\leftrightarrow$ queue size $= 0$.

- If $\lambda > \mu$, water gets drained out slower than it flows out of the tap, there is accumulation of water which increases up over time $\leftrightarrow$ queue size increases over time.

- If λ varies over time, the water level varies with time $\leftrightarrow$ queue size varies over time. This is one of the most important applications of the fluid approximation. It has the ability to handle uncertain arrivals which is absent in many other models.

7

1.4.1 Some General Relations

Let

$A(t)$: cumulative number of arrivals by time t.

$$A(t) = \int_0^t \lambda(u)du \quad [7]$$

$D(t)$: cumulative number of departures by time t.

Since the fluid is assumed to be infinitely divisible, the values of $A(t)$ and $D(t)$ are continuous. They are also non-decreasing and $A(t) \geq D(t)$ because water can not drain out before it comes out of the tap. Thus, queue length $= A(t) - D(t) \geq 0$. Also, if FCFS queue discipline is in place, the time n^{th} customer spends in the queue $= D^{-1}(n) - A^{-1}(n)$. This formula holds even if n is not a whole number since we consider customers to be infinitely divisible. So,

Time n^{th} customer spends in the queue $= D^{-1}(n) - A^{-1}(n)$ [7]

Total time spent in queue by all the customers $= \int_0^N [D^{-1}(n) - A^{-1}(n)]dn$ [7]

This can also be looked at in terms of the queue lengths.

Total time spent in queue by all the customers $= \int_0^T [A(t) - D(t)]dt$ [7]

Now,

$$\int_0^N [D^{-1}(n) - A^{-1}(n)]dn = \int_0^T [A(t) - D(t)]dt \quad [7]$$

$$=> \frac{N}{T} \cdot \frac{1}{N} \int_0^N [D^{-1}(n) - A^{-1}(n)]dn = \frac{1}{T} \int_0^T [A(t) - D(t)]dt \quad [7]$$

$$=> \lambda.W = L \quad [7]$$

8

where L is the average queue length, $\lambda(= \frac{N}{T})$ is the average arrival rate over the time period under observation and W is the average delay per customer.

Remark 1.4.1. We have assumed that the system starts and ends in an empty state.

To know the length of the queue as a function time, we need to know the function $D(t)$

Remark 1.4.2. We know that, the fluid cannot be accumulated at a rate greater than μ , so

$$\frac{dD(t)}{dt} \leq \mu$$

Remark 1.4.3. When the queue is non-empty, we have $A(t) > D(t)$ and the server serves at maximum capacity , so

$$\frac{dD(t)}{dt} = \mu$$

Remark 1.4.4. When the queue is empty, we have $A(t) = D(t)$ and the rate of departure would not be greater the arrival rate $\lambda(t)$, so

$$\frac{dD(t)}{dt} = min(\lambda(t), \mu)$$

Accumulating all the above remarks , we have

$$\frac{dD(t)}{dt} = \begin{cases} \mu & A(t) > D(t) \\ min(\lambda(t), \mu) & A(t) = D(t) \end{cases} \qquad [7]$$

Using the above function , we can integrate the differential and find $D(t)$ us-

9

ing graph method and then find the length of queue as a function of time. For further results and explanation refer *Fundamentals of Queueing Systems*[7]

1.5 The Square Root Staffing Law

In queuing theory, the square root staffing law is a rule-of-thumb used to compute the capacity that would be required to meet an increased amount of traffic in the queue. The law is widely used to help in capacity planning in the QED (Quality-and-Efficiency-Driven) regime. More formally, the question we are dealing with is that if the current quality of service is deemed to be acceptable and necessary then how much must the capacity be increased to serve the increased demand. Rule of thumb answers this question. It says that to hold quality of service constant, you must have a variability hedge equal to the square root of the load increase. This law dates back to Erlang's work in 1917 [3].

1.6 The Newsvendor Problem

The newsvendor problem is a familiar problem in operational research that has application in determining the optimal level of inventory one should stock. Fixed prices and uncertain demand are typical characteristics of the newsvendor problem. This problem gets its name because it mirrors the situation a newspaper vendor faces while deciding the number of copies of the daily paper to buy. Since the demand is uncertain, unsold copies will be wasted. Order is placed before demand materializes and there is a cost

incurred for ordering too much as well as for ordering too few items. These costs are analogous to the cost for idle servers and cost for poor customer service in the staffing problem.

1.6.1 Problem Formulation

Number of units bought is denoted by Q. The per unit selling price is denoted by P, the per unit buying price is denoted by C and the per-unit amount the newsvendor can get for an unsold unit, also known as the salvage price, is denoted by S. F is the cumulative distribution function of demand. The per-unit cost for any items that cannot be sold is called the overage cost (C_o) and the per-unit cost for not meeting demand is called the underage cost (C_u).

1.6.2 Solution of the Newsvendor Problem

The optimal Q is given by:

$$F(Q^*) = \frac{C_u}{C_u + C_o}$$

This is called the Critical Fractile Formula. Now,

$$C_o = C - S \quad and \quad C_u = P - C$$

So,

$$Q^* = F^{-1}(\frac{C_u}{C_u + C_o}) = F^{-1}(\frac{P - C}{P - C + C - S}) = F^{-1}(\frac{P - C}{P - S})$$

The solution we see here was originally proposed in the works of Harrison and Zeevi (2005) [5] and Whitt (2006) [8].

Chapter 2

Managing Supply Under Demand-Side Parameter Uncertainty

First, we solve a simpler problem that involves parameter uncertainty only on the demand side. We would later extend these findings to supply side parameter uncertainty. This problem is more complicated than it sounds. In every day life, the arrival rates are uncertain. However, most of the traditional literature assumes that all model primitives are known with certainty, and "noise" is restricted to stochastic variability.

2.1 Reviewing Existing Models

There are mainly 2 key factors considered to make the choice of capacity level:

1. **Efficiency:** This concerns itself with basic operating costs, e.g, factors like cost of hiring play an important role.

2. **Quality:** This concerns itself with quality of customer service, e.g., factors like abandonment costs play an important role.

Capacity planning is a trade-off between these 2 factors. The best quality of service would require a high amount of spare capacity which could be wasteful and the best efficiency would require preventing occurrence of idle servers this can greatly degrade the quality of service.

If the arrivals follow a Poisson process with rate λ and services are exponentially distributed with rate μ.

Some common regimes that are in use give the capacity C as:

1. **Efficiency-Driven (ED) Regime:**

$$C = \frac{\lambda}{\mu} - \gamma\frac{\lambda}{\mu} \ \ where \ \ 0 < \gamma < 1$$

This model is likely to be under-staffed.

2. **Quality-Driven (QD) Regime:**

$$C = \frac{\lambda}{\mu} + \delta\frac{\lambda}{\mu} \ \ where \ \ \delta > 0$$

This model is likely to be over-staffed.

3. **Quality- and Efficiency-Driven (QED) Regime:**

$$C = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} \quad where \quad -\infty < \beta < \infty$$

This regime is also known as the Halfin-Whitt regime [4].

In all the cases above, the term $\frac{\lambda}{\mu}$ is a base capacity used to meet mean demands. The second term is the variability hedge. The QED regime is popular because it takes into account both the quality and efficiency factors. It takes the form dictated by the The Square Root Staffing Law that was covered in the Key Concepts chapter.

However, the QED regime is not very good for practical implementation. The square-root-law expects a fairly accurate estimate of the service and arrival rates. However, in a practical setting, such parameters are estimated and predicted on the basis of historical data, and hence can be quite "noisy." That means there is expected to be ambiguity in the parameters, such as, arrival rate itself. Hence, we try to find a better solution to the problem.

2.2 Proposing an Approximate Solution based on the Newsvendor Problem

2.2.1 Formulating the Problem

- **Servers:** b identical servers
- **Customer Arrivals:** Doubly stochastic Poisson process, that is, the

arrival rate Λ is also a random variable. It has distribution F and mean λ. The arrival process is a homogeneous Poisson process with this rate.

- **Service Requirements:** i.i.d. exponential random variables. They are independent of the arrival process and rate and have mean $\frac{1}{\mu}$

- **Queue Discipline:** FCFS (First come first served) is used.

- **Waiting Policy:** Servers do not remain idle unless queue is empty. There is an infinite capacity buffer. The per-customer per-unit-time cost of a customer waiting in queue is called the holding cost and it is h per customer per unit time.

- **Abandonment Policy:** Customers have the impatience random variable τ, exponentially distributed with mean $\frac{1}{\gamma}$. A customer may abandon the system after getting impatient from waiting. The customer will abandon the system when her total waiting time in queue reaches τ time units. There is a cost incurred at a rate p per customer, this is called the abandonment cost.

- **Staffing Cost:** c per unit time per server.

- **Queuing Model:** $M/M/b + M$ with a random arrival rate.

- N is the random variable that represents the number of customers in the system in steady state.

- Length of the planning horizon is normalized to 1.

- Integrality constraints are ignored and b is assumed to be real-valued.

Putting things together,

Expected Steady State Queue length $= \mathbb{E}[N - b]^+$.

Cost related to Quality, or the total expected customer cost in steady state $= (h + p\gamma)\mathbb{E}[N - b]^+$.

Cost related to Efficiency, or the total staffing cost $= cb$.

The Optimization Problem:

$$Minimize \;\; \Pi(b) = (h + p\gamma)\mathbb{E}[N - b]^+ + cb$$

$$subject \;\; to \;\; b \geq 0$$

Let b^* denote the minimizer and $\Pi^* := \Pi(b^*)$ the corresponding minimum cost.

2.2.2 Proposing an Approximate Solution

The optimization problem looks simple. However, it is not possible to get an exact solution to it. This is because the distribution of N itself depends on b.

Approach: We ignore the stochastic variability in customer arrivals and service requirements. We only focus our attention on the uncertainty in the arrival rate. This relaxation makes things easy because now customer arrivals form a fluid queue with the fluid flowing at the rate Λ per unit time. The processing capacity is is equal to μb and is fixed.

$$Rate \;\; of \;\; Abandonment \approx \mathbb{E}[\Lambda - \mu b]^+$$

17

Also,

$$Rate \ \ of \ \ Abandonment = \gamma \mathbb{E}[N - b]^+$$

So,

$$\mathbb{E}[N - b]^+ \approx \frac{1}{\gamma}\mathbb{E}[\Lambda - \mu b]^+ \quad\quad [1]$$

This helps us get rid of N which was causing some problems. Now,

The Optimization Problem:

$$Minimize \ \ \bar{\Pi}(b) = (p + \frac{h}{\gamma})\mathbb{E}[\Lambda - \mu b]^+ + cb$$

$$subject \ \ to \ \ b \geq 0$$

Here, $\bar{\Pi}(.)$ is the new objective function that is being proposed.

This is an instance of the newsvendor problem that was visited in the Key Concepts chapter where $C = \frac{c}{\mu}$ and $P = p + \frac{h}{\gamma}$ and $S = 0$. Also, the Number of Units available is analogous to the Processing Capacity available. So, $Q = \mu b$. If F is the cumulative distribution function of Λ, then, $\bar{F} = 1$ - F. So,

$$\bar{Q} = F^{-1}(\frac{P - C}{P - S})$$

$$\bar{Q} = F^{-1}(\frac{P - C}{P - 0}) = F^{-1}(\frac{P - C}{P})$$

$$\bar{Q} = \bar{F}^{-1}(\frac{C}{P})$$

$$\mu\bar{b} = \bar{F}^{-1}(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}})$$

$$\bar{b} = \frac{1}{\mu} \bar{F}^{-1}\left(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}}\right)$$

Remark 2.2.1. Note that if the cost of capacity (per unit) is higher than the penalty charge (per unit), the optimal solution would not install any capacity. Our proposed solution would not install any capacity either. More formally, If,

$$\frac{c}{\mu} > p + \frac{h}{\gamma}$$

Then,

$$b^* = \bar{b} = 0$$

2.3 Analysing the Performance of the Proposed Solution

We will be using some numerical data to analyse the performance of the proposed solution relative to the optimal solution. For all the cases we consider $c = 1/3$, $p = 1$, $h = 1$, $\mu = 1$ and $\gamma = 1/3$. These values ensure that the condition of Remark 2.2.1 does not arise. In all the cases, the optimal solution has been picked from the internet and the proposed solution is computed using the formula derived in section 2.2.2. All values have been rounded to 1 decimal place.

After forming our intuitions based on the numerical data we mention some related mathematical results that justify our observations. These relations are derived in Bassamboo and Randhawa (2010) [1] and are stated here without proof.

Before moving further, a few terms need to be defined that will be used in this section.

Definition 2.3.1. The safety capacity (β) is defined as the difference between the optimal solution b^* and the proposed solution $\bar{b}$.

Definition 2.3.2. The accuracy gap $(\Delta(b))$ is defined as the difference between the value of the actual objective function at b, $\Pi(b)$, and the value of the proposed newsvendor objective function at b, $\bar{\Pi}(b)$.

Definition 2.3.3. The optimality gap is defined as the difference between the value of $\Pi(\bar{b})$ and the value of $\Pi(b^*)$.

2.3.1   Deterministic Arrival Rates

We first consider the simple case where there is no uncertainty in the arrival rate, that is, the case where arrival rate takes constant values.

Computing the Proposed Solution

$$\bar{b} = \frac{1}{\mu}\bar{F}^{-1}\left(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}}\right)$$

$$\bar{b} = \frac{1}{1}\bar{F}^{-1}\left(\frac{\frac{1/3}{1}}{1 + \frac{1}{1/3}}\right)$$

$$\bar{b} = \bar{F}^{-1}\left(\frac{1}{12}\right)$$

$$\bar{b} = F^{-1}\left(\frac{11}{12}\right)$$

Now, for the deterministic case,

$$F^{-1}(p) = \lambda \ \ where \ \ 0 < p < 1$$

So,

$$\bar{b} = \lambda$$

Comparing the Proposed Solution with the Optimal Solution using Numerical Data

Table 2.1 and Figure 2.1 show the performance of the proposed solution on deterministic arrival rates.

We see that the optimality gap increases as the value of the arrival rate increases. We see that the gap between b and b^* also increases with increasing arrival rate.

Remark 2.3.4. When the arrival rate increases by 4 times in the last 2 rows of Table 2.1, we see that the optimality gap approximately doubles. The value of the safety capacity is positive in all cases and it doubles in the aforementioned case. Looking carefully we see that β grows proportionally to $\sqrt{\lambda}$. This is very much in line with the square root staffing law, as one would expect in case of deterministic arrivals.

Related Mathematical Results

$$b^* \approx \bar{b} + C_0 \sqrt{\frac{\lambda}{\mu}} \quad and \quad \Pi(b^*) \approx \Pi(\bar{b}) - C_1 \sqrt{\frac{\lambda}{\mu}} \ \ [1]$$

Arrival Rate	Optimal Solution		Proposed Solution		Comparison	
λ	b^*	$\Pi(b^*)$	$\bar{b}$	$\Pi(\bar{b})$	$|b^* - \bar{b}|$	Optimality Gap
37.5	42	15.9	37	17.7	5	1.7
75	83	29.6	75	31.6	8	2.0
300	316	109	300	112.4	16	3.4

Table 2.1: Performance of the Proposed Solution on Deterministic Arrival Rates
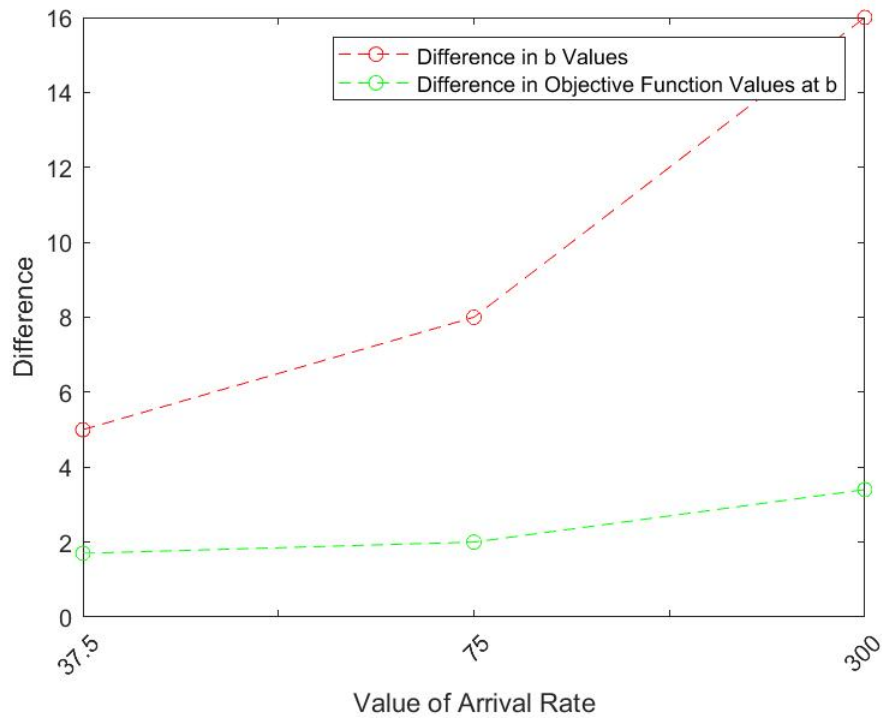

Figure 2.1: Performance of the Proposed Solution on Deterministic Arrival Rates

(a) This image shows the deviation of the proposed solution from the optimal one in the case of different arrival rate values.

(b) The red line shows the values of $|b^* - \bar{b}|$ and the green line shows the value of $\Pi(\bar{b}) - \Pi(b^*)$.

2.3.2   Uncertain Arrival Rates

We assume that the arrival rates follow a uniform distribution, U[a,b].

Computing the Proposed Solution

$$\bar{b} = \frac{1}{\mu} \bar{F}^{-1}\left(\frac{\frac{c}{\mu}}{p + \frac{h}{\gamma}}\right)$$

$$\bar{b} = \frac{1}{1} \bar{F}^{-1}\left(\frac{\frac{1/3}{1}}{1 + \frac{1}{1/3}}\right)$$

$$\bar{b} = \bar{F}^{-1}\left(\frac{1}{12}\right)$$

$$\bar{b} = F^{-1}\left(\frac{11}{12}\right)$$

Now, for U[a,b],

$$F^{-1}(p) = a + p(b - a) \quad where \quad 0 < p < 1$$

So,

$$\bar{b} = a + \frac{11}{12}(b - a) = \frac{11}{12}b - \frac{1}{12}a$$

Comparing the Proposed Solution with the Optimal Solution using Numerical Data

Table 2.2 and Figure 2.3 show the performance of the proposed solution on uncertain arrival rates.

We observe that the performance is quite bad when the mean arrival rate is small. It improves as we move towards larger mean arrival rates. The drop in

the value of optimality gap is, infact, quite drastic in the beginning. We see good performance after U[10,20]. The poor performance in case of smaller mean arrival rates can be ignored in a practical scenario because mostly large traffic is seen.

Remark 2.3.5. Observe that arrivals rate values in Table 2.1 are nothing but the mean of the distributions in last 3 rows of Table 2.2. The difference in the value of optimality gap in the two tables, however, is quite high. This shows that the proposed solution works better when the arrival rate has uncertainty.

Arrival Rate	Optimal Solution		Proposed Solution		Comparison	
Distribution	b^*	$\Pi(b^*)$	$\bar{b}$	$\Pi(\bar{b})$	$\|b^* - \bar{b}\|$	Optimality Gap
U[1,2]	3	1.4	1	5.3	2	3.8
U[2,4]	5	2.2	3	3.3	2	1.1
U[5,10]	11	4.3	8	5.0	3	0.7
U[10,20]	20	7.7	17	8.0	3	0.3
U[15,30]	29	11.0	26	11.1	3	0.2
U[20,40]	37	14.2	35	14.3	2	0.1
U[25,50]	46	17.3	43	17.5	3	0.2
U[50,100]	89	33.1	87	33.2	2	0.1
U[200,400]	351	127.1	350	127.1	1	0.0

Table 2.2: Performance of the Proposed Solution on Uncertain Arrival Rates

Related Mathematical Results

$$\Delta(b) = \Pi(b) - \bar{\Pi}(b) \approx K\mathbb{E}\left[\sqrt{\frac{\Lambda}{\mu}}exp(-\frac{(\frac{\Lambda}{\mu} - b)^2}{2\frac{\Lambda}{\mu}})\right] \quad [1]$$
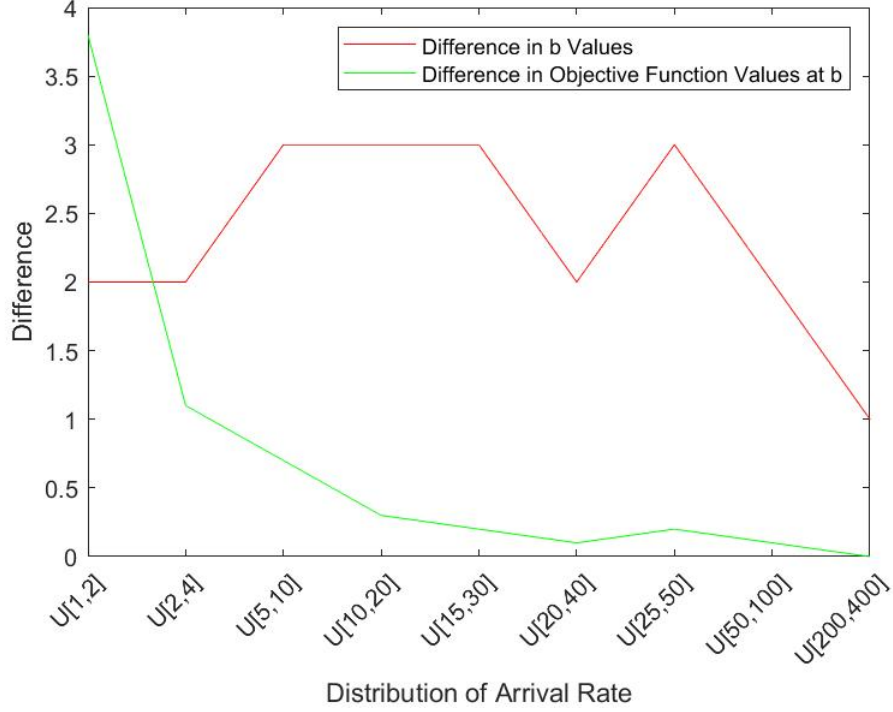
Figure 2.3: Performance of the Proposed Solution on Uncertain Arrival Rates

(a) This image shows the deviation of the proposed solution from the optimal one in the case of various arrival rate distributions.

(b) The red line shows the values of $|b^* - \bar{b}|$ and the green line shows the value of $\Pi(\bar{b}) - \Pi(b^*)$.

Remark 2.3.6. The exponential term dominates in RHS. This term is clearly small when there is some amount of uncertainty in the arrival rates. This means that minimising $\Pi(b)$ essentially means minimising $\bar{\Pi}(b)$. That is why, in case of uncertainty, the newsvendor solution works very well.

2.4 Some Important Results

This section mentions some important results derived in [1]. These are stated here without proof. In this section, the terms have been appended with an additional subscript λ where $\lambda = \mathbb{E}(\Lambda)$ is assumed to be large, as is the case in most practical scenarios. We assume that $\sigma_\lambda < \infty$. Some new terms must be defined before moving forward.

Definition 2.4.1. The coefficient of variation of the arrival rate (CV_λ) is defined as the ratio of σ_λ and λ.

Definition 2.4.2. The offered load (ε_λ) is defined as the ratio of λ and μ.

Theorem 2.4.3. *(Performance of the solution in different regimes)*
(a) (Uncertainty-dominated regime.) If $CV_\lambda \gg \frac{1}{\sqrt{\varepsilon_\lambda}}$, then,

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi_\lambda(b_\lambda^*) + \mathcal{O}\left(\frac{1}{CV_\lambda}\right)$$

(b) (Variability-dominated regime.) If $CV_\lambda \ll \frac{1}{\sqrt{\varepsilon_\lambda}}$, then,

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi_\lambda(b_\lambda^*) + \mathcal{O}(\sqrt{\varepsilon_\lambda})$$

(c) If the variability and uncertainty are balanced, then,

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi_\lambda(b_\lambda^*) + \mathcal{O}(\sqrt{\varepsilon_\lambda})$$

(d) (Deterministic regime). If $CV_\lambda = 0$, then,

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi_\lambda(b_\lambda^*) + \mathcal{O}(\sqrt{\varepsilon_\lambda})$$

Theorem 2.4.4. *(A special case of $\mathcal{O}(1)$ optimality)*

If CV_λ is bounded away from zero, then,

$$\Pi_\lambda(\bar{b}_\lambda) = \Pi_\lambda(b^*_\lambda) + \mathcal{O}(1)$$

Chapter 3

Optimizing the system management cost in the presence of self scheduling servers

(This chapter has been inspired from **Ibrahim(2018)** [**6**] , for all further explanations and results refer [6])

We have seen how the parameter side uncertainty can affect a system in the previous chapter, in this chapter, we will take a look at the system where the number of servers is random, because the servers can self-schedule for different shifts.

In a real life scenario, for many working systems, uncertainty in the availability of servers is inevitable .For models like Call Centres, Uber, Lyft, where

the employee can self-schedule their working hours, it is important for the system manager to maintain a optimal number of employees to keep up with the demand and keep the customers happy while also minimising the cost of maintaining the system.

So let us consider a multi period queuing period where the servers can self schedule whether to serve in a particular shift and the customers are impatient as in any real life system.

3.1 Modelling framework :

Let us consider a system with k shifts and the queuing models for a individual shift to be $G/G/N_j^n + GI$ in steady state. Here , N_j^n is a random variable that indicates the number of servers in j^{th} shift which depends on the total number of servers n. We also assume that all the servers are identical and service are identically distributed independent random variables and they follow a general distribution with mean $\frac{1}{\mu}$, and W.L.O.G , $\mu = 1$. Also there is no restriction on any server in successive shifts or multiple shifts.

When the customer is waiting in the queue, after a random time, the customer leaves without being served, we call this time as patience time. Let us assume that the patience times are identically distributed independent random variables with a cdf F, complementary cdf ccdf $\bar{F}$, density function f, hazard-rate function h_a and mean $\frac{1}{\theta}$. We assume the arrival, service, abandonment processes are mutually independent and also independent of the number of servers in a shift N_j^n. Also the queue is $FIFO$ and the capacity is unlimited. The customer arrival rate is stationary process of λ_j for the jth

shift.

3.2   System Manager's Problem

Being consistent with *Bassamboo and Randhawa*[1] , let's consider two costs
in the system

- a delay cost h_j incurred per unit time a customer waits in the queue
  for being served in the jth shift

- a penalty p_j incurred per every customer who abandons the system in
  the jth shift

Let $Q_{N_j^n}$ be the queue length and $\alpha_{N_j^n}$ be the net abandonment. The system
manager can decide on the staffing level n , the problem becomes

$$\min_{n \in \mathbb{N}} \Pi(n) = \sum_{1 \le j \le k} (c_j \mathbb{E}[N_j^n] + p_j \mathbb{E}[\alpha_{N_j^n}] + h_j \mathbb{E}[Q_{N_j^n}]) \tag{3.1}$$

where c_j is the compensation for each server in j^{th} shift.

Since an exact analysis of (3.1) is difficult, let us consider a fluid approxima-
tion of the above problem, we will later show that the fluid approximation
works good in case of a binomial distribution for the number of servers.

"For an $G/G/s + GI$ system, $\bar{q}_{\rho_s}$ and $\bar{\alpha}_{\rho_s}$ are the fluid approximations of
queue length and net abandonment rates respectively, with traffic intensity
$\rho_s \equiv \frac{\lambda}{\mu s}$. This result is obtained from *Whitt(2006)* [9]"

So the fluid approximation for (3.1) is

$$\min_{n \in \mathbb{N}} \Pi(n) = \sum_{1 \leq j \leq k} nG(c_j)c_j + p_j.\bar{\alpha}_{\rho_s/G(c_j)} + h_j.\bar{q}_{\rho_s/G(c_j)}[6] \qquad (3.2)$$

Here, $\mathbb{E}[N_j^n] = G(c_j)$, since $G(c_j)$ is a constant , let it be r_j , which we define as availability of a server in j^{th} shift. In the fluid approximation, we can see that the optimal staffing level is only dependent on the expectation of the number of servers in a particular shift , but in reality , the variance is a important factor as well which we will discuss later in the paper.

3.3 Benchmark Case : No Self Scheduling

Before we proceed to the problem, let us take a benchmark case where there is no self scheduling which means r_j=1 , for all the shifts. Clearly , in a system where self scheduling is absent the manager can choose to staff each shift optimally individually , where as in a system where the servers can self schedule they have to choose the pool of servers before hand and allow them to reschedule.

The queue length in j^{th} shift can be written as

$$q_j = \int_0^{w_j} D_j(u) \, du \qquad (3.3)$$

where $D_j(u)$ is the density of fluid that has been waiting for u units of time in j^{th} shift and it can be written as

$$D_j(u) = \lambda_j \bar{F}(u) \qquad (3.4)$$

31

Combining (3.3) and (3.4) , we have

$$q_j = \int_0^{w_j} \lambda_j \bar{F}(u) \, du \tag{3.5}$$

The net abandonment rate in j^{th} shift can be written as

$$\alpha_j = \lambda_j F(w_j) \tag{3.6}$$

where w_j denotes the waiting time in given shift j. Clearly, when there is no self-scheduling the individual optimal staffing in each shift

$$n_j^* = \lambda \bar{F}(w_j^*) \tag{3.7}$$

where w_j^* would be the optimal waiting time. Also clearly the individual staffing level for each shift is less than the λ_j for that shift since $\bar{F}(w_j) \leq 1$, so staffing more than λ_j in any shift would not be optimal.

With the results in (3.5) and (3.6) the system managers problem becomes

$$min_{w_j \geq 0} \left(\lambda_j c_j \bar{F}(w_j) - \lambda_j p_j \bar{F}(w_j) + \lambda_j h_j \int_0^{w_j} \lambda_j \bar{F}(u) \, du \right) \tag{3.8}$$

Lets look at some assumptions and propositions

Assumption 3.1. *For all j , $c_j < \min(h_j/h_a(0) + p_j, h_j/\theta + p_j)$*

The above assumption tells us that this source of servers are economical enough for the manager to staff from them and not avoid them altogether, this is consistent with the assumption in [1] , also if the abandonment process is exponential , we have $h_a(t) = \theta t$, so $h_a(0) = 0$, the first term becomes

32

inf , so the assumption becomes

$$c_j < h_j/\theta + p_j \tag{3.9}$$

Using this , we will try and show the next proposition (4.1.)

Proposition 4.1. *Under the above assumption , to optimise the costs in a system with no self scheduling , we should load all the shifts critically , i.e.* $n_j^* = \lambda_j$ *for all* j

Since we know that the staffing costs are reasonably inexpensive , it is in the interest of the system manager to match the demand as there will be no reneging in the system and it is optimal, where if there was no upper bound on the staffing cost , it would have been advisable to leave some customers unattended to optimize the overall problem.

3.4   The actual problem : Servers with Self-Scheduling capability

In this case , the r_j factor is no longer 1, so we define a new arrival rate called the *augmented arrival rate* Γ_j.

Definition 3.4.1. The augmented arrival rate is defined as the ratio of actual arrival rate and the showup probability *i.e* $\Gamma_j = \lambda_j/r_j$

Also W.L.O.G , we will assume that the shifts are ordered in the increasing order of Γ_j *i.e* $\Gamma_{j-1} \leq \Gamma_j$ for $j=1, 2, ..., k$, with this ordering , we can have the following:

Remark 3.4.2. If we take a staffing level n , such that $\Gamma_{j-1} < n < \Gamma_j$, then we can see that the shifts with $i \geq j$ are overloaded and the shifts with $i < j$ are under loaded.

Remark 3.4.3. If we take a staffing level n , such that $n = \Gamma_j$, then we can see that the shifts with $i > j$ are overloaded and the shifts with $i < j$ are under loaded , while the shift j is critically loaded.

Remark 3.4.4. In an overloaded shift, we have $\Gamma_i \bar{F}(w_i) = n$, i.e , $w_i = \bar{F}^{-1}(n/\Gamma_i)$.

As we observed in the benchmark case ,underloading all the shifts can result in a congested system , while on the other hand overloading might increase the cost of maintaining the system, so the system manager's problem in terms of the augmented arrival rate Γ_j becomes

$$min_{0 \leq n \leq \Gamma_k} C(n) = \left(\sum_{j=1}^{k} I(\Gamma_{j-1} \leq n < \Gamma_j) u_j(n) \right) \qquad (3.10)$$

where $I(n \in S)$ is a indicator random variable that tells us whether the value n is in the set S and $u_j(n)$ is given by

$$u_j(n) = \sum_{i=1}^{k} c_i n r_i + \sum_{i=j}^{k} \left(p_i(\lambda_i - nr_i) + h_i \lambda_i \int_0^{\bar{F}^{-1}(n/\Gamma_i)} \bar{F}(u)\, du \right) \qquad (3.11)$$

this u_j represents the cost of the system if n is chosen in the $[\Gamma_{j-1}, \Gamma_j)$, also we can see that only one of the indicator variables is non-zero and the rest of them are zero.So if we choose it in the $[\Gamma_{j_0-1}, \Gamma_{j_0})$ we have the cost of system is u_{j_0} and all the shifts below j_0 the system is under staffed and the manager faces customer related losses, where as for the shifts j_0 and above it is over

34

staffed and the staffing costs are high.

We will now derive a proposition which compares this system to the benchmark and tells us in which condition, maintaining a self scheduling system is not more expensive for the manager when compared to a system with no self scheduling

Proposition 3.2. *If the calculated Γ_j ("resulting augmented arrival rates") are equal across all shifts then and only then, the self scheduling system is not more expensive than a system with no self scheduling*

Proof. Let all the augmented arrival rates be equal *i.e* $\Gamma_j = \Gamma$, then assuming that $n^* = \Gamma$ gives, $n_i^* = n^* r_i = \lambda_i$ for all shifts, as we have seen in proposition (4.1), this is the optimal staffing cost in the case of the benchmark, so if all the augmented arrival rates are equal , we can eliminate the cost of self scheduling. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

The above proposition shows us the importance of having multiple shifts in the system. Because in the case of a single shift, the manager could just meet the demand in that shift by having a big enough staffing level like λ/r and have no cost of self scheduling, it is to accommodate multiple shifts, manager has to take in to consideration the cost of self scheduling.

In light of this proposition , let us consider a system with 2 shifts: shifts A and B, let us assume that the arrival rates in both the system is known to the manager and they are λ_A and λ_B and without loss of generality let us assume that $\lambda_A < \lambda_B$, then while selecting the staffing pool for this shifts, the manager could select a pool such that the probability of showing up is higher in shift A than B *i.e* $r_A < r_B$ and the ratios are equal and by doing the cost of self scheduling can be avoided. But as we know in a real life

scenario it may not be possible for the system manager to know these values beforehand. They can use the estimated historic values to make a decision.

Now let us dive into a case where the augmented arrival rates are not equal and the hazard rate for the abandonment process is increasing, specifically exponential.

3.5 Exponential Abandonment and Non-equal Augmented Arrival Rates

In this case, we actually see that it is efficient for the system to match the demand in one of the shifts with all of the shifts instead of matching it in every shift and leaving the remaining shifts underloaded or overloaded.

Proposition 3.3. *For exponential abandonment rates :*

1. *The optimization function in (3.10) is piece-wise linear.*

2. *There exists one shift s, such that by matching the demand in that shift we can optimize the system i.e. $n^* = \Gamma_s$.*

3. *i_0 satisfies the following condition:*

$$\sum_{j=1}^{k} c_j r_j - \sum_{j=i_0}^{k} \left(p_j + h_j/\theta\right) r_j < 0 \ and \sum_{j=1}^{k} c_j r_j - \sum_{j=i_0+1}^{k} \left(p_j + h_j/\theta\right) r_j > 0$$

$$(3.12)$$

Proof. If the cumulative distribution function $F(x) = 1 - e^{-\frac{x}{\theta}}$ i.e the abandonment process is exponential, then for $\Gamma_{i-1} \leq n \leq \Gamma_i$

$$u_i(n) = \sum_{j=1}^{k} c_j r_j n + \sum_{j=i}^{k} (p_j + h_j/\theta)(\lambda_j - nr_j) \qquad (3.13)$$

Clearly, under condition (3.12), C(n) is piece wise linear with

$$\frac{dC(n)}{dn} = \begin{cases} negative & n \leq \Gamma_{i_0} \\ positive & n > \Gamma_{i_0} \end{cases}$$

Hence,the minimum must occur at some Γ_s. □

CONCLUSION AND FUTURE SCOPE

The detailed survey has given us a comprehensive understanding of working with parameter uncertainty and self-scheduling servers. We will extend these learnings to understand and adapt Dong and Ibrahim's model which forgoes the fluid approximations in *Ibrahim (2018)* [6] and gives a better understanding of optimizing the operation costs in a blended workforce .The parameter uncertainty knowledge will be put to use to model a system with supply-side parameter uncertainty and the self-scheduling servers will help us model the flexible workers as well as the freelancers.

Bibliography

[1] Achal Bassamboo, Ramandeep Randhawa, and Assaf Zeevi. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56:1668–1686, 10 2010.

[2] Jing Dong and Rouba Ibrahim. Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Oper. Res.*, 68(4):1238–1264, July 2020.

[3] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, 10:189–197, 1917.

[4] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 06 1981.

[5] J. Michael Harrison and Assaf Zeevi. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Operations Management*, 7(1):20–36, January 2005.

[6] Rouba Ibrahim. Managing queueing systems where capacity is random and customers are impatient. pages 9–16, 8 2017.

[7] John F. Shortle, James M. Thompson, Donald Gross, and Carl M. Harris. *Fundamentals of Queueing theory*, volume 5. Wiley Series In Probability And Statistics, 2018.

[8] W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15:88–102, 2006.

[9] Ward Whitt. Fluid models for multiserver queues with abandonments. 1 2006.