On the Equilibrium Behaviour of Heterogenous Customers in Differentiated Service Systems

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Tejas P. Bodas
(Roll No. 08307R09)

Under the guidance of
Prof. D. Manjunath

Department of Electrical Engineering
INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
April 2016
Dissertation Approval

The thesis entitled

On the Equilibrium Behaviour of Heterogenous Customers in Differentiated Service Systems

by

Tejas P. Bodas
(Roll No. 08307R09)

is approved for the degree of

Doctor of Philosophy

Prof. Vishnu Narayanan
(Internal Examiner)

Prof. Sandeep Juneja
(External Examiner)

Prof. D. Manjunath
(Supervisor)

Prof. N. Hemachandra
(Chairman)

Date: 22/4/11
Place: IIT Bombay
Declaration

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

[Signature]

Tejas Boda

(Name of the student)

08307RD9

(Roll no)
Abstract

An important aspect of many queueing systems is that the customers that use the system are heterogeneous in nature, i.e., they do not all have identical preference to the quality of service (QoS) received and their sensitivity to the cost of service. Some examples of such service systems include road and transport systems, health-care systems, computer systems, call centers and communications systems. In such situations, providing differentiated service (different QoS to different customers), can improve social optimality and also provide higher revenues to the service provider. In this thesis, we consider situations when such customers are strategic and choose the service options, and hence the QoS, to optimize a private utility function which could also depend on the price of service. Specifically, we consider queueing systems that offer different service classes to the heterogeneous customer and the latter choose the service class selfishly. In this thesis, we model heterogeneity as different customers having different monetary cost per unit of expected delay.

We begin by considering a parallel server system where each server has an associated queue for the customers to wait. We define a social welfare function for the customers and characterize the optimal routing of the heterogeneous customers to these servers. Our first result shows that the optimal routing satisfies a certain threshold property. We then consider the case when customers are strategic and the servers charge an admission price to monetize the service offered. We perform a game-theoretic analysis of this system and characterize the Nash-equilibrium (also known as Wardrop equilibrium in this setting) for the game. We find that for a given set of admission prices, the equilibrium routing satisfies the same threshold property as that of the routing that maximizes the social welfare. This leads us to a mechanism design for this system where we determine the set of admission prices such that the routing under Wardrop equilibrium will coincide exactly with the routing that maximizes the social welfare.

In the above problem, the aim of the service system is to set admission prices such that the Wardrop equilibrium routing maximizes the social welfare function. For such service systems,
an alternative objective could be to maximize the revenue rate from the system. If all the servers belong to the same service system, then such a market is called as a monopoly. We consider revenue maximization in such a monopoly system. We observe that since the arriving customers are heterogeneous, characterizing the revenue rates as a function of the admission prices is not straightforward. To simplify the analysis, we use the threshold property of the equilibrium routing to provide an equivalent formulation for the problem. Using the equivalent formulation, we obtain the revenue maximizing admission prices to be set at the servers.

We next consider single server systems that can offer service quality differentiation using the Discriminatory Processor Sharing (DPS) scheduling policy. The DPS scheduling policy provides multiple service classes where each class has an associated weight that determines the proportion in which the server is shared among the service classes. For such a system, we consider the problem of routing of heterogeneous customers to the different service classes to maximize a social welfare function. We then provide an equilibrium analysis for the case when the service classes are charged an admission price and the customers are strategic in choosing them. Finally, we identify the similarity in the structure for the optimal and equilibrium routing and solve the corresponding mechanism design problem.

In the last part of the thesis, we consider a highest bidder first (HBF) system where heterogeneous customers are strategic and place bids to obtain service. In an HBF server, the position of a customer in the queue is determined by the bid placed for service. This is clearly a single server system with the priority scheduling discipline. We additionally assume that a first-in first-out (FIFO) server offering free service is provided as an alternative service facility to these customers. We first analyze the equilibrium routing of the customers to the HBF and the FIFO server and show that this routing is of the threshold type. We then analyze the effect on the system revenue (sum of the bids) with and without the presence of the free FIFO server. We show that while addition of a free FIFO server leads to decrease in the revenue, sharing the total service capacity between the two servers increases the revenue. We conclude with an analysis of the equilibrium routing for the case when the FIFO service is not free but comes with an admission price. The equilibrium structure that we obtain for this case is novel and interesting.
Contents

List of Figures vii

1 Introduction 1

1.1 Background and Motivation 1
1.2 Performance Measures for Service Systems 3
1.3 Scheduling and Routing for Control of Queues 4
1.4 Centralized and Decentralized Decision Making 9
1.5 Mechanism Design 12
1.6 Overview of the Thesis 13
1.7 Contribution of the Thesis 15

2 Optimal and Equilibrium Routing for Parallel Queues 19

2.1 Introduction 19
2.2 Model and Problem Formulation 23
2.3 Social Welfare Optimization 26
2.4 Admission Prices and Wardrop Equilibria 38
2.5 Admission Prices and Pigouvian Taxes for Welfare Optimization 42
2.6 Examples 45
2.7 Discussion 47

3 Revenue Maximization in a Monopoly System 49

3.1 Introduction 49
3.2 Notations, Preliminaries and Problem Formulation 51
3.3 Monopoly Market 56
3.4 Estimating the distribution \( F \) 68
3.5 Discussion 73
## 4 Heterogeneous Customers in a Discriminatory Processor Sharing System

4.1 Introduction ........................................ 75
4.2 Preliminaries ........................................ 78
4.3 Optimal Allocation of Customer Types to Service Classes .............. 80
4.4 Admission Prices, Selfish Users and Wardrop Equilibrium ................. 86
4.5 Discussion ........................................ 89

## 5 A Choice of Highest-Bidder-First and FIFO Service

5.1 Introduction ........................................ 91
5.2 Notation and Preliminaries ....... .......................... 94
5.3 A System with a FIFO and a HBF Server ........................ 97
5.4 When Arrivals Balk .................................. 105
5.5 When FIFO server is not free .............................. 108
5.6 Discussion ........................................ 113

## 6 Conclusion

References ............................ 117
List of Figures

1.1 Scheduling policies offering a single class of service ........................................ 5
1.2 Scheduling policies offering multiple class of service ........................................ 6
1.3 A multiple server system with two arriving classes ........................................... 8
2.1 Illustrating the feasible set for $(p_1^*, p_2^*)$ .................................................. 27
2.2 Lines $L_1$ and $L_2$ intersect at the stationary point $(0.5, 0.5)$ ............................... 28
2.3 $(p_1, p_2)$ minimizing Program $P_2$ for each $\gamma_i$ ......................................... 29
2.4 The operating regimes as $c$ is increased from 0 .................................................. 41
3.1 Representation of $K^W$ when $c_1 > c_2$ .......................................................... 53
3.2 Representation of $K^W$ when $c_1 < c_2$ .......................................................... 54
3.3 Illustrating $\gamma_1 \in [0, \gamma^+]$ and $c_1 > c_2$ ...................................................... 57
3.4 Illustrating $\gamma_1 \in (\gamma^+, \lambda]$ and $c_1 < c_2$ .................................................. 58
3.5 Illustrating $\beta_1(\gamma_1)$ when the servers are not identical ................................. 60
3.6 Illustrating $\beta_1(\gamma_1)$ for the case of identical servers ....................................... 61
3.7 Illustrating $q_1(\gamma_1)$ when the servers are not identical ..................................... 62
3.8 $R_T$ as a function of $\gamma_1$ when $D_j(\gamma_j) = \frac{2\gamma_j}{\mu_j}$ ................................. 66
3.9 $R_T$ as a function of $\gamma_1$ when $D_j(\gamma_j) = \frac{1}{\mu_j}$ ........................................ 67
3.10 Comparing the estimate of $f(.)$ with the true density function ............................. 72
3.11 Estimating the discrete distribution $F$ ............................................................ 72
4.1 $U(q(t))$ as a function of $t$ .................................................................................. 84
5.1 System with a HBF server and a FIFO server ...................................................... 98
5.2 Revenue as a function of $\mu_1$ for three different example $F(\beta)$ and different $\lambda$. 102
Chapter 1

Introduction

1.1 Background and Motivation

For most of us, queues have become an essential part of the daily life. Right from the early morning queue to buy a metro ticket to waiting for a dinner table at a popular restaurant, most service systems that we deal with involve queues. Queues are used at service systems to maintain a sequence in which the service is to be provided to customers. This maintains a discipline among customers and the service system is also better organized.

When customers or jobs wait for their turn in a queue, they experience a delay. Service systems in which the quality of the service depends on the delay experience of the customers are called delay systems. Some examples of service systems that are relevant to our discussion are transport and road systems, healthcare systems, computer systems, call centers and communication systems. We shall now briefly see a few examples of service systems that are in fact queueing systems because of various queueing and scheduling operations involved.

In case of public transport systems, one is often required to first stand in a queue before a ticketing kiosk to buy the travel tickets. In case of road networks, queues of vehicles are seen as a manifestation of congestion. An example from the airline industry is the queueing of airplanes near the runway for take-off during peak-hour periods. In healthcare systems, queueing theory plays an important role in managing the allocation of beds to patients, maintaining sufficient stocks of medicines and other medical equipment. Healthcare managers are also required to efficiently design waiting rooms for patients who are seeking appointments with doctors. In a computer system, queuing theory helps in organizing access to important units like the processor or the memory for use by various programs. While some concurrent programs may need
simultaneous access of the processor, this may not be possible. Various scheduling and queueing principles are then used to allocate the processor or the memory unit in an efficient manner. In service systems such as call centers, the number of call center operators are fixed and if all operators are busy then the new calls are put on hold. A customer on hold experiences delay in obtaining the required service and this delay affects its overall experience at the call center.

In service systems such as the present day communication systems, the number of simultaneous calls or connections that can be handled by any base station is finite in number. This determines the capacity of the system and new connections are lost if the number of active connections exceed this capacity. While the customers do not experience delay in such systems, they may however be denied service due to the absence of a buffer. Such systems are therefore known as loss systems. The popularity of the theory of queues among communication engineers can be attributed to the works of Erlang on such loss systems. Erlang B is a formula to determine the blocking probability of the calls in a loss system where the blocking probability is defined as the probability that a call or a customer finds the server(s) busy and is lost. The Erlang C formula is used for delay systems and it determines the probability that an arriving customer finds the server(s) busy and hence is required to wait for service in the queue. The Erlang B and the Erlang C formula are important performance measures for telephone systems and call centers respectively.

It is important to note that in most service systems, the customers, jobs or packets that use the system need not have identical preferences for the quality of the service received. In this thesis, we shall use the terms customers and jobs interchangeably. Similarly, customers with different preferences will be referred to as heterogeneous customers or multiclass customers. In service systems where the arriving customers are humans, it is known that such customers have different preferences for different service quality metrics based on the delay. While some customers may be tolerant to large delays, some others may be averse to long delays. It may be of advantage to the system if such heterogeneity of customers is accounted in any optimization of the use of the system resources. Many service systems have emerged that exploit the heterogeneous nature of customers and use it to their advantage. For example, these days airlines offer priority queues for boarding earlier. With the payment of an additional fee, a priority customer is allowed to board the plane prior to non-priority customers and this minimizes the delay experienced in a queue. Another example is that of healthcare emergency such as an epidemic or a terror attack. In such emergencies the number of patients that require immediate treatment...
is very large. In such cases, different patients are assigned different priorities based on their medical condition and are treated accordingly. This method to determine the sequence of the patients is part of the triage.

In computer systems, multiple programs or processes often require simultaneous access to system resources such as the processor or the memory unit. The processes need not be identical and primarily differ in their deadlines for completion. The operating system allocates the available resources by means of multitasking, i.e., time sharing the resources between the processes. The operating system maintains a program called the scheduler that assigns resources to various processes that require them. By prioritizing the processes based on their deadlines, the scheduler ensures that the processes meet their respective deadlines. This is again an example of a service system where the heterogeneous jobs are treated differently, in this case based on their deadlines. In the Internet, voice and video packets are more sensitive to delay as compared to packets that are generated from emails and web browsing. Such packets require a different end-to-end delay guarantee compared to other data packets. To ensure that these packets meet their respective delay guarantees, the routing and scheduling algorithms in the Internet Protocol suite treat them with high priority. For example, the Cisco routers that implement low latency queueing (LLQ) provide a separate priority queue for video and voice packets.

The preceding discussion motivates us to look at service systems as queueing systems where the customers of the system do not have identical preferences. We have already seen a few examples where treating customers of different classes differently can be of advantage. This is the primary motivation for the different problems considered in the thesis.

1.2 Performance Measures for Service Systems

Associated with any service system are performance measures that determine the quality of service provided. These measures are also known as utility functions or welfare functions or cost functions. A cost (resp. welfare) function captures the notion of the cost (resp. quality) of service that is experienced by a customer of the system. While welfare and utility functions mean the same thing, a cost function can be considered as the negative quantity of the welfare (or utility) function. In this thesis, we shall use these three names interchangeably when the context is clear.

A customer in a delay system experiences a cost for waiting in the system. There could
also be a benefit from obtaining the required service. An individual cost function measures the net cost of service for an individual customer i.e., the difference in the waiting cost and the service benefit for that customer. The individual cost function may also depend on the class of the customer and hence may be different for customers of different types. In a delay system where customers are charged an admission price, a customer’s individual cost is considered to be the sum of the admission price and the delay it experiences.

Social cost functions on the other hand measure the net benefit (or cost) derived by the entire system. This includes the cost to the customers as well as to the server(s). Social costs that are a functions of the individual costs of the customers are commonly considered in the literature. As an example, if the individual cost to a customer is the mean delay, then a measure for the social cost could be the aggregate of the mean delay of all customers in the system. Some other social welfare functions could be the mean number of customers in the system, higher moments of the delay or even the probability of the mean delay of any customer exceeding a specific value. For a queueing system with multiple queues, a social cost could be defined by first calculating the expected delay per customer in each queue and then taking the maximum over all the queues. One possible social cost in loss systems could be the blocking probability that was defined earlier. Note that in service systems with admission prices, while the admission price is part of the individual cost, it is not included in the social costs for the system. This is because the admission price is considered as a transfer of payment from the customer to the server and the money does not leave the system.

Having defined the performance metrics for queueing systems, the goal in many service systems is to optimize the system with respect to such metric. This can be achieved by controlling either the sequencing of customers in the queues or by controlling the access of the queues and servers to various customers. This is known as scheduling and routing in queueing systems which will be reviewed in the following section.

### 1.3 Scheduling and Routing for Control of Queues

Consider a service system with a single server and an associated queue for the customers to wait. When possibly heterogeneous jobs arrive in this system, the performance of the system may depend on the order in which the jobs are served. Scheduling of the job determines the order in which the queued jobs are to be processed by the server. Various scheduling rules or policies
have been proposed in the literature. Some important ones are first-in first-out (FIFO), last-in first-out (LIFO), processor sharing (PS), random order service, priority scheduling, generalized processor sharing (GPS), discriminatory processor sharing (DPS) and discriminatory random order service (DROS). These scheduling policies can be divided into two groups — (1) single service class policies and (2) multiple service class policies. Single service class policies offer an identical class of service to all its customers. On the other hand, policies with multiple service classes can offer different grades of service to different customers. We shall now review some of these policies in detail.

FIFO, LIFO, PS and random service order are examples of scheduling policies that offer a single class of service. A FIFO policy provides service to the arriving customers in order of their arrival times. This is the most commonly used policy in queueing systems that we encounter. The LIFO policy on the other hand serves those customers first that have arrived latest. Some examples of the use of LIFO are the order in which pages in the printer tray are inserted and printed, or the push and pop operations of a stack pointer for memory allocation and access. In a PS policy, all customers present in the queue are served simultaneously and are offered equal share of the processor or server. These policies are represented in Fig. 1.1 where a connection from the server to the queue indicates the customer currently in service. Since the processor sharing policy serves all customers simultaneously, the server has connection to all customers in the queue.
Examples of scheduling policies for a single server queue that offer multiple service classes are priority scheduling, GPS, DPS and DROS. See Fig. 1.2 for a representation of these policies. In the figure, the different service classes are indicated by the vertical partitions to the queue while the type of service is indicated by the interconnection between the server and the service classes.

In a priority scheduling policy, each service class has an associated priority level and the customers either choose a suitable service class or are assigned to one of them. In this policy, customers are served in the decreasing order of their priority class. This is indicated by a switch interconnection in Fig. 1.2. After each service completion, the switch toggles to the non-empty service class of the highest priority and serves a customer of that class before switching again. When the number of service classes are finite, the mean waiting time per customer of a service class was analyzed by Cobham [20]. This was further extended to the case of (uncountably) infinite service classes by Kleinrock [40].

Scheduling policies such as GPS and DPS are variants of the processor sharing policy and can serve multiple customers from the system simultaneously. In GPS and DPS, different service classes are assigned non-negative numbers or weights that determine the share of the total service capacity to the different service classes. In case of the GPS scheduling policy, a
separate (typically FIFO) queue is maintained for each class of service and arriving customers are directed to one of them, either according to a system defined rule or is chosen suitably by the customer itself. The total service capacity of the server is shared among the non empty queues in proportion to the weights and the share is independent of the number of customers in the queue. From each queue, only the head-of-line customer is served while the others wait. This is indicated in Fig. 1.2 by connections from the server to only the head-of-line customers.

A deterministic analysis for a GPS queueing system was first developed by Parekh and Gallager [54], while Zhang et al. [66] developed an early stochastic model. In DPS, an arriving customer is assigned to a class of service like in the GPS. Each service class has an associated weight and the total service capacity is shared among all the customers present in the system in proportion to their weights. This is indicated by the connections from the server to all customers in Fig. 1.2.

Thus, in DPS there is no fixed service guarantee to a service class or to a customer. This implies that any customer that is arriving into the system or departing from it will affect the service rate and hence the sojourn time, of every other customer in the busy period. The DPS system was first introduced by Kleinrock [41] and subsequently analyzed by several authors [28, 36, 35, 38].

In a discriminatory random order service (DROS) policy, each service class has an associated weight as in GPS or DPS. However DROS is not a processor sharing policy and hence the server can serve only one customer at a time. Refer Fig. 1.2 for a representation of DROS. The choice of the next customer is a random choice and the probability of choosing a customer of a service class depends on the weights and the number of customers of the different classes in the queue. DROS policy is also known as relative priority policy and was first introduced by Haviv and van der Wal [35]. For more analysis of this policy refer [34, 39]. For a detailed analysis of both deterministic and randomized scheduling policies see the book by Conway et al. [22].

Scheduling in queues is considered in most standard books on queueing theory [9, 21, 49].

Now consider a service system that has multiple servers and where the multiclass customers can be served at any of these parallel servers. Here parallel server means each server has an associated queue and the customers have to wait in one of the queue to obtain service. Such a service system with three servers and two customer classes where each customer class arrives according to a Poisson process is illustrated in Fig. 1.3. For simplicity assume that the scheduling policy at each server has a single service class. In such service systems, either each arriving customer is assigned to a server or the choice of the server is made by the customer itself. Determining the queue that an arriving customer must choose is known as routing. Routing policies
for such service systems can be either deterministic or randomized. In a deterministic routing policy, customers are routed to servers based on a predefined rule. For example, in round-robin routing, the customers that arrive in the system are routed to the servers in a round-robin order. If the customers are heterogeneous, then the routing rules could be different for customers of different classes. For example in delay systems, a routing policy could be to send the most delay sensitive customers to the fastest server. In a randomized routing policy, an arriving customer is routed to one of the servers based on a routing probability matrix. Now recall a single server system where the scheduling policy offers multiple service classes. Here routing is required to assign a customer to one of the service class. If customers are heterogeneous, then the routing rule could be different for customers of different types.

Figure 1.3: A multiple server system with two arriving classes. The arrival rates for each class is $\lambda_1$ and $\lambda_2$ respectively.

Note that the routing that we consider in this thesis is different from the graph routing that is associated with service systems such as data networks (Internet), web server farms or even telephone networks. Such networks are complex networks that are made up of multiple servers or nodes and the packets originating at the source node are to be transmitted to the destination node via the intermediate nodes. The task of finding a route optimizing a suitable metric while transferring a packet between a source node and a destination nodes is known as (graph) routing. We do not consider such routing in this thesis and shall not discuss this any further.
1.4 Centralized and Decentralized Decision Making

In the preceding discussions, we have seen that most practical service systems are queueing systems where the arriving customers need not have identical preferences. Further, a variety of scheduling and routing policies are known that can be used in these systems to achieve a desired performance metric. The obvious question that arises is who decides the routing and scheduling of customers in the service system? It is also interesting to see how the choices of the routing and scheduling decisions are different for different utility functions.

The most common model for decision making in service systems is the centralized model. In this model, a system operator or a manager is responsible for performing the routing and scheduling decisions and the decision must optimize a well-defined utility function. The notion of the system operator is often considered abstract and the problem is formulated as a linear or a nonlinear program where the optimization is over class of routing or scheduling policies. As an example of centralized scheduling, consider a single server queue where the arriving customers belong to a multiclass population. Let \( \lambda_i, c_i \) and \( \mu_i \) denote the Poisson arrival rate, the waiting cost per unit time and the mean service requirement for a customer of Class \( i \). It is easy to see that this single server is an \( M/G/1 \) queue. An optimal scheduling policy minimizing the total queueing cost in this \( M/G/1 \) queue is a class-wise priority policy determined by the ratio \( \frac{c_i}{\mu_i} \). In particular, if \( \frac{c_i}{\mu_i} > \frac{c_j}{\mu_j} \) then Class \( i \) customers have a priority over Class \( j \) customers. This policy is known as the \( c/\mu \) rule and the optimality of this was shown by Cox and Smith [23].

Now consider a centralized routing model where the service systems consists of parallel servers and the arriving delay sensitive customers are required to obtain service at one of the servers. Recall that a suitable choice for the utility function for such systems is the mean waiting time per customer, the aggregate delay of all customers, the mean number of customers in the system or the maximum delay among the set of servers available. Borst [16] analyzes this problem when the objective is to minimize the aggregate waiting cost per customer. The scheduling discipline at each queue is a priori fixed as FIFO. The instantaneous queue lengths are not available and an optimal probabilistic routing minimizing the objective function is obtained. Sethuraman and Squillante [57] additionally consider the scheduling problem at each queue and shows that as in the case of a single server queue, the optimal scheduling policy in the parallel queues is also the \( c/\mu \) policy.

---

1. Throughout this thesis, we shall use the popular Kendall’s notation.
A slightly less centralized routing model is the dispatcher based model that has been analyzed when there are a finite number of customer classes. In this model, each arriving class has an associated dispatcher and the dispatcher has to decide the routing strategy for the customer of its class. Refer Altman et al. [6] and Ayesta et al. [10] for such models. The centralized models typically achieve a globally optimal routing or scheduling using an algorithm; the task of the algorithm would be to determine the optimal routing scheme and have a central entity perform this algorithm on the arriving customers. Such central implementations are not always feasible and distributed solutions are sought to achieve the allocation that optimizes the system objective.

In a decentralized queueing model, the arriving customer is required to make a routing or scheduling decision to optimize a local objective function. The social outcome of such decision may or may not be optimal from the system point of view. A type of decentralized routing is where the arriving customer is assumed to be strategic or selfish and makes an individually optimal decision (also called as its strategy) that minimizes its individual cost. The strategy of a customer depends on the available information about the queueing system such as the instantaneous queue lengths, the speed of the servers, the priority levels in a priority scheduling policy or even the expected arrival rate of customers in the system. Since customers are selfish, they may impose a negative externality on the system and the system performance could be bad in comparison to a centralized model. The decentralized model can be viewed as a non-atomic game; non-atomic because the impact of any individual customer and its strategy on the performance of the system is infinitesimally small (refer Chapters 17 to 19 of the book by Nisan et al. [51]). The analytical interest in such models is to characterize the (Nash) equilibrium strategy chosen by the customers.

Decentralized routing and scheduling problems are well studied in the literature. An important example and an early work in decentralized queueing models is of Naor [50]. The model considers a single server queueing system where homogeneous customers obtain a reward after service completion. The queue is observable and the customers either choose to join the queue or balk, i.e., not join the system at all. It is shown that at equilibrium, an arriving customer balks if the number of customers that it sees in the queue exceeds a threshold \( n_e \). On the other hand, an optimal threshold \( n_o \) maximizing a social welfare function is shown to satisfy \( n_o \leq n_e \). The objectives of the customers and the society do not coincide and the strategic or selfish decisions made by the customers impose negative externality on the system. See Edelson and Hildebrand
for a generalization to the case of non-observable queues and Larsen [43] for a generalization with heterogeneous customers. There is a lot of work by several authors [44, 12, 13] for the case when queues are not observable and the customers are heterogeneous.

In case of decentralized routing in service systems with multiple parallel servers, the arriving customers have to choose one of the server to obtain service. Depending on the service system, customers may additionally be allowed to balk. Refer Bell and Stidham [14] for decentralized routing in parallel heterogeneous servers when the customers are identical and are not allowed to balk. Also see Bradford [17] and Masuda and Whang [46] for a generalization of Bell and Stidham [14] for the case when customers differ in their value for service.

Decentralized scheduling models are meaningful only for the case of multiple service class policies such as priority scheduling, GPS, DPS and DROS. Since multiple service class policies can offer different grades of service, an arriving selfish customer will choose a grade of service that optimizes its individual utility. As an example, consider a single server queue with a DPS scheduling policy offering two service classes. Each service class has an associated weight and suppose that the two weights are not equal. Further assume that the higher of the two weights has an associated non zero admission price. In this model, the objective of each arriving customer is to minimize its expected waiting cost in the system. The strategy is to either choose the free service class with a lower weight or choose the other service class after paying a non-zero price. Refer Chapter 4, Theorem 4.7 in the book by Hassin and Haviv [33] for the equilibrium analysis of this model. There are several alternative models for decentralized scheduling in service systems with a single DPS server [35, 63, 24].

Similar to the model discussed above, consider a priority scheduling policy with two service classes. Customers that belong to one of the service class have a preemptive priority over those of the other class. The high priority class has an associated admission price and the arriving customers must choose to either pay this price and join the high priority class or pay nothing and remain a low priority customer. This decentralized model was first analyzed by Adiri and Yechiali [2] and later in Hassin and Haviv [32]. There are some alternative decentralized priority pricing models in the literature [11, 61, 48]. Now consider an alternate priority pricing model where instead of a fixed price, the arriving customers can purchase their priority level in the queue after paying a self determined price. Different customers depending on their type may offer different prices and customers with higher prices are accorded better position in the queue. In other words, the arriving customers are placing a bid for the priority level that
they want. This decentralized priority model is used in several practical systems where bids are placed and their bid value is used to determine the order of service for the customers waiting in a queue. For a detailed analysis of this model refer [40, 45, 29]. We refer the readers to the books by Hassin and Haviv [33] and Stidham [59] for a survey of decentralized models in a variety of queueing systems.

1.5 Mechanism Design

In the previous section, we have seen some centralized and decentralized models for service systems with queues. While the aim in a centralized model is to achieve an optimal system performance, finding an optimal scheduling or routing policy may not be easy. For example, the social utility function in the optimal routing problem in Borst [16] is not convex and hence finding the optimal routing matrix is difficult. There may even be practical difficulties involved in implementing the optimal policies. For example while the $c/\mu$ rule is an optimal policy minimizing the mean number of customers in the system, the policy requires customers to reveal private information about their type or class. This policy will clearly be suboptimal if customers can lie about their type and there is no mechanism to ensure that the information revealed is truthful. Decentralized policies on the other hand are easy to implement since every arriving customer makes an individually optimal decision for itself and a centralized scheduling or a routing rule is not required. However, the disadvantage to this model is a possible degradation in the quality of service in the system. Often, the interest is to design a mechanism by which a decentralized system has the same or identical behavior and performance measure as that of a centralized system. This technique is known as mechanism design.

Recall the example of an observable single server queue from Naor [50] that was described earlier. It was observed that the optimal and the equilibrium threshold satisfies $n_o \leq n_e$. Now suppose that in this decentralized model, each arriving customer is required to pay an admission price $p$ before joining the queue of the server. It is shown in Naor [50] that the equilibrium threshold now depends on the value of $p$ and that $n_e(p)$ decreases with increase in $p$. Further there exists an admission price $p^*$ such that $n_o = n_e(p^*)$. The price payment is just an internal transfer of money between the customers and the server. The prices internalize the negative externality imposed by an arriving customer and therefore the decentralized system has the same social utility as that of a centralized system. The example shows how admission pricing
can be used for mechanism design. Hassin [30] provides an alternative mechanism design for this example that does not require setting an admission price. It is shown that by using a LIFO - Preemptive scheduling policy, a customer’s individually optimal queue join decision coincides with the socially optimal one. Because of the LIFO -PR policy, a customer’s decision to join the queue imposes no externality on subsequent arrivals and the objective of the society and the customer coincides. Mechanism design using pricing has been a popular approach in literature. Pigou [55] was among the first to use admission prices as a mechanism to internalize the negative externality imposed by an arriving customer. Hence these prices are commonly known as Pigouvian prices. Several similar models have been considered that use admission pricing for mechanism design [44, 27, 47, 48, 17, 46].

1.6 Overview of the Thesis

In the preceding discussion, we have seen examples of centralized and decentralized routing as well as scheduling in various service systems. We observed that when the customers belong to a heterogeneous population, the scheduling or routing decision for a customer certainly depend on the customer type. The aim of the thesis is to study such routing and scheduling schemes (for centralized and decentralized decisions) in queueing systems with heterogeneous customers.

We first consider a centralized service system with parallel servers. Here the arriving customers belong to a finite multiclass population and each class differs in its cost for unit delay. Parallel servers are seen in service systems such as web server farms, grid computing clusters, communication networks, road networks and health-care systems. An important problem in such systems is to optimally route the arriving customers (jobs or connections) to the servers to optimize a global objective function. In this problem, the expected sojourn time of a customer is the quality of service of interest. We then consider a decentralized system where each server charges an admission fee, independent of the class, to each customer. The customers are selfish and each arriving customer will randomly choose the server that minimizes its expected cost. The equilibrium behavior of such a system is investigated in this problem. Finally our interest is to assign the admission prices in such a decentralized system to achieve a specific equilibrium load distribution; in particular, the optimal load distribution of the centralized scheme.

Continuing with the decentralized system of parallel servers and heterogeneous customers, we next consider a revenue maximization problem where the objective of the service system is
to maximize the revenue rate that it receives. The revenue rate of a server is defined as the product of the admission price and the aggregate arrival rate of customers joining the queue of the server. To monetize from the heterogeneous customers, the objective of the service system is to determine the revenue maximizing admission price at the servers. We analyze such a revenue maximization problem in the first part of Chapter 3. In the second part of this chapter, we consider the problem of estimating the probability distribution function that characterizes the delay sensitivity of the heterogeneous customers. For most practical scenarios, such a distribution function need not be known to the operator of the service system. Hence to optimize its revenue, it is useful for the service system to be able to characterize the delay sensitivities of the customers. We provide a simple method to estimate the delay sensitivities for the customers and illustrate the method with numerical examples.

In the two problems discussed above, the queueing system consists of parallel servers and the customers are either routed centrally or are required to make an individually optimal choice of the server. Often, service systems comprise of a single server and all customers are required to obtain service from this server. In such system, the choice of the scheduling discipline plays an important role in the underlying optimization for system resources. Since we assume throughout the thesis that customers are heterogeneous, it is important to choose a scheduling discipline that can offer different service classes. In our next problem, we analyze a single server system with the discriminatory processor sharing policy (DPS). Specifically, we consider a DPS server with finite number of service classes and consider the problem of routing the heterogeneous customers to the service classes in such a way that a certain social utility function is maximized. We then analyze a decentralized system where the arriving customers make an individually optimal choice of the service class. Each service class has an associated admission price and the cost to customer is the sum of the admission price and the waiting cost. Taking inspiration from the solution structure in the first problem, we provide a corresponding mechanism design method based on admission prices.

While the first two problems in the thesis consider parallel server systems the third problem analyzes a single server system with multiple service classes. As an alternative, one could have a service system that is a combination of the two systems considered above. In the last chapter of the thesis, we consider such a two server system where one of the server has a priority scheduling discipline while the other server has a FIFO scheduling discipline. A customer that chooses the priority server is required to pay a bid that determines its priority in the queue. We
call such queues with bidding as highest-bidder-first (HBF) queues. Service systems using a highest-bidder-first policy have been studied in queueing literature for various applications and in economics literature to model corruption [45]. Such systems have applications in modern problems like scheduling jobs in cloud computing scenarios or placement of ads on web pages. However, using a HBF service is like using a spot market and may not be preferred by many users. For such users, it may be good to provide a simple scheduler, e.g., a FIFO service. Further, in some situations it may even be necessary that a free service queue operates alongside a HBF queue. Motivated by such a scenario, we propose and analyze a service system with a FIFO server and a HBF server in parallel. We assume that the service system is decentralized, i.e., the heterogeneous customers are required to choose one of the two servers for service. In Chapter 5, we characterize the equilibrium routing of customers to this two server system. We also analyze the impact of a free FIFO server on the revenue made from the HBF server. Finally, we conclude with the equilibrium analysis for the case when the FIFO server now has an admission price.

1.7 Contribution of the Thesis

The problems considered in Chapters 2, 3, 4 and 5 of the thesis analyze the routing or scheduling of customers in differentiated service systems. An important aspect of such service systems is the heterogeneous nature of the arriving customers. The contributions of the main chapters is provided below.

1.7.1 Chapter 2

In Chapter 2, we consider a queueing system with multiple heterogeneous servers serving a multiclass population. The classes are distinguished by time costs and we assume that all customers have i.i.d. service requirements. We assume that the arriving customers do not see the instantaneous queue lengths. While the arrivals are randomly routed to one of the servers, the routing probabilities are determined centrally to optimize the expected waiting cost. This is, in general, a difficult optimization problem and we obtain the structure of the routing matrix. In the second part of the problem, we consider a system in which each queue charges an admission price. The arrivals are routed randomly to minimize an individual objective function that includes the expected waiting cost and the admission price. Once again, we obtain the structure
of the equilibrium routing matrix for this case. Finally, we determine the admission prices to make the equilibrium routing probability matrix equal to a given optimal routing probability matrix.

1.7.2 Chapter 3

In this chapter, we consider a parallel server system where the objective of the service system is to maximize its revenue rate. To simplify the analysis we assume that there are two parallel servers and that the revenue made by the service system is the sum of the revenue from the two servers. Such a service system is an example of a monopoly. To ensure that the monopoly cannot increase the prices at the two servers to infinity, we assume that the service system fixes the admission price at one of the server. Due to the heterogeneity of the customers, we first identify the difficulty in analyzing the revenue maximization problem. The main difficulty lies in characterizing the equilibrium arrival rates to the servers as a function of the admission prices. To be able to analyze the problem, we provide an equivalent formulation that simplifies the analysis for the problem. In this alternative formulation, we treat the equilibrium arrival rate at one of the servers as a variable and obtain the required admission price and hence the revenue rate as a function of this arrival rate. We then perform the revenue maximization under different assumptions on the delay cost function and the delay sensitivity of the customers. Towards the end, we also provide a preliminary method to estimate the distribution that determines the heterogeneity of customers.

1.7.3 Chapter 4

In this chapter, we consider a single server queue with the DPS scheduling policy. As earlier, customers belong to a heterogeneous population and have to obtain service from this server. A DPS scheduling policy is considered because it allows for differential treatment to different customers. We consider the problem of assigning heterogeneous customers to the DPS server with $M$ service classes. In the first part of this chapter, we consider the problem of optimal assignment of the heterogeneous customers to the service classes to minimize a social cost function. We show that when there is a continuum of customer types, the optimal allocation is of a threshold type. For the special case of $M = 2$ we also explicitly characterize the threshold. When the customer types are finite, we show that at the optimal allocation, customers with the
highest (resp. lowest) delay cost will always have to be routed to the service class with highest (resp. lowest) weight. Further, when there are three customer types, we provide condition under which the middle class is indifferent to the choice of service class. Taking inspiration from the analysis of Wardrop equilibrium in Chapter 2, we then consider a DPS system that has different admission charges for different service classes and the customers can choose their service class for individual optimization. Here we show that if the prices are fixed, the individually optimally routing policy is a threshold policy. Finally, we return to mechanism design as seen in Chapter 2 and investigate the use of admission price so that the resulting equilibrium allocation coincides with the optimum allocation that minimizes the social cost.

1.7.4 Chapter 5

Finally in Chapter 5, we consider a system of FIFO and a highest-bidder-first server (HBF) and the heterogeneous customers have to choose between the two servers to obtain service. The HBF queues were first introduced by Kleinrock [40] and subsequently analyzed in detail by Lui [45] and Glazer and Hassin [29]. For such a system, we assume that the arriving customers are from a heterogeneous population with different valuations of their delay costs. The customers strategically choose between FIFO and HBF service and if HBF is chosen, they also choose the bid value to optimize an individual cost. We first characterize the Wardrop equilibrium in such a system. Since one of the server is an HBF server, we also analyze the revenue made by the system. We see that when the total capacity is fixed and is shared between the FIFO and HBF servers, revenue is maximized when the FIFO capacity is non zero. However, if the FIFO server is added to an HBF server, then the revenue decreases with increasing FIFO capacity. We also discuss the case when the FIFO service is not free and analyze the resulting Wardrop equilibrium.
Chapter 2

Optimal and Equilibrium Routing for Parallel Queues

2.1 Introduction

In this chapter, we study service systems involving customers of multiple types or classes, any of whom can be served by any one of several parallel heterogeneous servers. We assume that each queue has a FIFO scheduling discipline. Customers arrive into the system according to a random process, reside in a queue while waiting for service or being served, and then depart. Customer classes differ in their aversion to some metric based on waiting time, delay or congestion in the queue. We seek to determine how customers may be assigned to servers in such a way as to optimize a social welfare function, and also how pricing may be used to incentivize selfish customers to achieve the same social optimum.

There are several examples of service systems with parallel servers. Supermarkets have multiple payment counters and customers are required to join the queue for one of them to make their payment. In case of road networks, toll plazas with several toll gates can also be treated as systems with parallel servers. Mass transit systems usually have several parallel ticket counters for customers to buy their tickets. Other service systems that fall within our framework include healthcare systems, web server farms, cloud and grid computing clusters and communication networks. These are examples of delay systems where the users experience a congestion or delay before obtaining service. The customers that use the system however need not have identical preferences for the service quality metric. Typical such metrics for delay systems are the mean sojourn time of a customer in the system, the number of customers
waiting for service in a queue or some moments of the sojourn time. We assume that the customers belong to a multiclass population and that customers from different classes have strictly different preferences for the underlying metric. We capture customer preferences in the form of a cost function that is based on the mean delay or waiting time per customer in the system. We distinguish between customer classes by applying suitable multipliers to these functions. These multipliers are different for different classes and the multiplication converts the delay function into a corresponding delay cost function. Since the multipliers are distinct for the classes, a unit amount of delay corresponds to a distinct delay cost for customers of a class. The multipliers capture the delay aversion for the different classes quantitatively and shall enable us to define social and individual utility functions suitably.

We begin our analysis of the problem by first defining a social utility function. This function is defined as the sum of the delay costs of all customers of various classes that are present in the different queues. We consider an optimization problem to minimize this utility function and provide structure on the optimal routing rule for the problem. Next, we consider selfish customers who route themselves so as to optimize their individual expected utility. In this setting, we allow each server to charge a fixed, class-independent, admission price to each customer using it, and model the interaction between customers as a game. The cost to a customer at a queue is the sum of the admission price and the corresponding delay cost at that queue. We study properties of the Nash equilibria of this game, which are termed Wardrop equilibria in this context [62]. Finally, we obtain admission prices for which the corresponding Wardrop equilibrium coincides with the socially optimal routing.

2.1.1 Previous Work

There is a substantial literature on the allocation of multi-class customers to parallel servers or queues in both centralized and decentralized settings, including a variety of pricing schemes and game-theoretic formulations. Below, we describe some of the work that is related to the approach taken in this chapter. We use Kendall’s notation for queueing models throughout.

Borst [16] studied the probabilistic allocation of multiclass traffic to parallel $M/G/1$ queues so as to minimize a specific social cost function, namely the total mean waiting cost per unit of time. The arriving customers are from a multiclass population and a class $j$ is characterized by the two tuple $\left( \frac{a_j}{\tau_j}, \frac{\tau_j}{\tau_j} \right)$ where $a_j$ is the waiting cost per unit time while $\tau_j$ and
\( \tau_j^{(2)} \) denote the first and the second moment of the service requirement. It is shown that in an optimal allocation, customer types must be clustered according to their \( \left( \frac{\alpha_j}{\tau_j}, \tau_j^{(2)} \right) \) values and there can be at most one customer class between any two different servers. It is shown that while each server handles a traffic mix which is as homogeneous as possible, different servers have traffic mixes of a different composition. A special case is then analyzed where the service rates at the servers are identical and when the customer classes form an ordered set. For the case when the customer classes are all identical, it is shown that load balancing will lead to an optimal allocation minimizing the total waiting cost. However when the customer classes are not identical, it is shown that an optimal allocation will never have the load at all the servers to be equal or balanced. Sethuraman and Squillante [57] considered a variant of this problem where in addition to optimal routing, one is required to also choose an order in which the multiclass customers are served in the queues. It was shown that an optimal sequencing policy for the servers is to prioritize customers of higher classes. While the routing problem could not be solve explicitly, it was shown that it had an interior solution. There are significant differences between our work and that of [16, 57]. The structure that we obtain for our optimal routing is not the same as in Borst [16]. Further, our results apply to a very general class of queueing models and cost functions. We also consider a game-theoretic setting of selfish optimization and a pricing mechanism that will achieve social optimality with selfish optimization. While Sethuraman and Squillante [57] also consider the optimal sequencing in a queue, we assume that the servers cannot discriminate between classes in our model, which may be more realistic depending on the application.

The equilibrium allocation of customers in a multiqueue system is studied by Bell and Stidham [14] and Haviv and Roughgarden [34]. Bell and Stidham [14] studied a single nonbalking traffic class served by a set of parallel \( M/G/1 \) queues. The servers differ in their holding cost per unit time denoted by \( h_i \) and the service rate that is denoted by \( \mu_i \) for server \( i \). With the assumption that \( h_1/\mu_1 \leq h_2/\mu_2 \ldots \leq h_M/\mu_M \), the social cost minimization problem is solved using the generalized Lagrangian technique while the Wardrop conditions are used to identify the structure of the individually optimal routing. To compare the two allocations, it is shown that if a server is not used in an optimal allocation, then it will also not be utilized in the individually optimal routing. Further, it is shown that an individually optimal routing overloads the lower indexed servers, as compared to the socially optimal one. Restricting their attention on this problem to parallel \( M/M/1 \) queues, Haviv and Roughgarden [34] obtain an
upper bound on the price of anarchy (PoA), defined as the ratio of the total cost at the Wardrop 
equilibrium to that at the social optimum. In addition to allowing a general cost function we 
also consider a multiclass population of customers.

There are several works that use admission prices to reduce congestion [50, 27, 44, 17, 46]. In Naor [50] and Edelson and Hilderbrand [27], customers, who belong to a single class, have 
to choose between paying an admission price to enter the queue, incurring a delay cost and 
receiving a fixed reward for service, or balking (leaving without being served). Admission 
prices are set by an operator who seeks to maximize revenue. If customers can observe the 
queue length on arrival and base their balking decision on it, then the revenue maximizing 
admission price exceeds the one that maximizes social welfare ([50]). However, if customers 
cannot observe the queue and have to base their decision on only the known arrival and service 
rates, then these two admission prices coincide ([27, 44]). In the latter setting, Littlechild [44] 
obtains the admission fee as a Pigouvian tax and shows that this will induce a socially optimal 
arrival rate. While Edelson and Hilderbrand [27] and Littlechild [44] consider a single class 
of customers and a single M/M/1 queue with the delay as the cost, we consider multiclass 
customers, multiple servers and also a general cost function. More importantly, we do not allow 
balking and all customers have to join one of the queues. Bradford [17] extends some of the 
work ([50, 27, 44]) to multiclass customers each with its own delay cost and reward for service 
and obtains the Pigouvian admission charge for each class that achieves the socially optimal 
allocation. The admission charge is independent of the queue from which the customer receives 
service. This, and the fact that the system needs to elicit information of the customer class, 
we believe, makes their model inapplicable in many situations. More importantly, like in the 
preceding references, the objective is to reduce congestion in the system by allowing customers 
to balk. Masuda an Whang [46] considers a similar model except that customers of different 
classes have the same delay cost.

In some applications, it is natural to associate a dispatcher with each class that seeks to 
allocate customers of that class to servers in such a way as to minimize their expected waiting 
time. This setting, with M/G/1-processor sharing queues, is studied by Altman et al. [6] and 
Ayesta et al. [10]; the authors obtain bounds on the price of anarchy in the resulting game. 
Such a class-based routing problem is also seen in communication networks where each class 
is required to route its flow from an origin to the destination using different routes of the given 
network. The routes are composed of network links where the multiclass flow experiences a
delay due to congestion from the other flows. In such models, the interest is in obtaining the Nash equilibrium routing such that no class has an incentive to deviate from its flow allocation. The existence and uniqueness of such Nash equilibria for class based routing in communication networks is well studied [53, 42, 17]. Such a class based routing model is different from our decentralized routing at Wardrop equilibrium. We shall not consider such a class based routing model any further in the thesis.

An alternate use of admission prices is in purchasing priorities [11, 45, 56, 3]. This will be of interest to us in Chapter 5. See Hassin and Haviv [33] for a comprehensive survey of these models. More recently, there have been papers proposing the use of differentiated prices in the Internet and studying the resultant user strategies and equilibria [52, 37, 25, 15]. In the next section we develop the notation and describe the system model. Throughout the chapter, we will use a two-server, two-class system as a running example.

### 2.2 Model and Problem Formulation

We consider a system with $M$ classes of customers and $N$ queues. Class $m$ customers arrive according to a Poisson process of rate $\lambda_m$, independent of other classes. The allocation of arriving customers to queues has to be made with no knowledge of current or past queue occupancies, or past arrival times or routing decisions. Such an assumption may be less realistic for centralized allocation than when customers have to make individual decisions. Nevertheless, imposing this assumption uniformly permits clearer comparison of the two settings. Under this assumption, it is natural to restrict attention to Markovian routing policies, i.e., to policies which route customers of class $i$ to queue $j$ with some fixed probability $p_{ij}$. This is also the class of policies considered in Borst [16] and Sethuraman and Squillante [57]. Under Markovian routing, the aggregate arrival process into queue $j$ is a Poisson process of rate

$$\gamma_j = \sum_{i=1}^{M} \lambda_i p_{ij}. \quad (2.1)$$

We assume that customers of all classes have the same job size distributions, and that, once they join a queue, they are treated identically within it. Associated with queue $j$ is a cost function $D_j(\cdot)$ that specifies a cost associated with a given aggregate arrival rate. For example, the cost could be the mean sojourn time, or some higher moment of it, or the probability of the sojourn time exceeding a specified threshold. Our only assumption is that each function $D_j$ be
monotone increasing, and continuously differentiable in the interior of its domain (the set of arrival rates for which \( D_j \) is finite) with strictly positive derivative.

Finally, with each class \( i \), we associate a positive parameter \( \beta_i \) that quantifies its sensitivity to delay or congestion by multiplying the cost incurred by a class \( i \) customer by \( \beta_i \). The only distinction between customer classes is in applying different multipliers \( \beta_i \) to their costs in any queue. Without loss of generality, we take \( \beta_1 > \beta_2 > \ldots > \beta_M \); if \( \beta_i = \beta_j \), we can collapse them into a single class, as customers are otherwise assumed to be identical.

The assumptions above are rather mild. We do not restrict the number of servers at a queue or the service discipline. Indeed, different queues may have different numbers of servers and employ different service disciplines. They can also employ different cost functions, for example the mean sojourn time at one queue and the second moment at another. The only requirement is that each queue treat all customers alike, irrespective of their class. In addition to traditional queueing models, our set-up also encompasses transport models for example, where the mean journey time on a road may be some function of the traffic intensity on it. The main motivation for the assumption of Poisson arrivals is that it makes the \( D_j \) functions of a single real variable. It is not obvious how the monotonicity and differentiability assumptions would generalize if \( D_j \) were to be a function of the law of a stochastic process.

We are now ready to state the social welfare maximization problem. The objective is

\[
\inf_P U(P) = \sum_{i=1}^M \sum_{j=1}^N \beta_i \lambda_i p_{ij} D_j(\gamma_j),
\]

where the infimum is taken over all right stochastic matrices \( P \) (defined as matrices with non-negative entries whose row sums are unity), and the \( \gamma_j \) depend on \( P \) through (2.1), though this dependence has not been made explicit in the notation. Thus, the social cost is defined as the sum of the expected costs incurred by customers of different classes at different queues, weighted by the corresponding flow rates denoted henceforth by \( \gamma \).

Next, we consider the formulation of a game between customers. Here, we allow the queues to charge admission prices, denoted by \( c_j \) at queue \( j \). Without loss of generality, we take \( c_1 > c_2 > \ldots > c_N \); if \( c_i = c_j \), then we can collapse these two queues into a single queue whose delay function is the inf-convolution of the delay functions of its constituent queues, i.e.,

\[
D(\gamma) = \inf\{D_i(\gamma_1) + D_j(\gamma_2) : \gamma_1, \gamma_2 \geq 0, \gamma_1 + \gamma_2 = \gamma\}.
\]

The goal of a class \( i \) customer entering the system is to choose a queue \( j \) so as to minimize \( c_j + \beta_i D_j(\gamma_j) \) where \( \gamma_j \) is determined through the strategies of all customers. We assume that
the rates $\lambda_1, \ldots, \lambda_M$, the cost functions $D_j(\cdot)$ and the parameters $\beta_i$, $i = 1, \ldots, M$ and $c_j$, $j = 1, \ldots, N$ are all common knowledge. Under the additional assumptions, noted earlier, that a customer does not have access to current or past queue occupancies, or the history of arrival times or decisions, its strategy is necessarily restricted to choosing a server according to a fixed probability distribution, albeit one that may depend on its class. Thus, again, the joint strategies may be represented by a right stochastic routing matrix, $P$. We recall the condition for such a routing matrix $P$ to be a Wardrop equilibrium:

$$\forall i, j, k \quad p_{ij} > 0 \quad \Rightarrow \quad c_j + \beta_i D_j(\gamma_j) \leq c_k + \beta_i D_k(\gamma_k).$$

(2.2)

In words, the condition says that if a customer has positive probability of using a queue, then its expected cost in that queue must be no higher than its expected cost in any other queue. In particular, if a customer has positive probability of using both queues $j$ and $k$, then the above inequality must hold in both directions, Hence, the expected total cost in the two queues is equal in this case. A Wardrop equilibrium is simply a mixed strategy Nash equilibrium, and hence is guaranteed to exist. The Wardrop equilibrium is well studied for the class of non-atomic selfish routing games. The term non-atomic refers to the fact that the impact of any individual customer and its strategy on the performance of the system is infinitesimally small. For an overview of the non-atomic selfish routing, refer [51](Chapters 18).

The model described in the preceding discussion considers the case when the customers belong to one of the $M$ classes and all customers of class $i$ have the same parameter $\beta_i$. An alternative model to characterize the delay sensitivity of the arriving customers is by assuming that such sensitivities $\beta$ are from a continuum of values. Such a model is more realistic in service systems concerning human beings (such as healthcare systems or transport systems) where the cost for a unit delay for any individual need not be restricted to a set of $M$ finite values. In this paper, in addition to discrete classes we also consider this case with a continuum of customer classes. We model such a continuum of customer classes more formally as follows. Suppose that customer arrivals constitute a marked Poisson process of intensity $\lambda \times F$ on $\mathbb{R} \times \mathbb{R}_+$, where $F$ is a cumulative distribution function supported on the positive reals. Here, the marks, $\beta$, denote the delay sensitivities of the customers. The marks are i.i.d. random variables, independent of the arrival time of the customer, and of the past (arrival times and marks) of the process. We assume that the marks are drawn from a distribution $F$ which is absolutely continuous with respect to Lebesgue measure, with density $f$. We further assume that for every $\beta$ in the interior of its support, $f(\beta)$ is bounded away from zero and $\infty$. 

25
In the next section, we consider the welfare optimization problem and describe the structure of a matrix $P^*$ solving (P1). In Section 2.4, we show that a solution $P^W$ of (2.2) also has a similar structure, for any admission prices. We then show how to choose the admission prices so as to make $P^W$ coincide with $P^*$, and illustrate our general results with a numerical example. In each of these sections, we explain the details with the help of a running example. We conclude in Section 2.7 with a discussion of some open problems.

2.3 Social Welfare Optimization

In this section we will analyze the social welfare optimization problem (P1). We begin with a two class - two server service system and analyze the properties of the optimal routing for this example. The two class, two server system will be used throughout this chapter as a running example to aid better understanding of the discussion.

2.3.1 A Two Class - Two Server Example

Consider a two class - two server service system where the Class $m$ customers arriving according to a stationary Poisson process of rate $\lambda_m, m = 1, 2$. The service requirement for customers of either class are assumed to be i.i.d. exponential with unit mean and the two servers have a unit service rate. To simplify the discussion we assume that $1 > \lambda_1 + \lambda_2$, and $\beta_1 > \beta_2$. The arrivals of class $m$ are routed to queue $n$ with probability $p_{mn}$ independent of the routing of other customers. Since $p_{m1} = 1 - p_{m2}$, $U(\cdot)$ may be seen to be a function of only $(p_{11}, p_{21})$ taking values in $[0, 1]^2$. To simplify notation, we will use $p_1 = p_{11}$ and $p_2 = p_{21}$.

Since the arrivals are Poisson, the routing is a Bernoulli sampling of the arrivals and the service time is exponential, each of the two queues is an M/M/1 queue with $D_n(\gamma_n) = 1/(1 - \gamma_n)$. Therefore the social cost is

$$U(p_1, p_2) = (p_1\lambda_1\beta_1 + p_2\lambda_2\beta_2)(D_1(\gamma_1) - D_2(\gamma_2)) + (\lambda_1\beta_1 + \lambda_2\beta_2)D_2(\gamma_2). \tag{2.3}$$

We now obtain $(p_1^*, p_2^*)$, a minimizer of $U(p_1, p_2)$. Specifically, we show the following: (i) there exists a stationary point of $U(p_1, p_2)$ in the interior of $[0, 1]^2$, and (ii) $(p_1^*, p_2^*)$ is on the boundary of $[0, 1]^2$. These two together imply that the stationary point is not the global optimum and hence $U(p_1, p_2)$ is not a convex function. More formally, we have the following theorem.
Figure 2.1: Illustrating the feasible set of \((p_1^*, p_2^*)\), the minimizer of \(U(p_1, p_2)\) as determined in Theorem 2.1.

**Theorem 2.1.** \((p_1^*, p_2^*)\) is on the boundary of \([0, 1]^2\) and belongs to only the following sets.

1. \(\mathcal{R}_1 = \{p_1, p_2 : (p_1, p_2) \in S_1, p_1 \lambda_1 + p_2 \lambda_2 \leq (\lambda_1 + \lambda_2)/2\}\)

2. \(\mathcal{R}_2 = \{p_1, p_2 : (p_1, p_2) \in S_2, p_1 \lambda_1 + p_2 \lambda_2 \geq (\lambda_1 + \lambda_2)/2\}\)

where

- \(S_1 = \{p_1, p_2 : p_1 = 1, 0 \leq p_2 \leq 1 \text{ or } 0 \leq p_1 \leq 1, p_2 = 0\}\).

- \(S_2 = \{p_1, p_2 : p_1 = 0, 0 \leq p_2 \leq 1 \text{ or } 0 \leq p_1 \leq 1, p_2 = 1\}\).

These regions are illustrated in Fig. 2.1.

**Proof.** We first show that there exists an interior stationary point of \(U(p_1, p_2)\). The partial derivative of \(U(p_1, p_2)\) with respect to \(p_m\) is as below.

\[
\frac{\partial U(p_1, p_2)}{\partial p_m} = \phi_m(p_1, p_2) + \lambda_m \beta_m (D_1(\gamma_1) - D_2(\gamma_2)) \quad (2.4)
\]
Here

\[ \phi_m(p_1, p_2) = \lambda_m (p_1 \lambda_1 \beta_1 + p_2 \lambda_2 \beta_2) \left( \frac{\partial D_1(\gamma_1)}{\partial p_m} - \frac{\partial D_2(\gamma_2)}{\partial p_m} \right) + \lambda_m (\lambda_1 \beta_1 + \lambda_2 \beta_2) \frac{\partial D_2(\gamma_2)}{\partial p_m}. \] (2.5)

\[ \frac{\partial D_1(\gamma_1)}{\partial p_m} = \frac{\lambda_m}{(1 - \gamma_1)^2} \] (2.6)

\[ \frac{\partial D_2(\gamma_2)}{\partial p_m} = -\frac{\lambda_m}{(1 - \gamma_2)^2} \] (2.7)

for \( m = 1, 2 \). It can be verified that \( p_1 = p_2 = 0.5 \) is a stationary point of \( U(p_1, p_2) \).

Define \( \gamma^+ := \{ \gamma_1 : D_1(\gamma_1) = D_2(\gamma_2) \} \) and \( \mathcal{L}_1 := \{(p_1, p_2) : \gamma_1 = \gamma^+ \} \). Observe that \( \mathcal{L}_1 \) is a straight line. Clearly, \( \gamma^+ = (\lambda_1 + \lambda_2)/2 \) and is unique; further, \((0.5, 0.5) \in \mathcal{L}_1 \).

On \( \mathcal{L}_1 \), \( U(p_1, p_2) = (\lambda_1 \beta_1 + \lambda_2 \beta_2) D_2(\gamma^+) \), a constant. Now consider the straight line \( \mathcal{L}_2 \) defined by

\[ p_1 \lambda_1 \beta_1 + p_2 \lambda_2 \beta_2 = (\lambda_1 \beta_1 + \lambda_2 \beta_2)/2. \]

\( \mathcal{L}_1 \) and \( \mathcal{L}_2 \) intersect at \((p_1, p_2) = (0.5, 0.5)\). Since \( \beta_1 > \beta_2 \), the absolute value of the slope of \( \mathcal{L}_2 \) is higher than that of \( \mathcal{L}_1 \). For any \((p_1, p_2)\) on \( \mathcal{L}_1 \) and to the left of \((0.5, 0.5)\), the following is
Figure 2.3: For each $\gamma_1 < \gamma^+$, $(p_1, p_2)$ minimizing Program $\mathcal{P}_2$ is in $R_1$ while for $\gamma_1 > \gamma^+$, $(p_1, p_2)$ minimizing $\mathcal{P}_2$ is in $R_2$.

satisfied for $m = 1, 2$.

\[
p_1 \lambda_1 \beta_1 + p_2 \lambda_2 \beta_2 < \frac{(\lambda_1 \beta_1 + \lambda_2 \beta_2)}{2} = \frac{-(\lambda_1 \beta_1 + \lambda_2 \beta_2) \frac{\partial D_2(\gamma_2)}{\partial p_m} - \frac{\partial D_1(\gamma_1)}{\partial p_m} \frac{\partial D_2(\gamma_2)}{\partial p_m}}{\left( \frac{\partial D_1(\gamma_1)}{\partial p_m} - \frac{\partial D_2(\gamma_2)}{\partial p_m} \right)} \quad (2.8)
\]

The equality is obtained by evaluating the partial derivatives from (2.6) and (2.7). Rearranging the terms of (2.8) we have

\[
(p_1 \lambda_1 \beta_1 + p_2 \lambda_2 \beta_2) \left( \frac{\partial D_1(\gamma_1)}{\partial p_m} - \frac{\partial D_2(\gamma_2)}{\partial p_m} \right) + (\lambda_1 \beta_1 + \lambda_2 \beta_2) \frac{\partial D_2(\gamma_2)}{\partial p_m} < 0 \quad (2.9)
\]

This implies that for $(p_1, p_2)$ on $L_1$ and to the left of $(0.5, 0.5)$, we have $\phi_m(p_1, p_2) < 0$ and hence $\partial U(p_1, p_2)/\partial p_m < 0$. Using identical arguments, we can show that for $(p_1, p_2) \in L_1$ and to the right of $(0.5, 0.5)$, $\partial U(p_1, p_2)/\partial p_m > 0$. Refer Fig. 2.2 for details.

Clearly, $\partial U(p_1, p_2)/\partial p_m \neq 0$ on the either side of $(0.5, 0.5)$ on the line $L_1$. This implies that there exists $(p_1, p_2)$ on either side of the line $L_2$ such that $U(p_1, p_2) < U(0.5, 0.5)$ implying that the stationary point $(0.5, 0.5)$ is not a global optimum.

We will now show that $(p_1^*, p_2^*) \in R_1 \cup R_2$. For $\gamma_1 < \gamma^+$ we see that $D_1(\gamma_1) < D_2(\gamma_2)$,
and for $\gamma_1 > \gamma^+$ we see that $D_1(\gamma_1) > D_2(\gamma_2)$. Now for a fixed $\gamma_1 < \gamma^+$, $D_1(\gamma_1)$ and $D_2(\gamma_2)$ are constant and since $D_1(\gamma_1) < D_2(\gamma_2)$, we see that $U(p_1, p_2)$ is minimized by the following linear program.

$$\arg \max_{p_1, p_2} \quad p_1 \lambda_1 \beta_1 + p_2 \lambda_2 \beta_2 \quad \text{subject to} \quad p_1 \lambda_1 + p_2 \lambda_2 = \gamma_1$$

In $P_2$, since $\beta_1 \neq \beta_2$, the slope of the objective function and that of the constraint are different. Thus there exists a unique $(p_1, p_2)$ at the boundary of the feasible space which is a solution to $P_2$. For $\gamma_1 > \gamma^+$, the arg max in $P_2$ is replaced by arg min and the same argument applies. The above is illustrated in Fig. 2.3 where the dark dotted lines represent the solution to $P_2$ for different values of the equality constraint $\gamma_1$. This proves the theorem.

The following two corollaries are now obvious from the theorem.

**Corollary 2.2.** The cost function $U(p_1, p_2)$ is not convex.

*Proof.* The stationary point $(1/2, 1/2)$ is not a global optimum.

**Corollary 2.3.** There are at most four minimizers of $U(p_1, p_2)$.

*Proof.* On each edge of $[0, 1]^2$ one of $p_1$ and $p_2$ is a constant, $U(\cdot)$ is a function of the other. It can be verified that $U(p_1, 0)$ and $U(p_1, 1)$ (resp. $U(0, p_2)$ and $U(1, p_2)$) is a convex function of $p_1$ (resp. $p_2$) and the result follows.

### 2.3.2 Characterizing the Structure of $P^*$.

From the previous section, we now know that in general, $P_1$ need not be a convex program. Also, except for very simple forms of $D(\gamma)$, calculating even the local optima using the KKT conditions becomes messy. However as the maps $P \mapsto \gamma_j$ and $D_j$ are continuous, the problem is one of minimizing a continuous function over the compact set of right stochastic matrices. Hence, there is a matrix $P^*$ achieving the infimum. We now characterize some properties of the matrix $P^*$. We begin with the following definition.

**Definition 2.4.** An allocation matrix $P$ satisfies property $P_1$ if the following is true: $p_{mn} > 0$ for all $n : n_m \leq n \leq n^m$, and $p_{mn} = 0$ elsewhere with $n_1 \leq n^1 \leq n_2 \leq n^2 \leq \ldots \leq n_M \leq n^M$. 


If $P$ satisfies Property $\mathcal{P}_1$, then each class uses a contiguous set of servers, adjacent classes have at most one common server and adjacent servers have at most one common class. Note that a particular class may use more than one server and a server may serve more than one class. We now have the following theorem that shows the existence of a $P^*$ satisfying property $\mathcal{P}_1$.

**Theorem 2.5.** Consider an allocation matrix $P$, with the servers indexed such that $D_1(\gamma_1) \leq D_2(\gamma_2) \ldots \leq D_N(\gamma_N)$. There exists a $\hat{P}$ that has property $\mathcal{P}_1$ and with $U(\hat{P}) \leq U(P)$. This implies that there exists an optimal allocation $P^*$ that has property $\mathcal{P}_1$.

**Proof.** We consider the case when the given $P$ does not satisfy property $\mathcal{P}_1$. The case when $P$ satisfies property $\mathcal{P}_1$ is trivial in which case $\hat{P} = P$. When the given $P$ does not satisfy property $\mathcal{P}_1$, it is easy to see that there exist classes $1 \leq a < b \leq M$ and servers $1 \leq k < l \leq N$ such that (i) $D_k(\gamma_k) \leq D_l(\gamma_l)$, (ii) $\beta_a > \beta_b$, and (iii) $p_{a,l}, p_{b,k} > 0$. We first describe a swap of traffic between servers $k$ and $l$ of traffic classes $a$ and $b$ such that the $\gamma_k$ and $\gamma_l$ remains same and hence does not change $D_k(\gamma_k)$ and $D_l(\gamma_l)$; however, such a swap can decrement $U$. We then use the swap operation and provide an algorithm to obtain a $\hat{P}$ that has property $\mathcal{P}_1$ and with $U(\hat{P}) \leq U(P)$.

Increase $p_{a,k}$ and correspondingly decrease $p_{a,l}$ by $\epsilon_a > 0$. Simultaneously, decrease $p_{b,k}$, and correspondingly increase $p_{b,l}$ by $\epsilon_b$ where

$$\epsilon_b := \epsilon_a \lambda_a / \lambda_b.$$ 

We have thus moved $\delta := \epsilon_a \lambda_a$ volume of Class $a$ traffic from Server $l$ to Server $k$ and the same volume of class $b$ traffic in the opposite direction. This ensures that $\gamma_k$ and $\gamma_l$ remain unchanged and so does $D_k(\gamma_k)$ and $D_l(\gamma_l)$; however the matrix $P$ is now different. The swap decreases $U(P)$ by $\delta(\beta_a - \beta_b)(D_k(\gamma_k) - D_l(\gamma_l))$. Such a swap is feasible when $p_{a,l}, p_{b,k} > 0$. Denote this swap in $p_{a,l}, p_{b,k} > 0$ by $F_{k,l}(a, b, \delta)$.

If $p_{a,k} > 0$, we will say that class $a$ is using server $k$. Let $C^n$ be the set of classes using Server $n$ and $S^m$ be the set of servers serving Class $m$. Further, define

$$A_{m,n} := \{n' : n' \in S^m, n' > n\},$$

to be the set of servers with index greater than $n$ and used by Class $m$. Similarly, let $B_{m,n} := \{m' : m' \in C^n, m' > m\}$ be the set of classes with index higher than $m$ and using Server $n$.

From the preceding definitions, $\forall n' \in B_{m,n}$ and $\forall n' \in A_{m,n}$ we have (i) $p_{m,n'}, p_{m',n} > 0$, (ii) $D_n(\gamma_n) \leq D_{n'}(\gamma_{n'})$ and (iii) $\beta_m > \beta_{m'}$ and hence $F_{n,n'}(m, m', \delta)$ is feasible (i.e., feasible when $\{A_{m,n} \neq \emptyset, B_{m,n} \neq \emptyset\}$.)
Algorithm 1 uses the elementary $F$ swaps to convert an allocation matrix $P$ in which the servers are indexed such that $D_n(\gamma_n) \leq D_{n+1}(\gamma_{n+1})$ for $n = 1, \ldots, N-1$, and returns an allocation matrix $\hat{P}$ for which $\gamma_n$, and hence $D_n(\gamma_n)$, is unchanged from that of $P$, has property $\mathcal{P}_1$, and $U(\hat{P}) \leq U(P)$. This clearly implies that there exists an optimal allocation $P^*$ that has property $\mathcal{P}_1$.

We begin with Class 1 and Server 1 and perform the $F$ operation (described earlier) in which we swap out lower class traffic from Server 1 and bring in an equal amount of Class 1 traffic from a higher delay server. We continue doing this till either Server 1 has only Class 1 traffic (indicated by $B_{1,1} = \emptyset$) or all of Class 1 traffic is in Server 1 (indicated by $A_{1,1} = \emptyset$). In the former case, $n_1 = 1, n_1, n_2 > 1$ and we begin swapping Class 1 traffic into Server 2 (i.e., keeping $m = 1$ and incrementing $n$) indicating that $n_1, n_2 \geq 2$. In the latter case $n_1 = n_1 = 1$ and we begin swapping Class 2 traffic into Server 1 (i.e., incrementing $m$ and keeping $n = 1$) indicating that $n_2 = 1, n_2 \geq 2$. In either case, we have $1 = n_1 \leq n_1 \leq n_2$.

The process described above is continued till we exhaust all classes and all servers (indicated by $m = M$ and $n = N$). Property $\mathcal{P}_1$ is incrementally achieved by the classes and servers as we sweep through $m$ and $n$. Also note that $\delta$ is chosen such that for a given $(m, m')$ and $(n, n')$, the maximum possible amount is swapped; $\delta = p_{n'1} \lambda_{m'}$ implies moving all of Class $m'$ traffic from Server 1 to Server $n'$ while $\delta = p_{1n'} \lambda_1$, implies moving the entire Class 1 traffic from Server $n'$ to Server 1. Thus in each iteration of the while loop of line 4, either $A_{m,n}$ or $B_{m,n}$ is decremented by 1.

To show that Algorithm 1 terminates, it is sufficient to show that the while loop starting on line 4 terminates after a finite number of iterations. This is true because in each iteration either $A_{m,n}$ or $B_{m,n}$ is decremented; hence this loop is executed at most $|A_{m,n} + B_{m,n} - 1|$ times.  

\textbf{Remark 2.6.} If $P$ is such that $D_k(\gamma_k) \neq D_l(\gamma_l)$ for all $1 \leq k < l \leq N$ Algorithm 1 returns a $P$ with a reduced $U$. This is because the decrement in $U$ in every $F$ swap is non zero.

\textbf{Remark 2.7.} If $P^*$ is such that $D_k(\gamma_k^*) = D_l(\gamma_l^*)$ for some $k \neq l$ and that queues $k, l$ are non empty with at least two distinct classes between them, then it can be shown that property $\mathcal{P}_1$ is not necessary for $P^*$. However, in the following theorem, we shall now show that such a $P^*$ with $D_k(\gamma_k^*) = D_l(\gamma_l^*)$ is not possible when there are at least two distinct classes using the two queues $k$ and $l$. This implies that property $\mathcal{P}_1$ is indeed a necessary condition for $P^*$.

\textbf{Theorem 2.8.} Let $P^*$ achieve the minimum in $\mathcal{P}(1)$ and let $\gamma^*$ denote the arrival rates
Algorithm 1: Using the elementary exchange step $F$, $P$ is changed to $\hat{P}$ such that $U(P) \leq U(\hat{P})$ and $\hat{P}$ has Property $\mathcal{P}_1$.

1: Input $P$;
2: $m \leftarrow 1; n \leftarrow 1$;
3: Determine sets $A_{m,n}$ and $B_{m,n}$;
4: while $\{A_{m,n} \neq \emptyset, B_{m,n} \neq \emptyset\}$ do
5: Choose any $m' \in B_{m,n}$ and any $n' \in A_{m,n}$;
6: $\delta \leftarrow \min(p_{m',n}\lambda_{m'}, p_{m,n'}\lambda_m)$;
7: Perform $F_{n,n'}(m, m', \delta)$;
   Update $P, A_{m,n}, B_{m,n}$;
8: end while
9: if $\{n = N, m = M\}$ then
10: exit
11: end if
12: if $\{A_{m,n} = \emptyset, m \neq M\}$ then
13: $m \leftarrow m + 1$;
14: Jump to line 3
15: end if
16: if $\{B_{m,n} = \emptyset, n \neq N\}$ then
17: $n \leftarrow n + 1$;
18: Jump to line 3
19: end if
20: Return $P$
Moreover, it assigns non-zero amounts of class $i_j$ at two queues $P_i$ for optimality at $D_1$. We now apply the Karush-Kuhn-Tucker (KKT) conditions needed for the KKT necessary conditions to be applicable hold because the $P_i$ constraints on the matrices $j$ contradicting the minimality of $P_i$ now, $\beta$ social cost as

The proof of the first part is by contradiction. To lighten notation, we shall write $\beta_1, \lambda_1$ for $\beta_{i_1}, \lambda_{i_1}$. We write $p_{11}$ for $p_{i_1,j_1}$, $p_{m1}$ for $p_{m,j_1}$, $D_1^*$ for $D_1(\gamma_{j_1}^*)$ and so on.

Suppose first that $D_1^* > D_2^*$. For sufficiently small $\epsilon > 0$, it is clear that we can find a routing matrix $P$ such that

$$
\lambda_1 p_{11} = \lambda_1 p_{11}^* - \epsilon, \quad \lambda_1 p_{12} = \lambda_1 p_{12}^* + \epsilon,
$$

$$
\lambda_2 p_{22} = \lambda_2 p_{22}^* - \epsilon, \quad \lambda_2 p_{21} = \lambda_2 p_{21}^* + \epsilon,
$$

and all other elements of $P$ are the same as the corresponding elements of $P^*$. In words, we have shifted a quantity $\epsilon$ of the flow of class $i_1$ customers from queue $j_1$ to queue $j_2$, and an equal quantity of class $i_2$ customers from $j_2$ to $j_1$, leaving all others unaffected. Clearly, the total flow rates $\gamma$ under $P$ are exactly the same as $\gamma^*$. Consequently, we can compute the change in social cost as

$$
U(P) - U(P^*) = \epsilon (\beta_1 - \beta_2)(D_2^* - D_1^*).
$$

Now, $\beta_1 > \beta_2$, while we assumed that $D_1^* > D_2^*$. But this implies that $U(P) < U(P^*)$, contradicting the minimality of $P^*$. This establishes that $D_1^* \leq D_2^*$.

Suppose next that $D_1^* = D_2^*$. Consider the routing matrix $P$ described above, and define the matrices $P^\alpha = \alpha P + (1 - \alpha)P^*$. Then $P^\alpha$ is a right stochastic matrix for all $\alpha \in [0,1]$ and it assigns non-zero amounts of class $i_1$ and class $i_2$ traffic to each of the queues $j_1$ and $j_2$. By the same argument as above, the total flow rates $\gamma^\alpha$ induced by $P^\alpha$ are the same as $\gamma^*$. Moreover, $U(P^\alpha) = U(P^*)$ as $P^\alpha$ only differs from $P^*$ in changing the composition of traffic at two queues $j_1$ and $j_2$ of equal cost, while keeping the total flows unchanged. Hence, $P^\alpha$ also achieves the minimum in (P1). We now apply the Karush-Kuhn-Tucker (KKT) conditions for optimality at $P^\alpha$, where $p_{11}$, $p_{12}$, $p_{21}$ and $p_{22}$ are all strictly between 0 and 1. (The regularity conditions needed for the KKT necessary conditions to be applicable hold because the constraints on $P^*$ are affine.) The KKT conditions imply that

$$
\frac{\partial U(P^\alpha)}{\partial p_{11}} = \frac{\partial U(P^\alpha)}{\partial p_{12}}, \quad \frac{\partial U(P^\alpha)}{\partial p_{21}} = \frac{\partial U(P^\alpha)}{\partial p_{22}}.
$$
Using the definitions of $U$ and $\gamma_j$, we can rewrite the first equality above as

$$
\beta_1 \lambda_1 D_1^\alpha + \beta_1 \lambda_1^2 p_{11}^\alpha (D_1^\alpha)' + \lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m1}^\alpha (D_1^\alpha)' = \beta_1 \lambda_1 D_2^\alpha + \beta_1 \lambda_1^2 p_{12}^\alpha (D_2^\alpha)'
$$

$$
+ \lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m2}^\alpha (D_2^\alpha)',
$$

where we write $(D_1^\alpha)'$ and $(D_2^\alpha)'$ to denote the derivatives of $D_{j1}$ and $D_{j2}$ evaluated at $\gamma_{j1}^\alpha$ and $\gamma_{j2}^\alpha$ respectively. But $D_1^\alpha = D_2^\alpha$ by assumption, and $\gamma_1^\alpha$ coincides with $\gamma^*$ for all $\alpha$ in $[0, 1]$ by construction. Further as $P^\alpha$ only differs from $P^*$ in changing the composition of traffic from Classes $i_1$ and $i_2$ at two queues $j_1$ and $j_2$, the terms $\lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m1}^\alpha (D_1^\alpha)'$ and $\lambda_1 \sum_{m \neq 1} \beta_m \lambda_m p_{m2}^\alpha (D_2^\alpha)'$ are constant for any $\alpha \in (0, 1)$. Denoting these by $\eta_1$ and $\eta_2$ respectively, we obtain that

$$
\beta_1 \lambda_1^2 p_{11}^\alpha (D_1^\alpha)' = \beta_1 \lambda_1^2 p_{12}^\alpha (D_2^\alpha)' + \eta_2 - \eta_1,
$$

for all $\alpha \in (0, 1)$. But this is impossible because the $D'$ are non-zero by assumption, one of $p_{11}^\alpha$ and $p_{12}^\alpha$ is an increasing function of $\alpha$ while the other is a decreasing function. Hence, it is also not possible for $D_1^\alpha$ and $D_2^\alpha$ to be equal. This completes the proof of the first part of the theorem.

The second part will also be proved by a contradiction. Suppose that $P^*$ satisfies $p_{i1j1}^*, p_{i1j2}^* > 0$ and $p_{i2j2}^*, p_{i2j1}^* > 0$ simultaneously. Now since $p_{i1j1}^* > 0$ and $p_{i2j2}^* > 0$, from the first part of the theorem we have $D_{j1}(\gamma_{j1}^*) < D_{j2}(\gamma_{j2}^*)$. However from the assumption that $p_{i2j1}^* > 0$ and $p_{i1j2}^* > 0$, the first part of the theorem also implies $D_{j1}(\gamma_{j1}^*) < D_{j2}(\gamma_{j2}^*)$. Now $D_{j1}(\gamma_{j1}^*) < D_{j2}(\gamma_{j2}^*)$ and $D_{j1}(\gamma_{j1}^*) > D_{j2}(\gamma_{j2}^*)$ cannot be simultaneously possible leading to a contradiction. This implies that $P^*$ with both $p_{i1j1}^*, p_{i1j2}^* > 0$ and $p_{i2j2}^*, p_{i2j1}^* > 0$ simultaneously is not possible.

For sake of completeness, we will now use Theorem 2.8 to establish the necessity of Property $P_1$ for any optimal allocation of customers to queues, i.e., of any solution of (P1). 

**Corollary 2.9.** Suppose $P^*$ solves the optimization problem (P1), and let $\gamma_j^*$ denote the resulting flow rates, as given by (2.1). Consider a re-ordering of the queues such that $D_1(\gamma_1^*) \leq D_2(\gamma_2^*) \leq \ldots \leq D_N(\gamma_N^*)$. Then, $P^*$ satisfies Property $P_1$.

In words, the corollary says that each customer class uses a nearly dedicated set of queues in the above ordering, with a possible overlap only at the boundaries of the sets. Note that it is possible for more than two classes of customers to use the same queue. For example, if $n_1 = n_2 = n_3$, then customer classes 1, 2 and 3 certainly use the same queue. It is possible to have $n_M < N$, in which case there are some queues that aren’t used by any customer class.
This would be the case if, even at zero load, the delay in these queues is larger than in the alternatives.

Proof. Let \( P^* \) solve the welfare optimization problem (P1). Define \( Q_i = \{ j : p^*_{ij} > 0 \} \) to be the set of queues used by class \( i \) under \( P^* \), and order these queues in non-decreasing order of delays. Clearly, each \( Q_i \) is non-empty. Define

\[
D^\text{max}_i = \max_{j \in Q_i} D_j(\gamma^*_j), \quad D^\text{min}_i = \min_{j \in Q_i} D_j(\gamma^*_j).
\]

By Theorem 2.8, \( D^\text{max}_i \leq D^\text{min}_{i+1} \), with equality only if the same queue \( j \) attains the maximum in the first case and the minimum in the second. Moreover, in this case, every queue \( k \neq j \) in \( Q_i \) has \( D_k(\gamma^*_k) \) strictly smaller than \( D_j(\gamma^*_j) \), while every queue \( k \neq j \) in \( Q_{i+1} \) has \( D_k(\gamma^*_k) \) strictly larger than \( D_j(\gamma^*_j) \). The claim of the corollary now follows. \( \square \)

2.3.3 A Continuum of Customer Classes

The preceding results for a finite number of customer classes can be extended to the case when the delay sensitivities (\( \beta \)) are from a continuum of values. We first reformulate the social welfare maximization problem in this setting. Let \( K(\cdot, \cdot) \) denote a kernel on \( \mathbb{R}_+ \times \{1, 2, \ldots, N\} \), where \( N \) is the total number of queues. In other words, \( K(\beta, \cdot) \) is a probability distribution on \( \{1, 2, \ldots, N\} \) for each \( \beta \in \mathbb{R}_+ \), and \( K(\cdot, i) \) is a Borel-measurable function on \( \mathbb{R}_+ \) for each \( i \in \{1, 2, \ldots, N\} \). We interpret \( K(\beta, i) \) as the probability that a customer with delay sensitivity \( \beta \) is allocated to queue \( i \). Thus, the class of static routing policies in the continuous setting can be identified with the set of kernels described above. Now, the welfare optimization problem is

\[
\inf_K U(K) = \lambda \sum_{j=1}^{N} \int_{\beta=0}^{\infty} \beta K(\beta, j) D_j(\gamma^*_j) dF(\beta) \tag{P3}
\]

where \( \gamma_j = \lambda \int_{\beta=0}^{\infty} K(\beta, j) dF(\beta) \). We can characterize the optimal solution in a manner analogous to the setting with finitely many customer classes.

**Theorem 2.10.** Let \( K^* \) achieve the minimum in (P3) and let \( \gamma^* \) denote the arrival rates corresponding to \( K^* \). Suppose \( \beta_1 > \beta_2 > 0 \), and suppose \( i \) and \( j \) are distinct queues such that

\[
\int_{\beta_1}^{\infty} K(\beta, i) dF(\beta) > 0 \quad \text{and} \quad \int_{0}^{\beta_2} K(\beta, j) dF(\beta) > 0.
\]

Then \( D_i(\gamma^*_i) < D_j(\gamma^*_j) \).
In words, the theorem says that if queue $i$ is used by customers with $\beta_1$ or higher and queue $j$ is used by customers of $\beta_2$ or lower then at optimal load distribution, the cost in queue $i$ will be less than that in queue $j$.

**Proof.** The proof is similar to that of Theorem 2.8 so we only sketch it briefly.

The proof is again by contradiction. Suppose first that $D^*_i > D^*_j$. We can modify $K^*$ so as to swap a small but non-zero volume of traffic with delay sensitivity $\beta \leq \beta_2$ in queue $j$ with an equal volume of traffic with delay sensitivity $\beta \geq \beta_1$ in queue $i$. As the traffic intensities at the two queues are left unchanged by this swap, so are the delays $D^*_i$ and $D^*_j$. But the cost corresponding to these delays has strictly decreased for the swapped traffic (as traffic with higher delay sensitivity $\beta \geq \beta_1 > \beta_2$ has been moved to the queue with lower delay, and replaced with an equal quantity of less delay sensitive traffic), while remaining unchanged for all other traffic. Consequently, the total cost $U(\cdot)$ of the routing has been decreased. This contradicts the optimality of $K^*$.

Suppose next that $D^*_i = D^*_j$. Consider the swap described above, and let $K^1$ denote the kernel corresponding to the resulting routing. For $\alpha \in [0, 1]$, define $K^\alpha = (1 - \alpha)K^* + \alpha K^1$. The volumes of traffic, $\gamma^\alpha$, at any queue are exactly the same for every $K^\alpha$, and hence so are the delays at each queue. Consequently, the total cost $U(K^\alpha)$ does not depend on $\alpha$. Consequently, every $K^\alpha$ must be optimal.

Now consider modifying $K^\alpha$ by moving an $\epsilon$ quantity of traffic from queue $i$ to queue $j$. Such a change causes the total cost to increase by the quantity

$$\Delta U = \epsilon \left( D^*_j(\gamma^*_j) \int_0^\infty \beta K^\alpha(\beta, j)dF(\beta) - D^*_i(\gamma^*_i) \int_0^\infty \beta K^\alpha(\beta, i)dF(\beta) \right) + o(\epsilon).$$

Note in particular that, to first order, the change in cost does not depend on the composition of the traffic moved between the queues, but depends only on the externalities imposed by the move on the rest of the traffic in the queues. Now, the optimality of $K^\alpha$ requires that $\Delta U = o(\epsilon)$, i.e., that the expression in brackets be zero. But it is impossible that this can hold simultaneously for all $\alpha \in [0, 1]$ since $\int_0^\infty \beta K^\alpha(\beta, j)dF(\beta)$ increases with $\alpha$ while $\int_0^\infty \beta K^\alpha(\beta, i)dF(\beta)$ decreases with $\alpha$. \[\square\]

Now define two kernels $K^1$ and $K^2$ to be equivalent if the set $\{\beta : K^1(\beta, \cdot) \neq K^2(\beta, \cdot)\}$ has $F$-measure zero. We now have the following corollary which characterizes the structure of
any welfare maximizing allocation.

**Corollary 2.11.** Suppose \( K^* \) solves the optimization problem (P3), and let \( \gamma_j^* \) denote the resulting flow rates. Consider a re-ordering of the queues such that \( D_1(\gamma_1^*) \leq D_2(\gamma_2^*) \leq \ldots \leq D_N(\gamma_N^*) \). Then, there is an \( m \in \{1, 2, \ldots, N\} \) and \( \beta_1 > \beta_2 > \ldots > \beta_m = 0 \) such that \( K^* \) is equivalent to the allocation \( K \) given by

\[
K(\beta, \cdot) = \delta_i \quad \text{for all} \quad \beta \in [\beta_i - 1, \beta_i) \quad \text{and} \quad 1 < i \leq m,
\]

\[
K(\beta, \cdot) = \delta_1 \quad \text{for} \quad \beta \in [\beta_1, \infty).
\]

Here \( \delta_i \) denotes the probability distribution that puts unit mass on \( i \).

Note that if \( m = k < N \), then the servers with index greater than \( k \) are not used for allocation in the kernel \( K \).

### 2.4 Admission Prices and Wardrop Equilibria

We now consider the same queueing model, but generalized to include admission prices \( c_1 > c_2 > \ldots > c_N \) at queues \( 1, 2, \ldots, N \). Each customer seeks to join a queue that minimizes the sum of the admission price, which is common to all classes, and the expected delay cost, which is weighted by a class-specific sensitivity. In Section 2.2, we modeled the resulting interaction as a game, and wrote down the conditions for a routing matrix \( P \) to be a Wardrop equilibrium in 2.2. We shall now show that a Wardrop equilibrium has the same structure that we demonstrated for a social optimum in the previous section.

**Theorem 2.12.**  
- Consider two customer classes \( i_1 < i_2 \), so that \( \beta_{i_1} > \beta_{i_2} \), and two queues \( j_1 < j_2 \), so that \( c_{j_1} > c_{j_2} \). There is no Wardrop equilibrium \( P^W \) in which class \( i_1 \) uses queue \( j_2 \) while class \( i_2 \) simultaneously uses queue \( j_1 \), i.e., \( p_{i_1,j_2}^W > 0 \) and \( p_{i_2,j_1}^W > 0 \).

- For some Class \( m \), let \( p_{m,n} = 0 \) for servers \( n \) satisfying \( n_1 < n < n_2 \) and let \( p_{m,n_1}^W, p_{m,n_2}^W > 0 \). Then \( p_{m,n} = 0 \) for all \( m \) and for \( n \) satisfying \( n_1 < n < n_2 \).

**Proof.** We shall continue to use the lighter notation of the previous section.

The proof of the first part is by contradiction. Suppose such a Wardrop equilibrium exists. Since \( p_{i_2}^W > 0 \) and \( p_{i_1}^W > 0 \), we have by (2.2) that

\[
c_2 + \beta_1 D_2^W \leq c_1 + \beta_1 D_1^W, \quad c_1 + \beta_2 D_1^W \leq c_2 + \beta_2 D_2^W.
\]

Re-arranging these inequalities, we get

\[
\beta_1 (D_2^W - D_1^W) \leq c_1 - c_2 \leq \beta_2 (D_2^W - D_1^W).
\]

(2.10)
Since $c_1 > c_2$, the second inequality implies that $D_2^W - D_1^W$ is strictly positive. But $\beta_1 > \beta_2$, so the two inequalities together imply that $D_2^W - D_1^W \leq 0$. This is a contradiction, so such a Wardrop equilibrium cannot exist.

The second part of the theorem states that if a traffic class does not use a server but uses a succeeding one and a preceding one (when indexed in decreasing order of the admission prices), then no other class will use that server. Suppose by contradiction, there exists a class using this unused server. Then it is easy to see that this violates the first part of the theorem.

We now have the following corollary, which is analogue of Corollary 2.9. The proof is omitted as it is straightforward.

**Corollary 2.13.** Define $K := \{k : p_{m,k} > 0, \text{for some } m\}$, i.e., $K$ denotes the set of servers that are used by the classes at equilibrium. Re-index the servers in the set $K$ in decreasing order of the admission prices. Now let $P^W$ be the equilibrium routing matrix on $K$. Then $P^W$ satisfies Property $P_1$ (Definition 2.4).

The main difference from Corollary 2.9 is that we do not guarantee that all routing probabilities are strictly positive inside these ranges. Whereas, in the welfare-optimizing setting, any unused queues were necessarily those with the largest delays, now either a large delay or a high admission price or a combination of the two could result in a queue not being used by any customer class.

### 2.4.1 Continuum of customer classes

Following exactly the same arguments as in Theorem 2.12 and Corollary 2.13, Corollary 2.14 follows for the case when the delay costs of the customers are from an absolutely continuous distribution $F$. In the following corollary we have assumed for convenience that all the $N$ servers are being used. If this is not the case, then one needs to re-index the servers on a suitable set $K$ as defined earlier.

**Corollary 2.14.** Suppose $K^W$ satisfies the Wardrop equilibrium condition. There exist thresholds $\beta_1, \beta_2, \ldots, \beta_N$ such that $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_N \in \mathbb{R}_+$ and $K^W(\beta, \cdot) = \delta_i$ for all $\beta \in (\beta_{i-1}, \beta_i)$ and $1 < i \leq N$ while $K^W(\beta, \cdot) = \delta_1$ for $\beta \in (\beta_1, \infty)$ with $\delta_i$ denoting the probability distribution that puts unit mass on $i$.

Since $F$ is absolutely continuous, at each threshold $\beta_i$ in the interior of its domain, for
\[ i = 1 \ldots N - 1, \text{the following is satisfied} \]
\[ c_i + \beta_i D_i(\gamma_i) = c_{i+1} + \beta_i D_{i+1}(\gamma_{i+1}) \quad (2.11) \]

### 2.4.2 The Two Class, Two Server Example

We consider the running example described in Section 2.3.1 with two class of customers making an individually optimal choice of the server. Let \( \mu_i \) denote the service rate at Server \( i \) for \( i = 1, 2 \). We assume \( \mu_1 \geq \mu_2 \) and define \( c := c_1 - c_2 \). With \( c_1 > c_2 \), from Theorem 2.12, one of the following is true at equilibrium.

- \( p_1 = 1 \) and \( 0 \leq p_2 \leq 1 \).
- \( 0 \leq p_1 < 1 \) and \( p_2 = 0 \).

This gives us the following five equilibrium regimes.

1. Regime 1 for which \( p_1 = p_2 = 1 \),
2. Regime 2 for which \( p_1 = 1 \) and \( 0 < p_2 < 1 \),
3. Regime 3 for which \( p_1 = 1 \) and \( p_2 = 0 \),
4. Regime 4 for which \( 0 < p_1 < 1 \) and \( p_2 = 0 \),
5. Regime 5 for which \( p_1 = 0 \) and \( p_2 = 0 \).

Fixing the arrival rates \( \lambda_i \) and the service rates \( \mu_i \), we describe the change in the equilibrium traffic as \( c = c_1 - c_2 \) is increased from 0. Of course, a regime is feasible only if both the queues are stable, i.e., \( \mu_i > \gamma_i \) for \( i = 1, 2 \).

In Regime 1, \( p_1 = p_2 = 1 \) and hence \( \gamma_1 = \lambda_1 + \lambda_2 \) and \( \gamma_2 = 0 \). Here, we will need \( c < \beta_2 (D_2(0) - D_1(\lambda_1 + \lambda_2)) \). Let \( c_a := \beta_2 (D_2(0) - D_1(\lambda_1 + \lambda_2)) \), i.e.,

\[ c_a := \frac{\beta_2}{\mu_2} - \frac{\beta_2}{\mu_1 - \lambda_1 - \lambda_2} \quad (2.12) \]

Thus, if \( \mu_1 > \lambda_1 + \lambda_2 \) and \( 0 < c < c_a \), then the equilibrium traffic will be in Regime 1.

Regime 3 requires \( p_1 = 1 \) and \( p_2 = 0 \), i.e., \( \gamma_1 = \lambda_1 \) and \( \gamma_2 = \lambda_2 \). This leads to the following condition on \( c \).

\[ \beta_2 (D_2(\lambda_2) - D_1(\lambda_1)) < c < \beta_1 (D_2(\lambda_2) - D_1(\lambda_1)) \]

Let

\[ c_b := D_2(\lambda_2) - D_1(\lambda_1) = \frac{1}{\mu_2 - \lambda_2} - \frac{1}{\mu_1 - \lambda_1} \quad (2.13) \]

Thus if \( \mu_1 > \lambda_1, \mu_2 > \lambda_2 \) and \( 0 < \beta_2 c_b < c < \beta_1 c_b \), then the equilibrium will operate in Regime 3. The condition \( c_b > 0 \) is possible only if \( \mu_1 - \lambda_1 > \mu_2 - \lambda_2 \) which thus becomes a necessary condition for this regime to be possible.
Figure 2.4: The operating regimes as $c$ is increased from 0. In addition to the requirement on $c$, the stability conditions also need to be satisfied. $c_a$, $c_b$ and $c_c$ are as in Eqs. 2.12, 2.13 and 2.14 respectively.

If the equilibrium is in Regime 2, then $p_1 = 1$ and $0 < p_2 < 1$ and $\gamma_1 = \lambda_1 + p_2 \lambda_2$ and $\gamma_2 = (1 - p_2) \lambda_2$. For Regime 2 to be possible for all $p_2 \in [0, 1]$, we need $\mu_1 > \lambda_1 + \lambda_2$ and $\mu_2 > \lambda_2$. Since $p_1 = 1$ and $0 < p_2 < 1$, for the queue to be in this regime we require

$$c = \frac{\beta_2}{\mu_2 - (1 - p_2) \lambda_2} - \frac{\beta_2}{\mu_1 - \lambda_1 - p_2 \lambda_2}.$$  

Observe that as we move from Regime 1 ($p_2 = 1$) to Regime 3 ($p_2 = 0$) through Regime 2, $p_2$ decreases from 1 to 0. This happens as $c$ is increased from $c_a$ to $\beta_2 c_b$. We can show that $c_a \leq \beta_2 c_b$ with equality if $\lambda_2 = 0$.

For the equilibrium to be in Regime 5, we need $p_1 = p_2 = 0$, i.e., $\gamma_1 = \lambda_1 + \lambda_2$. This requires $c > \beta_1((D_2(\lambda_1 + \lambda_2) - D_1(0)))$. Define $c_c := \beta_1(D_2(\lambda_1 + \lambda_2) - D_1(0))$, i.e.,

$$c_c := \frac{\beta_1}{\mu_2 - \lambda_1 - \lambda_2} - \frac{\beta_1}{\mu_1}.$$  

Thus the equilibrium will be in Regime 5 if $c_c > 0$, $\mu_2 > \lambda_1 + \lambda_2$ (stability condition), and $c > c_c$.

Operating in Regime 4 requires $p_2 = 0$ and $0 < p_1 < 1$ which in turn means $\gamma_1 = p_1 \lambda_1$ and $\gamma_2 = (1 - p_1) \lambda_1 + \lambda_2$. Thus, a particular $(p_1, p_2)$ satisfying $0 < p_1 < 1$ and $p_2 = 0$ is achieved with

$$c = \beta_1 \left( \frac{1}{\mu_2 - (1 - p_1) \lambda_1 - \lambda_2} - \frac{1}{\mu_1 - p_1 \lambda_1} \right).$$  

Thus, as $c$ increases from $\beta_1 c_b$ to $c_c$, $p_1$ decreases from 1 to 0. The equilibrium will be in Regime 4 if $0 < \beta_1 c_b < c < c_c$ and (stability condition) $\mu_1 > \lambda_1$ and $\mu_2 > \lambda_1 + \lambda_2$.

Figure 2.4 summarizes the preceding discussion.
Remark 2.15. If $c = 0$, then at equilibrium one of the following is true: (1) $p_1 = p_2 = 1$ when $D_1(\lambda_1 + \lambda_2) < D_2(0)$; (2) $p_1 = p_2 = 0$ when $D_1(0) > D_2(\lambda_1 + \lambda_2)$; (3) $(p_1, p_2)$ is such that $D_1(\gamma_1) = D_2(\gamma_2)$.

Remark 2.16. If $c_1 < c_2$ then using similar arguments as above, we can show that at equilibrium, one of the following is true.

- $p_1 = 0$ and $0 \leq p_2 \leq 1$.
- $0 \leq p_1 \leq 1$ and $p_2 = 1$.

2.5 Admission Prices and Pigouvian Taxes for Welfare Optimization

A natural mechanism design problem suggested by the above results is whether we can set admission prices in queues in such a way that selfish users reacting to these prices would assign themselves to queues in the proportions required for optimizing social welfare. Pigou [55] proposed the use of a charge or levy to internalize the congestion externality in transport networks, thereby guiding the system to a social optimum. Such charges are known as Pigouvian taxes and they have been studied in several contexts including transportation networks [58, 64]. However, these have typically focused on managing demand or guiding route choice, whereas here we study their use to achieve service differentiation in a multi-class setting.

Let $P^*$ denote the routing matrix solving the social welfare optimization problem and $\gamma^*$ the corresponding vector of traffic flow rates at the different queues. A marginal customer at queue $n$ increases the delay of each customer at this queue by $D_n'(\gamma^*_n)$. This imposes a cost of $\beta_m D_n'(\gamma^*_n)$ on each class $m$ customer using this queue, of whom there are $\lambda_m p_{mn}^*$ per unit time. Thus, the total congestion externality caused by a marginal unit of traffic at queue $n$, which is the Pigouvian tax for this queue, is given by

$$c_n = \sum_{m=1}^{M} \beta_m \lambda_m p_{mn}^* D_n'(\gamma^*_n). \quad (2.15)$$

We shall show that, if the admission price at each queue is set equal to the Pigouvian tax at that queue, then the optimal allocation $P^*$ is a Wardrop equilibrium.
Theorem 2.17. Let $P^*$ be a routing matrix solving the social welfare optimization problem. Assume that the resulting flows $\gamma^*$ are such that $\gamma^*_n$ is in the interior of the domain of $D_n(\cdot)$ for each queue, $n$. Let the admission prices $c_1, c_2, \ldots, c_N$ at queues $1, 2, \ldots, N$ be set according to (2.15). Then $P^*$ is a Wardrop equilibrium of the resulting game.

Proof. Since $P^*$ solves the constrained optimization problem in (P1), the constraints on $P$ are affine, and $U$ is continuously differentiable at $P^*$, it follows that $P^*$ must satisfy the KKT conditions: these state that there exist $A \in \mathbb{R}^{M \times N}$ and $b \in \mathbb{R}^M$ such that

$$\frac{\partial U(P)}{\partial p_{in}} \bigg|_{P=P^*} = a_{in} + b_i, \quad a_{in} \geq 0, \quad a_{in} = 0 \text{ if } p_{in}^* > 0. \quad (2.16)$$

Here, $a_{in}$ is the Lagrange multiplier on the constraint $p_{in} \geq 0$, and is zero if the constraint is slack, and $b_i$ is the Lagrange multiplier on the constraint $\sum_{m=1}^N p_{im} = 1$.

If class $i$ uses queue $n$ in the optimal allocation $P^*$, then (2.16) implies that, for all $m = 1, \ldots, N$, we have

$$\frac{\partial U(P)}{\partial p_{in}} \bigg|_{P=P^*} \leq \frac{\partial U(P)}{\partial p_{im}} \bigg|_{P=P^*}.$$

Differentiating $U(P)$ defined in (P1), we can rewrite this as

$$\beta_i \lambda_i D_n(\gamma^*_n) + \sum_{j=1}^M \beta_j \lambda_j p_{jn}^* D'(\gamma^*_n) \lambda_i \leq \beta_i \lambda_i D_m(\gamma^*_m) + \sum_{j=1}^M \beta_j \lambda_j p_{jm}^* D'(\gamma^*_m) \lambda_i.$$

Substituting (2.15) in the above, we get

$$\beta_i D_n(\gamma^*_n) + c_n \leq \beta_i D_m(\gamma^*_m) + c_m,$$

for all $m = 1, \ldots, N$. This holds for every $i$ and $n$ such that class $i$ uses queue $n$. Comparing this with (2.2), we see that this is exactly the condition for $P^*$ to be a Wardrop equilibrium. This completes the proof of the theorem.

When the $\beta$ of the arriving customer is a i.i.d. random variable from an absolutely continuous distribution, then the Pigouvian price at queue $n$ is

$$c_n = \int_0^\infty \beta K^*(\beta, n) D_n'(\gamma^*_n) \lambda dF(\beta) \quad (2.17)$$

Setting this as the admission price to the queues and arguing as above we show below that the optimum $K^*(\cdot, \cdot)$ corresponds to a Wardrop equilibrium.

Theorem 2.18. Let $K^*$ be a kernel solving the social welfare optimization problem. Let the admission prices $c_1, c_2, \ldots, c_N$ at queues $1, 2, \ldots, N$ be set according to (2.17). Then $K^*$ is a Wardrop equilibrium of the resulting game.
Proof. Since $K^*$ solves the optimization problem in (P3), modifying $K^*$ by moving a non zero volume of traffic from queue $i$ to queue $j$ must not cause the total cost to decrease. Let $\beta_j$, for $j = 1, \ldots, \beta_m$ be as in Corollary 2.11. Specifically, consider $\beta^1$ and $\epsilon$ such that $\beta_{j-1} < \beta^1 < \beta^1 + \epsilon < \beta_j$. Modify $K^*$ to $K^*_\epsilon$ by reallocating customers with delay cost $\beta \in [\beta^1, \beta^1 + \epsilon]$ to server $i$. This increases the arrival rate to server $i$ by

$$h(\epsilon) := \int_{\beta^1}^{\beta^1+\epsilon} \lambda dF(\beta).$$

The change in the social utility due to this reallocation is

$$\Delta U = D_i(\gamma_i + h(\epsilon)) \int_0^{\infty} \lambda \beta K^*_\epsilon(\beta, i)dF(\beta)$$

$$+ D_j(\gamma_j - h(\epsilon)) \int_0^{\infty} \lambda \beta K^*_\epsilon(\beta, j)dF(\beta)$$

$$- D_i(\gamma_i) \int_0^{\infty} \lambda \beta K^*(\beta, i)dF(\beta)$$

$$- D_j(\gamma_j) \int_0^{\infty} \lambda \beta K^*(\beta, j)dF(\beta)$$

$$= h(\epsilon) \frac{(D_j(\gamma_j - h(\epsilon)) - D_j(\gamma_j))}{h(\epsilon)} \int_{\beta^1}^{\infty} \lambda \beta K^*_\epsilon(\beta, j)dF(\beta)$$

$$- D_j(\gamma_j - h(\epsilon)) \int_{\beta^1}^{\beta^1+\epsilon} \lambda \beta dF(\beta)$$

$$+ h(\epsilon) \frac{(D_i(\gamma_i + h(\epsilon)) - D_i(\gamma_i))}{h(\epsilon)} \int_0^{\infty} \lambda \beta K^*_\epsilon(\beta, i)dF(\beta)$$

$$+ D_i(\gamma_i + h(\epsilon)) \int_{\beta^1}^{\beta^1+\epsilon} \lambda \beta dF(\beta).$$

As $\epsilon \to 0$,

$$h(\epsilon) \to 0$$

$$\frac{h(\epsilon)}{\epsilon} \to \lambda f(\beta^1)$$

$$\frac{1}{\epsilon} \int_{\beta^1}^{\beta^1+\epsilon} \beta dF(\beta) \to \beta^1 f(\beta^1).$$

Dividing throughout by $\epsilon$ and taking limits as $\epsilon \to 0$, the limiting value of $\Delta U$ is

$$\lambda^2 f(\beta^1) D_i(\gamma_i) \int_0^{\infty} \beta K^*(\beta, i)dF(\beta) + \lambda D_i(\gamma_i) \beta^1 f(\beta^1)$$

$$- \lambda^2 f(\beta^1) D_j(\gamma_j) \int_0^{\infty} \beta K^*(\beta, j)dF(\beta) - \lambda D_j(\gamma_j) \beta^1 f(\beta^1).$$
Optimality of $K^*$ requires that this be non-negative. Hence dividing throughout by $\lambda f(\beta^1)$ and using (2.17), we have
\[
\beta^1D_j(\gamma_j) + c_j \leq \beta^1D_i(\gamma_i) + c_i.
\]
The choice of $\beta^1$ in the interval $(\beta_{j-1}, \beta_j)$ was arbitrary. Further, the choice of $i$ was also arbitrary. Hence for all $\beta \in (\beta_{j-1}, \beta_j)$, we have
\[
\beta D_j(\gamma_j) + c_j \leq \beta D_i(\gamma_i) + c_i.
\]
for all $i = 1, \ldots, N$.

This is the Wardrop condition and hence completes the proof.

2.6 Examples

In this section we first illustrate our results with some numerical examples. We then explain some example scenarios where the assumptions may be generalized.

The first example is of a system serving five customer classes with five servers i.e., $M = N = 5$ with mean delay as the cost function. Customers of class $i$ arrive according to a stationary unit rate Poisson process and have $\beta_i = (6 - i)$. All classes have exponentially distributed job sizes with unit mean. Server $j$ has service rate $\mu_j$ with $\mu_1 = 2$, $\mu_2 = 3$, $\mu_3 = 2.5$, $\mu_4 = 1.1$ and $\mu_5 = 1.5$. Thus the five queues are all $M/M/1$ and the mean delay at queue $j$ is $D_j(\gamma_j) = \frac{1}{\mu_j - \gamma_j}$.

$P^*$ that minimizes $U(P)$ is obtained numerically and is as follows.

\[
P^* = \begin{pmatrix}
Q_2 & Q_3 & Q_1 & Q_5 & Q_4 \\
Cl 1 & 1 & 0 & 0 & 0 & 0 \\
Cl 2 & 0.528 & 0.472 & 0 & 0 & 0 \\
Cl 3 & 0 & 0.788 & 0.212 & 0 & 0 \\
Cl 4 & 0 & 0 & 0.786 & 0.214 & 0 \\
Cl 5 & 0 & 0 & 0 & 0.517 & 0.483
\end{pmatrix}
\]

This corresponds to $\gamma_1 = 0.998$, $\gamma_2 = 1.528$, $\gamma_3 = 1.26$, $\gamma_4 = 0.483$, and $\gamma_5 = 0.731$ and $D_1 = 0.998$, $D_2 = 0.679$, $D_3 = 0.806$, $D_4 = 1.62$, and $D_5 = 1.3$. Observe that in $P^*$ the servers are reordered in increasing order of the mean delays with $Q_i$ denoting the server with service rate $\mu_i$. We also see that this $P^*$ satisfies Corollary 2.9. With the above allocation we see that $U(P^*) = 12.47$.  

45
Next, let the admission prices for queues 1, . . . , 5 be, respectively, 2.57, 1.53, 0.7, 0.42, and 0. A Wardrop equilibrium allocation at these prices will be

\[
P^W = \begin{pmatrix}
Q_1 & Q_2 & Q_3 & Q_4 & Q_5 \\
Cl 1 & 0.4 & 0.6 & 0 & 0 & 0 \\
Cl 2 & 0 & 1 & 0 & 0 & 0 \\
Cl 3 & 0 & 0.2 & 0.8 & 0 & 0 \\
Cl 4 & 0 & 0 & 0.8 & 0.2 & 0 \\
Cl 5 & 0 & 0 & 0 & 0.1 & 0.9
\end{pmatrix}.
\]

Clearly, \(P^W\) satisfies Corollary 2.13. The Pigouvian admission prices are 3.28, 2.77, 2.194, 1.59, and 1.27. It can be verified that for these prices, the Wardrop equilibrium condition is satisfied by the social welfare maximizing allocation, \(P^*\).

Next we consider an alternate cost function with \(D_n(\gamma_n)\) denoting the probability that the waiting time experienced by an arriving customer in queue \(n\) is more than a fixed amount \(T\). Since the queues are \(M/M/1\), We now have

\[
D_n(\gamma_n) = \rho_n \exp((\gamma_n - \mu_n)T)
\]

where \(\rho_n = \gamma_n/\mu_n\). For \(T = 1\) and the other parameters as in the preceding example, the optimal routing probabilities will be

\[
P^* = \begin{pmatrix}
Q_2 & Q_3 & Q_1 & Q_5 & Q_4 \\
Cl 1 & 1 & 0 & 0 & 0 & 0 \\
Cl 2 & 0.42 & 0.58 & 0 & 0 & 0 \\
Cl 3 & 0 & 0.61 & 0.39 & 0 & 0 \\
Cl 4 & 0 & 0 & 0.59 & 0.41 & 0 \\
Cl 5 & 0 & 0 & 0 & 0.37 & 0.63
\end{pmatrix}.
\]

Further, the Pigouvian prices such that \(P^W = P^*\) for this case are 1.5789, 0.9809, 0.5555, 0.2813, and 0.1489.

We remark here that the price vector that induces a \(P^W\) that is the same as \(P^*\) is not unique. For example, changing the Pigouvian prices of all the queues by a fixed amount will also have a \(P^W\) that is the same as \(P^*\).
2.6.1 \(M/G/1 - LIFO\) and \(M/G/1 - PS\) queues

Recall that in all the preceding analyses, we have assumed that customers of all classes have the same job size distributions, and that, once they join a queue, they are treated identically within it. We now describe some systems where we can relax this assumption and allow different classes to have a different job size distributions. Throughout this example, let \(S_i\) denote the mean job size of a Class \(i\) customer with the assumption that \(S_1 \geq S_2 \geq \ldots \geq S_N\). We assume that the service discipline at the servers is either processor-sharing (PS) or last-come-first-serve with preemption (LCFS-PR) and the service rate of each server is unity. Now define \(\hat{\beta}_i := \beta_i S_i\) and note that \(\hat{\beta}_1 > \hat{\beta}_2 > \ldots > \hat{\beta}_N\). The mean sojourn time of a Class \(i\) customer at server \(j\) is now given by

\[
D_{ij}(\rho_j) = S_i D_j(\rho_j)
\]

where

\[
D_j(\rho_j) = \frac{1}{1 - \rho_j}
\]

and \(\rho_j = \sum_{i=1}^{M} p_{ij} \lambda_i S_i\). The social welfare objective now is

\[
U(P) = \sum_{i=1}^{M} \sum_{j=1}^{N} \hat{\beta}_i \lambda_i p_{ij} D_j(\rho_j).
\]  

(2.20)

Note that replacing \(\lambda_i\) in \(\text{P1}\) by \(\lambda_i S_i\) gives (2.20) and the analysis for the structural properties of \(P^*\) and \(P^W\) remains unaffected. The theorems and corollaries discussed in the preceding sections should be suitably modified with \(\hat{\beta}_i\) now assuming the role of \(\beta_i\).

2.7 Discussion

In this chapter, we have considered the problem of routing multiclass traffic to multiple servers such that the routing minimizes an aggregate cost function. We note that the cost function \(U(\cdot)\) is not a convex function; this is even true for the two-class two-server example. Instead of analyzing the KKT conditions, we identify the structural properties of \(P^*(K^*)\), the optimal routing matrix (kernel). In Section 2.4, we consider a system where servers charge an admission price and where each arrival makes an individually optimal queue-join decision. The resulting equilibrium is called as the Wardrop equilibrium with \(P^W(K^W)\) denoting the equilibrium matrix (kernel). In Theorem 2.12, we show that for a given set of admission prices, \(P^W\) also satisfies
the same structure as that of the optimal routing matrix. Now due to the similarity in the structure of $P^*(K^*)$ with that of $P^W(K^W)$ the interest is in finding $c_1, c_2 \ldots c_N$ such that $P^* = P^W (K^* = K^W)$. This is exactly the result proved in Theorem 2.17 (Theorem 2.18). The results are verified using the running example of two identical servers with two traffic classes. Clearly, with proper choice of admission prices, a system operator can maximize social welfare without requiring a centralized control.

Our results raise a number of questions for future research. Firstly, as mentioned in Section 2.3 the computational complexity of determining the optimal allocation is unknown, though we showed that the optimization problem is non-convex. Likewise, the computational complexity of determining Wardrop equilibria is also unknown.

A second question concerns the informational constraints on the model. We have assumed that the parameters $\lambda_i$ and $\beta_i$ are known, and available as input to determining the socially optimal allocation or setting admission prices. In practice, this information is unlikely to be available, but needs to be inferred from observation. If customer classes are known upon arrival, then the arrival rates $\lambda$ can easily be measured, but eliciting $\beta$ truthfully can still be a challenge. The problem is much harder if customers are heterogeneous but there are either no clearly defined classes, or that class membership is unobservable, as is often likely to be the case. In such a situation, is it still possible to set admission prices in such a way as to ensure that the Wardrop equilibrium either coincides with the welfare optimizing allocation, or approximates it to within some factor?

Finally, we have assumed that a benevolent mechanism designer sets admission prices to maximize social welfare; it is interesting to ask what happens if the admission prices are set by a revenue maximizing service provider. Further, in such a revenue maximizing scenario it would be interesting to see if competing service providers can sustain differentiated services. We shall attempt to answer some of these questions in the following chapter.
Chapter 3

Revenue Maximization in a Monopoly System

3.1 Introduction

In the previous chapter, we have seen how admission prices can be used to achieve a desired routing of multi-class customers in a parallel server system. In particular, the admission prices were chosen to optimize the social welfare function. Alternatively, the admission prices at the servers can also be chosen to maximize the revenue rate from the servers where the revenue rate at a server is the product of the admission price and the arrival rate of customers to the server.

In this chapter, we consider the revenue maximization problem in a service system with parallel servers. With a monopolistic scenario, we assume that all the servers available to the customers for service belong to a single service system. The objective of this service system is to maximize the total revenue rate, i.e., the sum of the revenue rate at all its servers. We assume, like in the previous chapter, that the arriving customers have heterogeneous preferences to their delay cost. Further a customer cannot balk from the system without obtaining service. A customer’s choice of the server is an individually optimal choice that is characterized by the Wardrop condition described in the previous chapter.

Now consider the scenario of a monopoly market discussed above. As an example, consider a single service system with two servers. Each server has an associated queue for customers to wait for service and an admission price is charged to each customer joining the queue. We do not make any assumption on the maximum admission price a customer is willing to pay. In the absence of balking, it is not difficult to see that a revenue maximizing strategy for the
monopoly is to keep both the admission prices at infinity. This is because as customers cannot balk, they are required to choose one of the server for service. This observation is indeed true irrespective of the delay cost functions at the queues. Therefore one has to consider a more meaningful model for the monopoly market. Towards this, we make the following assumption. We assume that the admission price at one of the server, say Server 2 is a-priori fixed. This dissuades the service provider from fixing the admission price at Server 1 to unreasonably high values. If indeed the price at Server 1 was very high, then all the customers may choose Server 2 for service. This may be a sub-optimal strategy when the service provider is interested in maximizing its total revenue rate. Our interest for this model is to characterize the revenue maximizing admission price at Server 1 for different examples of the delay functions at the queue and of the delay cost distribution for the customers.

Classical monopoly models have been well studied for the case of single server queues. One of the first work to analyze monopoly models is Naor [50]. This model considers a single server queueing system where homogeneous customers obtain a reward after service completion. The queue is observable and the customers either choose to join the queue or balk. For such a system, the revenue maximizing admission price was first obtained in [50]. Subsequent to this, there have been several works analyzing the revenue maximization problem under different assumptions such as multiserver queue [26], GI/M/1 queue [65], customers with heterogeneous service valuations [43] and queue length dependent prices [18]. Note that the monopoly models considered above are for the case when the queue lengths are observable. Edelson and Hilderbrand [27] were the first to consider the profit maximization problem for the case when such queues are not observable. A generalization of [27] for the case of nonlinear preferences is [19]. See [47, 48, 17, 46] for some other single server profit maximization models with different assumptions on the system model.

The key difference of our model with that of the literature discussed above is as follows. In our model, customers are not allowed to balk from the system. Secondly, the customers have to obtain service at either of the two servers and the admission price at one of the server is fixed. Finally, the customers have heterogeneous preference for the delay experienced in the queue. This assumption makes our model interesting but also difficult to analyze. We use this structure of the equilibrium routing from the previous chapter to solve the monopoly maximization problem in the subsequent section. As in the case of the previous chapter, we make minimal assumptions on the distribution $F(\cdot)$ and on the delay cost function $D_j(\cdot)$. Our
key contribution is that our analysis holds for a wide class of $F(\cdot)$ and $D_j(\cdot)$ function. In our main result, we determine the admission price that should be charged at the first server such that the revenue rate from the system is maximized.

For most practical scenarios, the distribution function $F(\cdot)$ may not be known to the service system. The revenue maximizing strategy on the other hand depend on the distribution $F(\cdot)$. Without any knowledge of $F(\cdot)$, it is not possible to ascertain a revenue optimal admission price at the servers and in such cases, the service system is required estimate this distribution function. Towards the end of this chapter, we shall provide an elementary method to estimate this distribution $F(\cdot)$ by varying the admission prices and observing the change in the equilibrium traffic routing. The service system can then use this estimate to perform the necessary revenue maximization.

Rest of the chapter is organized as follows. In the next section, we shall formalize the notation, provide some preliminaries and formulate the problems. In Section 3.3 we consider the monopoly problem for revenue optimization and provide some numerical examples for the same. Finally in Section 3.4 we illustrate a mechanism based on admission prices to estimate the distribution function $F$.

### 3.2 Notations, Preliminaries and Problem Formulation

We will now recall some of the relevant notation that were introduced in the previous chapter. In the monopoly model, we assume that the system has two servers. Let $c_j$ denote the admission price at Server $j$ where $j = 1, 2$. The customers arrive according to a homogeneous Poisson process with rate $\lambda$ and have a service requirement that is i.i.d with exponential distribution and unit mean. We assume that customers of all classes have the same job size distributions. Let $D_j(\gamma_j)$ denote the cost function associated with queue $j$ when the aggregate arrival rate is $\gamma_j$, where $j = 1, 2$. Recall that $D_j$ is monotone increasing and continuously differentiable in the interior of its domain with a strictly positive derivative. Additionally we assume that the cost function at the two server satisfies the following two conditions: (1) $0 \leq D_1(0) < D_2(\lambda) < \infty$ and (2) $0 \leq D_2(0) < D_1(\lambda) < \infty$. For M/M/1 type delay functions where $D_i(\gamma_i) = \frac{1}{\mu_i - \gamma_i}$ for $i = 1, 2$, the assumption that $D_i(\lambda) \in (0, \infty)$, implicitly implies that $\mu_i > \lambda$. In words, the assumption implies that each server must have sufficient server capacity to serve the entire population of customers.
We associate with each customer a random variable $\beta$ that quantifies its sensitivity to delay or congestion. We assume that the delay sensitivity $\beta$ is a continuous random variable with an absolutely continuous cumulative distribution function $F$ supported on the interval $[a, b]$ of positive reals. We shall additionally assume $F(\cdot)$ is strictly increasing in its argument and hence $f(x) \neq 0$ for any $x$ in its domain where $f(\cdot)$ is the corresponding density function. Now let $K^W(\cdot, \cdot)$ denote the kernel satisfying the Wardrop equilibrium condition. We interpret $K^W(\beta, i)$ as the probability that a customer with delay sensitivity $\beta$ chooses queue $i$ at equilibrium. The arrival rate of customers to Server $j$ for $j = 1, 2$ is given by

$$\gamma_j = \lambda \int_{\beta=a}^{b} K(\beta, j) dF(\beta)$$

with $\gamma_1 + \gamma_2 = \lambda$. Let $\theta$ denote the collection of the system parameter, i.e.,

$$\theta = \{c_1, c_2, F(\cdot), D_1(\cdot), j = 1, 2.\}.$$ 

By $D_j(\cdot), j = 1, 2$ we mean the function form of the delay cost $D_j$ for $j = 1, 2$.

We first provide the following theorem that is a restatement of Corollary 2.14. It also provides a generalization of Remark 2.15 and Remark 2.16 for the case of a continuum of customers. More precisely, this theorem characterizes the Wardrop equilibrium kernel for our problem.

**Theorem 3.1.** Define $\delta_i$ as the probability distribution that puts unit mass on $i$ and suppose that the kernel $K^W$ satisfies the Wardrop equilibrium condition. Then there exists a threshold $\beta_1$ with $\beta_1 \in [a, b]$ such that

- when $c_1 > c_2$ (resp. $c_1 < c_2$),
  
  $$K^W(\beta, \cdot) = \begin{cases} 
  \delta_1 (\text{ resp. } \delta_2) & \text{for } \beta \in (\beta_1, b], \\
  \delta_2 (\text{ resp. } \delta_1) & \text{for } \beta \in [a, \beta_1].
  \end{cases}$$ (3.1)

  Further if $\beta_1 \in (a, b)$ then,

  $$c_1 + \beta_1 D_1(\gamma_1) = c_2 + \beta_1 D_2(\gamma_2).$$ (3.2)

- When $c_1 = c_2$, $K^W$ is not unique. In fact, any kernel $K$ with the corresponding arrival rates $\gamma_1$ and $\gamma_2$ satisfying $D_1(\gamma_1) = D_2(\gamma_2)$ is a valid Wardrop equilibrium kernel $K^W$. 

52
Refer Figures 3.1 and 3.2 for a representation of the Wardrop equilibrium kernel for the case when $c_1 > c_2$ and $c_1 < c_2$ respectively. Here $f(\cdot)$ denotes the underlying density function of the random variable $\beta$ while the shaded region identifies the delay cost parameter of those customers that choose Server 1.

**Proof.** The first part is simply a restatement of Corollary 2.14 for the case where $c_1 > c_2$. The proof for $c_1 < c_2$ is along similar lines. We now prove the second part.

Consider the case when $c_1 = c_2$ and recall the assumption that (1) $D_1(0) < D_2(\lambda)$ and (2) $D_2(0) < D_1(\lambda)$. Now $K^W$ must be such that $D_1(\gamma_1) = D_2(\gamma_2)$. To see why this must be true, suppose that this is not true and let $D_1(\gamma_1) \neq D_2(\gamma_2)$. Now for any kernel $K$ satisfying $\gamma_1 = \gamma^+$, since $c_1 = c_2$, the cost for any customer at the two servers is equal. Hence there is no incentive for any customer to deviate from its choice of the server. The Wardrop equilibrium kernel $K^W$ though not unique must however satisfy the following. $K^W \in \left\{ K : \lambda \int_{\beta=a}^{b} K(\beta, j)dF(\beta) = \gamma^+ \right\}$. 

53
Figure 3.2: Representation of $K^W$ when $c_1 < c_2$.

**Problem Formulations**

Having characterized the Wardrop equilibrium kernel $K^W$ for a two server system, we will now formulate the revenue maximization problem. Let $R_j(c_j, \gamma_j) := c_j \gamma_j$ denote the revenue rate at server $j$ when the arrival rate of customers due to the corresponding $K^W$ is $\gamma_j$ for $j = 1, 2$. Since $\gamma_2 = \lambda - \gamma_1$, it suffices to express the revenue rate as a function of only $\gamma_1$. For the monopoly model, let $R_T(c_1, \gamma_1)$ denote the revenue rate for the monopoly service system. We have

$$R_T(c_1, \gamma_1) := c_1 \gamma_1 + c_2 \gamma_2 = c_2 \lambda + (c_1 - c_2) \gamma_1.$$  

Note that the argument $\gamma_1$ is determined by the kernel $K^W$ which in turn depends on the admission prices $c_1$ and $c_2$. To indicate this dependence, we will for the time being denote $\gamma_1$ by $\gamma_1(c_1, c_2)$.

The revenue optimization problem for the monopoly can now be stated as follows.

$$\max_{c_1} \quad R_T(c_1, \gamma_1) = c_2 \lambda + (c_1 - c_2) \gamma_1(c_1, c_2)$$  \hspace{1cm} \text{(P4)}$$

subject to $0 \leq c_1 \leq c^1$

where $c^1 = \inf \{c : \gamma_1(c, c_2) = 0\}$. Here $c^1$ is a technical requirement to ensure a compact domain. This is without loss of generality since $\gamma_1(c, c_2) = 0$ for any $c > c^1$. To be able to solve program $\text{(P4)}$ using standard optimization techniques, a closed form expression for $\gamma_1(c_1, c_2)$...
would be convenient. Now when \( c_1 > c_2 \), and \( \beta_1 \in (a, b) \), from Theorem 3.1 we see that
\[
\gamma_1(c_1, c_2) = \lambda(1 - F(\beta_1)) \tag{3.3}
\]
where
\[
\beta_1 = \{ \beta : c_1 + \beta D_1(\lambda(1 - F(\beta))) = c_2 + \beta D_2(\lambda F(\beta)) \}.
\]
A similar condition follows when \( c_1 < c_2 \) and therefore obtaining explicit expression for \( \gamma_1(c_1, c_2) \) is difficult. Note that we have not yet assumed any functional form for \( D_j \) and the choice of this function can further complicate matters. As a result, we require an alternative approach to solve the program [P4]. One possible alternative is to let the equilibrium \( \gamma_1 \) (the value of \( \gamma_1 \) at Wardrop equilibrium) be the optimization variable and represent other variables of the system such as \( c_1, c_2, \beta_1 \) as a function of \( \gamma_1 \). With slight abuse of notation, we will use \( c_j(\gamma_j) \) to denote the admission price at Server \( j \) when the arrival rate to Server \( j \) at equilibrium is \( \gamma_j \) where \( j = 1, 2 \). Similarly, we shall use \( \beta_1(\gamma_1) \) to represent the threshold \( \beta_1 \) corresponding to a kernel \( K^W \) that has an equilibrium arrival rate of \( \gamma_1 \) to Server 1. Note that \( c_1(\gamma_1) \) is also a function of \( c_2 \). This is because the equilibrium \( \gamma_1 \) depends on the difference \( c_1 - c_2 \) rather than their individual value \( c_1 \) and \( c_2 \). This is clear from Theorem 3.1 (Eq. (3.2)). Therefore for a given \( c_2 \) and \( \gamma_1 \in (0, \lambda) \) one can determine \( c_1 \) using Eq. (3.2). We have suppressed this dependence on \( c_2 \) to simplify notation. For the monopoly model \( c_2(\gamma_2) = c_2 \) as \( c_2 \) is a constant. The equivalent revenue optimization problem that we solve for the monopoly is as follows.

\[
\begin{align*}
\max_{\gamma_1} & \quad R_T(c_1(\gamma_1), \gamma_1) = c_2 \lambda + (c_1(\gamma_1) - c_2)\gamma_1 \\
\text{subject to} & \quad 0 \leq \gamma_1 \leq \gamma^1(c_2) \leq \lambda 
\end{align*}
\tag{P5}
\]

where \( \gamma^1(c_2) \) determines the domain for the feasible values of \( \gamma_1 \) as a function of \( c_2 \). While we shall characterize \( \gamma^1(c_2) \) formally in Section 3.3, an intuitive explanation for the quantity is as follows. Consider \( c_1 = c_2 = 0 \). From Theorem 3.1, we have \( \gamma_1 = \gamma^+ \) where \( 0 < \gamma^+ < \lambda \).

Now using the notation \( \gamma_1(c_1, c_2) \), we have \( \gamma_1(0, 0) = \gamma^+ \). Now as one increases \( c_1 \) from zero, customers will find Server 1 to be more costly and the equilibrium \( \gamma_1 \) decreases with increase in \( c_1 \). Clearly, for any \( c_1 > 0, \gamma_1 \in (\gamma^+, \lambda) \) is not possible and the quantity \( c_1(\gamma_1) \) in program P5 cannot be defined for \( \gamma_1 \in (\gamma^+, \lambda) \) since no admission price at Server 1 can lead to this equilibrium \( \gamma_1 \). Hence the domain for the variable \( \gamma_1 \) should be restricted to \([0, \gamma^+]\) when \( c_2 = 0 \).

More generally, for an arbitrary \( c_2 \), the domain for equilibrium \( \gamma_1 \) in program P5 is defined using \( \gamma^1(c_2) \) that will be characterized explicitly in Section 3.3.
3.3 Monopoly Market

In this section, we will analyze the program $P_5$. Recall that the service system operates two parallel servers and an arriving customer has to choose one of this server for service. We assume that the admission price $c_2$ at Server 2 is a-priori fixed. To be able to solve program $P_5$ we need to characterize $c_1(\gamma_1)$ for a fixed value of $c_2$. An intuitive outline of how one can obtain $c_1(\gamma_1)$ is as follows. From Eq. (3.2) of Theorem 3.1, we know that when $\beta_1 \in (a, b)$, (and hence $\gamma_1 \in (0, \lambda)$) we have

$$c_1 - c_2 = \beta_1 (D_2(\gamma_2) - D_1(\gamma_1)).$$

To characterize $c_1(\gamma_1)$ it is sufficient to express the right hand side of the above equation as a function of $\gamma_1$. Define this right hand side as follows

$$g_1(\gamma_1) := \beta_1 (\gamma_1) (D_2(\lambda - \gamma_1) - D_1(\gamma_1))$$

where recall that $\beta_1(\gamma_1)$ now represents the threshold $\beta_1$ for a kernel $K^W$ that satisfies Theorem 3.1 and has an equilibrium arrival rate of $\gamma_1$. Now for a fixed $c_2$ and for a $\gamma_1$ satisfying $0 \leq \gamma_1 \leq \gamma^1(c_2) \leq \lambda$ (the domain of $\gamma_1$ in program $P_5$) we have $c_1(\gamma_1) = c_2 + g_1(\gamma_1)$.

We characterize $c_1(\gamma_1)$ in this section in the following manner. We first characterize $\beta_1(\gamma_1)$ using Lemma 3.2 and 3.3. Then in Lemma 3.5, we characterize $g_1(\gamma_1)$. For a fixed $c_2$, we then obtain $\gamma^1(c_2)$ that determines the domain of $c(\gamma_1)$ in Lemma 3.7. To prove this lemma we need to identify the relation between $\Delta := c_1 - c_2$ and $g_1(\gamma_1)$. This is part of Lemma 3.6. Finally we characterize $c(\gamma_1)$ in Theorem 3.9 once $g_1(\gamma_1)$ and $\gamma^1(c_2)$ have been obtained.

Recall that we have made very mild assumptions on the properties of the distribution $F(\cdot)$ as well as on the delay cost function $D_j(\cdot)$. For the sake of our numerical examples and also to numerically illustrate the properties of $\beta_1(\cdot), g_1(\cdot)$ and $c_1(\cdot)$, we consider $F(\cdot)$ and $D_j(\cdot)$ from one of the following examples.

- $F$ satisfies a uniform distribution over the range $[a, b]$.
- $F$ satisfies the exponential distribution with mean $\tau$.
- $F$ satisfies the Gamma distribution with shape $k$ and scale $\theta$.

For the delay cost function, we shall assume one of the following.

- $D_j(\gamma_j) = \frac{\gamma_j}{\mu_j}$. This corresponds to the case of linear delay.
Figure 3.3: Illustrating $\gamma_1 \in [0, \gamma^+]$ and $c_1 \geq c_2$.

- $D_j(\gamma_j) = \frac{1}{\mu_j - \gamma_j}$ and $\mu_j > \lambda$. This corresponds to $M/M/1$ type delay cost function that was already seen in the previous chapter.

We now begin with the following lemma that identifies the necessary and sufficient condition on the equilibrium $\gamma_1$ when either $c_1 \geq c_2$ or $c_1 < c_2$.

**Lemma 3.2.** $\gamma_1 \in [0, \gamma^+]$ iff $c_1 \geq c_2$ while $\gamma_1 \in (\gamma^+, \lambda]$ iff $c_1 < c_2$.

Refer Figure 3.3 and 3.4 for a representation of the lemma. Figure 3.3 consider the case when $\gamma_1 \in [0, \gamma^+]$ and $c_1 \geq c_2$. The other case in the lemma is illustrated in Figure 3.4.

**Proof.** We first prove that $\gamma_1 \in [0, \gamma^+]$ implies $c_1 \geq c_2$. Recall the definition of $\gamma^+$ that

$$\gamma^+ = \{\gamma_1 : D_1(\gamma_1) = D_2(\gamma_2)\}.$$

Since $D_j(\gamma_j)$ is monotonic and increasing in $\gamma_j$ for $j = 1, 2$ and that $\gamma_2 = \lambda - \gamma_1$ we have $D_1(\gamma_1) \leq D_2(\gamma_2)$ for $\gamma_1 \in [0, \gamma^+]$. Now let $\gamma_1 = 0$. From the Wardrop conditions, this implies that $c_1 + \beta D_1(0) > c_2 + \beta D_2(\lambda)$ for all $\beta$. Since $D_1(0) < D_2(\lambda)$ (assumption) $c_1 > c_2$ must indeed be true.

When $\gamma_1 = \gamma^+$, we will show that $c_1 = c_2$. Suppose this is not true, i.e., $\gamma_1 = \gamma^+$ while $c_1 \neq c_2$. Since $\gamma_1 = \gamma^+$, we have $D_1(\gamma_1) = D_2(\gamma_2)$. As $c_1 \neq c_2$, customers have an incentive to move from the server with a higher admission price to the one with a lower price. This implies that $\gamma_1 = \gamma^+$ is not an equilibrium condition.
Now consider \( \gamma_1 \in (0, \gamma^+) \) where \( D_1(\gamma_1) < D_2(\gamma_2) \). From Theorem 3.1, \( \gamma_1 \in (0, \gamma^+) \) implies \( \beta_1 \in (a, b) \) and hence \( c_1 + \beta_1 D_1(\gamma_1) = c_2 + \beta_1 D_2(\gamma_2) \). Since \( D_1(\gamma_1) < D_2(\gamma_2) \) we have \( c_1 \geq c_2 \).

We now prove that if \( c_1 \geq c_2 \), then \( \gamma_1 \in [0, \gamma^+] \). We first show that when \( c_1 = c_2 \), we have \( \gamma_1 = \gamma^+ \). Suppose that when \( c_1 = c_2 \), \( \gamma_1 \neq \gamma^+ \). From the definition of \( \gamma^+ \) we have \( D_1(\gamma_1) \neq D_2(\gamma_2) \) and hence customers have an incentive to move from the server with a higher mean delay to the one with lower mean delay. This implies that when \( c_1 = c_2 \), \( \gamma_1 \neq \gamma^+ \) is not a valid Wardrop equilibrium.

Now let \( c_1 > c_2 \). From Theorem 3.1 we have either (1) \( \beta_1 = a \), (2) \( \beta_1 = b \) or \( \beta_1 \in (a, b) \). The case \( \beta_1 = a \) cannot happen! This is because while \( c_1 > c_2 \), we have also assumed \( D_1(\lambda) > D_2(0) \). \( K^W \) with \( \beta_1 = a \) will be possible only when

\[
c_1 - c_2 \leq \beta(D_2(0) - D_1(\lambda))
\]

for all \( \beta \in [a, b] \). Now this is not possible as the left hand side is positive while the right hand side is negative. It is straightforward to see that when \( \beta_1 = b \), \( \gamma_1 = 0 \) while when \( \beta_1 \in (a, b) \) we have \( c_1 + \beta_1 D_1(\gamma_1) = c_2 + \beta_1 D_2(\gamma_2) \). Again, since \( c_1 > c_2 \), we have \( D_1(\gamma_1) \leq D_2(\gamma_2) \) and hence \( \gamma_1 \in (0, \gamma^+) \). The proof for \( \gamma_1 \in (\gamma^+, \lambda] \) follows along similar lines and will not be provided. This completes the proof.

Having seen the relation between the admission prices and the equilibrium \( \gamma_1 \) in the previous lemma, we shall now express the threshold \( \beta_1 \) of Theorem 3.1 as a function of \( \gamma_1 \). Recall
from the theorem that $K^W$ is characterized by $\beta_1$ when $c_1 \neq c_2$. From Lemma 3.2, $c_1 \neq c_2$ implies $\gamma_1 \neq \gamma^+$. We define $\beta_1(\gamma_1)$ to denote the value of the threshold $\beta_1$ that characterizes $K^W$ when $\gamma_1 \neq \gamma^+$. We have the following lemma.

**Lemma 3.3.**

$$
\beta_1(\gamma_1) = \begin{cases} 
F^{-1} \left( \frac{\lambda-\gamma_1}{\lambda} \right) & \text{for } 0 \leq \gamma_1 < \gamma^+, \\
F^{-1} \left( \frac{\gamma_1}{\lambda} \right) & \text{for } \gamma^+ < \gamma_1 \leq \lambda.
\end{cases}
$$

(3.5)

where $F^{-1}$ represents the quantile function or the inverse function of the distribution $F$.

**Proof.** From Lemma 3.2, $\gamma_1 \in [0, \gamma^+]$ implies that $c_1 > c_2$ while $\gamma_1 \in (\gamma^+, \lambda]$ implies $c_1 < c_2$. Now from Theorem 3.1, when $c_1 > c_2$, we have

$$
\gamma_1 = \lambda \int_{\beta_1}^{b} 1dF(\beta) = \lambda (1 - F(\beta_1)).
$$

Similarly, when $c_1 < c_2$ we have

$$
\gamma_1 = \lambda \int_{a}^{\beta_1} 1dF(\beta) = \lambda (F(\beta_1)).
$$

Now $\beta_1(\gamma_1)$ which represents the value of the threshold $\beta_1$ when the arrival rate to Server 1 is $\gamma_1$ can be represented as follows.

$$
\beta_1(\gamma_1) = \begin{cases} 
\beta : \int_{\beta}^{b} \lambda dF(\beta) = \gamma_1 & \text{for } 0 \leq \gamma_1 < \gamma^+, \\
\beta : \int_{a}^{\beta} \lambda dF(\beta) = \gamma_1 & \text{for } \gamma^+ < \gamma_1 \leq \lambda.
\end{cases}
$$

(3.6)

Now recall our assumption that $F(\cdot)$ is absolutely continuous and strictly increasing in its domain. Further, the support is $[a, b]$ and hence $F(\cdot)$ is a bijective function whose inverse exists. In fact $F^{-1}(\cdot)$ is continuous and strictly increasing in its domain. The statement of the lemma now follows. 

Now note that $\beta_1(\gamma_1)$ is undefined in Lemma 3.3 when $\gamma_1 = \gamma^+$. This is because the Wardrop kernel $K^W$ with $\gamma_1 = \gamma^+$ is not unique. For sake of convenience however we will choose a kernel $K^W$ which is characterized by a unique threshold $\beta_1$ and satisfies $\gamma_1 = \gamma^+$. As convention, we define $\beta_1(\gamma^+) = F^{-1} \left( \frac{\lambda-\gamma^+}{\lambda} \right)$. The modified $\beta_1(\gamma_1)$ is now as follows.

$$
\beta_1(\gamma_1) = \begin{cases} 
F^{-1} \left( \frac{\lambda-\gamma_1}{\lambda} \right) & \text{for } 0 \leq \gamma_1 \leq \gamma^+, \\
F^{-1} \left( \frac{\gamma_1}{\lambda} \right) & \text{for } \gamma^+ < \gamma_1 \leq \lambda.
\end{cases}
$$

(3.7)
Refer Fig. 3.5 for a numerical evaluation of Eq. (3.7) for the case when $F(\cdot)$ is an exponential distribution with $\tau = 20$. The delay functions are $D_j(\gamma_j) = \frac{2\gamma_j}{\mu_j}$ where $\mu_1 = 3.3$ and $\mu_2 = 4$.

Fig. 3.6 corresponds to the case when the two servers are identical, i.e., $\mu_1 = \mu_2 = 4$.

Remark 3.4. Note that since $F^{-1}(\cdot)$ is continuous in its arguments, $\beta_1(\gamma_1)$ is continuous when $0 \leq \gamma_1 < \gamma^+ \text{ and } \gamma^+ < \gamma_1 \leq \lambda$. However at $\gamma_1 = \gamma^+$, $\beta_1(\gamma_1)$ is in general not continuous (Refer Fig. 3.5). For the case when the servers are identical, i.e., $D_1(\gamma) = D_2(\gamma) = D(\gamma)$, we see from the definition of $\gamma^+$ that $\gamma^+ = \frac{\lambda}{2}$. For this case, it is easy to see that at $\gamma_1 = \gamma^+$, $\beta_1(\gamma_1)$ is continuous (but not differentiable). See Fig. 3.6.

Having obtained $\beta_1(\gamma_1)$, we shall now analyze $g_1(\gamma_1)$ that was defined earlier. This $g_1(\gamma_1)$ will be used later to obtain $c_1(\gamma_1)$. We have the following lemma.

Lemma 3.5. $g_1(\gamma_1)$ is continuous, monotonic decreasing in $\gamma_1$ for $0 \leq \gamma_1 \leq \lambda$. Further, $g_1(\gamma^+) = 0$.

Proof. Recall our assumption that $D_j(\gamma_j)$ is continuous and monotone increasing in $\gamma_j$ where $j = 1, 2$. Since $\gamma_2 = \lambda - \gamma_1$, $(D_2(\lambda - \gamma_1) - D_1(\gamma_1))$ is monotone decreasing in $\gamma_1$ for $0 \leq \gamma_1 \leq \lambda$.

Now consider $\gamma_1$ such that $0 \leq \gamma_1 < \gamma^+$. Recall Eq. (3.7) that determines $\beta_1(\gamma_1)$. For the particular range of $\gamma_1$, $\beta_1(\gamma_1)$ is continuous and strictly decreasing. The continuity follows...
from that of $F^{-1}(\cdot)$. Since $F^{-1}(\cdot)$ is strictly increasing in its arguments, $F^{-1}\left(\frac{\lambda - \gamma_1}{\lambda}\right) = \beta_1(\gamma_1)$ is decreasing in $\gamma_1$. Clearly, $g_1(\gamma_1)$ is monotone decreasing when $\gamma_1$ is such that $0 \leq \gamma_1 < \gamma^+$. When $\gamma_1$ is such that $\gamma^+ < \gamma_1 \leq \lambda$, from the definition of $\gamma^+$, we have $(D_2(\lambda - \gamma_1) - D_1(\gamma_1)) < 0$. In this range of $\gamma_1$, $\beta_1(\gamma_1)$ is continuous increasing in $\gamma_1$ as seen from Eq. (3.7). This again implies that $g_1(\gamma_1)$ is continuous decreasing when $\gamma_1$ satisfies $\gamma^+ < \gamma_1 \leq \lambda$.

$g_1(\gamma^+) = 0$ follows from the definition of $\gamma^+$ where $D_1(\gamma^+) = D_2(\lambda - \gamma^+)$. The continuity at $\gamma^+$ is obvious from the fact that $g_1(\gamma^+) = 0$ and $\lim_{\gamma_1 \to \gamma^+} g_1(\gamma_1) = 0$.

See Fig. 3.7 for a numerical evaluation of $g_1(\gamma_1)$ when $F(\cdot)$ is an exponential distribution with $\tau = 20$ and when the servers have a linear delay with $\mu_1 = 3.3$ and $\mu_2 = 4$.

To determine $c_1(\gamma_1)$, we also need to identify the domain over which the program $P_S$ is defined. As seen earlier, this is determined using $\gamma^1(c_2)$ for a given value of $c_2$. We shall characterize this in Lemma 3.7. But before doing so, we establish the following side result which is interesting in its own right. This result will also be used in the proof of Lemma 3.7.

Recall that Theorem 3.1 guarantees existence of a $\beta_1$ characterizing an equilibrium kernel $K^W$. However it does not guarantee the uniqueness of $\beta_1$ and hence of kernel $K^W$. In the following lemma, we show that when $c_1 \neq c_2$, $\beta_1$ and hence the kernel $K^W$ is unique.

**Lemma 3.6.** For a given $\Delta = c_1 - c_2$, the threshold $\beta_1$ characterizing the kernel $K^W$ in
Theorem 3.1 is as follows.

\[
\beta_1 = \begin{cases} 
    b & \text{if } \Delta \geq g_1(0) \text{ or } \Delta \leq g_1(\lambda), \\
    \beta_1(\gamma) & \text{if } g_1(\lambda) < \Delta < g_1(0) \text{ where } \gamma \text{ satisfies } \Delta = g_1(\gamma).
\end{cases}
\]  

(3.8)

For a fixed \(\Delta, \gamma_1\) and hence \(\beta_1\) is unique and this implies the uniqueness of \(K^W\).

Proof. Suppose \(\Delta \geq g_1(0)\). From the definition of \(\Delta\) and from Eq. (3.4), this implies that

\[
c_1 - c_2 \geq b \left( D_2(\lambda) - D_1(0) \right) \\
\geq \beta \left( D_2(\lambda) - D_1(0) \right)
\]

for all \(\beta \in [a, b]\). From the Wardrop equilibrium condition, this implies that \(K^W(\beta, \cdot) = \delta_2\) for \(\beta \in [a, b]\). This implies that \(\gamma_1 = 0\) and from Eq. (3.1) we have \(\beta_1 = b\). Similarly when, \(\Delta \leq g_1(\lambda) < 0\) we have

\[
c_1 - c_2 \leq b \left( D_2(0) - D_1(\lambda) \right) \\
\leq \beta \left( D_2(0) - D_1(\lambda) \right)
\]
where $\beta \in [a, b]$. Again, from the Wardrop equilibrium condition, this implies that $K^W(\beta, \cdot) = \delta_1$ for $\beta \in [a, b]$ and hence $\gamma_1 = \lambda$. From Eq. (3.1), we have $\beta_1 = b$.

Now suppose $g_1(\lambda) < \Delta < g_1(0)$. While $g_1(0) > 0$, we have $g_1(\lambda) < 0$. From Lemma 3.6 we know that $g_1(\gamma_1)$ is monotonically decreasing in $\gamma_1$. Therefore there exists a unique $\gamma$ with $0 < \gamma < \lambda$ such that $\Delta = g_1(\gamma)$. This proves the uniqueness of $\gamma_1$. To see how $\beta_1 = \beta_1(\gamma)$ note that $\Delta = g_1(\gamma)$ implies that

$$c_1 - c_2 = \beta_1(\gamma) \left( D_2(\lambda - \gamma) - D_1(\gamma) \right).$$

Now if $\gamma \leq \gamma^+$ we have $D_2(\lambda - \gamma) > D_1(\gamma)$. In this case,

$$c_1 - c_2 \leq \beta (D_2(\lambda - \gamma) - D_1(\gamma))$$

for $\beta \in [a, \beta_1(\gamma)]$. This means that $K^W(\beta, \cdot) = \delta_2$ for all $\beta \in [a, \beta_1(\gamma)]$. Similarly, we have

$$c_1 - c_2 \geq \beta (D_2(\lambda - \gamma) - D_1(\gamma))$$

and $K^W(\beta, \cdot) = \delta_1$ when $\beta \in [\beta_1(\gamma), b]$. Similar arguments hold when $\gamma > \gamma^+$ and hence $\beta_1 = \beta_1(\gamma)$ when $g_1(0) < \Delta < g_1(\lambda)$.

From Theorem 3.1, $K^W$ is characterized by $\beta_1$ and for a fixed $\Delta$, $\beta_1$ is unique. This implies uniqueness of $K^W$. It is important to mention that $K^W$ is actually not unique when $\Delta = 0$. It is unique because of the assumptions made to ensure $\beta_1(\gamma_1)$ well defined at $\gamma_1 = \gamma^+$.

We now characterize $\gamma^1(c_2)$ in the following lemma.

**Lemma 3.7.**

$$\gamma^1(c_2) = \begin{cases} 
\lambda & \text{for } c_2 \geq -g_1(\lambda), \\
\gamma : g_1(\gamma) = -c_2 & \text{for } c_2 < -g_1(\lambda)
\end{cases} \quad (3.10)$$

*In words, when $c_2 < -g_1(\lambda)$ we have $\gamma^1(c_2) = \{\gamma : g_1(\gamma) = -c_2\}$ and for any $c_1 \geq 0$, the equilibrium $\gamma_1 \notin (\gamma^1(c_2), \lambda)$. However when $c_2 \geq -g_1(\lambda)$, we have $\gamma^1(c_2) = \lambda$ in which case for suitable choices of $c_1, \gamma_1 \in [0, \lambda]$.***

**Proof.** Suppose $c_2$ satisfies $c_2 < -g_1(\lambda)$. Assume that $c_1 = 0$ so that we have $\Delta > g_1(\lambda)$. From Lemma 3.6 this implies that the equilibrium $\gamma_1$ satisfies $g_1(\gamma_1) = \Delta = -c_2$. Let us label this $\gamma_1$ as $\hat{\gamma}$. Now increase $c_1$ from $c_1 = 0$ by a small enough $\epsilon > 0$ such that there exists $\gamma_1$ that satisfies $\Delta = \epsilon - c_2 = g_1(\gamma_1)$. Now from the monotonicity of $g_1(\cdot)$ it is clear that the equilibrium $\gamma_1$ is decreasing as the $\Delta$ increases. This implies that a higher $\Delta$ caused by increasing $c_1$ will only
lead to a \( \gamma_1 \) satisfying \( \gamma_1 < \hat{\gamma} \). Clearly, for any choice of \( c_1 \geq 0 \), we have \( \gamma_1 \notin [\hat{\gamma}, \lambda] \) and hence for this case \( \gamma^1(c_2) = \hat{\gamma} \).

Now suppose that \(-c_2 \leq g_1(\lambda)\). When \( c_1 = 0 \), this implies \( \Delta \leq g_1(\lambda) \) and from Lemma 3.6 this implies \( \beta_1 = b \) with the corresponding \( \gamma_1 \) satisfying \( \gamma_1 = \lambda \). As we increase \( c_1 \), the equilibrium \( \gamma_1 \) decreases and hence \( \gamma_1 \) satisfies \( \gamma_1 \in [0, \lambda] \). The compact representation now follows.

The above lemma implies that, if the system parameters are such that \( c_2 \geq -g_1(\lambda) \), then for any \( c_1 \) such that \( c_1 \in (0, c_2 + g_1(\lambda)) \) we have \( \gamma_1 = \lambda \). On the other hand, if the parameters of the system are such that \( c_2 + g_1(\lambda) < 0 \), then for any set of admission prices \( c_1 \) at Server 1, we have \( \gamma_1 < \gamma^1(c_2) \).

**Remark 3.8.** In the above lemma, note that when \( c_2 < -g_1(\lambda) \), we have \( \gamma^1(c_2) = \gamma \) where \( g_1(\gamma) = -c_2 \). Now from Lemma 3.3 we know that \( g_1(\gamma) < 0 \) for \( \gamma > \gamma^+ \). Hence when \( c_2 \geq 0 \), we have \( \gamma^1(c_2) \geq \gamma^+ \) with strict equality when \( c_2 = 0 \).

Finally, we have the following theorem which expresses \( c_1 \) as a function of \( \gamma_1 \).

**Theorem 3.9.** \( c_1(\gamma_1) = c_2 + g_1(\gamma_1) \) for \( 0 < \gamma_1 < \gamma^1(c_2) \leq \lambda \). For \( \gamma_1 = 0 \), \( c_1(0) \) must be at least equal to \( c_2 + g_1(0) \), i.e., \( c_1(0) \geq c_2 + g_1(0) \). Similarly when \( \gamma^1(c_2) = \lambda \), we have \( c_1(\lambda) \leq c_2 + g_1(\lambda) \).

**Proof.** First consider a fixed \( \gamma_1 \) satisfying \( \gamma_1 \in (0, \gamma^1(c_2)) \) for a given value of \( c_2 \). The corresponding threshold \( \beta_1 \) is determined by Eq. (3.7) and hence we have \( \beta_1 \in (a, b) \) for \( \gamma_1 \in (0, \gamma^1(c_2)) \). Now recall that Lemma 3.6 relates the threshold \( \beta_1 \) with \( \Delta \). Since \( \beta_1 \neq b \), from Lemma 3.6 \( \Delta \) must satisfy \( \Delta = g_1(\gamma_1) \). Therefore for a fixed \( c_2 \), the admission price \( c_1(\gamma_1) \) resulting in the arrival rate of \( \gamma_1 \) at Server 1 is given by

\[
c_1(\gamma_1) = c_2 + g_1(\gamma_1).
\]

For the case \( \gamma_1 = 0 \), from Eq. (3.7), we have \( \beta_1 = b \). From Lemma 3.6 this implies that \( \Delta \geq g_1(0) \). From the definition of \( \Delta \), we have \( c_1(0) \geq c_2 + g_1(0) \). Similarly when \( \gamma_1 = \lambda \), from Eq. (3.7), we have \( \beta_1 = b \). From Lemma 3.6 this implies \( \Delta \leq g_1(\lambda) \) and hence \( c(\lambda) \leq c_2 + g_1(\lambda) \). This completes the proof.

In the above theorem, \( c_1(0) \) and \( c_1(\lambda) \) are not unique values. When \( \gamma_1 = 0 \), the revenue at Server 1 is zero and w.l.o.g we can define \( c_1(0) = c_2 + g_1(0) \). \( \gamma_1 = \lambda \) requires \( c(\lambda) \leq c_2 + g_1(\lambda) \). Since the revenue maximizing price in this case is \( c_2 + g_1(\lambda) \), we also define \( c(\lambda) = c_2 + g_1(\lambda) \).
Further, for $\gamma^1(c_2) < \gamma_1 < \lambda$, $c_1(\gamma_1)$ is undefined and hence can be set to arbitrarily low values. The function $c_1(\gamma_1)$ for the domain $0 \leq \gamma_1 \leq \gamma^1(c_2) \leq \lambda$ can now be expressed as follows.

$$
c_1(\gamma_1) = \begin{cases} 
  c_2 + g_1(\gamma_1) & \text{for } 0 < \gamma_1 \leq \gamma^1(c_2) < \lambda, \\
  c_2 + g_1(0) & \text{for } \gamma_1 = 0, \\
  c_2 + g_1(\lambda) & \text{for } \gamma_1 = \gamma^1(c_2) = \lambda.
\end{cases} \tag{3.11}
$$

Now recall the revenue maximization problem $P5$. Define $\gamma_1^*$ as the optimizer for this program and $c_1(\gamma_1^*)$ as the admission price to be set at Server 1 such that the revenue of the service system is maximized. Now since $R_T(c_1(\gamma_1), \gamma_1) = c_2 \lambda + (c_1(\gamma_1) - c_2) \gamma_1$, clearly $\gamma_1^*$ must be such that $c(\gamma_1^*) > c_2$. Now from Theorem 3.9 we know that $c_1(\gamma_1) - c_2 = g_1(\gamma_1)$ for $\gamma_1 \in (0, \gamma^1(c_2))$ and further from Lemma 3.5 we have $g_1(\gamma_1) > 0$ for $\gamma_1 \in (0, \gamma^\ast)$. Now note from Remark 3.8 that $\gamma^1(c_2) \geq \gamma^\ast$. This implies that $\gamma_1^* \in (0, \gamma^\ast)$. The term $c_2 \lambda$ in $R_T(c_1(\gamma_1), \gamma_1)$ is a constant and hence we have the following equivalent program for the revenue maximization problem.

$$\max_{\gamma_1} g_1(\gamma_1) \gamma_1 \tag{P6}$$
subject to $0 \leq \gamma_1 \leq \gamma^\ast$

where $g_1(\gamma_1)$ is given by Eq. (3.4).

Note from Lemma 3.5 that $g_1(\cdot)$ is a continuous function of its domain. Program $P6$ involves maximizing a continuous function over a compact set and hence a maximizer $\gamma_1^*$ exists. Since $g_1(\gamma_1)$ is strictly decreasing (and hence quasi-convex), $g_1(\gamma_1)\gamma_1$ is a product of two quasi-convex functions. Product of quasi-convex functions need not be quasi-convex function. Non-convex programs such as $P6$ are in general difficult to analyze. Further recall that our assumptions on the distribution function $F$ and the delay cost function $D_j$ for $j = 1, 2$ are very minimal. To understand Program $P6$ we perform a numerical evaluation of $g_1(\gamma_1)\gamma_1$ under a combination of assumptions on the distribution functions $F$ and the delay functions $D_j(\gamma_j)$ that were outlined earlier.

Example 1: In this example we shall assume that the $D_j(\gamma_j) = \frac{\tau}{\mu_j}$ for $j = 1, 2$. We assume that $\mu_1 = 3.3$ and $\mu_2 = 4. Further, the arrival rate $\lambda = 3$ and we consider the following three examples for the distribution $F(\cdot)$. (1) $F$ has a uniform distribution with support on $[2,6]$. (2) $F$ has an exponential distribution with mean $\tau = 4$ and (3) $F$ has a Gamma distribution with
The scale $k$ and shape $\theta$ parameters 2 and 2 respectively. Note that these three distributions have the same mean. We plot $R_T(c_1(\gamma_1), \gamma_1) = c_2\lambda + g_1(\gamma_1)\gamma_1$ as a function of $\gamma_1$ in Fig. 3.8 where we assume $c_2 = 1$. When $F$ has the uniform distribution, $\gamma_1^* = 0.62$. The optimal revenue rate $R_T(\gamma_1^*) = 4.306$ while the admission price $c_1(\gamma_1^*)$ maximizing $R_T$ is 3.106. The corresponding values for the exponential distribution are $\gamma_1^* = 0.44$, $R_T(\gamma_1^*) = 4.712$ and $c_1(\gamma_1^*) = 4.89$ while the values for gamma distribution are $\gamma_1^* = 0.51$, $R_T(\gamma_1^*) = 4.532$ and $c_1(\gamma_1^*) = 4$.

**Example 2:** In this example, we assume that $D_j(\gamma_j) = \frac{\mu_j}{\mu_j - \gamma_j}$ where again $\mu_1 = 3.3$ and $\mu_2 = 4$. Note that $\lambda < \mu_j$ for $j = 1, 2$. The choice of $F(\cdot)$ as in the previous example. A plot of $R_T(c_1(\gamma_1), \gamma_1)$ as a function of $\gamma_1$ is provided in Fig. 3.9. When $F$ has the uniform distribution, $\gamma_1^* = 0.48$. The optimal revenue rate $R_T(\gamma_1^*) = 3.83$ while the admission price $c_1(\gamma_1^*)$ maximizing $R_T$ is 2.72. The corresponding values for the exponential distribution are $\gamma_1^* = 0.33$, $R_T(\gamma_1^*) = 4.21$ and $c_1(\gamma_1^*) = 4.67$ while the values for gamma distribution are $\gamma_1^* = 0.38$, $R_T(\gamma_1^*) = 4.04$ and $c_1(\gamma_1^*) = 3.74$.

We conclude the analysis of the revenue maximization problem with the following observations made from the two examples given above.

- Firstly, we see that for the given choices of $F$, $R_T(c_1(\gamma_1), \gamma_1)$ is a unimodal function.

---

**Figure 3.8:** $R_T$ as a function of $\gamma_1$ when $D_j(\gamma_j) = \frac{\mu_j}{\mu_j - \gamma_j}$.
in $\gamma_1$. For the three choices of the distribution function, it can be shown that $F^{-1}(\cdot)$ is differentiable in its arguments. For such distribution functions with differentiable $F^{-1}(\cdot)$, this implies that $g_1(\gamma_1)$ and hence $R_T(c_1(\gamma_1), \gamma_1)$ is differentiable in $\gamma_1$ when $0 < \gamma_1 < \gamma^+$. Now from Rolle’s Theorem (Theorem 10.2.7 [60]), this implies that there exists a $\gamma_1 \in (0, \gamma^+)$ such that $\frac{dR_T}{d\gamma_1} = 0$. A $\gamma^*_1$ satisfying this equation is the revenue maximizing arrival rate to server 1. The admission price corresponding to this $\gamma^*_1$ can now be obtained using Eq. (3.11).

- For each of the three distributions, note that we have $E[\beta] = 4$. However, the revenue rate $R_T$ as a function of $\gamma_1$ is distinct in all the three cases. This implies that the revenue rate $R_T$ depends on the higher moments of the distribution $F$ and not just on its mean value.

- Note that $R_T$ depends on admission price through the addition factor of $c_2 \lambda$. For different values of $c_2$, the corresponding $\gamma^*_1$ does not change. However it is easy to see from Eq. (3.11) that $c_1(\gamma_1)$ increases linearly in $c_2$. 

3.4 Estimating the distribution $F$

Recall that $F$ denotes the distribution function for the delay sensitivity of the arriving customers. The knowledge of $F$ is necessary to determine the optimal and the equilibrium kernel $K^*$ and $K^W$ introduced in Chapter 2. Further, the kernel $K^W$ must be known for the revenue maximization problem seen in this chapter. However in most practical situations, the distribution function $F$ may not be known and due to the unobservable nature of the queues it may not be possible to even illicit such information from the arriving systems. In such situations the only alternative may be to estimate this distribution function. One possible method to do so is to vary the admission prices at the servers and then measure the change in the arrival rate of customers to the different server and then use the Wardrop equilibrium conditions to estimate $F$. In this section, we shall provide an elementary procedure to estimate the underlying continuous distribution function $F$. We also consider the case when there are discrete customer classes (such a multiclass model was considered in the previous chapter) in which case the aim is to identify the delay cost for the different classes as well as find the Poisson arrival rate of customers per class.

Throughout this section, we shall make the following assumptions. To fix a functional form for $D_j(\gamma_j)$, we shall assume that the two servers are modeled as $M/M/1$ queues with service rates $\mu_1$ and $\mu_2$ and admission prices $c_1$ and $c_2$ respectively. It goes without saying that our analysis will also hold for any delay cost $D_j(\cdot)$ that is monotonic and strictly increasing in its arguments. We assume that once the admission prices $c_1$ and $c_2$ at the servers are announced and that the Wardrop equilibrium is reached, each server $j$ will accurately determine or measure the equilibrium arrival rate $\gamma_j$ and the mean delay cost $D_j(\gamma_j)$ for $j = 1, 2$. Hence the measured values $\gamma_j$ and $D_j(\gamma_j)$ and the corresponding quantities at the Wardrop equilibrium will be assumed to be the same. We also assume that the total arrival rate of customers to the system denoted by $\lambda$ is known a priori and that $c_1 > c_2$, i.e., the admission price at the first server is higher than the second. Note that since the distribution $F(\cdot)$ is unknown, the functions $\beta_1(\cdot), g_1(\cdot), c_1(\cdot)$ also cannot be determined and used for our procedure.

We begin by estimating the distributions $F$ that belongs to a parameterized family, say for example the exponential distribution. Let the parameter for the exponential distribution be denoted by $\alpha$. Now when $c_1$ and $c_2$ at the two servers are fixed, the equilibrium $\gamma_1$ and $\gamma_2$ at the servers is measured immediately. We choose a $c_1, c_2$ such that $\gamma_j > 0$ for $j = 1, 2$. From
this, the mean delay cost \( D_j(\gamma_j) \) for \( j = 1, 2 \) is also calculated. Now since all the quantities (except \( \beta_1 \)) in Eq. (3.2) of Theorem (3.1) are known, the threshold \( \beta_1 \) can be determined as

\[
\beta_1 = \frac{c_1 - c_2}{D_2(\gamma_2) - D_1(\gamma_1)}.
\]

Now increase \( c_1 \) to \( c_1^\delta \) where \( c_1^\delta = c_1 + \delta \) for \( \delta > 0 \). This decreases the equilibrium \( \gamma_1 \) to say \( \gamma_1^\delta \). Let \( \beta_1^\delta \) denote the threshold when the arrival rate to Server 1 is \( \gamma_1^\delta \). Again, using the measurements of the arrival rates and the delay functions \( \beta_1^\delta \) can be determined from Eq. (3.2).

Now since \( \gamma_1^\delta < \gamma_1 < \gamma^+ \), from Lemma [3.3] we know that \( \beta_1(\gamma_1^\delta) > \beta_1(\gamma_1) \). This implies that \( \beta_1^\delta > \beta_1 \). Clearly, the ratio \( \frac{\gamma_1^\delta - \gamma_1}{\lambda} \) is the probability of an arriving customer with \( \beta \in [\beta_1, \beta_1^\delta] \) and hence

\[
\int_{\beta_1}^{\beta_1^\delta} \alpha e^{-\alpha x} dx = \frac{\gamma_1^\delta - \gamma_1}{\lambda}.
\]  

The only unknown quantity is the exponential parameter \( \alpha \) which can now be obtained from the above equation.

**Remark 3.10.** Since the exponential distribution has a single parameter, the parameter could be obtained using only Eq. (3.12). For a parameterized distribution with \( k \) parameters, we need \( k \) simultaneous equations in terms of the underlying parameters. These can be obtained by following the procedure above for \( k \) different admission price \( \{c_1^k\} \) at Server 1.

We will now describe a numerical method to obtain a piecewise constant approximation for the density function \( f \) that is not necessarily from a parameterized family of distribution functions. As an example, consider a random variable \( \beta \) supported on the range \([0, 4]\). Suppose the distribution function is

\[
P(\beta \leq x) = F(x) = \frac{x^2}{16}.
\]

The corresponding density function is denoted by \( f(x) = x/8 \) for \( x \in [0, 4] \). For this example assume that there are two \( M/M/1 \) servers with service rates \( \mu_1 = 5 \) and \( \mu_2 = 5 \), admission prices initially set to \( c_1 = c_2 = 5 \) and the total arrival rate \( \lambda = 5 \). As earlier, we assume that once the admission prices at the servers are announced, the Wardrop equilibrium is reached instantaneously and each servers can accurately determine the aggregate arrival rates and the mean delay per customer.

Now increase \( c_1 \) by \( \delta > 0 \) and for the admission price vector \( (c_1 + \delta, c_2) \), measure the equilibrium arrival rates and the mean delay in the queues and calculate the corresponding
threshold $\beta_1$ using Eq. (3.2). Repeat this for a finite number of times, each time increasing $c_1$ from its previous value by $\delta$. This experiment is denoted in Table 3.1.

<table>
<thead>
<tr>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$\gamma_1$</th>
<th>$\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>5</td>
<td>1.98</td>
<td>2.84</td>
</tr>
<tr>
<td>5.2</td>
<td>5</td>
<td>1.69</td>
<td>3.04</td>
</tr>
<tr>
<td>5.4</td>
<td>5</td>
<td>1.44</td>
<td>3.20</td>
</tr>
<tr>
<td>5.6</td>
<td>5</td>
<td>1.23</td>
<td>3.33</td>
</tr>
<tr>
<td>5.8</td>
<td>5</td>
<td>1.05</td>
<td>3.44</td>
</tr>
<tr>
<td>6.0</td>
<td>5</td>
<td>0.89</td>
<td>3.53</td>
</tr>
<tr>
<td>6.2</td>
<td>5</td>
<td>0.75</td>
<td>3.60</td>
</tr>
<tr>
<td>6.4</td>
<td>5</td>
<td>0.63</td>
<td>3.67</td>
</tr>
<tr>
<td>6.6</td>
<td>5</td>
<td>0.52</td>
<td>3.37</td>
</tr>
<tr>
<td>6.8</td>
<td>5</td>
<td>0.43</td>
<td>3.78</td>
</tr>
</tbody>
</table>

Table 3.1: The table indicates the price vector $(c_1, c_2)$, the measured value of $\gamma_1$ and the threshold $\beta$ obtained from Eq. (3.2).

Using the earlier notation, we observe from the table that as $c_1$ increases to, say $c_1 + \delta$, $\gamma_1$ decreases to $\gamma_1^\delta$ while the threshold $\beta_1$ increases (to $\beta_1^\delta$). As earlier, we have

$$\int_{\beta_1}^{\beta_1^\delta} f(x) dx = \frac{\gamma_1 - \gamma_1^\delta}{\lambda}$$

where the density function $f(x)$ is to be estimated. Now assume for all $x \in (\beta_1, \beta_1^\delta)$ that $f(x) = z$, where $z$ is a constant. By assuming this, we are approximating the density function $f(x)$ for $x \in (\beta_1, \beta_1^\delta)$ by a horizontal line of magnitude $z$ and thus approximating $f(x)$ by a piecewise constant function. As $\delta \to 0$, the approximation should converge to the true density function. We now have

$$z = \frac{\gamma_1 - \gamma_1^\delta}{\lambda(\beta_1^\delta - \beta_1)}.$$  \hspace{1cm} (3.13)

The value of $z$ for a fixed $c_1$ and $c_1 + \delta$ can be viewed as an estimate for the density function $f(x)$ and obviously $z \to f(x)$ as $\delta \to 0$. These values of $z$ for different values of $c_1$ are given in Table 3.2.

A plot comparing the true density function and the estimate is given in Fig. 3.10. The plot shows that the density function is reasonably accurately estimated and for better estimation, one naturally required more such measurement points.
<table>
<thead>
<tr>
<th>$c_1$</th>
<th>$c_1 + \delta$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.2</td>
<td>0.37</td>
</tr>
<tr>
<td>5.2</td>
<td>5.4</td>
<td>0.39</td>
</tr>
<tr>
<td>5.4</td>
<td>5.6</td>
<td>0.41</td>
</tr>
<tr>
<td>5.6</td>
<td>5.8</td>
<td>0.42</td>
</tr>
<tr>
<td>5.8</td>
<td>6.0</td>
<td>0.44</td>
</tr>
<tr>
<td>6.0</td>
<td>6.2</td>
<td>0.44</td>
</tr>
<tr>
<td>6.2</td>
<td>6.4</td>
<td>0.45</td>
</tr>
<tr>
<td>6.4</td>
<td>6.6</td>
<td>0.46</td>
</tr>
<tr>
<td>6.6</td>
<td>6.8</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 3.2: The estimates $z$ can be obtained from Eq. (3.13) from the successive changes in the admission prices and the corresponding measurements of the arrival rates.

There is however a limitation to this method. Note that when $c_1 = c_2$, the corresponding value of $\beta_1 = 2.84$. Any increase or decrease in either $c_1$ or $c_2$ cannot result in a $\beta_1$ such that $\beta_1 < 2.84$. This is because, for the underlying distribution we have from Eq. (3.7) that $\beta_1(\gamma^+) = 2.84$ and for any $\gamma \in [0, \lambda]$ with $\gamma \neq \gamma^+$, we have $\beta_1(\gamma) > \beta_1(\gamma^+)$. As a result, the density function $f(x)$ cannot be estimated for $x \leq 2.84$.

### 3.4.1 Estimating Discrete Distribution

For sake of completeness, we shall also consider the case where the distribution $F$ is a discrete distribution with $M$ customer classes as defined in the previous chapter. We will assume that for each Class $i$, the associated waiting cost $\beta_i$ and the arrival rate $\lambda_i$ are unknown. Further, $\beta_1 > \beta_2 \ldots > \beta_M$. We continue with the assumption that there are two servers each charging an admission price $c_1$ and $c_2$. We begin by setting $c_2 = 0$ and $c_1$ to an arbitrarily large value such that $\gamma_1 = 0$ while $\gamma_2 = \lambda$. This is represented in part (a) of Fig. 3.11. It goes without saying that we are assuming for now that $\mu_2 > \lambda$. Now start decreasing $c_1$ in steps of size $\delta$ and stop at the first instance when $\gamma_1$ increases to an arbitrarily small value $\epsilon$. We use the notation $c_1^j$ and $\gamma_1^j$ to denote the admission price and the arrival rate at Server 1 when $c_1$ is decreased $j$ times by $\delta$, i.e., when $c_1^j = c_1 - j\delta$. $\gamma_1 = \epsilon$ implies that the most sensitive delay class $\beta_1$ must now be using Server 1 along with Server 2. Now since the delay function at each queue can be measured, $\beta_1$...
Figure 3.10: Comparing the estimate of $f(\cdot)$ with the true density function.

Figure 3.11: Estimating the discrete distribution $F$
can be easily determined from the corresponding Wardrop condition
\[ c_1^j + \beta_1 D_1(\gamma_1^j) = \beta_1 D_2(\gamma_2^j). \]

We will now determine \( \lambda_1 \) corresponding to this \( \beta_1 \). Continue decreasing \( c_1 \). The proportion of Class 1 customers using Server 1 keeps increasing till all Class 1 customers use only Server 1. When this happens, the corresponding Wardrop equilibrium condition for some \( k > j \) satisfies
\[ c_k^j < \beta_1 \left( D_2(\gamma_2^k) - D_1(\gamma_1^k) \right) \]
and this is represented by part (b) in Fig. 3.11. For a Class 2 customer to start using Server 1, the Wardrop equilibrium condition is
\[ c_1^m = \beta_2 \left( D_2(\gamma_2^m) - D_1(\gamma_1^m) \right) \]
where \( m > k \). Further since \( m > k \), we have
\[ \beta_2 \left( D_2(\gamma_2^m) - D_1(\gamma_1^m) \right) < \beta_1 \left( D_2(\gamma_2^k) - D_1(\gamma_1^k) \right). \]
and hence for all \( l \) such that \( k < l < m \), we have
\[ \beta_2 \left( D_2(\gamma_2^m) - D_1(\gamma_1^m) \right) < c_l^j < \beta_1 \left( D_2(\gamma_2^k) - D_1(\gamma_1^k) \right). \]
This means that for any \( c_1 \) satisfying \( c_1^k < c_1^j < c_1^m \), \( \gamma_1^j \) and \( \gamma_2^j \) remain unchanged. Clearly in this case \( \lambda_1 = \gamma_1^j \). Fig. 3.11 part (c) represents the fact that for any \( c_1^j > c_1^m \), Class 2 customers use both the servers at Wardrop equilibrium. Now continue this process till all the \( \lambda_i, \beta_i \) as well as the number of customer classes \( M \) is determined. It should be noted that the accuracy of our method increases as \( \delta \to 0 \). A downside of a small \( \delta \) is that the procedure may take a very long time to discover the system parameters.

3.5 Discussion

In this chapter, we have considered the problem of revenue maximization in parallel server systems. We have specialized with the case of two servers and assumed that both the servers belong to the same service provider. Further, the admission price at one of the server is required to be fixed and the service system can change the admission price at the other server to maximize its revenue. The Wardrop equilibrium when customers are heterogeneous and strategic has
already been characterized in Chapter 2. We use this characterization to simplify the revenue maximization program to make it more amenable to analysis. The equivalent program is easy to interpret, analyze and provides more insight into the problem. While it is intuitive that for a fixed $c_2$, the revenue maximizing $c_1$ should always be greater than $c_2$, the equivalent program enables us to characterize the revenue maximizing $c_1^*$ as a function of $c_2$.

An important assumption that was made throughout our analysis was that the distribution function $F$ was assumed to be known to the service system. We relaxed this assumption in Section 3.4 and have provided a procedure to estimate this distribution function. The proposed method is elementary and assumes that one is allowed to change admission price any number of times to obtain several measurements of the arrival rate. Further, we also assume that there is no cost to make the arrival rate and the delay cost measurements. A more realistic estimation method incorporating these practical limitations may make the problem more relevant and this is part of future work.

As part of the future work, we also wish to consider a duopoly model where each server competes with the other one to maximize their individual revenue rate. This is a standard game-theoretic problem and the interest is to identify the Nash equilibrium set of prices.
Chapter 4

Heterogeneous Customers in a Discriminatory Processor Sharing System

4.1 Introduction

In this chapter, we consider single server systems which offer multiple service classes to a multiclass customer population. First suppose that the heterogeneous customer population is served by a single processor sharing (PS) server. Like in the previous chapter, we assume that different type of customers have different costs per unit sojourn time. Since the service discipline is PS, the service capacity is shared equally among all the customers present in the system. Computer and communication systems provide several examples for such scenario. In an e-commerce web server, several threads will be simultaneously active and the CPU capacity needs to be shared among all the active threads. Some of the threads may be performing a more delay sensitive transaction (e.g., payment) while others a lesser delay-sensitive transaction (e.g., browsing). In a communication system, a down link may be shared among several different types of traffic (e.g., real time streams and elastic flows) with each of them having different delay sensitivities. Such systems have been widely studied in the literature.

In the ‘vanilla’ PS system, the server provides egalitarian service in that all customers in the system have the same share of the processor capacity. If the total sojourn time cost per unit time is the cost of the system, then it can be shown that this cost can be reduced by giving a higher share of the service capacity to customers with a higher sojourn time cost. Two variants of such a processor sharing system where different customers obtain a different shares of the service capacity have been studied in the literature—the generalized processor
sharing (GPS) and the discriminatory processor sharing (DPS). Recall our discussion of these scheduling policies from Chapter 1.

In both GPS and DPS, a number of service classes are defined and each class is assigned a fixed weight. In GPS, a separate queue is maintained for each class of service and the total service capacity of the server is shared among the non empty queues in proportion to the associated weights and independent of the number of customers in the queue. From each queue, only the head-of-line customer is served and each class of service is thus assured a minimum service rate. In DPS, an arriving customer is assigned to a class of service and the total service capacity is shared among all the customers present in the system in proportion to their weights and not just one customer of the service class. Thus any customer that is arriving into the system or departing from it will affect the service rate, and hence the sojourn time, of every other customer in the busy period. The DPS system was introduced in Kleinrock [41]. Fayolle et al. [28] derived a set of simultaneous integro-differential equations whose solution obtains the mean sojourn time of a customer of each service class conditioned on its service time. For the case of exponential service times, Haviv and van der Wal [36] develop a system of linear equations whose solution obtains the mean sojourn time for each service class. An alternative analysis is presented in Haviv and van der Wal [35] where an arriving customer has to make a bid before observing the queue occupancy, and the weight is equal to the bid. For the case of homogeneous customers (all have the same waiting cost), Haviv and van der Wal [35] show that the Nash bidding strategy is one where everyone has the same bid; further, this equilibrium bid is unique. Kim and Kim [38] have recently provided the sojourn time distribution in an $M/M/1$ system with the DPS scheduling policy. Altman et al. [5] provides a more recent survey of the analytical models for DPS queues.

In this chapter our interest is in a DPS queueing system serving a heterogeneous population. As was mentioned before, different types of customers have different costs per unit sojourn time or delay. Consider a system with $M$ service classes with the weight of class $m$ being $w_m$. Since different customers have different delay costs, it is natural to assign customers of higher delay costs to classes offering higher share of service. Our first interest is in the optimal allocation of the customer types to service classes that minimizes the social cost, i.e., the sum of the waiting costs of the customers. For the case when there is a continuum of customer types, we show that the optimal policy is of the threshold type. We analyze the optimal policy when the number of customer types is finite. Next we consider a system with an admission charge,
different for the different service classes which is to be paid by the arriving customers. The weights and the price of each class is announced a priori and the arriving customers have to choose the class without observing the queue occupancy. The arriving customers make an individually optimal decision on the queue that minimizes the sum of the admission price and their delay cost. This leads to a Wardrop equilibrium routing. For fixed weights and prices, we show that the equilibrium routing is of threshold type, similar to the optimal allocation. However, the equilibrium threshold values need not coincide with that of the optimal allocation. Finally, we obtain admission price so that the resulting equilibrium allocation coincides with the optimum allocation minimizing the social cost.

A heavy traffic analysis of a DPS system with heterogeneous customers is also considered in literature \[63, 24\]. Wu et al. \[63\] consider a model where an arriving customer of a type makes a bid and its weight in the DPS system is proportional to this bid. Doncel et al. \[24\] consider a type level bidding decision where all customers of the same type make a common bid to minimize the aggregate cost of service for that type. The heavy traffic regime simplifies the analysis and allows for closed form expressions of the mean sojourn time in the DPS system for any customer type. This is not the case in our model and hence we do not consider such systems any further.

The rest of the chapter is organized as follows. In Section 4.2 we describe the notation used throughout the chapter. In Section 4.3 we first consider the case when there is a continuum of types and show that the optimal policy is of the threshold type. We then characterize this threshold for a special case with two service classes. Next, in Section 4.3.1 we consider the case when the number of customer types is finite with a lower index value corresponds to a lower waiting cost per unit time. We show that for the case of \(N\) such customer classes, Class \(N\) (resp. Class 1) customers will always use the service class with highest (resp. lowest) weight in an optimal allocation. This is independent of the arrival rates of that class or of the other classes. For the case with three customer classes, we show that when a specified condition on the system parameters is satisfied, the class 2 customers are indifferent in choosing either of the two service classes and the optimal value of the social utility remains constant for any choice of service class made by Class 2 customers. In Section 4.4 we consider a system with an admission charge and we show that the resulting equilibrium routing is also of the threshold type. Finally we provide a mechanism to determine the admission prices so that the resulting equilibrium allocation coincides with the optimum allocation minimizing the social utility.
Another difference between this work and that of Haviv and van der Wal [35], apart from
the key difference of heterogeneity of customers, is that while we restrict the DPS system to have
only $M$ service classes, Haviv and van der Wal [35] consider the case where each unique bid is
a service class in itself. Such a system with infinite service classes coupled with heterogeneous
customers turns out to be hard to analyze.

4.2 Preliminaries

In this section we first set up the notation. The DPS queuing system has $M$ service classes
and $w_m$ is the weight associated with class $m$, $m = 1 \ldots M$. The service requirement for
each customer is i.i.d with an exponential distribution of unit mean. The server serves at unit
rate. Customers arrive according to a homogeneous Poisson process of rate $\lambda$. Associated
with each customer is a random variable $\beta$ that quantifies its sensitivity to delay. A continuum
of customer types is modeled by making $\beta$ to be a continuous random variable. We assume
$0 \leq a \leq \beta \leq b < \infty$ and that $\beta$ has a distribution function $F(\cdot)$. We continue with the
assumption from the previous chapter that $F(\cdot)$ is absolutely continuous with a density function
denoted by $f(\cdot)$. For the case when there are a finite number of classes, we assume that $\beta$ is a
discrete random variable taking values $\beta_1, \beta_2, \ldots \beta_N$, where $a = \beta_1 < \beta_2 < \ldots < \beta_N = b$. In
this case $p_n := \Pr(\beta = \beta_n)$.

Throughout this chapter, $\beta$ will be used to denote the realization of the random variable
$\beta$. While we could continue using the kernel notation of the previous chapters to denote the
probabilistic routing of customers to different service classes, we shall in this chapter proceed
with a simplified description for the same. With this in mind, let $q_m(\beta)$ denote the probability
that an arriving customer with a delay cost per unit time of $\beta$ is routed to service class $m$ and
$q(\beta)$ denote the corresponding $M$ dimensional vector, i.e., $q(\beta) = [q_1(\beta), q_2(\beta), \ldots, q_M(\beta)]$.

We will let $Q$ denote the set of $q(\beta)$ such that $\sum_{m=1}^M q_m(\beta) = 1$ for all $\beta$. $Q$ will be called the
set of routing functions. We assume that $q_m(\beta) \in L^p([a, b])$ for arbitrary $1 \leq p < \infty$ where
$L^p([a, b])$ is the set of all equivalence classes of $p$-integrable functions. If $x(\phi)$ and $y(\phi)$ are two
$p$-integrable functions on the interval $[a, b]$, then we say that $x \sim y$ if

$$d_p(x, y) = \left( \int_{[a,b]} |x - y|^p d\phi \right)^\frac{1}{p} = 0.$$ 

For convenience, we shall henceforth restrict to $q_m(\beta) \in L^1([a, b])$ and hence $p = 1$. The arrival
rate of customers to service class \( m \) is \( \lambda_m := \lambda \int_a^b q_m(\beta) dF(\beta) \). The mean sojourn time of a
customer in service class \( m \) is denoted by \( S_m(\lambda), m = 1 \ldots M \); to simplify notation, we omit
\( \lambda \) since the dependence is clear. Without loss of generality, we further assume \( 0 \leq w_1 < w_2 < 
\ldots < w_M \). Using the preceding notation, we define the social cost \( U(\mathbf{q}(\beta)) \) as

\[
U(\mathbf{q}(\beta)) = \lambda \sum_{m=1}^{M} \int_{a}^{b} \beta q_m(\beta) S_m dF(\beta).
\]

\( U(\mathbf{q}(\beta)) \) is the sum of the expected costs incurred by customers of different type at different
service classes, weighted by the corresponding arrival rates.

From Haviv and van der Wal [36] the mean sojourn time for each s ervice class can be
obtained as a solution to the following system of equations.

\[
S_i = \frac{1}{(1 - \sigma_i)} + \sum_{j=1}^{M} \frac{\lambda_j w_j S_j}{w_i + w_j (1 - \sigma_i)},
\]

where \( \sigma_i = \sum_{j=1}^{M} \frac{\lambda_j w_j}{w_i + w_j} \).

Solving for the special case of two service classes, i.e., \( M = 2 \), it can be shown that

\[
S_1 = \frac{1}{(1 - \lambda)} \left( \frac{w_1(1 - \lambda_1) + w_2}{w_1(1 - \lambda_1) + w_2(1 - \lambda_2)} \right),
\]

\[
S_2 = \frac{1}{(1 - \lambda)} \left( \frac{w_1(1 - \lambda_2) + w_2}{w_1(1 - \lambda_1) + w_2(1 - \lambda_2)} \right).
\]

(4.1)

From the above, we see that since \( w_1 < w_2 \), \( S_1 > S_2 \). In fact this is can be generalized as the
property below.

**Property 4.1.** If \( w_i < w_j \) then \( S_i > S_j \) for all \( i, j \in [1, M] \).

**Proof.** Let \( n_m, 1 \leq m \leq M \) denote the number of jobs in service class \( i \) and define
\( \bar{n} := (n_1, n_2, \ldots, n_M) \). Let \( \bar{n} \) denote the state of server. Define \( S(i, \bar{n}) \) as the mean residual sojourn
time of a customer using service class \( m \) when the system is in state \( \bar{n} \). From [36] we have

\[
S(i, \bar{n}) = \frac{1}{1 - \sigma_i} \left( 1 + \sum_{m=1}^{M} \frac{w_m n_m}{w_i + w_m} \right)
\]

where \( \sigma_i = \sum_{m=1}^{M} \frac{w_m}{w_i + w_m} \). Since \( w_i < w_j \),

\[
\frac{w_m}{w_i + w_m} > \frac{w_m}{w_j + w_m}
\]

for all \( m \). This implies that \( \sigma_i > \sigma_j \), and hence

\[
\frac{1}{1 - \sigma_i} > \frac{1}{1 - \sigma_j}.
\]

79
Further, it can be seen that
\[ 1 + \sum_{m=1}^{M} \frac{w_m n_m}{w_i + w_m} > 1 + \sum_{m=1}^{M} \frac{w_m n_m}{w_j + w_m}. \]

Hence \( w_i < w_j \) implies \( S(i, \tilde{n}) > S(j, \tilde{n}) \). This is true for all \( \tilde{n} \) and hence the result follows.

\[ \square \]

### 4.3 Optimal Allocation of Customer Types to Service Classes

We begin with the case when there are a continuum of customer types, i.e., \( \beta \) is a continuous random variable taking values in \([a, b] \). The optimal allocation is the social cost minimization problem

\[
\inf_{q(\cdot) \in \mathbb{Q}} U(q(\cdot)), \tag{P7}
\]

Let \( q^*(\cdot) \) denote a minimizer of program (P7). The optimization problem (P7) in general is not easy to analyze. While the objective function \( U(q(\cdot)) \) is not necessarily a convex function, the space of feasible allocations \( \mathbb{Q} \) may not even be a compact set. In this section, we analyze the structural properties of any optimal allocation of customers to service classes i.e., \( q^*(\cdot) \). We begin with the following definition.

**Definition 4.2.** A routing function \( q(\cdot) \) is said to be threshold-type if there exist thresholds \( t_1, t_2, \ldots, t_{M+1} \) such that \( a = t_1 \leq t_2 \leq \ldots \leq t_{M+1} = b \) and \( q_i(\beta) = 1 \) for all \( \beta \in (t_i, t_{i+1}) \) and \( q_i(\beta) = 0 \) for all \( \beta \notin [t_i, t_{i+1}], i = 1 \) to \( M \).

In the following theorem, we show that any minimizer of program (P7) has an equivalent threshold-type policy.

**Theorem 4.3.** Let \( q^*(\cdot) \) be an optimal allocation minimizing the social cost. Then there exists a threshold-type allocation \( \bar{q}(\cdot) \) such that \( q^*(\cdot) \sim \bar{q}(\cdot) \).

**Proof.** Proof is by contradiction. Suppose \( q^*(\cdot) \) is such that it does not satisfy the hypothesis of theorem and has no equivalent threshold-type policy. This implies that for some \( m \) such that \( 1 \leq m \leq M \) and for any threshold-type policy \( \bar{q}(\cdot) \) we have

\[ d_1(q^*_m(\cdot), \bar{q}_m(\cdot)) > 0. \]

Now define \( \bar{T}(z) = \lambda \int_a^z q^*_M(\beta)dF(\beta) \), where \( \alpha \leq z \leq b \). \( \bar{T}(z) \) is a continuous non decreasing function in \( z \) and it is the aggregate arrival rate of customers for service class \( M \) having a delay.
cost lower than $z$. Similarly, define $T(z) = \lambda \sum_{m=1}^{M-1} \int_z^b q^*_m(\beta)dF(\beta)$, where $a \leq z \leq b$. $T(z)$ is the aggregate arrival rate of customers to service classes other than $M$ and having a delay cost above $z$. Clearly, $T(z)$ is continuous non-increasing function in $z$.

For $z = a$, $\bar{T}(z) = 0$ while at $z = b$, $\bar{T}(z) = 0$. Hence there exist a $z = v_1$ such that

$$\int_{v_1}^a q^*_M(\beta)dF(\beta) = \int_{v_1}^b q^*_M(\beta)dF(\beta).$$

We have the following three cases for $v_1$.

1. If $v_1 = a$, we have

$$\lambda \sum_{m=1}^{M-1} \int_{v_1}^b q^*_m(\beta)f(\beta)d\beta = 0.$$  

This is possible only if $q^*_m(\beta) = 0$ for $m = 1$ to $M - 1$ and $\beta \in [a, b]$. This implies that all the customers choose service class $M$ and hence $t_1 = t_2 = \ldots = t_M = a$ and $t_{M+1} = b$. This is a contradiction as $q^*(\cdot)$ was assumed not equivalent to any threshold-type policy.

2. If $a < v_1 < b$ find an $x$ such that for some class $m$

$$\int_{v_1}^b q^*_m(\beta)dF(\beta) = \int_x^{v_1} q^*_M(\beta)dF(\beta) \quad (4.2)$$

Now swap the customers of type $\beta > v_1$ in service class $m$ with customers in service class $M$ of type $\beta < v_1$. Let $\hat{q}(\beta)$ denote the modified allocation function as a result of the swap operation. We have

$$\hat{q}_m(\beta) = \begin{cases} 0 & \text{for } \beta > v_1 \\ q^*_m(\beta) + q^*_M(\beta) & \text{for } x \leq \beta \leq v_1 \end{cases}$$

$$\hat{q}_M(\beta) = \begin{cases} q^*_m(\beta) + q^*_M(\beta) & \text{for } \beta > v_1 \\ 0 & \text{for } x \leq \beta \leq v_1 \end{cases}$$

and $\hat{q} = q^*$ elsewhere. Thus the difference in social cost from the swap is as follows.

$$U(q^*(\beta)) - U(\hat{q}(\beta)) = S_m \left( - \int_v^{v_1} \beta q^*_M(\beta)dF(\beta) + \int_{v_1}^b \beta q^*_m(\beta)dF(\beta) \right)$$

$$+ S_M \left( \int_x^{v_1} \beta q^*_M(\beta)dF(\beta) - \int_{v_1}^b \beta q^*_m(\beta)dF(\beta) \right)$$

$$= (S_m - S_M) \times \left( \int_{v_1}^b \beta q^*_m(\beta)dF(\beta) - \int_x^{v_1} \beta q^*_M(\beta)dF(\beta) \right) \quad (4.3)$$

81
From Eq. (4.2) it can be seen that
\[ \int_{v_1}^{b} \beta q^*_m(\beta) dF(\beta) > \int_{v_1}^{v_1} \beta q^*_M(\beta) dF(\beta). \] (4.4)
Substituting Eq. (4.4) in Eq. (4.3) we have \( U(q^*(\beta)) - U(\hat{q}(\beta)) > 0 \). This implies \( U(q^*(\beta)) > U(\hat{q}(\beta)) \) and any \( q^*(\cdot) \) not satisfying the theorem hypothesis is not an optimal allocation.

3. If \( v_1 = b \) then
\[ \lambda \int_{a}^{v_1} q^*_M(\beta) dF(\beta) = 0. \]
This implies that \( q^*_M(\beta) = 0 \). In that case check for the existence of a \( v' \) such that \( a \leq v' < b \) and satisfies
\[ \lambda \int_{a}^{v'} q^*_{n+1}(\beta) dF(\beta) = \lambda \sum_{m=1}^{n} \int_{v_1}^{v'} q^*_m(\beta) dF(\beta) \] (4.5)
for some \( n, 1 \leq n < M - 1 \). If such a \( v' \) for \( a \leq v' < b \) exist then from cases 1 and 2 above we have contradiction. If there exist no \( n \) such that \( a \leq v' < b \), then it implies that \( v_1 = b \) for all \( n, 1 \leq n < M - 1 \). Now for \( n = 1 \) in Eq. (4.5), we have \( \lambda \int_{a}^{v_1} q^*_{n+1}(\beta) dF(\beta) = \lambda \sum_{m=1}^{n} \int_{v_1}^{b} q^*_m(\beta) dF(\beta) = 0 \). This implies \( q_1(\beta) = 1 \) for all \( \beta \in [a, b] \) and the hypothesis of the theorem is satisfied. This contradicts our assumption on \( q^*(\cdot) \).

This completes the proof. \( \square \)

We now specialize the preceding result for the special case when \( M = 2 \) and obtain a more explicit characterization of the routing function for this case. From Theorem 4.3, \( q^*(\beta) \) is a threshold routing policy such that \( a = t_1 \leq t_2 \leq t_3 = b \). Therefore, the routing function for this case has the form \( q_1(\beta) = 1 \) for all \( \beta \in [a, t_2) \) and \( q_1(\beta) = 0 \) elsewhere. Substituting this in the Eq. (4.1), the mean sojourn time per customer at each of the two service classes are
\[ S_1 = \frac{1}{(1-\lambda)(w_1(1-\lambda F(t_2)) + w_2(1-\lambda(1-F(t_2)))}, \] (4.6)
\[ S_2 = \frac{1}{(1-\lambda)(w_1(1-\lambda F(t_2)) + w_2(1-\lambda(1-F(t_2)))}, \] (4.7)
and the social cost is
\[ U(q^*(\beta)) = \lambda \left( S_1 \int_{a}^{t_2} \beta dF(\beta) + S_2 \int_{t_2}^{b} \beta dF(\beta) \right). \]
Differentiating w.r.t. $t_2$ and simplifying we get
\[
\frac{dU}{dt_2} = \lambda t_2 f(t_2)(S_1 - S_2) \\
+ \lambda \left( \frac{dS_1}{dt_2} \int_a^{t_2} \beta dF(\beta) + \frac{dS_2}{dt_2} \int_{t_2}^b \beta dF(\beta) \right)
\]
where
\[
\frac{dS_1}{dt_2} = -\frac{\lambda f(t_2)(w_2 - w_1)}{(1-\lambda)} \frac{w_1(1 - \lambda) + w_2}{(w_1(1 - \lambda F(t_2)) + w_2(1 - \lambda(1 - F(t_2))))^2} \tag{4.9}
\]
\[
\frac{dS_2}{dt_2} = -\frac{\lambda f(t_2)(w_2 - w_1)}{(1-\lambda)} \frac{w_1 + w_2(1 - \lambda)}{(w_1(1 - \lambda F(t_2)) + w_2(1 - \lambda(1 - F(t_2))))^2} \tag{4.10}
\]

We are now ready to state the following corollary.

**Corollary 4.4.** When $M = 2$, the threshold $t_2$ is such that $a < t_2 < b$. Further,
\[
\frac{U(q^*(\beta))}{\lambda} = \frac{t_2}{1 - \lambda}. \tag{4.11}
\]

**Proof.** To prove that $a < t_2 < b$, it is sufficient to show that $\left. \frac{dU}{dt_2} \right|_a < 0$ and $\left. \frac{dU}{dt_2} \right|_b > 0$. From Eq. (4.8) at $t_2 = a$ we have
\[
\left. \frac{dU}{dt_2} \right|_a = \frac{\lambda f(a)(S_1 - S_2) + \lambda \frac{dS_2}{dt_2} \int_a^b v dF(v)}{\lambda^2 f(a)} = \frac{\lambda^2 f(a)(w_2 - w_1)}{1 - \lambda} \frac{(w_1 - w_2) + \lambda \frac{dS_2}{dt_2} \int_a^b v dF(v)}{(w_1(1 - \lambda F(t_2)) + w_2(1 - \lambda(1 - F(t_2))))^2}
\]
\[
= \frac{\lambda^2 f(a)(w_2 - w_1)}{(1-\lambda)(w_1 + w_2(1-\lambda))} (a - \int_a^b v dF(v)) < 0.
\]

Along the same lines, it can be shown that $\left. \frac{dU}{dt_2} \right|_b > 0$.

For $t_2$ to satisfy $a < t_2 < b$, we apply the first order necessary conditions for optimality, substitute the values of $S_1$ and $S_2$, and simplify as follows.
\[
\frac{dU(q^*(\beta))}{dt_2} = \frac{(w_1 - w_2)\lambda f(t_2)}{(w_1(1 - \lambda F(t_2)) + w_2(1 - \lambda(1 - F(t_2))))} \times \left( S_1 \int_a^{t_2} \beta dF(\beta) + S_2 \int_{t_2}^b \beta dF(\beta) \right)
\]
\[
- \frac{t_2}{1 - \lambda} = 0
\]

Thus the threshold must satisfy
\[
\frac{t_2}{1 - \lambda} = \frac{U(q^*(\beta))}{\lambda}.
\]
From the preceding we see that any optimal allocation must route a non-zero fraction of arriving customers to both the service classes. Further, the threshold $t_2$ is such that the delay cost per customer at optimal routing is the same as the delay cost for type $t_2$ customer in a M/M/1 queue. We now illustrate the corollary with the help of the following numerical example.

Consider a customer population with Poisson arrival rate $\lambda = 0.7$. Assume that the delay sensitivity $\beta$ is a uniform random variable with supports $a = 0$ and $b = 10$ and hence $f(\cdot) = \frac{1}{10}$. Assume that the service system offers two service classes with weights $w_1 = 10$ and $w_2 = 20$. Now since $M = 2$, the optimal allocation is characterized by the threshold $t_2$. Define $q(t)$ as a threshold routing policy such that $q_1(\beta) = 1$ for $\beta \in (a, t)$ and $q_1(\beta) = 0$ elsewhere. For the given parameters we first plot $U(q(t))$ as a function of $t$ in Fig. 4.1 We see that, $U(q(t))$ is minimized when $t = 4.5476$ for which case, $U(q(4.5476)) = 10.6111$. Now recall from the corollary that, the threshold $t_2$ must satisfy the first order optimality condition $\left. \frac{dU}{dt} \right|_{t_2} = 0$, i.e, $t_2$ is a stationary point of $U(q(t))$. We see from Fig. 4.1 that $U(q(t))$ is unimodal with $t$ as the stationary point and hence $t_2 = t = 4.5476$. Now $U(q^*(\beta)) = \frac{M_2}{1-\lambda} = 10.6111$ and this validates the corollary for this example.

### 4.3.1 Finite Customer Types

We now consider the case when there are a finite number of customer types, i.e., $\beta$ is a discrete random variable. For ease of discussion, customers of type $\beta_n$ will be denoted as type $n$ and
to simplify notation, $q_i(\beta_n)$ will be denoted by $q_{i,n}$. Since the expressions for the mean sojourn times are known only for the case of two service classes, we restrict the analysis of this section for the case when $M = 2$. The social cost in this system is

$$C(q(\beta)) = \sum_{n=1}^{N} \lambda p_n q_{1,n} \beta_n S_1 + \sum_{n=1}^{N} \lambda p_n q_{2,n} \beta_n S_2.$$  \hspace{1cm} (4.12)

In the following theorem, we will show that in a DPS system with two service classes, the optimal allocation is such that all the customers of type $N$ are allocated to the service class with the highest weight, while all the customers of type 1 are allocated to the service class with lower of the two weights.

**Theorem 4.5.** When $M = 2$, $q^*(\cdot)$ is such that $q_{1,1}^* = q_{2,N}^* = 1$.

**Proof.** We first differentiate Eq. (4.12) w.r.t. $q_{1,n}$ to obtain

$$\frac{dU(q)}{dq_{1,n}} = \sum_{i=1}^{N} p_i q_{1,i} \beta_i \frac{d(S_1 - S_2)}{dq_{1,n}}$$

$$+ \sum_{i=1}^{N} p_i \beta_i \frac{dS_2}{dq_{1,n}} + p_n \beta_n (S_1 - S_2).$$  \hspace{1cm} (4.13)

Differentiating the expressions for $S_1$ and $S_2$ w.r.t. $q_{1,n}$, we get

$$\frac{dS_1}{dq_{1,n}} = \frac{-\lambda p_n (w_2 - w_1)}{(1 - \lambda)} \frac{w_1 (1 - \lambda)}{(w_1 (1 - \lambda_1) + w_2 (1 - \lambda_2))^2}$$

$$\frac{dS_2}{dq_{1,n}} = \frac{-\lambda p_n (w_2 - w_1)}{(1 - \lambda)} \frac{w_1 + w_2 (1 - \lambda)}{(w_1 (1 - \lambda_1) + w_2 (1 - \lambda_2))^2}$$

Substituting the above in Eq. (4.13) and simplifying, we have

$$\frac{dU}{dq_{1,n}} = -\lambda p_n (w_2 - w_1) \times$$

$$\sum_{i=1}^{N} p_i (\beta_i - \beta_n) \frac{(q_{1,i} S_1 + (1 - q_{1,i}) S_2)}{(w_1 (1 - \lambda_1) + w_2 (1 - \lambda_2))}.$$  \hspace{1cm} (4.14)

For $n = 1$, $\beta_i - \beta_n$ is positive for all $i$ and hence $\frac{dU}{dq_{1,1,n}} \leq 0$. $U(q(\beta))$ is a decreasing function in $q_{1,1}$ and hence $q_{1,1}^* = 1$. Similarly, for $n = N$, $\beta_i - \beta_n$ is negative for all $i$ and hence $\frac{dU}{dq_{1,N}} \geq 0$. This implies that $C(q(\beta))$ is an increasing function in $q_{1,N}$. Therefore $q_{1,N}^* = 0$. \hfill $\square$

Now let us analyze the special case of $M = 2$ and $N = 3$. From the preceding theorem, we know that $q_{1,1}^* = 1$, $q_{1,3}^* = 0$. Substituting $N = 3$, $q_{1,1} = 1$ and $q_{1,3} = 0$ in Eq. (4.14), we have
\[
\frac{dU}{dq_{1,2}} = -\lambda p_2 (w_2 - w_1) \times \frac{p_1 (\beta_1 - \beta_2)(w_1 + w_2 - w_1\lambda) + p_1(\beta_1 - \beta_2)(w_1 + w_2 - w_2\lambda)}{(w_1(1 - \lambda_1) + w_2(1 - \lambda_2))^2}
\]  
(4.15)

For \(0 < q^*_1, q^*_2 < 1\), we require \(\frac{dU}{dq_{1,2}} = 0\). From Eq. (4.15), this is equivalent to

\[
p_1(\beta_2 - \beta_1)(w_1 + w_2 - w_1\lambda) = p_3(\beta_3 - \beta_2)(w_1 + w_2 - w_2\lambda).
\]  
(4.16)

The above equality is independent of \(q_{1,2}\) and hence any \(q_{1,2} \in [0, 1]\) satisfies \(\frac{dU}{dq_{1,2}} = 0\) when the parameters satisfy Eq. (4.16). Thus the social cost remains constant for any choice of \(q_{1,2}\) and hence type 2 customers can be assigned to any service class without increasing the social cost.

Note that when \(p_1(\beta_2 - \beta_1)(w_1 + w_2 - w_1\lambda) > p_3(\beta_3 - \beta_2)(w_1 + w_2 - w_2\lambda)\), \(q_{1,2} = 1\) and when the inequality is reversed, \(q_{1,2} = 0\).

As an illustration of this fact, consider the following example. Let \(\beta_1 = 1, \beta_2 = 2\) and \(\beta_3 = 3.25\). Let \(w_1 = 1\), and \(w_2 = 2\). Further, \(\lambda = 0.5\). The service requirement of all customers is still identical with an exponential distribution of unit rate. For these set of parameters, we have

\[
p_1(\beta_2 - \beta_1)(w_1 + w_2 - w_1\lambda) = p_3(\beta_3 - \beta_2)(w_1 + w_2 - w_2\lambda) = 0.8333,
\]

i.e., Eq. (4.16) is satisfied. Now it can be checked numerically that for these set of parameters, \(U(q)\) is a constant irrespective of the choice of \(q_{1,2}\) and thus the customers of type 2 can be routed in any fraction to the two different service classes without affecting the optimal cost function.

### 4.4 Admission Prices, Selfish Users and Wardrop Equilibrium

In this section we consider a non-cooperative game between the arriving customers that are now required to pay a service class dependent admission fee—a customer joining service class \(m\), pays an admission price \(c_m\). Recall that such a game theoretic model was considered earlier in Chapter 2 for service systems with parallel servers. The analysis for the present case of a single server system with multiple service classes will be similar. Further, we shall present the analysis only for the case of a continuum of customer classes. The case with discrete classes can be handled in an identical manner.
Recall our assumption that the weights for the different service classes satisfy $0 \leq w_1 < w_2 < \ldots < w_M$. We further make a reasonable assumption that $0 \leq c_1 < c_2 < \ldots < c_M$, i.e., service classes with higher weights have a higher admission price. Now the total cost to a customer of type $\beta$ that chooses service class $m$ is $c_m + \beta S_m$. An arriving customer of type $\beta$ will have routing policy $q_i(\beta)$ that minimizes its individual total cost. It is easy to see that an equilibrium routing policy will result in a Wardrop equilibrium which has the following properties—(1) service classes that receive non zero flow of class $\beta$ will all have the same total cost for $\beta$ while all the other classes will have higher or equal total cost for type $\beta$. In other words,

$$\forall m, n, \beta \quad q_m(\beta) > 0 \quad \Rightarrow \quad c_m + \beta S_m \leq c_n + \beta S_n.$$  

Recall that these Wardrop equilibrium conditions are the same as the Wardrop conditions of Eq. (2.2) of Chapter 2. Let $\tilde{q}(\beta)$ denote the Wardrop equilibrium routing function. In this section we will analyze the properties of $\tilde{q}(\beta)$ and subsequently show that $\tilde{q}(\beta)$ is also characterized by a set of thresholds that determine the equilibrium choice of customers to the service classes. We begin with the following theorem.

**Theorem 4.6.** Let $\tilde{q}(\cdot)$ be a Wardrop equilibrium routing function. There exist thresholds $t_1, t_2, \ldots, t_{M+1}$ such that $a = t_1 \leq t_2 \leq \ldots \leq t_{M+1} = b$ and $\tilde{q}_i(\beta) = 1$ for all $\beta \in (t_i, t_{i+1})$ and $\tilde{q}_i(\beta) = 0$ for all $\beta \notin [t_i, t_{i+1}]$, $i = 1$ to $M$.

**Proof.** Proof is by contradiction. Suppose the Wardrop equilibrium routing function $\tilde{q}(\cdot)$ is such that it does not satisfy the hypothesis of theorem. Then there exist customers with cost $\beta_1$ and $\beta_2$ with $\beta_1 < \beta_2$ such that $\beta_1$ chooses class $j$ and $\beta_2$ chooses class $i$ and $i < j$. The total cost for $\beta_1$ is $c_j + \beta_1 S_j$ and for $\beta_2$ it is $c_i + \beta_2 S_i$. However, at equilibrium $\beta_1$ choosing class $j$ implies

$$c_j + \beta_1 S_j \leq c_i + \beta_1 S_i,$$

or

$$\beta_1 (S_j - S_i) \leq c_i - c_j$$

or

$$\beta_1 \geq \frac{c_j - c_i}{S_i - S_j}. \quad (4.17)$$

Arguing similarly we have

$$\beta_2 \leq \frac{c_j - c_i}{S_i - S_j} \quad (4.18)$$

87
Now since $\beta_1 < \beta_2$, the Wardrop conditions of Eqs (4.17) and (4.18) cannot be simultaneously true and hence $\tilde{q}(\cdot)$ is not a Wardrop equilibrium. This contradicts our assumption that $\tilde{q}(\cdot)$ is a Wardrop policy. Hence the theorem.

Observe from the above theorem that a Wardrop equilibrium has the same structure as the socially optimal allocation. A natural mechanism design question like the one in Chapter 2 suggests itself—can we charge admission prices to the service classes in such a way that selfish users would choose an equilibrium routing policy that is the same as the optimal routing policy. This is answered in the affirmative by the following theorem.

**Theorem 4.7.** Let $q^*(\cdot)$ be the threshold-type routing policy for social optimization. There exist admission prices $c_1, \ldots, c_M$ for service classes $1, \ldots, M$ such that $\tilde{q}(\cdot) = q^*(\cdot)$.

**Proof.** First, we assume that all service classes are used, i.e., for every $m = 1, 2, \ldots, M$ there is a $\beta$ such that $q^*_m(\beta) > 0$. If not, we redefine $M$ to be largest index of service classes that are being used and ignore those that are not used.

From the structure of $q^*(\cdot)$ in Theorem 4.3, there can be at most one customer type that uses both $m$ and $m + 1$ service classes in the social optimal solution. In fact, since $\beta$ is a continuous random variable, it is easy to see that there exists exactly one customer class say $\beta$ that uses both $m$ and $m + 1$ service classes. Hence from the definition of a Wardrop equilibrium routing we have

\[
c_m + \beta S_m = c_{m+1} + \beta S_{m+1}. \tag{4.19}
\]

The quantities $S_m$ and $S_{m+1}$ are known being the sojourn times at service classes $m$ and $m + 1$. Hence, the above equation explicitly yields a relation in $c_m$ and $c_{m+1}$.

We will now verify that the admission prices obtained by the outlined procedure so indeed satisfy the conditions of Wardrop equilibrium. Towards this, suppose that for some customer type $\beta$ and some service class $m$, $q^*_m(\beta) > 0$. Then, with the prices determined as above, we need to show that

\[
c_m + \beta S_m \leq c_{\hat{m}} + \beta S_{\hat{m}}, \tag{4.20}
\]

for all $\hat{m} \neq m$, i.e., no other service class is strictly better for this customer type.

Suppose $\hat{m} > m$. From the structure of $q^*(\cdot)$ from Theorem 4.3 and from Eq. (4.19) we know that if $\hat{m} \geq m + 1$, then there exist a customer type $\hat{\beta} \geq \beta$ such that

\[
c_m + \hat{\beta} S_m = c_{m+1} + \hat{\beta} S_{m+1}
\]
which implies

\[ c_m + \beta S_m < c_{m+1} + \beta S_{m+1}. \]

Thus we have shown Eq. (4.20) for \( \hat{m} = m + 1 \). Now suppose \( \hat{m} \geq m + 2 \); there exist a class \( \bar{\beta} \geq \hat{\beta} \geq \beta \) for which \( q^*_m(\bar{\beta}) \) and from our price determination procedure

\[ c_{m+1} + \bar{\beta} S_{m+1} \leq c_{m+2} + \bar{\beta} S_{m+2} \]

which implies

\[ c_{m+1} + \bar{\beta} S_{m+1} \leq c_{m+2} + \beta S_{m+2}. \]

Continuing this argument, we can show that Eq. (4.20) holds for all \( \hat{m} > m \).

Similarly we can show that Eq. (4.20) holds for all \( \hat{m} < m \) to complete the proof. \( \square \)

**Remark 4.8.** The above theorem guarantees existence of admission prices at service classes such that \( \tilde{q}(\cdot) = q^*(\cdot) \). In fact, the proof is constructive—we determine the required admission prices satisfying the Wardrop condition. Further the construction is recursive—\( c_{m-1} \) is determined from \( c_m \). Without loss of generality, we can start with \( c_M = 0 \). Note the similarity between this theorem and Theorem 2.18 of Chapter 2 which characterizes the admission prices leading to the optimal routing kernel at equilibrium. The primary difference however is that while specific admission prices known as Pigouvian prices were used for the mechanism design in Theorem 2.18, the admission prices constructed using Theorem 4.7 can be more general. In fact, this constructive method can also be used to determine the admission prices for the model with parallel servers. It is easy to see in that case, the Pigouvian prices will be one of the several admission prices that can be constructed using this proof technique.

### 4.5 Discussion

Several additional questions present themselves for possible future work. First, given \( M, \lambda, \) and \( f(\beta) \), what is the optimum set of weights that minimizes the social cost. A second issue relates to the setting of prices. If we fix \( M, \lambda, f(\beta) \), what \( w \) would maximize the revenue to the server. In the latter we are assuming that the admission price is the weight for the service class. For both social optimality and revenue maximization, another obvious question would be \( M \). While it appears that larger \( M \) would yield lower cost or higher revenue, when the weights are suitably set, a proof is elusive. We believe that the marginal gains would start to decrease rapidly with increasing \( M \).
Chapter 5

A Choice of Highest-Bidder-First and FIFO Service

5.1 Introduction

Consider a service system with a single server and its own queue into which customers from a heterogeneous population arrive, pay an admission price, wait in a queue if the server is busy and depart after receiving service. As in the preceding chapters, the heterogeneity of customers is captured by different costs per unit delay. In such a scenario the server can increase its revenue by providing differential service with different prices for different grades of service. We have seen one such scheduling discipline in the previous chapter, namely the DPS discipline that can provide differential grades of service. The DPS scheduling discipline is characterized by a finite number of service classes and due to its processor sharing nature, all customers present in the queue were served simultaneously. Further, a customer’s share of the service rate depends not only on the type of the service class that it belongs to, but also on the number of customers in the different service classes. From the point of view of analysis, DPS however has its disadvantages. The expressions for the mean waiting time of a customer in a service class is known only for the case of two service classes. Further, a finite number of service classes offer a limited (finite) amount of differentiation to the quality of service offered.

An alternative approach to service quality differentiation is to let arriving customers purchase their priority and the server use priority scheduling with higher prices being accorded higher priorities. The arriving customers are essentially placing a bid for the priority level that they want. Such a model has applications to several systems where bids can be placed and
a scheduling mechanism based on the bid values can be used, e.g., Abhishek et al. [1] and Altmann et al. [8] suggest such a model for the spot market in cloud computing services. In this chapter, we will consider such a decentralized approach for service quality differentiation where arriving customers have to place a bid and in return obtain a service quality that is in some sense proportional to its bid value.

Such bidding for priorities has also been called bribing in some of the early literature. Queueing systems in which customers can purchase priorities is well studied in the literature [56, 29, 45, 2, 11, 40]. The bribing model for priority in queues, introduced by Kleinrock [40] is as follows. Customers arrive according to a Poisson process of rate $\lambda$ and each arrival, without observing the queue occupancy, offers a bid for service that is independent of all other offers. Service discipline is highest-bidder-first (HBF) preemptive (or non preemptive) priority. For this system, the expected delay as a function of the offered price was derived by Kleinrock [40]. An immediate extension is when the bid value of each customer depends on its delay cost, i.e., there is a bidding policy mapping delay cost to a bid value. In this scenario, since customers have different valuations for their delays, a natural model would be to have selfish customers that will choose their bids to minimize their total cost, i.e., the sum of delay cost and price paid for priority. This leads to a game theoretic situation in which equilibrium (or stable) bidding policies, policies from which a customer has no incentive to deviate, are of interest.

In a single server system, Kleinrock [40] shows that in an equilibrium bidding policy the bids are an increasing function of delay cost; closed form expressions for equilibrium bidding policies are obtained by Lui [45] and Glazer and Hassin [29]. Providing prioritized service requires additional resources and, especially in cloud computing like systems, this may be at the cost of doing ‘useful work.’ Further, it is not always feasible to expect that customers will bid; they may prefer obtaining a FIFO service for free or at a fixed price. This motivates the first system that we consider—a two server system in which one server uses FIFO scheduling and arrivals to the second server have to bid for priorities. An arriving customer selfishly makes two decisions—the server to use and, if it chooses the HBF server, the value of its bid. For this system, we first obtain equilibrium strategies (routing and bidding) and show that the equilibrium routing policy is of threshold type; customers with delay cost above a threshold choose the prioritizing server while those below the threshold choose to wait for their turn in the FIFO queue and obtain a free service.

In the above system, an obvious interest is in the impact of having a parallel FIFO server on
the total revenue. Numerical results show that if the total capacity is fixed and is split between the FIFO and HBF servers, then having a FIFO server increases revenue substantially; a formal proof has been elusive. On the other hand, if additional capacity is allocated to a FIFO server, then we formally show that the revenue will be less than that with a single HBF server for any type profile.

Finally, we briefly consider the case when arriving customers can balk, i.e., not join the queue. This can happen when the reward for receiving service is lower than the cost of obtaining it (sum of delay cost and the bid). Such a system with just the HBF server is analyzed by Lui [45] and Glazer and Hassin [29] where it was shown that if customers bid strategically, those with delay costs above a threshold will balk. Our interest is to try to retain these balking customers and possibly increase the revenue by adding a free server; such an option is motivated by the argument that customers with lower delay costs will prefer the FIFO server which decreases congestion at the HBF server. This in turn incentivises some high delay cost customers to not balk and instead join for service at the HBF server. We present a preliminary analysis of such a balking system and compare its revenue to a system with only the HBF server.

The work of Hassin [31] is closely related to the work of this chapter. In Hassin [31], bidding is a mechanism to self-regulate the arrival process into a single-server HBF queue and the focus is on homogeneous customers. The social welfare has customers balking if the value of service is lower than the total cost, sum of bid and waiting costs. It is shown when customers balk, at equilibrium, the social welfare objective coincides with the server profit maximizing objective. Further, when the service times are assumed exponential, it is shown that the service rate that maximizes the server revenue is lower than the socially optimal service rate. These were shown to also be true when there are a discrete set of customer types. There are two main differences between the model of Hassin [31] and that of this chapter. First, we focus much of our attention on the case when there is no reward for service and hence the customers cannot balk; the customer objective is to minimize its total cost. Second, we always assume heterogeneous customers with a continuum of classes for this chapter.

There are also similarities between the model that we consider and that of Abhishek et al. [1]. The latter has an infinite server queue and a $K$-server queue in parallel. The $K$-server queue is a preemptive HBF server like the one that we have described and is the spot market. Since the infinite server queue has no waiting time, the charge for service from that queue can be high and hence attract customers with high waiting time costs. Further, the customers have
a value for the service received and can balk (not join the queue) if this value is lower than the cost (sum of delay cost and charge for service). Thus in the revenue maximizing regime of Abhishek et al. [1] the arrival rate to the HBF queue is zero. In the model considered in this chapter, customers with low delay costs use the free FIFO service and the revenue maximizing regime shares the total capacity between the two queues suitably.

Finally, we remark that the highest-bid-first discipline allows for a continuum of prices; alternatively, one can allow a fixed number of priorities and fix the price of each of the priorities. Such systems are widely considered in the literature [2, 56, 4]. Systems with multiple FIFO queues, each with its own server were considered in [16, 10]. We will not be interested in such systems in this chapter.

The rest of the chapter is organized as follows. In Section 5.2 we describe the notation, recap some results from literature and characterize properties of the revenue for a single HBF server. In Section 5.3 we first characterize the equilibrium routing and bidding policy when a FIFO and a HBF server are in parallel. Numerical results for the case when the total capacity is shared between the FIFO and the HBF server are presented. We then analyze a system in which the FIFO server is added to the HBF server. In Section 5.4 we analyze the system in which customers balk and in Section 5.5 we analyze the case when the FIFO server charges and admission price. We conclude with a discussion in Section 5.6.

5.2 Notation and Preliminaries

In this section we set up the notation and recap some results from literature for the single server system using the highest-bidder-first (HBF) priority discipline. For such a system we also characterize the revenue rate.

Customers arrive to a service system according to a homogeneous Poisson process of rate \( \lambda \). Service times of customers are i.i.d. random variables with distribution \( G(\cdot) \) and unit mean. Associated with each arriving customer is a random variable \( \beta \), \( 0 \leq a \leq \beta \leq b < \infty \), representing its cost per unit delay. \( \beta \) are i.i.d. with distribution \( F(\beta) \) that is absolutely continuous in \((a, b)\). \( \beta \) is also called type of the customer and \( F(\beta) \) is called the type profile. As in the previous chapter, \( \beta \) denotes the realization of the random variable \( \beta \). A single server serves with rate \( \mu \) and utilization \( \rho := \frac{\lambda}{\mu} \) using HBF non-preemptive priority discipline. The type of a customer is private information but \( G(\cdot), F(\beta), \lambda \) and the service rates of the servers
are assumed to be common knowledge. Further, the customer does not know its service time. Each customer is assumed to be infinitesimally small and does not affect the system dynamics on its own. Customers do not balk, i.e., all arriving customers receive service. Further, once a customer has made the choice of the queue and the bid if joining the HBF queue, then it cannot change either of these; it also does not renege and it leaves the system only after receiving the service.

Like in the literature \cite{29, 45, 11, 40} we assume oblivious bidding, i.e., an arriving customer bids for its priority without observing the queue occupancy; service is non preemptive with higher priorities for higher bids. Thus on a service completion, the next customer is the one in the queue with the highest bid. Preemptive service can be analyzed identically to the non preemptive case and we do not discuss it further.

\( X(\beta) \) is the bidding policy, i.e., customers of type \( \beta \) bid \( X(\beta) \). \( W(x) \) is the expected waiting time (time in queue excluding service time) of a customer that bids \( x \). The expected total cost of receiving service for a customer of type \( \beta \) is

\[
C(\beta) := X(\beta) + \beta \left( W(X(\beta)) + \frac{1}{\mu} \right).
\]

Customers behave strategically and choose their bids to minimize their individual costs. A bidding policy is an *equilibrium policy* if no individual customer can unilaterally deviate from it and lower its total cost \( C(\beta) \).

Let \( B^X(x) \) denote the distribution of the bids under bidding policy \( X \), \( W^X(x) \) the expected waiting time (time in queue excluding service time) of a customer that bids \( x \), and \( C^X(\beta) \) the total cost to customer of type \( \nu \) under policy \( X \). Let \( W_0 \) be the expected waiting time added to that of an arriving customer due to residual service time of a customer in service. \( W_0 \) is the product of the residual service time and probability that an arriving customer sees a busy server, i.e.,

\[
W_0 = \frac{\lambda}{2} \int_0^\infty \tau^2 dG(\mu\tau).
\]

Following \cite{29}, we obtain

\[
X^E(\beta) = \int_0^\beta \frac{2\rho W_0 y}{(1 - \rho + \rho F(y))^3} dF(y)
\]

(5.1)
as an equilibrium bidding policy. Property 5.1 below summarizes the properties of \( X^E(\beta) \) that have been derived in the literature \cite{40, 45, 29}. Note that the \( X(\beta) \) in the following is actually \( X^E(\beta) \) but we drop the superscript \( E \) to simplify notation. Further, \( B(\cdot), W(\cdot), \) and \( C(\cdot) \) depend on \( X(\beta) \). Once again, we do not explicitly capture this dependence in the notation.
Property 5.1. 1. $X(\beta)$ is continuous and strictly increasing in $\beta$. Further $B(X(\beta)) = F(\beta)$.

2. $W(\beta)$ is strictly decreasing in $\beta$. Further, $W(\beta) = \frac{\mu^2 W_0}{(\mu - \lambda (1 - F(\beta)))^2}$ where $W_0$ is as above.

3. $C(\beta) := \min_x \left\{ x + \beta \left( W^X(x) + \frac{1}{\mu} \right) \right\}$ is continuous and strictly increasing concave.

4. $\frac{dC(\beta)}{d\beta} = W(x) + \frac{1}{\mu}$ and $\frac{dX(\beta)}{d\beta} + \beta \frac{dW(\beta)}{d\beta} = 0$

Let $R^X(\lambda, F(\beta))$ be the revenue rate for a system with arrival rate $\lambda$, type profile $F(\beta)$, and an equilibrium bidding policy $X(\beta)$. Also, wherever applicable, to indicate the dependence on the service rate of the server, we also use the alternative notation of $R^X(\lambda, F(\beta), \mu)$ to denote the revenue rate when the server serves with rate $\mu$. Clearly,

$$R(\lambda, F(\beta)) = \lambda \int F(\beta) dF(\beta).$$

The following lemma characterizes the equilibrium revenue function where we obtain the direction of change of the revenue when one of arrival rate, service rate and type profile is changed while keeping the other parameters unchanged.

Lemma 5.2. 1. If $\lambda_1 < \lambda_2$, then

$$R(\lambda_1, F(\beta)) < R(\lambda_2, F(\beta)).$$

2. Let $\mu_1$ and $\mu_2$ be two service rates with $\mu_1 < \mu_2$. Then

$$R(\lambda, F(\beta), \mu_1) > R(\lambda, F(\beta), \mu_2).$$

3. Let $F_1(\beta)$ be a type profile and let $F_2(\beta) = F_1(\beta - \beta_0)$ with $\beta_0 > 0$. For this case

$$R(\lambda, F_1(\beta)) < R(\lambda, F_2(\beta)).$$

4. Let $F_2(\beta) = F_1(\beta / c)$ with $c > 1$. Then

$$R(\lambda, F_1(\beta)) < R(\lambda, F_2(\beta)).$$

5. Let $F_1(\beta)$ be a type profile and define

$$F_2(\beta) := \begin{cases} 
0 & \beta < a_1 \\
\frac{f_0^a dF_1(w)}{f_{a_1}^a dF_1(w)} & \beta \geq a_1
\end{cases}$$

Then $R(\lambda, F_1(\beta)) < R(\lambda, F_2(\beta))$.  

96
We provide informal arguments and omit formal proofs as they are quite straightforward from the analysis in Kleinrock [40], Lui [45] and Glazer and Hassin [29]. The first and second statements are complimentary—increasing the arrival rate or decreasing the service rate increases the revenue. This happens because increasing $\lambda$ (or decreasing $\mu$) introduces more congestion and more competition. Thus all customers have to bid higher to ‘stay in place.’ This is also evident in (5.1) where the denominator of the integrand increases with increasing $\lambda$ and decreasing $\mu$. For the case of increased $\lambda$ there are also more arrivals. The third statement says that if the type profile is ‘shifted to the right,’ i.e., the delay cost of every customer is increased by a fixed quantity $\beta_0$, then the revenue rate is increased. This is because, every one has an increased delay cost and hence every one has to bid higher than before to, once again, stay in place. This is also seen in (5.1) where there is a $y$ in the integrand and the integration range is over larger values of $y$. The fourth statement says that ‘stretching’ the type profile increases the revenue rate. Once again, every one bids higher than before and the reasoning is similar to the preceding case. The last statement concerns the case when arrival rate is kept constant but the types below a cutoff are removed from the population. This also causes every one to bid higher because they are all competing with more customers of the same type.

5.3 A System with a FIFO and a HBF Server

We now analyze a service system with two servers. One server uses the non-preemptive HBF discipline and serves at rate $\mu_1$. The second server uses the FIFO discipline and serves at rate $\mu_2$. Customers choosing the HBF server will have to bid at least $M$, $M \geq 0$, while the FIFO service is free. Customers arrive according to a homogeneous Poisson process of rate $\lambda$ and the type profile is $F(\beta)$. In this section we will assume that all arrivals will have to receive service from one of the two servers and they cannot balk. Thus an arriving customer now has to make the following decisions on arrival. Which server to use, and, if it chooses the HBF server, then the value of its bid. We will assume oblivious decisions, i.e., the arrivals make the choices without observing the queue occupancy. We now characterize the structure of these choices.

Consider the choices made by a customer of type $\beta$. First, let $p(\beta) : [a, b] \rightarrow [0, 1]$ be the routing policy, i.e., the probability that it chooses the FIFO server. Next, let $M + X(\beta)$ be the bid if it chooses the HBF server. With this, the expected total cost of receiving service for a
customer of type $\beta$ is

$$C(\beta) := X(\beta) + M + \beta \left( W(X(\beta)) + \frac{1}{\mu_1} \right).$$

The pair $(p(\beta), X(\beta))$ is the strategy of a customer of type $\beta$ and it is an *equilibrium* strategy if no customer can unilaterally deviate and lower its $C(.)$.

Given $(p(\beta), X(\beta))$, we see that $\lambda_2 := \lambda \int_0^\infty p(\beta) dF(\beta)$ is the arrival rate to the FIFO server and $\lambda_1 = \lambda - \lambda_2$ is the arrival rate to the HBF server. Define $\rho_i := \lambda_i / \mu_i$. All customers that choose the HBF server experience a bid-dependent waiting time, denoted by $W_1(\beta)$ for a customer of type $\beta$, and all those that choose the FIFO server experience the same expected waiting time, denoted by $W_2$. Let $D_1(\beta) := W_1(\beta) + \frac{1}{\mu_1}$ and $D_2 := W_2 + \frac{1}{\mu_2}$ be the expected sojourn times in, respectively, the HBF and the FIFO servers. Fig. 5.1 illustrates this notation.

Let us now consider the equilibrium strategy for this system. Since this is a non atomic system and all customers choose individually optimal strategies, the equilibrium is a Wardrop equilibrium, first described in [62] and used extensively in transportation systems. This means that at equilibrium the following is true for all $\beta$.

$$\text{If } p(\beta) > 0, \text{ then } \beta D_2 \leq M + X(\beta) + \beta D_1(X(\beta)).$$

(5.2)

The following theorem characterizes the equilibrium strategy for this system.
Theorem 5.3. Using $\beta_1$ determined below, define $p^E(\beta)$, $F_1(\beta)$, and $W_0$ as follows.

\[
p^E(\beta) = \begin{cases} 
0 & \text{for } \beta > \beta_1, \\
t & \text{for } \beta = \beta_1, \\
1 & \text{for } \beta < \beta_1.
\end{cases}
\]

(5.3)

\[
F_1(\beta) := \begin{cases} 
\int_{\beta}^{b} \frac{dF(x)}{F_1(b)} & \beta_1 \leq \beta \leq b, \\
1 & \beta > b,
\end{cases}
\]

(5.4)

\[
W_0 = \frac{\lambda_1}{2} \int_{0}^{\infty} \tau^2 dG(\mu_1 \tau),
\]

(5.5)

\[
X^E(\beta) = \int_{0}^{\beta} \frac{2\rho_1 W_0 y}{(1 - \rho_1 \rho_1 F_1(y))^3} dF_1(y)
\]

(5.6)

For the routing and bidding policy $(p^E(\beta), X^E(\beta))$ determined as above, let $D_1(\beta)$ be the bid-dependent expected sojourn time in the HBF server and $D_2(\lambda_2)$ be the expected sojourn time in the FIFO server when the arrival rate to it is $\lambda_2$.

$\beta_1$ is determined as follows.

- If using $\beta_1 = a$ in (5.3)–(5.6) satisfies $M + aD_1(a) < aD_2$, then set $\beta_1 = a$.

- Else if using $\beta_1 = b$ in (5.3)–(5.6) satisfies $M + bD_1(b) > bD_2$ then set $\beta_1 = b$.

- Else find $\beta_1$ which when used in (5.3)–(5.6) satisfies

\[
M + \beta_1 D_1(\beta_1) = \beta_1 D_2.
\]

(5.7)

$(p^E(\beta), X^E(\beta))$ is an equilibrium strategy with $\beta_1$ defined as above. Further, $\beta_1$ is unique.

Proof. We will prove that the Wardrop conditions of (5.2) are satisfied for the choice of $p^E(\beta)$ and $X^E(\beta)$ as described in (5.3)–(5.6).

With $p^E(\beta)$ as in (5.3), the arrival rate to the HBF server is $\lambda_1 = \lambda \int_{\beta_1}^{b} dF(\tau)$. The type profile of the customers choosing the HBF server will be as in (5.4). From the previous section (and from [40, 45, 29]) we know that those that decide to join the HBF server will have to use the bidding policy as in (5.6) for individual optimization.

With bidding policy as in (5.6) for those that choose the HBF server, we will now verify that $p^E(\beta)$ of (5.3) satisfies the Wardrop condition of (5.2). First consider the case when $\beta_1 \neq \ldots$
a, b. In this case, for any $\beta < \beta_1$ we see that

\[
M = \beta_1(D_2 - D_1(\beta_1)) > \beta(D_2 - D_1(\beta_1)) = \beta(D_2 - D_1(\beta))
\]

where the last equality follows from the absolute continuity of $F$ and the fact that while customer with $\beta_1$ occupies the end of the queue in the routing function $p^E(\beta)$, so would the deviating $\beta$ customer. This verifies the Wardrop condition of (5.2) for $\beta < \beta_1$ i.e., $p^E(\beta) = 1$ is the optimum choice for $\beta < \beta_1$.

For a customer of type $\beta_1 + \epsilon$ for some $\epsilon > 0$,

\[
C(\beta_1 + \epsilon) \leq C(\beta_1) + \epsilon \frac{dC(\beta_1)}{d\beta} = C(\beta_1) + \epsilon D_1(\beta_1)
\]

The inequality is from the concavity of $C(\beta)$ (item (3) of Property 5.1) and the equality is from item (4) in Property 5.1. For $\beta = \beta_1$, $X^E(\beta_1) = 0$ and the preceding inequality leads to

\[
C(\beta_1 + \epsilon) \leq M + (\beta_1 + \epsilon)D_1(\beta_1) = \beta_1 D_2 + \epsilon D_1(\beta_1) < \beta_1 D_2 + \epsilon D_2.
\]

Both the equality and inequality above follow from (5.7). After rearranging the terms we have

\[
M + (\beta_1 + \epsilon)D_1(\beta_1) < (\beta_1 + \epsilon)D_2
\]

which is substituted in (5.8) to get

\[
C(\beta_1 + \epsilon) < (\beta_1 + \epsilon)D_2.
\]

Hence a customer of type $\beta_1 + \epsilon$ will have a lower cost with the HBF server than with the FIFO server. This verifies that $p(\beta_1 + \epsilon) = 0$ is optimum for $\beta = \beta_1 + \epsilon$. Since $\epsilon$ is arbitrary, $p^E(\beta) = 0$ for all $\beta > \beta_1$ is also verified.

For the cases when $\beta_1 = a$ (resp. $\beta_1 = b$) we can use similar arguments to show that $p^E(\beta) = 0$ (resp. $p^E(\beta) = 1$) is an equilibrium policy.

We will now show that $\beta_1$ satisfying (5.7) is unique. We need only consider the case when $\beta_1 \neq a, b$. Observe that $D_1(\beta_1)$ is a strictly decreasing function of $\beta_1$ while $D_2$ is strictly increasing in $\beta_1$. Further as $\beta_1 \neq a, b$ the following complementary conditions must be true.

\[
D_1(a) + \frac{M}{a} \geq D_2
\]

\[
D_1(b) + \frac{M}{b} \leq D_2
\]
Hence there exists a unique $\beta_1$. \hfill \Box

**Corollary 5.4.** In the class of routing policies $p(\beta)$ that results in $F_1(\beta)$ (the profile of the customers choosing the HBF server) being absolutely continuous, a Wardrop equilibrium routing policy is of the threshold type.

**Proof.** From Theorem 5.3, $C(\beta_1) \leq \beta_1 D_2$ implies $C(\beta_1 + \epsilon) < (\beta_1 + \epsilon) D_2$ i.e., if a customer of type $\beta$ joins the HBF server, then for a $\epsilon > 0$ a customer of type $\beta + \epsilon$ also joins this queue. \hfill \Box

Theorem 5.3 shows that the arrival rate to the HBF server can be less than $\lambda$ and that the type profile of these customers is truncated on the left. From Lemma 5.2 the former reduces the revenue while the latter increases it. This prompts us to investigate the effect of the parallel FIFO server on total revenue. Two scenarios present themselves immediately. First, we can let the total service rate be fixed (say unity) and share it between the HBF and the FIFO servers, i.e., $\mu_1 + \mu_2 = 1$. The second scenario of interest is when $\mu_1 = 1$ and we add additional service capacity in the form of a FIFO queue of service rate $\mu_2$. In both cases, we investigate the revenue as a function of $\mu_2$.

### 5.3.1 HBF and FIFO Servers Sharing Capacity

We present a sample of the numerical results that we have obtained. We assume $M = 0$ and that the total service capacity of unity is shared between the HBF and FIFO servers, i.e., $\mu_1 + \mu_2 = 1$. Service time distribution is exponential with unit mean and HBF server is non-preemptive. From (5.5), $W_0 = \frac{\lambda}{\mu_1}$. We consider the following three different $F(\beta)$, all of which correspond to $a = 0$ and $b = 100$.

1. $F(\beta) = \frac{\beta}{100}$ for $0 \leq \beta \leq 100$.
2. $F(\beta) = \frac{\beta^{0.5}}{10}$ for $0 \leq \beta \leq 100$
3. $F(\beta) = \frac{\beta^2}{10^5}$ for $0 \leq \beta \leq 100$

For these three examples, in Fig. 5.2 we plot the revenue as a function of $\mu_1$ for $\lambda = 0.3, 0.5, 0.7$. The graphs are rather self explanatory and we will not dwell on the details, except pointing out the two observations the stand out from these examples.

1. The revenue can actually increase if some of the capacity is allocated to a FIFO server.

This in itself is not very surprising because there is no balking in the system and all
Figure 5.2: Revenue as a function of $\mu_1$ for three different example $F(\beta)$ and different $\lambda$. 

$$F(v) = \frac{v^{0.3}}{10}$$

$$F(v) = \frac{v}{100}$$

$$F(v) = \frac{v^2}{10^3}$$
arriving customers have to take the service. That the magnitude of the increase was substantial seems surprising.

2. The HBF server needs a minimum $\mu_1$ before it can generate any revenue. Once again, it appears that this minimum is substantial.

5.3.2 Adding a FIFO Server to an HBF Server

We now compare the revenue when a FIFO server is added to the HBF server. Let $S_1$ be a system with a single HBF server of rate $\mu_1$ and $S_2$ be a system with a HBF server of rate $\mu_1$ and a FIFO server of rate $\mu_2$. Customers arrive according to a Poisson process into the system, they have a type profile $F(\beta)$ and make selfish choices of the server and the bid. In $S_1$, customers follow the strategy described in [29, 45] and in $S_2$ they follow that outlined in Theorem 5.3. We need additional notation.

In all quantities of interest a superscript $S_i$ will indicate the system that is being referred. For example, $R^{S_i}(\cdot, \cdot)$, will be the revenue in system $S_i$. $F^{S_1}(\beta)$ and $F^{S_2}(\beta) = F_1(\beta)$ will be the type profile of customers choosing the HBF server in systems $S_1$ and $S_2$ respectively. $\lambda_i$ will denote the arrival rate to the HBF server in $S_i$. Clearly, $\lambda_1 = \lambda$ and $\lambda_2 = \lambda(1 - F(\beta_1))$ where $\beta_1$ is as obtained from Theorem 5.3. $D^{S_2}(\beta)$, $C^{S_2}(\beta)$ will denote the sojourn time and the total cost respectively for a customer joining the HBF server of $S_2$.

We first show that the mean waiting time for customers of type $\beta > \beta_1$ is lower in $S_2$ than in $S_1$ and then characterize the revenue.

**Lemma 5.5.** For a customer of type $\beta > \beta_1$, we have $W^{S_1}(\beta) > W^{S_2}(\beta)$. Further

$$\frac{dW^{S_1}(\beta)}{d\beta} < \frac{dW^{S_2}(\beta)}{d\beta}.$$  

**Proof.** From item 1 of Property 5.1 the bids are monotonic in $\beta$. Now consider a customer of type $\beta \geq \beta_1$. In both $S_1$ and $S_2$, the arrival rate of customers of priority $\beta > \beta$ to the HBF queue is the same. Thus the denominator in the expression for $W(\beta)$ (see item (2) in Property 5.1) is the same for both $S_1$ and $S_2$ but $W^{S_1}_0 > W^{S_2}_0$. Hence $W^{S_1}(\beta) > W^{S_2}(\beta)$.

Now from the expression for $W(\beta)$ in item (2) in Property 5.1 we have for $i = 1, 2$

$$\frac{dW^{S_i}(\beta)}{d\beta} = \frac{-2\mu_1^2 \lambda_i W^{S_i}_0 dF^{S_i}(\beta)}{(\mu_1 - \lambda_i(1 - F^{S_i}(\beta)))^3}.$$
As earlier, the arrival rate of customers of priority $\beta > \beta_1$ to the HBF queue is the same i.e.,

$$\lambda_1(1 - F(\beta)) = \lambda_2(1 - F_1(\beta)).$$

Since $W_{01}^S > W_{02}^S$, and the remaining terms are the same for both $S_1$ and $S_2$ the lemma is true. □

This sets us up for the main result of this subsection.

**Theorem 5.6.** Adding a FIFO server of service rate $\mu_2$ does not increase the revenue i.e.,

$$R^{S_1}(\lambda_1, F(\beta)) \geq R^{S_2}(\lambda_2, F_1(\beta)).$$

**Proof.** In the first part of the proof, we will prove by contradiction that for all $\beta > \beta_1$, $X^{S_1}(\beta) \geq X^{S_2}(\beta)$. Suppose the claim is not true and there exists a $\beta_2 > \beta_1$ such that

$$X^{S_1}(\beta_2) < X^{S_2}(\beta_2).$$

(5.10)

Since the bidding policy $X^{S_i}(\beta)$ is increasing in $\beta$,

$$X^{S_1}(\beta_1) \geq X^{S_2}(\beta_1) = 0.$$  

(5.11)

Therefore from (5.10) and (5.11), there exists a $\beta^*$ such that

$$X^{S_1}(\beta^*) = X^{S_2}(\beta^*)$$

(5.12)

where for all $\beta < \beta^*$,

$$X^{S_1}(\beta) > X^{S_2}(\beta).$$

Recall that for all $\beta > \beta_1$, from Lemma 5.5 we have

$$\frac{dW^{S_1}(\beta)}{d\beta} < \frac{dW^{S_2}(\beta)}{d\beta}.$$  

Therefore, for all $\beta > \beta_1$, from item (4) in Property 5.1 we get

$$\frac{dX^{S_1}(\beta)}{d\beta} > \frac{dX^{S_2}(\beta)}{d\beta}.$$  

(5.13)

From (5.12) and (5.13), for all $\beta \geq \beta^*$ we have

$$X^{S_1}(\beta) \geq X^{S_2}(\beta).$$

(5.14)

But this contradicts the assumption of (5.10) on $\beta_2$ as $\beta_2 > \beta^*$. This implies that for all $\beta$,

$$X^{S_1}(\beta) \geq X^{S_2}(\beta).$$

(5.15)
The revenue in $S_2$ is given by

$$R^{S_2}(\lambda_2, F_1(\beta)) = \lambda_2 \int_{\beta_1}^{b} (M + X^{S_2}(\beta)) \, dF_1(\beta)$$

$$= \frac{\lambda_2}{1 - F(\beta_1)} \int_{\beta_1}^{b} (M + X^{S_2}(\beta)) \, dF(\beta)$$

$$= \lambda \int_{\beta_1}^{b} (M + X^{S_2}(\beta)) \, dF(\beta)$$

$$\leq \lambda \int_{\beta_1}^{b} (M + X^{S_1}(\beta)) \, dF(\beta)$$

$$\leq \lambda \int_{a}^{b} (M + X^{S_1}(\beta)) \, dF(\beta)$$

$$= R^{S_1}(\lambda_1, F(\beta)).$$

The second equality is from the definition of $F_1(\beta)$ in (5.4), the third equality is from the definition of $\lambda_2$ and the first inequality is from (5.15). This completes the proof.  

5.4 When Arrivals Balk

In this section we assume that each customer receives a fixed reward for obtaining the service from the system. In this case, a type $\beta$ customer whose total cost of receiving service, $C(\beta)$, exceeds $P$ has no incentive to join the system and will balk. In such a system the $\beta^*$ satisfying $P = X(\beta^*) + \beta^*W(\beta^*) + M$ is the highest type of customer receiving service; all customers with $\beta > \beta^*$ will balk.

As in the previous section, we will compare two systems $S_1$ and $S_2$ and to simplify the analysis, we will assume $F(\beta) = \beta/b$ and $a = 0$. Let $\beta_{u,1}$ and $\beta_{u,2}$ be the highest type customer joining the HBF server of, respectively, $S_1$ and $S_2$. Let $F_1(\beta)$ and $F_2(\beta)$ denote the type profile and $\lambda_1$ and $\lambda_2$ denote the arrival rate to the HBF server in $S_1$ and $S_2$ respectively. Since the arrivals balk, we have

$$F_1(\beta) = \frac{F(\beta) - F(a)}{F(\beta_{u,1}) - F(a)}$$

$$F_2(\beta) = \frac{F(\beta) - F(\beta_1)}{F(\beta_{u,2}) - F(\beta_1)}$$

$$\lambda_1 = \lambda(F(\beta_{u,1}) - F(a))$$

$$\lambda_2 = \lambda(F(\beta_{u,2}) - F(\beta_1)).$$
We first state the following lemma that compares $\beta_{u,1}$ and $\beta_{u,2}$.

**Lemma 5.7.** $\beta_{u,1} \leq \beta_{u,2}$.

**Proof.** The proof is by contradiction. Suppose $\beta_{u,1} > \beta_{u,2}$. Recall the balking condition

$$P = X(\beta^*) + \frac{\beta^*}{\mu_1} + M.$$

Since $P$, $M$ and $\mu$ are constant and $\beta_{u,1} > \beta_{u,2}$ we have

$$X^{S_1}(\beta_{u,1}) < X^{S_2}(\beta_{u,2}). \quad (5.16)$$

At $\beta = \beta_1$,

$$X^{S_1}(\beta_1) > X^{S_2}(\beta_1) = 0. \quad (5.17)$$

Therefore from (5.16) and (5.17), there should exist a $\beta_2 \leq \beta_{u,2}$ which satisfies

$$X^{S_1}(\beta_2) = X^{S_2}(\beta_2). \quad (5.18)$$

From item (4) of Property 5.1, we have for $i = 1, 2$

$$\frac{dW^{S_i}(\beta)}{d\beta} = \frac{-2\mu_1^2 \lambda_i W_0^{S_i} dF_i(\beta)}{(\mu_1 - \lambda_i(1 - F_i(\beta)))^3}. \quad (5.19)$$

From the definition of $F_1, F_2, \lambda_1$ and $\lambda_2$ we have

$$\frac{dW^{S_1}(\beta)}{d\beta} = \frac{-2\mu_1^2 \lambda W_0^{S_1} dF(\beta)}{(\mu - \lambda(F(\beta_{u,1}) - F(\beta)))^3} \quad (5.19)$$

$$\frac{dW^{S_2}(\beta)}{d\beta} = \frac{-2\mu_1^2 \lambda W_0^{S_2} dF(\beta)}{(\mu - \lambda(F(\beta_{u,2}) - F(\beta)))^3}. \quad (5.20)$$

Since $\beta_{u,1} > \beta_{u,2}$ and $a < \beta_1$, we have $\lambda_1 > \lambda_2$ and from the definition of $W_0$ we have

$$W_0^{S_1} > W_0^{S_2}. \quad (5.21)$$

Also $\beta_{u,1} > \beta_{u,2}$ implies

$$F(\beta_{u,1}) - F(\beta) > F(\beta_{u,2}) - F(\beta). \quad (5.22)$$

Therefore from (5.19), (5.20), (5.21) and (5.22),

$$\frac{dW^{S_1}(\beta)}{d\beta} < \frac{dW^{S_2}(\beta)}{d\beta}.$$

Substituting the above in item (4) of Property 5.1 we have

$$\frac{d(X^{S_1}(\beta))}{d\beta} \geq \frac{d(X^{S_2}(\beta))}{d\beta}. \quad (5.23)$$
From (5.18) and (5.23), we have, for all $\beta > \beta_2$

$$X^{S_1}(\beta) \geq X^{S_2}(\beta).$$

Therefore

$$X^{S_1}(\beta_{u,2}) \geq X^{S_2}(\beta_{u,2}).$$

Now since $X(\beta)$ is increasing, for $\beta_{u,1} > \beta_{u,2}$,

$$X^{S_1}(\beta_{u,1}) \geq X^{S_1}(\beta_{u,2}).$$

Hence

$$X^{S_1}(\beta_{u,1}) \geq X^{S_2}(\beta_{u,2}).$$

However this contradicts (5.16).

Hence $\beta_{u,1} \leq \beta_{u,2}$.

Note that this lemma is true for arbitrary choice of $F(\beta)$ since the proof technique does not make use of $F(\beta) = \beta/b$. Having characterized the threshold $\beta^*$ for the two system, we will now analyze their revenue for the case when $F(\beta) = \beta/b$.

First consider the revenue for $S_1$.

$$R^{S_1}(\lambda_1, F_1(\beta)) = \lambda_1 \int_0^{\beta_{u,1}} (M + X^{S_1}(\beta)) dF_1(\beta)$$

$$= \lambda_1 \int_0^{\beta_{u,1}} \frac{(M + X^{S_1}(\beta))}{F(\beta_{u,1}) - F(a)} dF(\beta)$$

$$= \frac{\lambda}{b} \int_0^{\beta_{u,1}} (M + X^{S_1}(\beta)) d\beta.$$

where the first and the second equality follow from the definition of $F_1(\beta)$ and $\lambda_1$ respectively.

Similarly the revenue for $S_2$ is given by

$$R^{S_2}(\lambda_2, F_2(\beta)) = \frac{\lambda}{b} \int_{\beta_1}^{\beta_{u,2}} (M + X^{S_2}(\beta)) d\beta.$$

The preceding arguments prove the following theorem that compares the revenue from the two systems.

**Theorem 5.8.** If

$$\int_0^{\beta_{u,1}} (M + X^{S_1}(\beta)) d\beta > \int_{\beta_1}^{\beta_{u,2}} (M + X^{S_2}(\beta)) d\beta,$$

then

$$R^{S_1}(\lambda_1, F_1(\beta)) > R^{S_2}(\lambda_2, F_2(\beta)).$$

otherwise

$$R^{S_1}(\lambda_1, F_1(\beta)) \leq R^{S_2}(\lambda_2, F_2(\beta)).$$


5.5 When FIFO server is not free

In the preceding analysis, we have always assumed that the FIFO server is a free server and customers joining this queue do not pay any admission price. From the point of view of maximizing its revenue, operating a free server may not be revenue optimal. It is therefore of practical interest to consider the system with an admission price at the FIFO server. In this section, we relax this assumption of free FIFO service and characterize the equilibrium routing and bidding function when this server charges an admission price henceforth denoted by \( c \). To simplify the analysis and restrict the number of cases arising due to the heterogeneity of servers, we shall assume that \( \mu_1 = \mu_2 = \mu \), i.e., the servers are identical with service rate \( \mu \). We shall further assume that \( c > 0 \) and that the minimum bid \( M = 0 \). Note that for the case of \( c > 0 \) and \( M = 0 \), the Wardrop equilibrium conditions can be obtained from Eq. (5.2) by replacing \( M \) with \(-c\). Precisely, the conditions are as follow.

\[
\text{If } p(\beta) > 0 \text{ then } c + \beta D_2 \leq X(\beta) + \beta D_1(X(\beta)). \tag{5.24}
\]

A possible candidate for the Wardrop equilibrium routing and bidding policy is the one that was identified in Theorem 5.3. Such a policy was characterized by a threshold \( \beta_1 \in [a, b] \) and customers with \( \beta > \beta_1 \) choose the HBF server while those with \( \beta < \beta_1 \) choose the FIFO server at the Wardrop equilibrium. We begin the analysis of this section by first investigating whether the routing and bidding policy \( p^E(\beta) \) and \( X^E(\beta) \) as described in Eq. (5.3)–(5.6) of Theorem 5.3 holds for some \( \beta_1 \in [a, b] \) for the case when \( M = 0 \) and the FIFO server charges and admission price \( c > 0 \). In the following lemma, we will show that when \( c > 0 \) and \( M = 0 \), \( p^E(\beta) \) and \( X^E(\beta) \) as described in Eq. (5.3)–(5.6) with \( \beta_1 \in (a, b] \) is not possible. This implies that if \( p^E(\beta) \) and \( X^E(\beta) \) satisfy Eq. (5.3)–(5.6), then the underlying threshold \( \beta_1 \) must satisfy \( \beta_1 = a \).

**Lemma 5.9.** Consider a routing and bidding policy \((p^E(\beta), X^E(\beta))\) as described in Eq. (5.3)–(5.6) of Theorem 5.3 and assume that \( c > 0 \) and \( M = 0 \). Then such a \((p^E(\beta), X^E(\beta))\) with \( \beta_1 \in (a, b] \) and satisfying the Wardrop equilibrium conditions of Eq. (5.24) does not exist.

**Proof.** We first prove that \( \beta_1 \notin (a, b) \) using contradiction. Assume that when \( c > 0 \) and \( M = 0 \), the corresponding \((p^E(\beta), X^E(\beta))\) satisfy Eq. (5.3)–(5.6) with \( \beta_1 \in (a, b) \). From Eq. (5.3), we
have

\[ p^E(\beta) = \begin{cases} 
  0 & \text{for } \beta > \beta_1, \\
  t & \text{for } \beta = \beta_1, \\
  1 & \text{for } \beta < \beta_1
\end{cases} \tag{5.25} \]

where \( t \) is arbitrary. Since \( \beta_1 \in (a, b) \), the equilibrium routing function \( p^E(\beta) \) requires that \( c + \beta_1 D_2 = \beta_1 D_1(\beta_1) \). This implies that \( D_2 < D_1(\beta_1) \) and hence for any \( \beta < \beta_1 \), we have

\[ c = \beta_1(D_1(\beta_1) - D_2) > \beta(D_1(\beta_1) - D_2) = \beta(D_1(\beta) - D_2). \tag{5.26} \]

To see how the last equality is true, consider a customer with \( \beta < \beta_1 \) (that deviates from the prescribed \( p^E(\beta) \)) and chooses the HBF server instead of the FIFO server. For the marginal \( \beta_1 \) customer, we know from the definition of \( X^E(\cdot) \) that \( X^E(\beta_1) = 0 \). Since \( \beta < \beta_1 \) we have \( X^E(\beta_1) = 0 \) and this customer does not pay a bid and occupies the end of the queue. For such a customer we have \( D_1(\beta) = D_1(\beta_1) \) and this follows from the absolute continuity of \( F \) and the fact that \( \beta_1 \) occupies the end of the queue in the routing function \( p^E(\beta) \). From Eq. (5.26), we have \( c + \beta D_2 > \beta D_1(\beta) \). Thus for this customer, the cost at the FIFO queue is more than the cost it would experience at the HBF server. This customer has an incentive to deviate from \( p^E(\beta) \) and hence \( p^E(\beta) \) is not an equilibrium for any \( \beta_1 \in (a, b) \).

Now consider the case when \( \beta_1 = b \). This implies that none of the customers choose the HBF server i.e., when \( c + \beta D_2 < X(\beta) + \beta D_1(\beta) \) for all \( \beta \). Note that since \( \beta_1 = b \), \( \lambda_1 = 0 \) and hence \( X(\beta) = 0 \) and \( D_1(\beta) = \frac{1}{\mu} \). Further \( \lambda_2 = \lambda \) and hence \( D_2 = \frac{1}{\mu - \lambda} \). Now \( \beta_1 = b \) if \( c + \beta \frac{1}{\mu - \lambda_1} < \beta \frac{1}{\mu} \) for all \( \beta \). However this is not possible for any \( \beta \) as \( c > 0 \) and \( \frac{1}{\mu - \lambda_1} > \frac{1}{\mu} \).

Next, we outline the conditions on the parameters for \( \beta_1 = a \). From the definition of \( p^E(\beta) \), this corresponds to the case when none of the customers choose the FIFO server i.e., when \( c + \beta D_2 > X(\beta) + \beta D_1(\beta) \) for all \( \beta \). It is straightforward to see that this case is possible when \( c \) is set to an arbitrarily very large value such that for all \( \beta \) we have \( c + \beta D_2 > X(\beta) + \beta D_1(\beta) \). Clearly, the routing and bidding function \( (p^E(\beta), X^E(\beta)) \) as specified in Theorem 5.3 with \( \beta_1 \in (a, b] \) is not possible. Further, \( \beta_1 = a \) is possible only for arbitrarily large values of \( c \).
An alternative candidate for the equilibrium routing and bidding policy is as follows.

\[ p(\beta) = \begin{cases} 
1 & \text{for } \beta > \beta_1, \\
t & \text{for } \beta = \beta_1, \\
0 & \text{for } \beta < \beta_1.
\end{cases} \]  

(5.27)

with the threshold \( \beta_1 \in (a, b) \) and the corresponding distribution \( F_1(\beta) \) satisfying

\[ F_1(\beta) := \begin{cases} 
0 & \beta < a \\
\frac{\int_a^\beta dF(x)}{\int_a^{\beta_1} dF(x)} & a \leq \beta \leq \beta_1, \\
1 & \beta > \beta_1.
\end{cases} \]  

(5.28)

It is easy to see that \( p(\beta) \) for \( \beta_1 = a \) (resp. \( b \)) is equal to \( p^E(\beta) \) (Eq. (5.3)) with \( \beta_1 = b \) (resp. \( a \)) and the conditions for such equilibrium is already outline in the previous lemma. In the following lemma, we shall now show that when \( c > 0 \) and \( M = 0 \), an equilibrium satisfying Eq. (5.28) with \( \beta_1 \in (a, b) \) is also not possible.

Lemma 5.10. Let \( c > 0, M = 0 \) and consider a routing policy that satisfies Eq. (5.27) with \( \beta \in (a, b) \) and has the type profile of customers to the HBF server given by Eq. (5.28). Such a routing policy does not satisfy the Wardrop equilibrium conditions of Eq. (5.24).

Proof. The proof is by contradiction. Suppose \( p(\beta) \) as in Eq. (5.27) satisfies the Wardrop conditions of Eq. (5.24). Since \( \beta \in (a, b) \), we have

\[ c + \beta_1 D_2 = X(\beta_1) + \beta_1 D_1(\beta_1). \]

Now since \( p(\beta) = 0 \) for \( \beta < \beta_1 \) we have \( X(\beta_1) > 0 \) and \( D_1(\beta_1) = \frac{1}{\mu} \). Further, \( \lambda_2 > 0 \) implies \( D_2 > \frac{1}{\mu} \) and hence \( X(\beta_1) > c \). For \( \beta > \beta_1 \), we therefore have the following.

\[ X(\beta_1) - c = \beta_1 \left( D_2 - \frac{1}{\mu} \right) < \beta \left( D_2 - \frac{1}{\mu} \right). \]

Thus for any customer with \( \beta > \beta_1 \) we have

\[ X(\beta_1) + \frac{\beta}{\mu} < c + \beta D_2. \]  

(5.29)
Now consider a customer with \( \beta > \beta_1 \) that deviates from \( p(\beta) \) and chooses the HBF server instead of the FIFO server. Suppose the bid paid by this customer is \( X(\beta) = X(\beta_1) \) resulting in \( D_1(\beta) = \frac{1}{\mu} \). Now from Eq. (5.29), the cost at the HBF server for such a customer \( (X(\beta_1) + \beta \mu) \) is lower than the corresponding cost \( (c + \beta D_2) \) at the FIFO server. This customer now has an incentive to deviate and the therefore the routing policy \( p(\beta) \) of Eq. (5.27) is not an equilibrium policy.

We now come to the main result of this section. Having discarded two candidate policies, we shall provide the routing and bidding policy satisfying the Wardrop equilibrium condition of Eq. (5.24) under certain conditions on its parameters. This policy is characterized by two thresholds \( \beta_1 \) and \( \beta_2 \) that satisfy \( a < \beta_1 < \beta_2 < b \). In this policy, customers with \( \beta \) satisfying \( \beta_1 < \beta < \beta_2 \) choose the FIFO server while the rest choose the HBF server. We characterize the conditions for such an equilibrium policy in the following theorem.

**Theorem 5.11.** Let \( \beta_1 \) and \( \beta_2 \) be two thresholds satisfying the conditions outlined below. Define \( p^E(\beta) \), \( F_1(\beta) \), and \( W_0 \) as follows.

\[
p^E(\beta) = \begin{cases} 
1 & \text{for } \beta_1 < \beta < \beta_2, \\
t & \text{for } \beta \in \{\beta_1, \beta_2\}, \\
0 & \text{elsewhere}.
\end{cases}
\]

\[
F_1(\beta) := \begin{cases} 
\frac{\int_0^\beta dF(x)}{(1-\int_{\beta_1}^{\beta_2} dF(x))} & \beta < \beta_1 \\
\frac{\int_{\beta_1}^{\beta_2} dF(x)}{(1-\int_{\beta_1}^{\beta_2} dF(x))} & \beta_1 \leq \beta \leq \beta_2, \\
\frac{\int_{\beta_1}^{\beta_2} dF(x) + \int_{\beta_2}^\beta dF(x)}{(1-\int_{\beta_1}^{\beta_2} dF(x))} & \beta_2 < \beta \leq b, \\
1 & \beta > b,
\end{cases}
\]

\[
W_0 = \frac{\lambda_1}{2} \int_0^\infty \tau^2 dG(\mu \tau),
\]

\[
X^E(\beta) = \int_0^\beta \frac{2 \rho_1 W_0 y}{(1 - \rho_1 + \rho_1 F_1(y))^3} dF_1(y)
\]

The conditions that need to be satisfied by \( \beta_1 \) and \( \beta_2 \) are as follows.

1. \( a < \beta_1 < \beta_2 < b \).
2. \( X^E(\beta_1) = c \).
3. \( X^E(\beta_1) + \beta_1 D_1(\beta_1) = c + \beta_1 D_2(\lambda_2) \) where \( \lambda_2 = \int_{\beta_1}^{\beta_2} \Lambda dF(x) \).
Then, \((p^E(\beta), X^E(\beta))\) is an equilibrium strategy satisfying the Wardrop conditions of Eq. \((5.24)\).

**Proof.** As in the case of Theorem \([5.3]\), we will prove that \(p^E(\beta)\) and \(X^E(\beta)\) as described in \((5.30)–(5.33)\) satisfy the Wardrop conditions of Eq. \((5.24)\). Note that with \(p^E(\beta)\) as in \((5.30)\), the arrival rate to the FIFO server is \(\lambda_2 = \Lambda \int_{\beta_1}^{\beta_2} dF(\tau)\). The arrival rate to the HBF server is \(\Lambda - \lambda_2\) and the type profile of customers choosing this server will be as in Eq. \((5.31)\). As seen in Theorem \([5.3]\), customers that join the HBF server will use the equilibrium bidding policy as in Eq. \((5.33)\). We will now verify that \(p^E(\beta)\) of \((5.30)\) satisfies the corresponding Wardrop condition.

First note that second and third conditions of the theorem imply that \(D_1(\beta_1) = D_2(\lambda_2)\). Further, since \(F_1(\beta)\) is a constant for \(\beta \in (\beta_1, \beta_2)\), we have \(X(\beta_1) = X(\beta_2)\) and \(D_1(\beta_1) = D_1(\beta_2)\). Now consider a customer with \(\beta < \beta_1\) that chooses the HBF server for the routing policy of Eq. \((5.30)\). For this customer, we know from (3) in Property \([5.1]\) that \(X(\beta)\) is the optimal bid minimizing \(C(\beta)\), i.e.,

\[
X(\beta) + \beta D_1(\beta) < X(\beta_1) + \beta D_1(\beta_1).
\]

This implies that

\[
X(\beta) + \beta D_1(\beta) < c + \beta D_2(\lambda_2)
\]

and hence any customer with \(\beta < \beta_1\) has no incentive to choose the FIFO server. Similarly, for a customer with \(\beta > \beta_2\) we have

\[
X(\beta) + \beta D_1(\beta) < X(\beta_2) + \beta D_1(\beta_2)
= c + \beta D_2(\lambda_2)
\]

and clearly customers with \(\beta > \beta_2\) have no incentive to choose the FIFO server.

Now consider any customer with \(\beta \in (\beta_1, \beta_2)\). If such a customer were to choose the HBF server instead of the prescribed FIFO server, its cost at the HBF server will be \(X(\beta_1) + \beta D_1(\beta_1)\). This follows from (1) and (3) of Property \([5.1]\) and the fact that a deviation by a marginal customer does not change Eq. \((5.30)\) and \((5.31)\) due to the absolute continuity of \(F\). Now \(X(\beta_1) + \beta D_1(\beta_1) = c + \beta D_2(\lambda_2)\), i.e., the cost at the two server is the same and there is no incentive for this customer to deviate from \(p^E(\beta)\) given by Eq. \((5.30)\). This completes the proof. \(\square\)
Remark 5.12. In the above discussion, we have assumed $c > 0$ and $M = 0$. W.l.o.g the observations are true even for the case when $c > M$ and $M \neq 0$. For this case, let $\hat{c} = c - M$. The above theorem and the previous lemma’s will hold with $c$ being replaced by $\hat{c}$. It is easy to see that the case $c < M$ and $M \neq 0$ corresponds to the Wardrop equilibrium as described in Theorem 5.3.

5.6 Discussion

The primary motivation for our models in this chapter are from the need for new pricing and auction mechanisms in on-demand resource provisioning, e.g., cloud computing systems. An additional interest is, like in [40] and in [45], the economic aspects of bribing. While [45] assumed that the full capacity was auctioned, we investigate the economics of ‘partial corruption’ in which some of the population is provided an ‘honest service’ via the FIFO queue. In this setting $M$ (the minimum bid) is interpreted as the ‘social reward’ to a customer for doing the right thing, i.e., for not bribing. The results of Fig. 5.2 provides the intuitively appealing interpretation that it might be more rewarding to be partially corrupt than to be fully corrupt. Further, ‘small’ levels of corruption does not pay.

There is of course more work to be done. Formal proofs for the numerical findings of Fig. 5.2 a better understanding of the system of Section 5.3.2 and analyzing the balking system in more detail are on our immediate agenda.

An alternate system where the bid value determines the share of the server capacity in a processor sharing system has also been analyzed in the previous chapter. A comparative understanding of these systems with direct application to service provisioning cloud computing and wireless communication systems are also of interest.
Chapter 6

Conclusion

In this thesis, we have analyzed two types of queueing systems that can be used as models for service systems with heterogeneous customers. Such service systems include and are not limited to road and transport systems, health-care systems, computer systems, call centers and communications systems. The two queueing systems that are used to model the different problems considered in the thesis are (1) a parallel server system and (2) a single server system with a scheduling policy that offers service differentiation (such as in the DPS and HBF systems).

In Chapter 2, we have considered multiple parallel queues serving heterogeneous customers, and have studied the problem of routing such non-identical customers to queues so as to minimize the aggregate delay cost of all the customers. We characterize the structural properties of the routing minimizing this cost function and show that this optimal routing satisfies a certain threshold property. We then consider strategic customers that make selfish routing decisions and characterize the structure of Wardrop equilibria. Our main result showed that, if queues charge admission prices, and these are set equal to the congestion externalities at the optimal allocation, then the resulting Wardrop equilibrium coincides with the social optimum. We have presented these results for the case of finite number of customer classes as well as for the case of a continuum of customer classes indexed by a delay sensitivity parameter $\beta$ with a known distribution. It must be noted that our analysis makes minimal assumptions on the distribution $F$ and the delay function $D(\cdot)$ and this makes the model applicable of a variety of scenarios.

While the system objective in the previous problem was to maximize a certain social welfare function, an alternative objective in certain other service systems could be to maximize the revenue earned from the admission prices. We have considered this revenue maximization
problem in Chapter 3 where for simplicity, the service system is assumed to have two parallel servers. Since the servers belong to the same service system, this is an example of a monopoly. We note that the heterogeneous nature of the customers makes characterization of the revenue maximizing prices very difficult. In this chapter, we have simplified the revenue maximization programs into equivalent program by making use of the threshold property of the Wardrop equilibrium. The equivalent formulation has simplified our analysis for the problem while still being applicable to a wide class of distribution functions as well as delay cost functions.

In Chapter 4, we have analyzed the single server queueing system with a discriminatory processor sharing (DPS) scheduling policy. The DPS system is characterized by $M$ service class weights that determine the share of the service rate for that service class. In the first part of this chapter, we considered the problem of optimally assigning heterogeneous customers (both a continuum and finite set of delay costs) to a DPS system with $M$ service classes and given weights. When the delay cost is a continuous random variable, we show that the optimal allocation is of a threshold type. When the delay cost is a discrete random variable we have shown that customers with the highest (resp. lowest) delay cost will always use the service class with highest (resp. lowest) weight. Interestingly, when there are three customer types we identify conditions when the optimal solution is indifferent to the service class allocated to Type 2 customers. In the second part of the chapter, we considered differential admission prices in which arriving customers must pay a price for the chosen class of service. We again show that the individually optimal routing policy is also of a threshold type. Finally, we show that there exist admission prices for which the resulting equilibrium allocation coincides with the optimum allocation minimizing the social utility.

Finally in Chapter 5, we have considered a Highest bidder first (HBF) service system placed in parallel with a FIFO server. We first assume that the FIFO server offers free service to the customers while a minimum bid $M$ is a requirement in the HBF server. We have analyzed the equilibrium routing of the customers to the HBF and the FIFO server. This routing is characterized by a unique threshold where customers with delay cost above the threshold choose the HBF server while those below it choose the free FIFO service. We then analyzed the effect on the revenue made by the system due to the presence of this free FIFO server. We have shown that when there is capacity addition in the form of a free FIFO server, there is a decrease in the revenue from the system. This is because, the addition of capacity not only reduces the congestion in the HBF server, but it also reduces the value of the optimal bids received. We
next perform a numerical analysis of the situation where the total service capacity is kept fixed and is shared between the HBF and the FIFO servers. This sharing of capacity on the contrary increases the revenue from the system. The reason for this is that sharing capacity decreases the service rate of the HBF server and this increases the congestion effect and hence increases the value of the optimal bids made by all the customers. We conclude with an analysis of the case where the FIFO server now charges an admission price \( c \) at the FIFO server such that \( c > M \). The equilibrium routing for this case is also of the threshold type, albeit with two thresholds \( \beta_1 \) and \( \beta_2 \). Customers with \( \beta \in [\beta_1, \beta_2] \) choose the FIFO service while the others choose the HBF server. This two-threshold equilibrium routing was not intuitive to us and was a very interesting observation.
Bibliography


Publications


Acknowledgments

First and foremost, I would like to thank my advisor, Prof. D. Manjunath for his sincere guidance all along the course of my Ph.D. I have enjoyed every minute of the technical and non-technical discussions that I have had with him over the past eight years. I have learned a lot from him over these years and I hope to continue doing so over the future. I would also like to thank my research committee members, Prof. D. Chakraborty and Prof. V. Narayanan for the numerous technical feedbacks during the course of the thesis. I would also like to thank my thesis reviewers, Prof. Urtzi Ayesta, Prof. Sandeep Juneja and Prof. Joy Kuri for the constructive feedbacks and suggestions that have helped me improve the quality of the thesis considerably.

I was fortunate to work with Prof. A. Ganesh on one of my thesis problem. The discussions with him have always helped me to have a better clarity on the problem. I was glad to be able to collaborate with Prof. Kavitha from IEOR, IIT Bombay. Apart from the technical discussions, I have enjoyed her informative lectures very much. I would also like to thank my colleagues-cum-collaborators Tanmay Sarkar, Sripada K, Murtaza Ali and Dileep Kumar with whom I have had the opportunity to working closely.

I feel grateful to have been a member of the Networking Lab and the Bharti Centre for Communication in the EE Department. I thank all the members of these labs, both past and present, for all the wonderful memories shared together. I would also like to thank Vibha madam for all the administrative assistance she has provided over the period. I would like to thank all my M.Tech and Ph.D batch mates for all the fun time spent. My hostel days in IIT Bombay were a special experience and the memories are hard to forget. When it comes to memories, I can’t leave my badminton buddies behind. I thank all of them for all the competitive matches and tournaments played together. Badminton is something that kept my feet grounded all these years. I would thank the sport itself for being there during boring times.

I have had two wonderful Internship opportunities during the course of my Ph.D. I would like to thank Prof. Konstantin Avratchenkov for hosting me twice at INRIA, Sophia Antipolis. I
have thoroughly enjoyed working with him and hope to continue the collaboration. I would also like to thank Theja Thulabandula for hosting me at Xerox India. With him, Prabu and Deepak, I have enjoyed discussing MDP’s, puzzles among other things.

I have had over half a dozen trips to Europe in the past few years. Some for leisure, some for conferences and internship. I would like to thank CEFIPRA, IFCAM, Bharti Centre and the various other funding agencies and grants that have made the trips possible. I could not have asked for more trips and each trip was worth it.

Most importantly, I would like to thank my family for being there for me. I would like to express my sincere thanks to my parents for their understanding and support. Their unconditional love and faith helped and inspired me during the course of my Ph.D. I would also like to thank my parents-in-law for their support and faith in me.

Finally, I would like to thank my soulmate Arti for her infinite love and understanding over these years. She was always there when I needed her the most with her unwavering support. We have enjoyed so much together — the Eurotrips, the shopping trips, baddy sessions, cooking experiments and now Game of Thrones! This thesis would have been impossible without her support.

Date: 22/4/2016

Tejas P. Bodas