MARGINAL SOCIAL COST AUCTIONS FOR CONGESTED AIRPORT FACILITIES

by

Raphael Avram Schorr

B.A., Physics
Harvard University, 2000

Submitted to the Department of Civil and Environmental Engineering and the Sloan School of Management in Partial Fulfillment of the Requirements for the Degrees of

Master of Science in Transportation and SM in Operations Research

at the Massachusetts Institute of Technology September 2002

© Raphael Schorr. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part.

Signature of Author.......................................................... Department of Civil and Environmental Engineering

Certificate by.................................................... Yosef Sheffi

Director, Center for Transportation & Logistics

Professor, Department of Civil and Environmental Engineering

Accepted by.................................................... Oral Buyukozturk

Chairman, Departmental Committee on Graduate Studies

Department of Civil and Environmental Engineering

Accepted by.................................................... James B. Orlin

Codirector, Operations Research Center
MARGINAL SOCIAL COST AUCTIONS FOR
CONGESTED AIRPORT FACILITIES

by

Raphael Avram Schorr

Submitted to the Department of Civil and
Environmental Engineering and the Sloan School of
Management on August 16, 2002 in Partial Fulfillment of
the Requirements for the Degrees of Master of Science
in Transportation and SM in Operations Research

ABSTRACT

Airport congestion costs airlines and the traveling public over $5 billion a year in the United States alone. With growing demand and limited room for capacity improvements, there is a pressing need for non-capacity-based solutions to reduce congestion. At the heart of the congestion problem is the fact that airlines and general aviation only partially bear the marginal costs of their operations. This study focuses on market-based solutions to airport congestion and the congestion externality problem.

Previous research on market-based solutions to airport congestion has advocated one of two approaches to maximize economic welfare. One approach is to auction some fixed number of slots, with the number of slots being determined using a judgmental approach. The other approach is to toll the airport operators for the Marginal Social Cost of their operations. Our research finds that neither of these methods is appropriate for maximizing economic welfare. We suggest a new approach, which we have called the Marginal Social Cost Auction. This new approach combines a congestion simulation model based on the one suggested by Koopman (1972), a combinatorial auction approach based on the one suggested by Rassenti, Smith, and Bulfin (1982), and a new objective function to maximize the economic welfare that an airport provides.

Thesis Supervisor: Yosef Sheffi
Title: Director, Center for Transportation & Logistics
ACKNOWLEDGMENTS

I would like to acknowledge and thank the following people for helping me complete my degrees over the past two years:

Professor Yosef Sheffi, my advisor, for providing me with guidance, direction, constructive feedback, and friendship over the last two years. Yossi's imagination, sense of humor, and research insights were invaluable in helping me write this thesis. Last, but not least, Yossi helped me practice speaking Hebrew, something for which I will always be grateful.

Other professors at MIT: Professor Amedeo Odoni for encouraging me to pursue this research topic and mentoring me; Professor Cynthia Barnhart for providing me with direction since the days when I was studying Physics and had not yet applied to MIT; and the many other professors from whom I have learned so much and whose friendship I continue to value including professors Arnold Barnett, Peter Belobaba, Moshe Ben-Akiva, Ismail Chabini, George Kocur, Andreas Schultz, and Nigel Wilson.

The National Defense Science and Engineering Graduate Fellowship of the US Government, the Rosenblith Fellowship, and the University Transportation Fellowship for providing me with generous support over the last two years. I can only hope that I give back as much as I have been given.

My teachers over the many years who not only taught me, but also encouraged me to really learn, had faith in me, and advocated for me every step of the way.

My parents, for encouraging me to pursue education over the last 23 years and for their enduring faith that I would eventually become a student. They have provided me with the intellectual, emotional, and financial support that only parents can.

Finally, I would like to thank my wife, Yonit, for supporting my educational goals. Yonit has been accommodating and flexible from the day we moved to Cambridge so that I could go to school. Also, thank you, as always, for being a great editor.
# TABLE OF CONTENTS

Abstract ................................................................................................................................. 3  
Acknowledgments .................................................................................................................. 5  
Table of Contents .................................................................................................................. 7  
List of Figures ......................................................................................................................... 9  
List of Tables ........................................................................................................................... 9  
Chapter 1: Introduction .......................................................................................................... 11  
  1.1 The Problem ................................................................................................................... 12  
  1.2 Proposed Solutions ....................................................................................................... 13  
  1.3 Research Scope ............................................................................................................. 16  
  1.4 Organization of Thesis .................................................................................................. 16  
Chapter 2: Background .......................................................................................................... 18  
  2.1 Airport Finances .......................................................................................................... 18  
  2.2 Queuing ......................................................................................................................... 19  
  2.3 Costs of Queuing .......................................................................................................... 21  
  2.4 Efficiency ...................................................................................................................... 23  
Chapter 3: Airport Congestion Models .................................................................................. 25  
  3.1 Six Models of Airport Congestion and Optimal Congestion Pricing ......................... 25  
  3.1.1 Busy Period Model ................................................................................................. 26  
  3.1.2 Regressive Model ................................................................................................. 27  
  3.2.3 Time-Dependent Stochastic Queuing Model ......................................................... 30  
  3.2.4 Time-Independent Stochastic Queuing Model ....................................................... 31  
  3.2.5 Deterministic Bottleneck Model ............................................................................. 33  
  3.2.6 Stochastic Bottleneck Model ................................................................................ 36  
  3.3 Model Comparison ....................................................................................................... 37  
  3.4 Conclusion .................................................................................................................... 39  
Chapter 4: Marginal Social Cost Tolls ................................................................................ 40  
  4.1 Proof: Existence of a MSC Toll Set for a Steady State Queue .................................... 41  
  4.1.1 Part I: Operators with Homogeneous Service Time Distributions ....................... 41  
  4.1.2 Part II: Operators with Heterogeneous Service Time Distributions .................... 47  
  4.2 Estimating Demand Functions ..................................................................................... 53  
  4.3 Discrete Nature of Demand .......................................................................................... 54  
  4.4 More Complex Models .................................................................................................. 56  
  4.5 Solution Degeneracy ..................................................................................................... 56  
  4.5.1 Why Steady State Queues can Have Multiple MSC Toll Sets ............................... 57  
  4.5.2 Example: M/M/1 Queue with Two User Types .................................................... 58  
  4.5.3 Significance of Multiple MSC Toll Sets .................................................................. 60  
Chapter 5: Slot Auctions ....................................................................................................... 63  
  5.1 Administrative Limits .................................................................................................... 63
### 5.2 Overview of Auctions

5.3 Vickrey Auctions

5.4 Complications of Slot Auctions

5.5 Combinatorial Auctions

5.6 Prices in Combinatorial Auctions


6.1 A New Objective Function

6.2 The Marginal Social Cost Auction

   6.2.1 Slots

   6.2.2 Bidding and Cost Estimates

   6.2.3 The Congestion Model

   6.2.4 Optimization

   6.2.5 Pricing

   6.2.6 Administration

6.3 Making the Marginal Social Cost Auction Viable Through Buy-in

   6.3.1 Case Study: El Toro Marine Corps Air Station

6.4 Objections to the Marginal Social Cost Auction

6.5 Operational Considerations

### Chapter 7: Summary, Contributions, and Future Research Directions

7.1 Summary and Contributions

7.2 Future Research Directions

### References
LIST OF FIGURES

Figure 1: c as a function of Toll ................................................................. 59
Figure 2: MSC vs Toll ............................................................................. 59
Figure 3: x as a function of toll ............................................................... 60
Figure 4: Welfare as a function of toll .................................................... 62

LIST OF TABLES

Table 1: Typology of Congestion Models .................................................. 38
Table 2: Comparison of Predictive Models ................................................. 39
Table 3: Two Attempted Tolls for an M/M/1 Queue .................................. 56
Table 4: Parameters for Two Queue Participant Types ......................... 58
Table 5: The relative advantages and disadvantages of Auctions and MSC Pricing .... 78
INTRODUCTION

In recent years, airlines and the flying public faced unprecedented congestion-related delays. In the late 1990s, a combination of over scheduling at the most popular airports, unevenness of scheduling associated with hub-and-spoke banks, and frequent reductions in capacity due to inclement weather created record delays. Flights were delayed by a total of more than 3 thousand hours per day, costing airlines and consumers an estimated $5 billion a year (ATA 2001). While the worst delays were at the nation's busiest airports, the tightly coupled nature of the air transportation system caused rippling delays throughout the entire system.

A look at New York LaGuardia (LGA), one of the most congested airports, shows both an example of how severe congestion has been in some instances and how bad it might become elsewhere if not adequately addressed. In 1969, the FAA adopted a High Density Rule (HDR) to temporarily address congestion at LGA and four other congested airports, limiting the number of scheduled operations (i.e. landings and takeoffs) (FAA 2001). This "temporary" measure remains in effect, but was relaxed in 2000 when the Aviation Investment and Reform Act for the 21st Century (2000) exempted certain flights that service non-hub or small-hub airports and some flights operated by new-entrant airlines from the HDR limits. During this period, delays at LGA grew by 238% and accounted for 25% of the nation's total delays (FAA 2002); by November of 2000, flights using LGA were delayed by an average of nearly one hour (Odoni 2001). These delays had an annualized cost to airlines and passengers of well over $1 billion.

It is useful to put these losses into perspective by comparing them to airline revenues and profits. In the US, the cost of airline delays is about 5% of the airlines' revenues and roughly

---

1 This figure may actually be a gross underestimate of the cost of congestion as it only accounts for delays, but does not account for the cost of schedule padding. Over the years, airlines have added significant buffers to their schedules because of the increasing congestion. For example, recent research shows that between 1995 and 1999, airlines added an average of five minutes to the scheduled block time of the 100 most frequent routes (Bratu and Barnhart 2000).
equal to the airlines’ net income in the best of years. Delay costs at LGA alone, during its worst congestion, were more than 1% of industry revenues. With costs on this order of magnitude, it is clearly in the airlines’ and passengers’ interest to find ways of reducing this deadweight cost and recapturing some of this lost time.

In the aftermath of September 11, 2001, industry schedule reductions resulted in delay subsidence. Nonetheless, airline traffic is expected to rebound over the next two years (FAA 2002), threatening to bring back the levels of delays that existed in 2000. Delays will continue to plague the airline industry and its passengers as long as no serious measures are taken to alleviate congestion.

1.1 The Problem

At the most basic level, congestion occurs because the demand for air travel exceeds the air transportation system capacity. More specifically, delays result when some component of the air transportation system does not have sufficient capacity to satisfy the demand for that system during some period of time. In the current aftermath of September 11, 2001, security checkpoints are perhaps the most visible example of such a component. Many checkpoints have long queues during peak travel times because the rate of people arriving to the security checkpoints exceeds the rate of people departing from them during those peak times. While security delays are clearly of interest, our focus is on the delays to aviation operations that occur when Air Traffic Control (ATC) is unable to accommodate all of the flights in a particular region, airports are unable to accommodate all of the flights that would like to access or egress the airport terminals, and/or airports are unable to accommodate all the flights that wish to use their runways to land and takeoff.

Our particular focus is on runway delays because they are the primary source of flight delays in the United States. Even in Europe, where the primary source of congestion is ATC (Golaszewki 2002), runway delays are a very serious problem. ATC delays dominate in Europe because of the lack of coordination between the ATC systems of the European countries, something that could eventually be fixed. Thus, the runway problem is a serious one that deserves our attention.
1.2 Proposed Solutions

There are three broad categories of commonly proposed solutions to runway congestion: capacity-based initiatives, administrative measures, and market-based solutions. Capacity-based initiatives call for reducing separation distances between successive operations, increasing capacity through building runways, and/or building new airports. Separation distances are safety guidelines governed by the sizes of the leading and trailing aircraft. Changing the legal requirements or efficiently sequencing the operations (e.g. handling a block of large jets and then a block of smaller planes rather than mixing them up) can reduce the average separation distance and increase capacity. Without compromising safety, these would not generate significant increases in capacity, but should not be discounted as, at the very least, worthwhile partial solutions. At some point in the future, there could be a technological advancement that reduces, or eliminates, the need for runways for takeoffs and/or landings, but such a solution is unlikely to come soon enough to deal with our current problems.

Building runways in the most congested airports could relieve much of the worst congestion. However, many of the busiest airports have physical barriers to expansion such as water and highways. Other airports that abut residential areas could theoretically acquire some of the residential land, but doing so is often exceedingly expensive. In addition to the issue of expense, building more runways is a contentious political issue in many cities. The neighborhoods surrounding airports are typically strongly opposed to any expansion because of the associated noise, pollution, and traffic.

Lastly, building new airports in the largest cities could help reduce runway congestion. However, building new airports is typically even more difficult than building more runways for the same physical, fiscal, and political reasons. The only way that this becomes a feasible option is if the airport can be built on the edge of the urban area with good transportation links to the city. In most cases, the edge of the urban area may now be too far for conventional transportation links. In fact, the only large US city to build an airport in the last 30 years in Denver, a city that has open land in relatively close proximity to the city. Even if such an airport is built, if it takes much longer to get to than other airports, as is the case with Montreal's Mirabel Airport, it may be a very unpopular facility and will have trouble diverting traffic from the congested facility.
Thus, none of the suggested methods for increasing the collective runway capacity are sufficient solutions. We can at best expect limited technological improvements that reduce separation distances, a small increase in the number of runways in existing airports, and rare instances of construction of entirely new airports. Even if some combination of these measures were able to increase capacity and alleviate current congestion-related delays, the expansion would have to keep pace with expected demand growth for air transportation to be an effective long-term solution. The continued emergence and growth of profitable, low-fare airlines provides reason to believe that demand will continue to grow in the future, even if not at the 4.5% annual rate (measured in enplanements) that it has since deregulation in 1978. Moreover, recent research suggests that, even without increased competition, per capita demand for high-speed (i.e. air) travel is expected to grow as long as the GDP per capita is increasing (Schafer 1998). Thus, even if US population growth subsides (as many predict that it eventually will), we expect demand growth for air travel to continue. Since potential capacity increases are limited, capacity-based solutions, on their own, will not be effective long-term solutions.

The second class of solutions is consists of administrative measures. Administrative measures set a limit on scheduled operations per time period and then distribute slots, rights of usage, to airport runway users. Since demand for slots exceeds their supply, they must be rationed. Rationing can be done on the basis of prorating airlines’ current (i.e. pre-administrative measure) airport use, a lottery, or one of a number of other ways. These measures have been used in a few US airports and in many European airports and have effectively reduced congestion. However, administrative methods do not encourage efficient use of these scarce resources in at least two important ways. First, they fail to allocate the resources to the users that value them most. Second, they choose the number of slots to distribute, effectively determining the level of congestion, without the necessary cost-benefit analysis.

The third category of solutions consists of market-based measures. These can be divided into two types: those that are based on auction principles and those that are based on Marginal Cost (MC) pricing. Auctions are a price-based competition used to efficiently award goods or services. They come in many varieties, but have in common the principle that the highest bid(s) is (are) the winning bid(s). In the context of airports, it has been proposed that auctions could be used with administrative limits to award some number of slots to the airlines that
value them most. Thus, auctions could have been classified under administrative limits instead of market-based solutions. However, since part of the auction process involves a market-based approach, we have decided to differentiate auctions from the rest of the administrative limit-based solutions that do not use any market-based approach.

MC pricing is based on the notion that welfare is maximized when the price of a good or service is equal to the MC for society to offer that good. Thus, in general, the MC price is a Pareto-efficient allocation because we can find no price that will make some people better off without making some segment of society worse off (Vickrey 1987). When the price is equal to MC, lowering the price would increase the cost to society more than the increase in value obtained from the good or service, while raising the price would decrease the cost to society less than the decrease in value obtained from the good or service. In our context, MC pricing means that airport users pay for both private and external costs of operating a flight. Private costs include both normal operating costs and the cost of delays that the user experiences, while external costs include the cost of increased delays to other users, which we refer to as the marginal social cost (MSC), and the increased operating costs to the airport authority and ATC. Airlines (in an approximate way) usually pay for the airport and ATC costs, but do not typically bear the MSC. Vickrey (1969) suggested that airlines pay a toll equal to the MSC as a way of achieving a Pareto-efficient allocation by only allowing those users who are willing to bear the costs of the congestion externality.²

Thus, market-based mechanisms promise efficiency in a way that capacity and administrative solutions cannot. Both auction mechanisms, which allocate a fixed number of slots efficiently, and MC pricing, which makes airlines bear the congestion externality, should improve upon the status quo. We focus our attention on such market-based mechanisms with the objective of finding a practicable, efficient solution for allocating scarce runway capacity.

² While, we devote much attention to the MSC, we do not address the question of whether the current airport/ATC fees properly reflect a marginal cost structure, because it should be a relatively trivial matter to adjust these fees to reflect this structure.
1.3 Research Scope

While this research is explicitly concerned with runway scarcity and congestion, it has potentially important insights into and lessons for other areas. There are similar congestion phenomena in other areas of transportation such as urban roadways, highways, transit facilities, ports, and waterways and in several areas of telecommunications. The combined cost of delays in these different areas is enormous, with automobile congestion in the US alone causing 4.3 billion hours of delay at an estimated cost of $70 billion (BTS 2000; TTI 2002). The fundamental results of this research should be of importance to many of these other applications.

1.4 Organization of Thesis

In Chapter 2, we provide some background information for the remainder of the thesis. The material serves as an introduction to some of the concepts and idioms that we use in the later chapters.

In Chapter 3, we review the literature on airport congestion models. We present six models for airport congestion and review their relative strengths and weaknesses. Given several different possible purposes for these models, we explore which models are most appropriate for which purposes.

In Chapter 4, we discuss the notion of the Marginal Social Cost Toll. We show a set of conditions under which we are guaranteed to be able to find a toll (or set of tolls) that cause the marginal social cost(s) of using the airport to be equal to the toll(s), a condition that is traditionally assumed to imply optimality. We discuss several of the difficulties with the notion of this toll. We further discuss whether the Marginal Social Cost Toll is, in general, unique and discuss the implications of whether or not it is unique.

In Chapter 5, we discuss auction theory, in general, and the traditional approach to using auctions to allocate scarce airport slots. We discuss the particular problems of using administrative limits to set the number of slots available for auction. We also discuss some of the challenges involved in designing an auction that can efficiently and fairly handle the challenges of simultaneously auctioning multiple slots types.
In Chapter 6, we present a new market-based approach to managing congestion at airports. The new approach, tentatively named the Marginal Social Cost Auction, combines a combinatorial auction with elements of MSC pricing to optimize welfare. We review this approach and discuss some of its details, merits, and challenges.

In Chapter 7, we conclude with a summary of the research in this thesis. In addition, we highlight some of the specific contributions that this thesis makes in furthering research on dealing with airport congestion and advancing the state of the art for efficiently dealing with this congestion. Lastly, we suggest areas that call for further research.
Background

The purpose of this chapter is to provide some background for the chapters that follow through the introduction of several important topics. The material presented here introduces some of the systems, jargon, and ideas that form the basis of this thesis. We begin by discussing airport finances, continue with an overview of runway queuing and the costs of queuing, and close with a discussion of efficiency in the context of congested airports.

2.1 Airport Finances

Levine (1969) provides an excellent analysis of the financial structure of the US air transportation system. While an exhaustive review of this system is beyond the scope of this thesis, we provide a summary of some of the most important features of the system. The system is comprised of components such as airports, air corridors connecting the airports, and an air traffic control (ATC) system. The airports serving commercial aviation are typically publicly owned and generally operate on a “First Come, First Served” basis.

Funding for the system comes from a variety of sources. The government pays for the cost of the ATC system and some portion of airport infrastructure improvements by collecting a fuel tax and a per-passenger-enplanement Federal Excise Tax. Airports pay for passenger facilities, some portion of airport infrastructure improvements, and ongoing operating costs through revenues collected from a per-passenger-enplanement Passenger Facility Charge; selling the monopoly rights to airport concessions (e.g. taxis and restaurants); and landing fees. The landing fees are typically based on weight and are only meant to recover “that part of the operator’s portion of the investment in the airport which cannot be recovered from charges on non-aeronautical uses” (Levine 1969, p. 79).

Beyond congestion related issues, which are the basis of our topic, there are several notable problems with this system. First, because some airport concessions are operated as monopolies, they are very expensive to use and are often underutilized as a result (Levine
Thus, airport customers seeking a car rental or a meal are often discouraged by monopolistically high prices. Second, the current subsidies that airports receive from the federal government limit their freedom in pricing landing fees. This makes it difficult to profit from expansion, a problem that we shall come back to later, and obviously makes it difficult to use any market-based initiatives to address congestion. This latter problem is critical in that all discussions of market-based approaches are moot without changing this rule; federal legislation removing this limitation is the sine qua non of market-based approaches to airport congestion that we discuss in this thesis.

2.2 Queuing

In Chapter 1, we noted that runway delays are the primary source of aviation delays in the US. Queues develop when runways cannot accommodate all the flights that wish to use them during some period of time. We can qualitatively describe these queues in terms of configuration, physical process, discipline, arrivals process, and departures process.

The number of queues at a given airport depends on the configuration of the airport. For example, one airport may use three different runways to separately service propeller aircraft operations, departing jet aircraft, and arriving jet aircraft and another airport may use one runway for all three of these types of service.

These queues are physical in the sense that they occupy space. A queue for takeoffs will usually be, at least in part, on the taxiways leading to the runway. However, this queue can also include aircraft at terminal gates and/or ramps waiting for permission to taxi. A queue for landings will usually have aircraft in holding patterns in the air. Because airplanes do not have enough fuel to circle an airport for hours, ATC will order flights that are planning on landing at that airport and have not yet departed to hold their positions on the ground. Thus, the queue can extend to other airports.

While airports generally follow a "First Come, First Served" policy, there are some exceptions. For instance, landing aircraft often get priority over departing aircraft because of safety concerns. While this thesis does not extensively deal with alternatives to the "First Come, First Served" policy, it is worth mentioning that there are several possible alternative policies that could be considered. An airport could decide to handle the operation with shortest service
time first. Alternatively, an airport could use nonpreemptive priority queuing, with priorities based on aircraft size, differentiated service fees, or another differentiating factor.

Arrivals to a queue occur when an airplane requests permission to either takeoff or land using an airport's runway(s). The pattern of such arrivals may be somewhat predictable based on the scheduled operations at an airport. However, the arrivals typically do not perfectly match the schedule for two reasons. The first reason is that flights do not operate exactly on schedule, something that stochastic queuing models capture well. The second reason is that some flights do not operate on a schedule; many airports allow flights to operate that are not scheduled far in advance. Thus, the arrivals process may follow some pattern, but will often deviate substantially from the deterministic arrival times in the airport's schedule of flights.

Departures from a queue occur when an airplane uses the runway to either takeoff or land. To better understand the pattern or rate of departures, we should briefly discuss the notion of runway capacity. The first thing that we must note is that airports typically operate under various configurations depending on the weather; the weather will effect which runways can be used and what purposes those runways will be used for. For instance, most airports will operate differently under Instrument Meteorological Conditions (IMC), which dictate Instrument Flight Rules (IFR), and Visual Meteorological Conditions (VMC), which allow Visual Flight Rules (VFR). They will also typically operate differently under different wind conditions.

Within a given operating configuration, the number of aircraft that can be serviced in a given amount of time will depend on several factors. First, landings and takeoffs take different amounts of time; deciding what proportion of operations is going to be landings and what proportion takeoffs will influence the number of operations that can be handled. Second, separation distances, which dictate times between successive operations (landings in particular), are dependent on the leading and trailing aircraft types. Thus, the sequencing of operations will largely influence how much time there is between successive operations on a particular runway. Finally, the actual service times have some inherent variability from conditions such as weather (i.e. temperature will affect the amount of runway used for an operation) and human factors (i.e. one pilot may take longer than another to perform an identical operation). Thus, the service rate of the airport will, broadly speaking, depend on the configuration and
weather conditions under which it operates, the sequence of airplanes that it services, and other stochastic factors. This is a key point in understanding the congestion phenomenon. The service rate (like the arrivals process) is stochastic and there is therefore no magic number of arrivals to schedule to ensure that a queue does not develop.

2.3 Costs of Queuing

Throughout this thesis, we refer to the various costs associated with congestion. Here, we briefly summarize these different costs so that they can be properly contextualized. First, the total cost of congestion should include both the cost of congestion to airlines and passengers. With regard to airlines, there are two cost components. The first is direct operating costs; when flights take longer, airlines spend more on wages, fuel, maintenance, etc. The second component is indirect costs from loss of productivity; when flights take longer, more pilots and flight attendants need to be trained and more equipment needs to be acquired to perform the same service. With regard to the cost to passengers, there are various ways to estimate the costs including estimating how much the passengers value their time or estimating what loss of productivity the congestion causes.

One very important question is who bears the cost of passenger delays. While this may seem obvious at first, it is not. In the most direct sense, passengers bear this cost. The passenger wanting to go from New York to Hong Kong clearly suffers if the flight is delayed significantly. However, to the extent that there is some elasticity in demand, the airline also suffers. This is especially true where there are competitive alternative modes of transportation, such as rail or automobile, or there are alternate airports to choose from. While airline economics are far too complex and enigmatic to provide a realistic example of what losses the airline suffers, we present a simple example for purposes of illustration alone. In this example, the demand does not vary by day (i.e. it is consistent from day-to-day), there is no marginal cost of carrying a passenger, the airline flies one flight a day with 110 seats, the airline may only use one fare, and the demand function is:

\[ D = 200 - \text{Price}. \] (1)

---

3 For an example of an econometric approach to measuring the value of time for air travelers, see Morrison and Winston (Morrison and Winston 1989).
Since the costs of operating this flight are fixed, the airline maximizes revenues by setting the price equal to $100, which results in a demand of 100 and collecting $10,000 in revenue. Let us consider what happens if the airport gets busier to the point where each passenger experiences a delay cost of $5. If passengers have alternatives and are aware of this delay, they should include it in the cost of the flight. If passengers now have a demand function of

\[ D = 195 - \text{Price}, \tag{2} \]

the airline now finds that it will maximize revenues by setting the price equal to $97.50, which results in a demand of 97.5 and collecting revenues of $9,506.25. That means that the airline loses $493.75 as a result of the passengers being delayed. Let us compare this for a moment to the traditional estimate of cost of delays to passengers. Since there are 97.5 passengers and each one bears a $5 cost of delay, it would seem that the passenger loss is some $487.50. That means that the airline actually bears the perceived passenger loss, and perhaps even a few dollars more! In the real world, the marginal cost of carrying passengers is positive, airlines make decisions about what size aircraft to operate, and the daily variation in demand creates the occasional spill of passengers, all of which complicate this model; however, this model suggests that when markets are elastic to air travel, airlines suffer from passenger delays. For purposes of simplification, we will generally assume that the airline bears the cost of passenger delays. While this may seem like an onerous assumption, it turns out not to be as we shall see later in Chapter 6.

The Marginal Cost (MC) of congestion is a measure of the change in the total cost of congestion from a marginal operation. This MC can be broken into two components, the Marginal Social Cost (MSC) and the private Delay Cost (DC). The MSC is the component of the MC that is not experienced by the marginal operation of interest, but rather experienced by the other flights. The DC is the component of MC that is experienced by the marginal operation.

Whereas the MC of congestion is the total price that society pays for the marginal operation, the operator faces a different price for this use. The congestion-related price (i.e. all other costs aside) that the operator faces is the sum of the DC that the operator experiences and the price of the congestion toll(s) for operating that flight. Finally, we mention that we will
sometimes refer to a toll set, which is a set of tolls with each member of the set corresponding to the toll for some specific operation type.

2.4 Efficiency

In this thesis, we make frequent reference to the notion of efficiency. It is worthwhile to define what we mean by this. First, we note that wherever we employ efficiency, it can be readily interchanged with Pareto efficiency. The second thing that we need to make clear is that there are two types of efficiency that we refer to in this thesis.

The first type of efficiency is particular to the congestion externality. In this context, efficiency implies that the level of congestion is efficient. This may sound strange since all congestion is undesirable, but that is not the case. For example, if an additional flight operating on a congested runway will increase congestion, it is efficient to allow such a flight if the flight operator would be willing to pay for the increased congestion to all the other parties using the runway. This is fairly intuitive. If the airline delays ten other flights at a cost of $100 a piece and the additional flight is willing to pay each of these other ten flights $100 (or more) a piece, it is clearly more efficient to allow this flight. Moreover, it is efficient to allow such a flight whether it ends up paying the other operators for their losses or not. Looking at this from another perspective, if there is an operator in the airport system that is unwilling to pay for the cost imposed on others from some operation, granting access for such an operation is not efficient. Whether or not we can devise a method for the effected parties to pay the operator to not perform the operation is not relevant; the mere fact that the benefit the operator gets is less than the cost it imposes on society makes the operation inefficient.

The second type of efficiency that we shall refer to is a market one. In an efficient market, price balances the supply of and demand for a good. Markets are assumed to be relatively efficient where information is plentiful and the markets have sufficient time to adjust. For instance, markets trading commodities or securities regularly bring supply and demand into balance through adjustment of price. Some markets do not have time to adjust. The primary example of such a market is a one-time sale. For instance, if the government is to distribute broadcast bands to TV broadcasters, the government cannot simply adjust the price as time goes on. Equilibrium will only exist if a price can be found to match the demand to the fixed supply. Such equilibrium would be efficient because the scarce goods are awarded to those
who value them most. Thus, our criterion for allocative efficiency is measuring to what extent scarce resources are allotted to those who value them most.
Chapter 3

AIRPORT CONGESTION MODELS

3.1 Six Models of Airport Congestion and Optimal Congestion Pricing

There are a variety of models of airport runway congestion in the literature. We review six such models that offer a variety of perspectives, meanings, and scientific abilities. These models can help us do one, or more, of the following:

- Evaluate historical congestion-related costs, such as the Marginal Social Cost (MSC) at an airport, given a record of its operating patterns.
- Predict queuing conditions and congestion-related costs (e.g. MSC) given a physical description of an airport and a schedule of the runway operations that the airport plans to accommodate.
- Describe how flight operators respond to congestion, with and without some form of congestion pricing.

In order to determine how well a model accomplishes one of these goals, we can establish criteria to evaluate the models. First, if a model evaluates historical congestion and its related costs, we can simply evaluate the model based on its accuracy and simplicity (or transparency). Second, if a model predicts congestion levels and congestion-related costs, we can evaluate it based on whether it:

- Allows for time-varying queue arrival rates or requires steady-state conditions,
- Treats the airplane arrival rate to the runway(s) queue(s) as deterministically following a schedule or includes a stochastic component,
- Treats the service rate of the runway(s) as deterministic or stochastic,
• Addresses the impact that inclement weather can have on an airport's service rate,

• Treats the possibility of delays rolling over from one period to another,

• Offers useful analytical expressions, and

• Explicitly recognizes the different service patterns for different types of operations (i.e. takeoffs and landings and different aircraft types).

Third, if a model describes user response to congestion, does the model assume that schedule response exists? That is, does the model treat the tendency of flight operators to avoid scheduling at the times with the longest expected wait times, all else being equal?

Lastly, in addition to how well a model incorporates each of the above three considerations, there are other qualitative issues of concern. We would like to know such things as whether the model derives from a proper physical understanding; is able to treat the case of multiple, parallel queues (i.e. multiple runways); and whether it is scalable to larger airport networks (i.e. modeling several airports with interactions between them)?

3.1.1 Busy Period Model

Carlin and Park (1970) studied congestion at La Guardia Airport (LGA) during a one-year period beginning in April of 1967. Their model starts with the observation that during a busy period in the airport (i.e. when a queue persists), each flight subsequently delays all flights that follow it during the remainder of the busy period. The impact that a marginal operation has on all flights that follow it during that busy period is none other than the MSC. Assuming that operations can be categorized on the basis of service times and delay costs per unit time into m (some finite number) types, the MSC of a type j operation is

\[
MSC_j(t) \equiv S_j \cdot \sum_{i=1}^{m} N_i(t) \cdot c_i, \quad \forall j = 1...m
\]

where:

4 An operation type is defined as either air carrier or general aviation and either takeoff or landing (Carlin and Park 1970). This model could obviously be extended to have more than these four operation types.

5 We have deviated slightly from the original notation, as we do elsewhere in this chapter, for the purposes of making the different models more consistent in notation and therefore more easily comparable.
$S_j$, the (deterministic) service time of a type $j$ operation, is determined by operating characteristics; $N_j(t)$ is the number of type $j$ operations that follow during the busy period; and $c_j$ is the delay cost per unit time for a type $j$ operation.

This equation is only particularly useful for calculating the MSC on a particular historical day at a particular time. If the operating schedule at the airport repeats on a regular basis, we can extend (3) to measure the expected MSC of an operation on a future date. Using information on average operating characteristics, Carlin and Park estimate the expected MSC of a type $j$ operation at time $t$ as

$$E[MSC_j(t)] = S_j \cdot \frac{\sum_{i=1}^{m} n_j c_i}{\sum_{i=1}^{m} n_j S_i}, \forall j=1...m \quad (4)$$

where:

$n_j$ is the average proportion of type $j$ operations during the busy period and $E[B(t)]$ is the expected remaining busy period at time $t$.

This model is simple and offers an accurate way for measuring historical MSC and predicting future MSC under repeating conditions. Moreover, its ability to measure MSC is not reliant on descriptions of queuing processes. However, it lacks any ability to predict MSC if airport conditions change. If the number of operations during some period were to change, this model would not be able to predict the effects on busy periods, estimated delays, and/or MSC. Lastly, we note that if we introduce a toll or toll set, this model would not be able to predict the impact of the toll(s) on congestion, an ability that is at the heart of the MSC toll set that we discuss in Chapter 4.

3.1.2 Regressive Model

Morrison (1983) provides a model for airports to simultaneously determine optimal prices and the optimal number of runways. At the heart of this model is a method for predicting delays during each time period of the day (e.g. hour-by-hour) based on the level of scheduled operations in that period. Morrison uses the demands in each period as the basis of the
schedule. This substitution is only appropriate for airports where any airport demanding service is admitted on a “First Come, First Served” basis; in an airport where only some of those demanding service are served (e.g. a slot-controlled airport), the demand will not be an appropriate substitute for the number of scheduled flights. The MSC of an operation of type \( j \) at time \( t \) is given by:

\[
MSC'_j = \sum_{i=1}^{m} Q'_i \frac{\partial D'_i}{\partial Q'_j} = \sum_{i=1}^{m} Q'_i \frac{\partial D'_i}{\partial Q'_j}, \quad \forall j = 1...m
\]  

where:

\( Q'_j \), the demand for a type \( j \) operation in time period \( t \), is a function of the operating cost;

\( D'_j \), the delay cost for a type \( j \) operation in period \( t \), is a function of the volume-capacity ratio;

and \( Q' = \sum_{i=1}^{m} Q'_i \).

The essence of this model is in its ability to estimate how the numbers of operations of each type affect delays. Morrison specifies a regression model for delays as a function of the number of operations and weather patterns:

\[
\log AD (GD) = \beta_1 CLR + \beta_2 PCLDY + \beta_3 CLDY + \beta_4 \\
\left( \frac{ACA}{ACRW} \right) + \beta_6 \left( \frac{ACD}{ACRW} \right) + \beta_6 \left( \frac{GAOPS}{GARW} \right)
\]  

where:

AD is the air (arrivals) delay,
GD is the ground (departures) delay,
CLR is the percentage of clear days,
PCLDY is the percentage of partially cloudy days,
CLDY is the percentage of cloudy days,
ACA is the number of air carrier arrivals,
ACD is the number of air carrier departures,
GAOPS is the number of general aviation operations (arrivals and departures),
ACRW is the number of runways for air carrier use, and
GARW is the number of runways for general aviation use.
Using delay data from one hundred US airports, Morrison (1983) uses Ordinary Least Squares Regression to estimate the parameters in the regression model. Morrison shows that the current airport fees do not cover the combination of marginal cost to the airport and the MSC of delay to other flights during peak times, but often exceed those costs during off-peak times.

Morrison and Winston (1989) extend this analysis by estimating the benefits of charging a toll to cover both the MC to the airport and the MSC of delay, with and without optimal capacity expansion. The model found that optimal pricing, with no capacity investment, could yield a $3.8 billion gain in societal welfare; however, this gain would largely come at the price of a net transfer of welfare of $11.5 billion from airlines to airports, possibly costing the traveling public in the form of higher fares. In contrast to this, optimal pricing along with optimal investment would cause an $11 billion gain in welfare at almost no expense to airlines (p. 93).

Like Carlin and Park’s model, Morrison’s is simple and not dependent on descriptions of the queuing process. However, it is significantly different in two ways. The first major difference is that this model specifies a relationship between the numbers of operations of each type and delays. Combined with an estimated demand function as in Morrison and Winston (p. 94), this could allow for finding a toll set that would result in usage levels consistent with MSC principles. The second major difference is that this model assumes that there are no inter-period interactions; delays from one period do not roll over into the next period. In some situations, this assumption may be reasonable, but in cases where there are frequently long busy periods, relative to the period of study, it will be entirely unreasonable and will not capture the essence of the congestion phenomenon. While extending this model to predict delays based on scheduled operations in that period and previous periods of the day would present difficulties because of excessive independent variables and multicollinearity, this model might be improved by using the predicted congestion in the previous period as a variable in the regression model.

---

6 A regression model is similarly specified for variable airport costs and finds that these costs exhibit constant return to scales and are small compared to congestion costs (Morrison and Winston 1989). This further justifies our focus on the congestion portion of airport tolling.

7 All of their figures are in 1988-dollar equivalents.

8 There appears to be an arithmetic error in arriving at this number and the appropriate number appears to be $8.5 billion, still an impressive gain.
3.2.3 Time-Dependent Stochastic Queuing Model

Koopman (1972) provides an early queuing-theoretic model. He assumes that arrivals to the queue can be treated as Poisson distributed with a time-varying rate, $\lambda(t)$. Koopman notes that a Poisson distribution for service times represents the extreme of randomness from the perspective of information theory and, as such, an opposite to deterministic service times. Koopman studies queuing under both of these extremes with the idea that realistic service time distributions are not deterministic, but less random than the Poisson distribution. Koopman finds the results from the two models to be fairly similar and concludes that either could be used, but argues that Poisson service times should be used because they have the added benefit of the Memoryless Property, which is useful in simulations and numerical methods. In standard queuing theory notation, this model can be written as $M(t)/M/1$.

Koopman (1972) adopts a discrete computational approach study the queues, which are completely described by a set of state probabilities, $\{P_0(t), P_1(t), \ldots, P_K(t)\}$, where $P_n(t)$ is the probability of having $n$ aircrafts in the queue at time $t$. Limiting the number of aircraft in the queue to be no more than $K$ (some relatively large number), this set of probabilities can be denoted by a $(K+1)\times 1$ vector for each time $t$. It is possible to calculate the values of this vector for any time $t$. To do so, time is divided into short intervals (relative to the expected inter-arrival and service times) and a $(K+1)\times (K+1)$ transition matrix is used for each of these intervals, where the transition matrix for time $t$ is based entirely on the arrival and service rates at time $t$. The state vector at any time is then found by multiplying a series of these transition matrices times a vector representing the initial boundary conditions of the queue being empty at some time $t=0$ (i.e. middle of the night). With full knowledge of these state probabilities, metrics such the average and standard deviation of queue length at time $t$ are easily computed. Lastly, Koopman extends the model to the case of multiple servers, or $M(t)/M/m$.

While this article does not address issues of MSC, it should be possible to use this model to estimate the effect that any one flight has on the flights that follow it. Since total delay costs are completely determined by the $\{P_0(t), P_1(t), \ldots, P_K(t)\}$, we can calculate the difference in expected total delay costs with and without any scheduled flight, which will yield the full MC.
We can also calculate the expected private delay costs for that flight. We can then find the MSC by subtracting the private delay costs from the MC.

By explicitly treating the arrivals process, service process, and inter-period interactions, this model provides a very compelling model of airport congestion that allows us to fully characterize the relationship between usage levels and queuing delays. However, the model is complex and does not lend itself to measuring actual congestion or the MSC of a current or historical airport situation. Also, because a simulation approach is adopted, there are no analytical expressions for quantities such as expected queue length or MSC, a fact that will be highly significant in our discussion in Chapter 6. Thus, the model is appropriate on the planning level, but cannot be used to report historical information such as MSC.

3.2.4 Time-Independent Stochastic Queuing Model

Jansson (1998) analyzes steady-state queuing models that primarily differ from the models studied by Koopman (1972) in that the arrivals rate is constant and does not vary by time of day and that there are no boundaries. While such a simplification is a less accurate representation of reality, it leads to models that are rich in analytical results and relationships. While Jansson looks at M/M/1 and M/G/1 systems with and without nonpreemptive priorities, we will focus our attention on the M/G/1 system without nonpreemptive priorities as an example of what types of results can be obtained. We note that M/M/1 is actually a special case of M/G/1. We provide some details of the derivation of useful expressions for the M/G/1 system for pedagogical purposes.

This model has m types of operations, with operation type j having arrival rate $\lambda_j$, service rate $\mu_j$ (service time $S_j$), and delay cost per unit time $c_j$. We find the total arrivals rate

$$\lambda = \sum_{j=1}^{m} \lambda_j, \quad (7)$$

an average value of delay time

$$\epsilon = \frac{\sum_{j=1}^{m} \lambda_j c_j}{\lambda}, \quad (8)$$

31
an average service time

\[ E[S] = \frac{1}{\mu} = \sum_{j=1}^{n} \left( \frac{\lambda_j}{\lambda} \times \frac{1}{\mu_j} \right), \]  

and an expectation for the service time squared

\[ E[S^2] = \sum_{j=1}^{n} \left( \frac{\lambda_j}{\lambda} E[S^2] \right) = \sum_{j=1}^{n} \left( \frac{\lambda_j}{\lambda} \left( \frac{1}{\mu_j^2} + \sigma_j^2 \right) \right). \]  

The utilization ratio, \( \rho \), is defined as:

\[ \rho = \frac{\lambda}{\mu}. \]  

With this information, we can define the marginal cost, private delay cost, and the marginal social cost of a type \( j \) operation:

\[
MC_j = \frac{\partial (TC)}{\partial \lambda_j} = \frac{\partial (eLq)}{\partial \lambda_j} = \frac{\partial (e\lambda \bar{w}_q)}{\partial \lambda_j} = DC_j + MSC_j, \quad \forall j=1...m
\]

where

\[
DC_j = e_j \bar{w}_q = e_j \frac{\lambda E[S^2]}{2(1-\rho)}, \quad \forall j=1...m \quad \text{and}
\]

\[
MSC_j = e\lambda \bar{w}_q = e\lambda \frac{(1-\rho)E[S^2] + \frac{\lambda}{\mu_j} E[S^2]}{2(1-\rho)^2}, \quad \forall j=1...m
\]

Equation (12) shows that we can calculate both the expected private delay cost and expected marginal social cost, for each operation type, for any possible set \( \{ \lambda \} \equiv \{ \lambda_1...\lambda_m \} \) if we know \( E[S_j], E[S^2_j] \), and \( e_j \) for each operation type \( j \).

In other words, we can compute two functions of \( \{ \lambda \} \):

\[
DC_j = f_{DC} (e_j, \{ \lambda \}) \quad \text{and} \quad MSC_j = f_{MSC} (e_1...e_m, \{ \lambda \}), \quad \forall j=1...m
\]
Jansson suggests that if the demands $\lambda_j$ are known functions of cost (i.e. fees plus $DC_j$), we can find a toll set that will result in usage levels where the MSC for each operation type is equal to the toll for that operation type. In other words, he suggests numerically solving for an optimal toll by taking into account how the airport users will respond to changes in tolls and delay costs. This is similar to the suggestion of Morrison and Winston (1989), however Jansson provides the details of how to solve this system of nonlinear equations (1998 p. 24-28). We discuss this notion of equilibrium conditions at greater length in Chapter 4.

As we previously noted, Jansson’s (1998) model sacrifices some of the detail of other models in favor of analytical expressions. This model is very useful as a pedagogical tool because it has expressions for the various costs and queue characteristics as functions of the number of each of the $m$ operation types. While this model may be too coarse for describing real situations or for evaluating the actual MSC at an airport, it is useful for demonstrating the properties of congestion, MSC, and MSC tolls in a stochastic environment.

3.2.5 Deterministic Bottleneck Model

Vickrey [1969] introduced the bottleneck model, which is based on the observation that a queue will develop when, during some period of time, the number of vehicles attempting to pass through a transportation link exceeds the maximum capacity. This idea that there is some fixed capacity at which things go from smooth to congested is a result of the deterministic nature of this model. Vickrey deals specifically with the problem of a rush-hour automobile commute, but the model has useful insights to our problem. To simplify matters, he assumes that commuters have identical valuations for the differing values of time associated with a commute: time at home, time in the office before work starts, and time in the office after work starts. Given this and the fact that commuters will minimize the overall inconvenience of commute, including both delay and schedule delay (i.e. arriving early or late), the overall cost of inconvenience must remain constant throughout the commute. If one departure time becomes more inconvenient than another, commuters will simply adjust their plans to take advantage of this. Thus, this model suggests that users will tradeoff schedule delay with queuing time and that the queue will be longest for those departing and arriving at the most desirable times. Vickrey notes that the queue will evolve according to:
\[
\frac{dq(t)}{dt} = \begin{cases} 
\frac{w_b - w_p}{w_b - w_q} = \frac{w_b - w_p}{w_b} = 1 - \frac{w_p}{w_b} & \text{if Early} \\
\frac{w_b - w_j}{w_b - w_q} = \frac{w_b - w_j}{w_b} = 1 - \frac{w_j}{w_b} & \text{if Late}
\end{cases}
\] (14)

where:

\( q(t) \) is the waiting time in the queue required to leave the bottleneck at \( t \),
\( w_b \) is the value of time at home,
\( w_p \) is the value of time at work before it starts,
\( w_j \) is the value of time at work after it starts, and
\( w_q \) is the value of time spent in the queue.

In this fully deterministic model, the queue length rises linearly from zero, peaks, and then linearly falls to zero. This model provides us with complete information such as when the \( n^{th} \) car will pass through the queue, what fraction of cars will pass through during the queue buildup and arrive early to work (vs. passing through during the queue shrinkage and arriving late), and the maximum queue waiting time.

The deterministic capacity in this model allows for a toll that causes the arrivals rate to and departures rate from the bottleneck point to equal capacity throughout the rush-hour, completely eliminating queuing. As we previously noticed, at equilibrium, the cost of the commute must remain constant throughout the commute. For the rate of commuters passing through the bottleneck to be equal to capacity, the cost of the commute must remain constant throughout the commute time and, as a result, the toll must change according to:

\[
\frac{dToll}{dt} = \begin{cases} 
\frac{w_b - w_p}{w_b} & \text{if Early} \\
\frac{w_b - w_j}{w_b} & \text{if Late}
\end{cases}
\] (15)

The net result of this toll is that users will be inconvenienced as much as before in terms of schedule delay and now pay tolls that cost exactly as much as the queuing time that they would otherwise face at the bottleneck. This toll changes nothing for the commuters, but allows society to turn the deadweight loss of the queuing into toll revenues. This toll can in turn benefit the commuters if the revenues are used to finance expansion, subsidize a preexisting flat toll, or reduce other taxes. One interesting outcome of this model is that if such a toll
were introduced, the net benefit of a capacity expansion in the bottleneck is solely found in the greater desirability of commuters’ travel times; congestion is entirely eliminated with or without the enhancement and the toll is merely a transfer of wealth from the commuters to the facility provider (Vickrey 1969).

Amott, De Palma, and Lindsey (1990) generalize Vickrey’s model to other transportation bottlenecks. Defining the costs per unit of time early and late as $\beta$ and $\gamma$, the number of commuters as $N$, and the capacity as $s$, they find that the optimal toll rises linearly from zero, when the rush hour begins, to:

$$p_p = \frac{\beta y}{(\beta + \gamma)} \frac{N}{s}$$  \hspace{1cm} (16)

at peak demand time, $t_p$, and then declines linearly to zero when rush hour ends.

Amott, De Palma, and Lindsey (1993) note that the total congestion cost in the bottleneck model is proportional to the number of users squared. Thus, the average cost that a user experiences and the MC is proportional to the number of users, with the latter being exactly twice the former in magnitude.

This group of deterministic bottleneck models is remarkable because costs are modeled *mutatis mutandis* by accounting for the fact that the system participants adjust their schedules in response to each other. As a result, the MSC is the same for all users no matter when they arrive to the queue.

The Deterministic Bottleneck Model is significantly different than the other models that we have thus far explored. Unlike the other models, this model explicitly models the behavior of those using the transportation facility; when applied to airports it can show how users trade off the inconvenience of queuing with the inconvenience of schedule adjustment. Like Jansson’s model, it establishes a clear, analytical relationship between delays and MSC on the one hand and the number of operations on the other hand.

However, this model is not without its challenges. One problem is that because it lacks a stochastic component it will have trouble modeling situations where the arrivals (or service) process is not too predictable. For example, an airport that expects sixty arrivals to the runway
between 4:00 PM and 5:00 PM and expects them to be spread out, may still find that in an interval from 4:00 PM to 4:10 PM there are sometimes only five arrivals and sometimes as many as fifteen. In other words, in a more realistic model, queuing would be possible even if the average arrivals rate to the queue did not exceed the average service rate. This kind of unpredictability leads to increased queuing that is not captured in this model. Another drawback of this model is that it may have difficulty assessing the MSC in situations where the airport usage is not driven by a series of hub-and-spoke connecting banks or by other sharp peaks in demand. In cases where the demand for scheduling flights is less peaked, the model will have difficulty predicting the delays and MSC. The extent to which this model will present this latter difficulty shall be a serious problem will depend on the particular characteristics of the airport being studied.

3.2.6 Stochastic Bottleneck Model

Daniel (1995) combines the time-dependent stochastic treatment of Koopman (1972) with the bottleneck treatment of Vickrey (1969) and others. Daniel models a hub-and-spoke airport system with a dominant airline that has several banks in a day, where each bank is characterized by a short time period when the airline tries to have many flights land and takeoff in as short a time interval to offer convenient connections to passengers. The queues for landing and takeoff are modeled with Poisson arrivals processes with a time varying rate (based on the underlying demand model) and deterministic service processes.

This model is a significant departure from the Deterministic Bottleneck Model because it assumes that the arrivals are probabilistic and that there will always be some probability of queuing unless the arrivals rate is zero. It is therefore the case that a toll will not completely eliminate queuing (unless it completely eliminates demand); the optimal toll induces some peak spreading to reduce the delays caused by a high arrivals-to-capacity ratio. An important implication is that price of using the runway(s) will go up for some operators since the toll cannot completely replace congestion costs as it does in the deterministic model.

Daniel (1995) studied airport data collecting at Minneapolis-St. Paul Airport (MSP) in May of 1990. One of Daniel’s more interesting results is that airline scheduling behavior is best described as atomistic. In other words, airlines do not seem to consider the delays that one of their flights imposes upon their other flights. This may stem from the fact that airlines believe
that if they do not schedule a flight at some specific time, some other airline might (and probably will). The airline would therefore think that it is not necessary to consider the impact that scheduling a flight will have on the airline’s other flights.

Daniel’s model treats some of the flaws in the Time-Dependent Stochastic Queuing Model (Koopman 1972) and in the Deterministic Bottleneck Model (Vickrey 1969). By combining the structure of the deterministic bottleneck model with a stochastic component, Daniel presents a very rich model. The primary drawbacks are its complexity and its assumption that Hub-and-Spoke connection banks drive scheduling, which is only sometimes appropriate.

3.3 Model Comparison

Clearly, these models offer a variety of perspectives, interpretations, and capabilities. Table 1 summarizes some of the key feature of these six models. One key point in this table is that the Busy Period model is good for measurement, but is the only model that lacks the ability to predict congestion conditions. Another point worth noting is that bottleneck models are the only models that explicitly address the behavioral adjustments to congestion.
Table 2 compares the five models that predict queuing conditions and congestion related costs. This table shows that each model has its advantage. The Time-Independent Stochastic Queuing Model is best for showing analytical results of stochastic queuing and the Deterministic Bottleneck Model is best for showing analytical results of schedule adjustment. Regression gives numerical relationships, making it almost as transparent as the analytical models, but also offers the advantages of taking into account the weather patterns and is able to capture stochastic effects by way of measurement. However, we should note that regression is severely problematic for heavily congested airports in not taking into account interperiod effects. Lastly, the two most realistic models, in as much as they incorporate the most features of the real situation, are the ones that are based on time-dependent queuing. The Time-Dependent Queuing Model offers nearly all that the steady-state one does, but is far more realistic in not assuming a consistent level of scheduled operations. This model offers the most promise for predicting queuing conditions given an actual schedule. The Stochastic Bottleneck Model offers nearly all that the deterministic one does, but is far more realistic in assuming that the real world has stochasticity. This model may be slightly less accurate than the Time-Dependent Queuing Model for predicting congestion conditions, but is able to start
from a structural demand based on a hub-and-spoke system and predict delays based on the predicted schedule, which takes into account schedule-adjusting behavior.

<table>
<thead>
<tr>
<th>Model</th>
<th>Time-varying Schedule</th>
<th>Stochastic Schedule</th>
<th>Stochastic Service</th>
<th>Weather</th>
<th>Interperiod Effects</th>
<th>Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
<td>Numerical</td>
</tr>
<tr>
<td>Steady State Queuing</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Difficult</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>Time-dependent Queuing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Possible</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Deterministic Bottleneck</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Stochastic Bottleneck</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Difficult</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 2: Comparison of Predictive Models

It is worth noting that all of the five models presented in Table 2 are based on a physical understanding of the congestion phenomenon and are therefore intuitive in some ways. Regression differs slightly from the others in that it does not explain the process of congestion, but rather assumes it to be a function of the various inputs; however, regression offers a direct relationship between congestion and the number of operations, a physical understanding that is quite appealing. Lastly, all of these five models could be used to treat cases of multiple runways, or even networks of several airports, but that the best equipped to deal with such complexities is an the Time-Dependent Queuing Model. If we have information on scheduled operations, even at many different airports, it should be possible to construct a computer simulation to predict conditions.

3.4 Conclusion

In our research, we pay particular attention to the stochastic models. We use the steady-state one for analytical results and for purposes of illustration/demonstration. The Time-Dependent Queuing model is the preferred model for realistic simulation of queuing conditions at airports. While these are our primary focus, from time to time, we make use of the other models that offer a wealth of perspectives.
Marginal Social Cost tolling has been proposed as a form of the Marginal Cost (MC) pricing for congested transportation facilities. The principle of MSC tolling is to charge for the cost of marginal delay that each operation causes for the other operations, forcing operators to consider both the internal and external costs of congestion when making usage decisions. In conjunction with charging for other marginal costs (e.g. runway wear-and-tear and air traffic control), this toll helps encourage efficient utilization.

In the context of airports, this type of toll promises several benefits. First, it should discourage over-scheduling at peak times. In the untolled environment, Levine (1969) notes that airlines schedule as many flights as possible during peak hours. Since “the airline will only experience the average, rather than the marginal, delay, measuring the cost to the [air]line of adding [an operation to] the schedule against the incremental revenue will yield a more favorable result than would be the case if the costs to all users were taken into account” (p. 91). Tolling so that the airline would experience the marginal delay, rather than the average delay, will lead airlines to reduce their schedules; profit-maximizing airlines will only schedule flights that are expected to have incremental revenue in excess of the incremental cost, including the cost to other airport users. A second notable benefit is that such tolling encourages the use of larger airplanes with greater capacity because of the economies of scale in MSC that do not exist in DC. In other words, introducing a MSC toll will shift some of the congestion related cost from DC to MSC and, whereas the former scales with aircraft size, the latter remains fairly consistent across aircraft types. The result is that the congestion cost per seat can be expected to decrease through the use of larger airplanes. Finally, airports may be able to accommodate more passengers after they implement MSC tolling through the airlines’ use of larger equipment. It is thus possible to reduce costly delays, while maintaining accessibility to high-demand facilities.
MSC tolling clearly offers great promise as a mechanism for promoting efficiency, but we must address whether this is merely an abstract concept or whether it is possible to find appropriate MSC tolls. The first part to answering this question is to show that, under some set of assumptions, we can prove that an appropriate MSC toll set exists. While there is much literature demonstrating the existence of a MSC toll in the Deterministic Bottleneck Model, we are unaware of any comparable proof for a MSC toll in the case of a stochastic queue. For this reason, we first demonstrate that for a steady-state queue, with a very limited set of assumptions, a MSC toll set must exist. Further research will be required to show whether a MSC toll set’s existence is guaranteed for more complicated queuing systems with multiple queues and/or non-steady-state conditions.

4.1 Proof: Existence of a MSC Toll Set for a Steady State Queue

We want to show that in a queuing system with continuous, downwardly sloping demand functions, a toll set exists that results in usage level consistent with MSC principles. Jansson (1998) provides a set of nonlinear, transcendental equations for finding such a toll set for an M/G/1 queue, where queue participants have heterogeneous valuations for performing an operation, values of time, and service time distributions. Jansson claims that this set of equations can be solved numerically and demonstrates results for specific scenarios, but offers no proof that, in general, this system of equations has a solution. We first demonstrate the existence of a (single) MSC toll for a queue with a convex delay function and operators with homogeneous service time distributions, but heterogeneous price sensitivities and values of time. We then demonstrate the existence of a MSC toll set for a queue with a convex delay function and operators with heterogeneous service time distributions, price sensitivities, and values of time.

4.1.1 Part I: Operators with Homogeneous Service Time Distributions

First, suppose that there is a steady-state queuing system with arrivals and service processes governed by some distributions (e.g. Poisson, deterministic, kth order Erlang, kth order hyperexponential, or some general distribution) and that there are m parallel servers. We further suppose that successive interarrival times and service times are independently and identically distributed. We further stipulate that the following conditions that must be satisfied:
1. The heterogeneity of users can be described by \( m \), a finite number, of different types of operations, with these types forming a set \( M \equiv \{1...m\} \). Operation type \( j \) has average arrival rate \( \lambda_j \) and value of time (i.e. cost per unit time delayed) \( \epsilon_j \). The service rate, \( \mu \), is homogeneous (i.e. same for all users).

The total arrival rate, \( \lambda \), and weighted average value of time, \( \epsilon \), are given by:

\[
\lambda = \sum_{j=1}^{m} \lambda_j \quad \text{and} \quad \epsilon = \frac{\sum_{j=1}^{m} \lambda_j \epsilon_j}{\lambda}.
\]

(17)

The expected queuing time is dependent on the arrivals rate and distribution, service rate and distribution, and the number of servers. The service distribution, average service rate, and the number of servers are fixed, whereas the arrivals rate is a variable in our system. We thus express the expected queuing time as a function of the arrivals rate:

\[
\overline{W}_q = \overline{W}_q(\lambda)
\]

(18)

where:

\[
\frac{\partial \overline{W}_q(\lambda)}{\partial \lambda_j} > 0, \quad \forall j \in M \quad \text{and}
\]

\[
\overline{W}_q(\lambda = 0) = 0.
\]

(19)

2. The expected queue length is convex in \( \lambda \) and given by Little’s Law:

\[
\overline{L}_q(\lambda) = \lambda \overline{W}_q(\lambda).
\]

(20)

3. The external cost of congestion (or MSC) for operation type \( j \) can be written as:

\[
MSC_j \equiv \epsilon \lambda \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda_j} = \sum_{i=1}^{m} \epsilon_i \lambda_i \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda_j}, \quad \forall j \in M.
\]

(21)

Because the service is homogeneous, we can simplify the above expression to:
MSC_j = MSC \equiv c_j \frac{\partial \overline{W}_q (\lambda)}{\partial \lambda} = \sum_{i=1}^{n} c_j \lambda_i \frac{\partial \overline{W}_q (\lambda)}{\partial \lambda}, \quad \forall j \in M . \quad (22)

4. Each operation type faces a price for using the facility, composed of the expected delay costs \(^9\) and the toll (\(\tau\)):

\[
\text{Price}_j \equiv DC_j (\lambda) + \tau = c_j \overline{W}_q (\lambda) + \tau, \quad \forall j \in M .
\] \quad (23)

5. The demand for operations of each type, and hence the average arrival rate for that operation type, can be expressed as a function of the price:

\[
\lambda_j \equiv \lambda_j (\text{Price}_j) = \lambda_j (DC_j + \tau) = \lambda_j (c_j \overline{W}_q (\lambda) + \tau), \quad \forall j \in M
\]

where:

1. \(\lambda_j (0) > 0, \quad \forall j \in M\),

2. \(\lim_{\text{Price}_j \to \infty} \lambda_j (\text{Price}_j) = 0, \quad \forall j \in M\), and

3. \(\frac{\partial \lambda_j (\text{Price}_j)}{\partial \text{Price}_j} \leq 0, \quad \forall j \in M\).

One thing worth noting about this system is that equation (24) is in transcendental form; \(\overline{W}_q (\lambda)\) is a function of \(\lambda\), but each of the demands, \(\lambda_j\), are a function of \(\overline{W}_q (\lambda)\). In order for the demand functions given in (24) to be proper functions, we will express them as a function of \(\tau\) exclusively. To do so, we need to show that these functions of \(\tau\) are uniquely valued or, in other words, pass the Vertical Line Test. The proof is as follows:

---

\(^9\) Operators may also weigh such metrics as estimated variance of delay costs, but we simplify the model by assuming that the only quantity of interest is the expected delay.
Suppose \( \exists \tau, j, \text{ s.t. } \lambda_j^1 \left( \overline{W}_q \left( \lambda^1(\tau) \right), \tau \right) \neq \lambda_j^2 \left( \overline{W}_q \left( \lambda^2(\tau) \right), \tau \right) \)

Case 1: \( \lambda^1(\tau) = \lambda^2(\tau) \),

Then \( \text{Price}^1_j(\tau) = \text{Price}^2_j(\tau), \forall i \in M \)

Therefore \( \lambda_j^1 \left( \text{Price}^1_j(\tau) \right) = \lambda_j^2 \left( \text{Price}^2_j(\tau) \right), \forall i \in M \) - Contradicting our assumption!

Case 2: \( \lambda^1(\tau) \neq \lambda^2(\tau) \) and without loss of generality, \( \lambda^1(\tau) > \lambda^2(\tau) \).

Then \( \text{Price}^1_j(\tau) > \text{Price}^2_j(\tau), \forall i \in M \)

Therefore \( \lambda_j^1 \left( \text{Price}^1_j(\tau) \right) \leq \lambda_j^2 \left( \text{Price}^2_j(\tau) \right), \forall i \in M \)

Therefore \( \lambda^1(\tau) \leq \lambda^2(\tau) \) - Contradicting our assumption!

We can now write the demand functions as implicit functions of toll:

\[
\lambda_j \equiv \lambda_j^1(\tau), \forall j \in M. \tag{25}
\]

With these assumptions and definitions, we want to prove that there exists some toll where the resulting MSC will be equal to the toll, satisfying the optimality conditions. To accomplish this, we will first demonstrate that, for \( \tau \in [0, \infty) \), the MSC is:

1. Continuous in \( \tau \),
2. Positive when there is no toll, and
3. Zero in the limit of \( \tau \) going to infinity.

Lemma 1: MSC is continuous in \( \tau \). This proof is in four steps:

The first step is to prove that \( \overline{W}_q(\lambda) \) is monotonically non-increasing in \( \tau \), which we do by way of contradiction:

Suppose \( \exists \tau, \varepsilon > 0, \tau' = \tau + \varepsilon \text{ s.t. } \overline{W}_q(\lambda(\tau')) > \overline{W}_q(\lambda(\tau)) \)

Then \( \text{Price}_j(\tau') > \text{Price}_j(\tau), \forall j \in M \)

Then \( \lambda_j(\tau') \leq \lambda_j(\tau), \forall j \in M \)

Then \( \overline{W}_q(\lambda(\tau')) \leq \overline{W}_q(\lambda(\tau)), \) contradicting our assumption!

44
The second step is to show that \( \overline{W}_q(\lambda) \) is continuous in \( \tau \), which we do by way of contradiction:

\[
\text{Suppose } \exists \tau, \epsilon > 0, \tau' = \tau + \epsilon \text{ s.t. } \lim_{\epsilon \to 0} \overline{W}_q(\lambda(\tau')) < \overline{W}_q(\lambda(\tau))
\]

Then \( \lim_{\epsilon \to 0} DC_j(\lambda(\tau')) < DC_j(\lambda(\tau)), \ \forall j \in M \)

Then \( \lim_{\epsilon \to 0} Price_j(\lambda(\tau')) = \lim_{\epsilon \to 0} DC_j(\lambda(\tau')) + \lim_{\epsilon \to 0} (\tau') = \lim_{\epsilon \to 0} DC_j(\lambda(\tau')) + \tau < DC_j(\lambda(\tau)) + \tau = Price_j(\lambda(\tau)), \ \forall j \in M \)

Then \( \lim_{\epsilon \to 0} \lambda_j(\tau') = \lambda_j(\tau), \ \forall j \in M \)

Then \( \lim_{\epsilon \to 0} \overline{W}_q(\lambda(\tau')) \geq \overline{W}_q(\lambda(\tau)), \) contradicting our assumption!

The third step is to prove that the demand for each operation type is continuous in \( \tau \):

\[ \overline{W}_q(\lambda(\tau)) \text{ is continuous in } \tau. \]

Therefore \( DC_j(\lambda(\tau)) \) is continuous in \( \tau \), \( \forall j \in M \).

Therefore \( Price_j(\tau) \) is continuous in \( \tau \), \( \forall j \in M \).

Since \( \lambda_j(Price_j) \) is continuous in \( Price_j \) and \( Price_j(\tau) \) is continuous in \( \tau \),

\[ \lambda_j(Price_j(\tau)) \text{ is continuous in } \tau, \ \forall j \in M. \]

We can now complete the proof:

For \( \overline{L}_q(\lambda) \) to be convex in \( \lambda \),

\[ \frac{\partial^2 \overline{L}_q(\lambda)}{\partial \lambda^2} = 2 \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda} + \lambda \frac{\partial^2 \overline{W}_q(\lambda)}{\partial \lambda^2} \]

must be positive and well defined.

Therefore \( \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda} \) must be continuous in \( \lambda \).

Since \( \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda} \) is continuous in \( \lambda \) and \( \lambda \) is continuous in \( \tau \),

\[ \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda} \text{ is continuous in } \tau. \]

Therefore \( \lambda_i \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda} \) is continuous in \( \tau \), \( \forall i \in M \).

Therefore \( \sum_{i=1}^{w} \epsilon_i \lambda_i \frac{\partial \overline{W}_q(\lambda)}{\partial \lambda} \) is continuous in \( \lambda \).
Lemma 2: MSC is positive when there is no toll. Proof:

Suppose  
\[ MSC(\tau = 0) = \sum_{i=1}^{n} c_i \lambda_i(\tau = 0) \frac{\partial W_q}{\partial \lambda}(\lambda(\tau = 0)) = 0 \]

Note that \( \frac{\partial W_q}{\partial \lambda}(\lambda) > 0, \ \forall j \in M \) by assumption.

Therefore  \( \lambda_i(\tau = 0) = 0, \ \forall j \in M \).

Therefore  \( Price_j(\tau = 0) = c_j W_q(\lambda = 0 + 0 = 0, \ \forall j \in M \).

But this contradicts our assumption that \( \lambda_j(Price = 0) > 0, \ \forall j \in M \). 

Lemma 3: MSC is zero in the limit of \( \tau \) going to infinity. Proof:

\[ \lim_{\tau \to \infty} Price_j(\tau) = \lim_{\tau \to \infty} DC_j(\lambda(\tau)) + \lim_{\tau \to \infty} (\tau) = \infty, \ \forall j \in M. \]

\[ \lim_{\tau \to \infty} \lambda_j(Price) = 0, \ \forall j \in M. \]

Therefore  \( \lim_{\tau \to \infty} \lambda_j(\tau) = 0, \ \forall j \in M. \)

Therefore  \( \lim_{\tau \to \infty} MSC(\tau) = \lim_{\tau \to \infty} \sum_{i=1}^{n} c_i \lambda_i(\tau) \frac{\partial W_q}{\partial \lambda}(\lambda(\tau)) = 0. \)

We have now shown that the MSC is continuous, approaches infinity as \( \tau \) approaches 0, and approaches 0 as \( \tau \) approaches infinity. We can define a new variable

\[ x \equiv x(\tau) = \frac{MSC(\tau)}{\tau} \]

where:

1. \( x \) is continuous in \( \tau \) for \( \tau \geq 0 \),
2. \( \lim_{\tau \to 0^+} x = \infty \), and
3. \( \lim_{\tau \to \infty} x = 0. \)

The Intermediate Value Theorem than guarantees that

\[ \exists \tau \in [0, \infty) \text{ s.t. } x = 1, \text{ or equivalently,} \]

\[ \exists \tau \in [0, \infty) \text{ s.t. } MSC(\tau) = \tau \]

(27)
We have thus proved the existence of an optimal toll in a steady-state queue with heterogeneous price sensitivities and values of time with only a few, mild assumptions. We now show that, when service time distributions are heterogeneous, there exists a set of m tolls to charge for the m types or operations that will result in operating levels with MSC for each operation type equal to the toll for that operation type.

4.1.2 Part II: Operators with Heterogeneous Service Time Distributions

Suppose that there is a queuing system like the one above, with a few modifications:

1. Service rates are operation-type specific, with operation type \( j \) having service rate \( \mu_j \). The average service rate is

\[
\mu = \left( \frac{\sum_{j=1}^{n} \left( \frac{\lambda_j}{\mu_j} \right) \mu_j}{\sum_{j=1}^{n} \lambda_j} \right)^{\frac{1}{4}}. \tag{28}
\]

2. We define \( \hat{\lambda} \equiv \{ \lambda_1, \ldots, \lambda_m \} \) as the set of demands from the m user types.

3. The expected queuing time is dependent on the arrival rates and distributions, service rates and distributions, and the number of servers. (28) shows that the service rate is dependent on the arrival rates of the different operation types. Thus, we can express the expected queuing time as a function of the set of demands from the m user types:

\[
\overline{W_q} = \overline{W_q}(\hat{\lambda}). \tag{29}
\]

where:

\[
\frac{\partial \overline{W_q}(\hat{\lambda})}{\partial \lambda_j} > 0, \forall j \in M \quad \text{and} \quad \overline{W_q}(\hat{\lambda} = 0) = 0. \tag{30}
\]

4. The expected queue length is

\[
\overline{L_q}(\hat{\lambda}) = \lambda \overline{W_q}(\hat{\lambda}), \tag{31}
\]
where \( \overline{L}_q(\hat{\lambda}) \) must be convex in each \( \lambda_j \). Moreover, the mixed second-order partial derivates of the form:

\[
\frac{\partial^2 \overline{L}_q(\hat{\lambda})}{\partial \lambda_i \partial \lambda_j} \text{ must be well defined, } \forall i, j \in M.
\] (32)

All of the basic steady state queuing models with which we are familiar (including M/M/1, M/M/1/K, M/M/m, M/M/m/K, and M/G/1) satisfy (32). This condition prevents having expressions for expected queue length of the form:

\[
\overline{L}_q(\hat{\lambda}) = f_1(\hat{\lambda}_1) + f_2(\hat{\lambda}_2)
\]

where:

1. \( \hat{\lambda}_1 \subset \hat{\lambda} \) and \( \hat{\lambda}_1 \neq \emptyset \),
2. \( \hat{\lambda}_2 \subset \hat{\lambda} \) and \( \hat{\lambda}_2 \neq \emptyset \), and
3. \( \hat{\lambda}_1 \cup \hat{\lambda}_2 = \hat{\lambda} \) and \( \hat{\lambda}_1 \cap \hat{\lambda}_2 = \emptyset \).

5. The price for performing an operation is composed of the expected delay costs\(^{10}\) and the type-specific toll \((\tau_j)\):

\[
\text{Price}_j = DC_j(\hat{\lambda}) + \tau_j = c_j \overline{W}_q(\hat{\lambda}) + \tau_j, \quad \forall j \in \{1...m\}.
\] (34)

The set of tolls, \( \{\tau_1, \tau_2...\tau_m\} \), is represented by the notation \( \hat{\tau} \).

6. The external cost of congestion, the MSC, can be written as:

\[
\text{MSC}_j = c \lambda \frac{\partial \overline{W}_q(\hat{\lambda})}{\partial \lambda_j}, \quad \forall j \in M.
\] (35)

We previously defined the variable \( x \) for the case of one MSC and one toll. In this model with m different values of MSC and m different tolls, we define the set

\(^{10}\) Users may also weigh such metrics as estimated variance of delay costs, but we will simplify this model by assuming that the only quantity of interest is the expected delay.
\[ \hat{x} \equiv \{x_1, x_m\} \], where:
\[ x_j \equiv x_j(\hat{\tau}) = \frac{MSC_j(\hat{\tau})}{\tau_j}, \quad \forall j \in M. \]  

(36)

The quantity of demand for each user type is expressed as a function of the price:

\[ \lambda_j \equiv \lambda_j(Price_j) = \lambda_j(DC_j(\hat{\lambda}) + \tau_j) = \lambda_j(c_j \overline{W}_q(\hat{\lambda}) + \tau_j), \quad \forall j \in M \]

where:

1. \( \lambda_j(0) > 0, \quad \forall j \in M \),
2. \( \lim_{Price_j \to \infty} \lambda_j(Price_j) = 0, \quad \forall j \in M \), and
3. \( \frac{\partial \lambda_j(Price_j)}{\partial Price_j} < 0, \quad \forall j \in M \).

(37)

Like (24), (37) gives an expression for the demand in transcendental form; \( \overline{W}_q(\hat{\lambda}) \) is a function of \( \hat{\lambda} \), but each of the demands, \( \lambda_j \), is a function of \( \overline{W}_q(\hat{\lambda}) \). In order for the demand functions to be proper functions, we will write them as implicit functions of \( \hat{\tau} \). To do so, we need to show that these functions of \( \hat{\tau} \) are uniquely valued or, in other words, pass the Vertical Line Test. The proof is as follows:

Suppose \( \exists \hat{\tau}, j \), s.t. \( \lambda^1_j(\overline{W}_q(\hat{\lambda}^1(\hat{\tau})), \tau_j) \neq \lambda^2_j(\overline{W}_q(\hat{\lambda}^2(\hat{\tau})), \tau_j) \)

Case 1: \( \overline{W}_q(\hat{\lambda}^1(\hat{\tau})) = \overline{W}_q(\hat{\lambda}^2(\hat{\tau})) \)

Then \( Price_i^1(\hat{\tau}) = Price_i^2(\hat{\tau}), \quad \forall i \in M \)

Therefore \( \lambda^1_i(Price_i^1(\hat{\tau})) = \lambda^2_i(Price_i^2(\hat{\tau})), \quad \forall i \in M - Contradicting our assumption! \)

Case 2: \( \overline{W}_q(\hat{\lambda}^1(\hat{\tau})) \neq \overline{W}_q(\hat{\lambda}^2(\hat{\tau})) \) and without loss of generality, \( \overline{W}_q(\hat{\lambda}^1(\hat{\tau})) > \overline{W}_q(\hat{\lambda}^2(\hat{\tau})) \).

Then \( Price_i^1(\hat{\tau}) > Price_i^2(\hat{\tau}), \quad \forall i \in M \)

Therefore \( \lambda^1_i(Price_i^1(\hat{\tau})) \leq \lambda^2_i(Price_i^2(\hat{\tau})), \quad \forall i \in M \)

Therefore \( \overline{W}_q(\hat{\lambda}^1(\hat{\tau})) \leq \overline{W}_q(\hat{\lambda}^2(\hat{\tau})) - Contradicting our assumption! \)

We can now write the demand functions as implicit functions of the toll set:

49
\( \lambda_j \equiv \lambda_j (\hat{\tau}), \forall j \in M. \) (38)

With these assumptions and definitions, we want to prove that there exists some set of tolls where the MSC for each operation type is equal to the toll for that type, thereby satisfying the optimality conditions. To accomplish this, we will demonstrate that, for \( \hat{\tau} \in [0, \infty) \), \( x_j (\hat{\tau}) \):

1. is continuous in \( \tau_k \), \( \forall j, k \in M \),
2. converges uniformly to infinity as \( \tau_j \) goes to zero, \( \forall j \in M \),
3. and converges uniformly to zero as \( \tau_j \) goes to infinity, \( \forall j \in M \).

Lemma 1: \( x_j (\hat{\tau}) \) is continuous in \( \tau_k \), \( \forall j, k \in M \). This proof is in 4 steps:

The first step is to show that \( \bar{W}_j (\hat{\lambda}) \) is monotonically non-increasing in \( \tau_j \), \( \forall j \in M \), which we show by way of contradiction:

Let \( \tau'_j = \tau_j + \delta_j \cdot \epsilon \) where \( \delta_j \) is the Kronicher delta.

Suppose \( \exists \hat{\tau}, j, \epsilon > 0, \hat{\tau}'_j = \{\tau'_1, \ldots, \tau'_m\} \) s.t. \( \bar{W}_j (\hat{\lambda}(\hat{\tau}')) > \bar{W}_j (\hat{\lambda}(\hat{\tau})) \)

Then \( Price_k (\hat{\tau}') > Price_k (\hat{\tau}), \forall j, k \in \{1 \ldots m\} \)

Then \( \lambda_k (\hat{\tau}') \leq \lambda_k (\hat{\tau}), \forall j, k \in \{1 \ldots m\} \)

Then \( \bar{W}_j (\hat{\lambda}(\hat{\tau}')) \leq \bar{W}_j (\hat{\lambda}(\hat{\tau})) - \text{Contradicting our assumption!} \)

The second step is to show that \( \bar{W}_j (\hat{\lambda}) \) is continuous in \( \tau_j \), \( \forall j \in M \), which we do by way of contradiction:

Suppose \( \exists \hat{\tau}, j, \epsilon > 0, \hat{\tau}'_j = \{\tau'_1, \ldots, \tau'_m\} \) s.t. \( \lim_{\epsilon \to 0} \bar{W}_j (\hat{\lambda}(\hat{\tau}')) < \bar{W}_j (\hat{\lambda}(\hat{\tau})) \)

Then \( \lim_{\epsilon \to 0} DC_k \left( \hat{\lambda}(\hat{\tau}') \right) < DC_k \left( \hat{\lambda}(\hat{\tau}) \right), \forall k \in \{1 \ldots m\} \)

Then \( \lim_{\epsilon \to 0} Price_k (\hat{\tau}') = \lim_{\epsilon \to 0} DC_k \left( \hat{\lambda}(\hat{\tau}') \right) + \lim_{\epsilon \to 0} (\tau_k + \delta_j \cdot \epsilon) = \lim_{\epsilon \to 0} DC_k \left( \hat{\lambda}(\hat{\tau}') \right) + \tau_k \)

Then \( \lim_{\epsilon \to 0} \lambda_k (\hat{\tau}') \geq \lambda_k (\hat{\tau}), \forall k \in M \)

Then \( \lim_{\epsilon \to 0} \bar{W}_j (\hat{\lambda}(\hat{\tau}')) \geq \bar{W}_j (\hat{\lambda}(\hat{\tau})) - \text{Contradicting our assumption!} \)
The third step is to show that \( \lambda_i(\mathbf{t}) \) is continuous in \( \tau_j, \forall i, j \in M \).

Proof:

\[
\overline{W}_q(\dot{\lambda}(\mathbf{t})) \text{ is continuous in } \tau_j, \forall j \in M.
\]

Therefore \( D_{C_i}(\dot{\lambda}(\mathbf{t})) \) is continuous in \( \tau_j, \forall i, j \in M \).

Therefore \( P_i(\dot{\lambda}(\mathbf{t})) \) is continuous in \( \tau_j, \forall i, j \in M \).

Since \( \lambda_i(\lambda_i) \) is continuous in \( \lambda_i \) and \( P_i(\dot{\lambda}(\mathbf{t})) \) is continuous in \( \tau_j, \forall i, j \in M \),

\[
\lambda_i(\lambda_i) \text{ is continous in } \tau_j, \forall i, j \in M.
\]

We can now complete the proof:

\[
\chi_j(\mathbf{t}) = \frac{MSC_j(\mathbf{t})}{\tau_j}
\]

\[
MSC_j(\mathbf{t}) = \sum_{i=1}^{n} c_i \lambda_i(\mathbf{t}) \frac{\partial \overline{W}_q(\dot{\lambda})}{\partial \lambda_i(\mathbf{t})} = c_i \lambda_i \frac{\partial \overline{W}_q(\dot{\lambda})}{\partial \lambda_i(\mathbf{t})}
\]

By assumption, \( \frac{\partial^2 \overline{W}_q(\dot{\lambda})}{\partial \lambda_j(\dot{\lambda}) \partial \lambda_k(\dot{\lambda})} = \frac{\partial \overline{W}_q(\dot{\lambda})}{\partial \lambda_j(\dot{\lambda})} \frac{\partial \overline{W}_q(\dot{\lambda})}{\partial \lambda_k(\dot{\lambda})} + 2 \lambda \frac{\partial^2 \overline{W}_q(\dot{\lambda})}{\partial \lambda_j(\dot{\lambda}) \partial \lambda_k(\dot{\lambda})} \) exists, \( \forall j, k \in M \).

Therefore \( \frac{\partial \overline{W}_q(\dot{\lambda})}{\partial \lambda_j(\dot{\lambda})} \) must be continuous in \( \lambda_k \), which is in turn continous in \( \tau_k, \forall j, k \in M \).

Therefore \( MSC_j(\mathbf{t}) = \sum_{i=1}^{n} c_i \lambda_i(\mathbf{t}) \frac{\partial \overline{W}_q(\dot{\lambda})}{\partial \lambda_i(\mathbf{t})} \) is continuous in \( \tau_k, \forall j, k \in M \).

Therefore \( \chi_j(\mathbf{t}) = \frac{MSC_j(\mathbf{t})}{\tau_j} \) is continuous in \( \tau_k, \forall j, k \in M \).

Lemma 2: \( \chi_j(\mathbf{t}) \) converges uniformly to infinity as \( \tau_j \) goes to zero, \( \forall j \in M \). Proof:
Equivalently, we show that for any finite \( C > 0 \),
\[
\exists \varepsilon_j > 0 \text{ s.t. } x_j (\hat{t}, \tau_j = \varepsilon_j) \geq C, \forall \{\hat{t} \setminus \tau_j\}, j \in M.
\]

Note that \( \frac{\partial \overline{W}_q (\hat{\lambda})}{\partial \lambda_j} > 0 \), \( \forall \hat{\lambda}, j \in M \). Therefore \( \exists \delta_j > 0 \) s.t. \( \frac{\partial \overline{W}_q (\hat{\lambda})}{\partial \lambda_j} > \delta_j, \forall \hat{\lambda}, j \in M \).

Let \( \xi \equiv \min \{\xi_j\} \); \( \lambda_{\text{min}} \equiv \lambda^{-1}_{\text{min}} (\lambda) = \min \{\lambda\} \) s.t. \( \overline{W}_q (\hat{\lambda}) = \lambda \); \( \overline{W}_{q_{\text{min}}} = \min \{\overline{W}_q (\hat{\lambda} (\hat{\tau}))\}; \hat{\lambda} \equiv \lambda_{\text{min}} (\overline{W}_{q_{\text{min}}}); \) and \( \eta_j \equiv \xi_j \cdot \hat{\lambda} \cdot \delta_j, \forall j \in M \).

Then \( MSC_j (\hat{\tau}) \geq \eta_j \) and \( x_j (\hat{\tau}, \tau_j) \geq \frac{\eta_j}{\tau_j}, \forall \hat{\tau}, j \in M \).

For any finite \( C > 0 \), let \( \varepsilon_j \equiv \frac{\eta_j}{C} \).
\( x_j (\hat{\tau}, \tau_j = \varepsilon_j) \geq C, \forall \{\hat{\tau} \setminus \tau_j\}, j \in M - \) satisfying our requirement.

Lemma 3: \( x_j (\hat{\tau}) \) converges uniformly to zero as \( \tau \) goes to infinity, \( \forall j \in M \). Proof:

Equivalently, we show that for any \( \varepsilon > 0 \), \( \exists C > 0 \) s.t. \( x_j (\hat{\tau}, \tau_j = C) \leq \varepsilon, \forall \{\hat{\tau} \setminus \tau_j\}, j \in M \).

Note that \( \frac{\partial \overline{W}_q (\hat{\lambda})}{\partial \lambda_j} \) is finite, \( \forall \hat{\lambda}, j \in M \).

Therefore \( \exists \Delta_j > 0 \) s.t. \( \frac{\partial \overline{W}_q (\hat{\lambda})}{\partial \lambda_j} \leq \Delta_j, \forall \hat{\lambda}, j \in M \).

Let \( \overline{\tau} \equiv \max \{\xi_j\} \); \( \lambda_{\text{max}} \equiv \lambda^{-1}_{\text{max}} (\lambda) = \max \{\lambda\} \) s.t. \( \overline{W}_q (\hat{\lambda}) = \lambda \);
and \( \overline{W}_{q_{\text{max}}} = \max \{\overline{W}_q (\hat{\lambda} (\hat{\tau}))\} \)

It is trivial to show that \( \overline{W}_{q_{\text{max}}} (\hat{\lambda}) = \overline{W}_q (\hat{\lambda} (\hat{\tau} = 0)) \), where \( \hat{\tau} = 0 \) is a set of \( m \) zeros.

Let \( \overline{\lambda} \equiv \lambda_{\text{max}} (\overline{W}_q (\hat{\lambda} (\hat{\tau} = 0))) \) and \( H_j \equiv \overline{\tau} \cdot \overline{\lambda} \cdot \Delta_j \).

Then \( MSC_j (\hat{\tau}) \leq H_j \) and \( x_j (\hat{\tau}, \tau_j) \leq \frac{H_j}{\tau_j}, \forall \hat{\tau}, j \in M \)

For any \( \varepsilon > 0 \), let \( C_j = \frac{H_j}{\varepsilon} \).
\( x_j (\hat{\tau}, \tau_j = C) \leq \varepsilon, \forall \{\hat{\tau} \setminus \tau_j\}, j \in M - \) satisfying our requirement.
We have now established that $\forall j \in M$, $x_j(\hat{\tau})$:
1. is continuous in $\tau_k$, $\forall j, k \in M$,
2. converges uniformly to infinity as $\tau_j$ goes to zero, $\forall j \in M$,
3. and converges uniformly to zero as $\tau_j$ goes to infinity, $\forall j \in M$.

With the Implicit Function and Intermediate Value Theorems, we have enough to show that:

$$\exists \hat{\tau} \in [0, \infty) \text{ s.t. } \hat{\tau} = \hat{t}, \text{ or equivalently,}$$
$$\exists \hat{\tau} \in [0, \infty) \text{ s.t. } MSC_j(\hat{\tau}) = \hat{t}, \forall j \in M. \quad (39)$$

We have thus, with a mild set of assumptions, proved the existence of a Marginal Social Cost toll set in a steady-state queue with heterogeneous service time distributions, price sensitivities, and values of time.

Now that we have shown the MSC toll set to exist in the steady-state queuing model, there are several important issues worth addressing with regard to these tolls. First, we discuss how difficulties in estimating demand functions might affect our ability to find MSC tolls. Second, we examine the implication of the fact that demand in the real world comes in discrete units. Third, we discuss the possibility of more complex congestion models that may not have MSC toll sets, even with a similar set of assumptions to the ones we made above. Lastly, we explore whether the MSC toll set might be degenerate and if so what the implications are. This last point is the most important, because if the MSC toll set is degenerate, it could present a problem that is both practical and theoretical in nature. Degeneracy, as we shall see, could undermine the basis underlying principles of MSC tolling.

4.2 Estimating Demand Functions

As we have shown above, finding a MSC toll set requires solving a set of nonlinear, transcendental equations. These equations are based on the congestion model, categorization of operation types, and demand functions for each of these operation types. Thus, in order to find a MSC toll set, all potential operations need to be divided into a discrete number of categories based on service distributions and sensitivity to time. It may be sufficient to do this on the basis of aircraft types and intended use (takeoff or landing) since aircraft of the same type performing the same runway operation should have similar service time distributions and
costs of delay. The next step is to estimate how the demand for each operation type responds to overall cost of using the facility. Even if econometric methods are used to estimate such a demand function, small errors can make finding a somewhat accurate MSC toll set quite difficult if the demand is highly elastic with regard to price. In other words, if demand is highly elastic, small deviations from the correct MSC toll will cause a significant inequality between the MSC and the toll. In such a case, it may take a lot of trial-and-error (i.e. set a toll set, observe resulting demand, adjust toll set, observe resulting demand...) to find a toll set that is consistent with MSC principles. Because the underlying demand for the airport changes on a seasonal basis and with changes in economic and demographic conditions, there may not be sufficient time to ever settle on an appropriate toll set. This problem could become even more difficult if a more realistic, but complex, congestion model were used.

4.3 Discrete Nature of Demand

Even assuming that it is possible to estimate a demand function, the assumption that demand is continuous may not be realistic. If the demand is of a discrete nature, there may be situations where a certain toll is below MSC, but raising the toll, even a little, would cause the MSC to drop to below the toll. To demonstrate why this is so, we again return to the analytically simple steady-state queuing model to provide an example.

In this example, there is one runway with consistently even, Poisson-generated demand throughout the day. The runway has a Poisson service process for all operations with a service rate of 100 operations per hour. Thus, the M/M/1 steady state model appropriately describes queuing for this runway. The airport authority charges an identical toll for each operation. Lastly, all operations have an identical delay cost rate of \( c = \$2000\) per hour and there is a downward sloping, discrete demand function that describes the demand response to cost. For the purposes of this example, we need not actually provide the demand function. Basic results from the M/M/1 queue tell us that the expected queue length and wait times are

\[
L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad \text{and} \quad W_q = \frac{\lambda}{\mu(\mu - \lambda)},
\]

where \( \lambda \) is the arrivals rate. If \( c \) is the cost per unit of queuing time, the total cost is
TC = cL_q = c \frac{\lambda^2}{\mu(\mu - \lambda)}. \quad (41)

Since \lambda is a discrete variable, the Marginal cost of the \lambda^{th} operation per hour is:

\[ MC = c \left[ \overline{L_q}(\lambda) - \overline{L_q}(\lambda - 1) \right] = c \frac{\lambda^2 (\mu - \lambda + 1) - (\lambda - 1)^2 (\mu - \lambda)}{\mu(\mu - \lambda)(\mu - \lambda + 1)} \]

\[ = c \frac{\lambda^2 \mu - \lambda^3 + \lambda^2 - (\lambda^2 \mu - \lambda^3 + 2\lambda^2 - 2\lambda\mu + \mu - \lambda)}{\mu(\mu - \lambda)(\mu - \lambda + 1)} \]

\[ = c \frac{-2\lambda\mu + \lambda - \lambda^2 - \mu}{\mu(\mu - \lambda)(\mu - \lambda + 1)} \quad (42) \]

The Marginal Cost can be broken into the usual components:

\[ MC = DC + MSC \quad \text{where} : \]

\[ DC = c\overline{W_q} = c \frac{\lambda}{\mu(\mu - \lambda)} = c \frac{\mu\lambda - \lambda^2 + \lambda}{\mu(\mu - \lambda)(\mu - \lambda + 1)} \]

and

\[ MSC = c(\lambda - 1)\left[ \overline{W_q}(\lambda) - \overline{W_q}(\lambda - 1) \right] \]

\[ = c(\lambda - 1) \frac{\lambda(\mu - \lambda + 1) - (\lambda - 1)(\mu - \lambda)}{\mu(\mu - \lambda)(\mu - \lambda + 1)} \]

\[ = c \frac{\mu\lambda - \mu}{\mu(\mu - \lambda)(\mu - \lambda + 1)} \quad (43) \]

Table 3 shows two tolls that the airport may try for this system and why neither works. The first toll is $1,618.18, which is equal to MSC(\lambda=90, \mu=100), and results in an equilibrium demand of 93. The private DC is $265.71, the MSC is $3,285.71, and the total cost of using the runway is $1,883.90 (DC and toll). Clearly, the toll is too low because the MSC is more than twice the toll. The second toll is $2,000, which is equal to MSC(\lambda=91, \mu=100), and results in an equilibrium demand of 85. The private DC is $113.33, the MSC is $700, and the total cost of using the runway is $2,113.33. Demand dropped considerably because of the substantial increase in cost of using the runway. Clearly, the second toll is too high, as it is nearly three times the MSC.
Clearly, neither of those tolls is appropriate. Instead, we might choose to charge the lowest possible toll that results in having $\lambda$ equal to 90, which will be somewhere between these two tolls. This too is problematic. First, we will only be able to find such a toll in cases where we have a very fine knowledge of the demand curve. Second, it apparently violates MSC principles by assessing a toll that is in excess of MSC. This topic is revisited in Chapter 6.

### 4.4 More Complex Models

Even if we assume that the demand functions are continuous and that we can have complete knowledge of the demand functions, we must address the question of whether there is always a MSC toll set. For instance, in more complex congestion models, the possibility remains that the mixed partial derivatives of the queue length will not be well defined, leading to the existence of some $j, k$ for which $x_j(\hat{r})$ is not continuous in $\tau_k$. In such a case, we could not guarantee the existence of a MSC toll set.

### 4.5 Solution Degeneracy

Lastly, there is the problem of solution degeneracy. Even under the relatively simple model with homogeneous service time distributions, which we previously explored, there may be cases where there are multiple MSC toll sets. Thus, using numerical methods to solve a set of non-linear transcendental equations, such as the ones that Jansson suggest, may not find all of the MSC toll sets. We will first show why multiple MSC toll sets can occur in steady state queues. We then provide a specific example of a queuing system with multiple MSC tolls. Lastly, we will discuss the implications of these findings.
4.5.1 Why Steady State Queues can Have Multiple MSC Toll Sets

When the service time distributions are homogeneous, finding a MSC toll set only requires finding a single MSC toll. We defined a quantity \( x(\tau) \equiv \frac{MSC(\tau)}{\tau} \), which (with few assumptions) is continuous in \( \tau \) and asymptotically approaches 0 as \( \tau \) approaches infinity and infinity as \( \tau \) approaches 0. These conditions and the Intermediate Value Theorem guarantee that a toll exists such that \( x(\tau)=1 \) or \( \tau = MSC(\tau) \). In such a case, to guarantee that there is only one toll satisfying these conditions, we would need to guarantee that \( x(\tau) \) is monotonically decreasing in \( \tau \). Let us explore why this may not be true in some cases.

We first recall the equation for MSC:

\[
MSC_j = MSC \equiv c \lambda \frac{\partial W_q(\lambda)}{\partial \lambda}, \quad \forall j \in M
\]  

(44)

We note that, in this case, Lemma 1 (p. 44) says that \( W_q(\lambda) \) is monotonically decreasing in \( \tau \). This implies that \( \lambda \) is also monotonically decreasing in \( \tau \) because \( W_q(\lambda) \) is strictly increasing in \( \lambda \). The next question we should ask is whether \( \frac{\partial W_q(\lambda)}{\partial \lambda} \) is decreasing in \( \lambda \), and therefore possibly increasing in \( \tau \), in some region. Alternatively, we ask whether \( W_q(\lambda) \) is concave in \( \lambda \) in some region. While we cannot rule out the possibility of a queuing model with such a feature, we can say that \( W_q(\lambda) \) is convex in the queuing models with which we are familiar (including M/M/1, M/M/1/K, M/M/m, M/M/m/K, and M/G/1). Thus, in queuing models with convex \( W_q(\lambda) \), if MSC is to be increasing in \( \tau \), we will need to find a case where \( c \) is increasing in \( \tau \). This requirement is actually satisfied in many models because of a common relationship between sensitivity to cost of accessing the runway (delay and toll) and sensitivity to time. It is often the case that those who are most sensitive to time (i.e. big commercial planes) are least sensitive to the cost of using the facility, while those who are least sensitive to time (i.e. leisure propeller planes) are most sensitive to the cost of using the facility. In such cases, there is a disproportionate reduction in operations by time-insensitive (i.e. cost-
sensitive) operators as the toll is increased. The result is that the average cost per unit of time delay rises. While we have shown that it is possible for \( c \), the average sensitivity to delay, to be increasing in \( \tau \) and provided an argument for why, as a result, MSC could be rising in \( \tau \), we have not proven that multiple MSC tolls can exist. To do so, we provide a specific example.

4.5.2 Example: M/M/1 Queue with Two User Types

In this M/M/1 model there are two types of operators, with identical service distributions. Operator type 1 has a high cost of delay and is relatively insensitive to price, whereas type 2 has a low cost of delay and is relatively sensitive to price. Both types have demand functions of the form \( \lambda_j = \max\{0, \alpha_j - \beta_j \cdot \text{Price}_j\} \). Table 4 gives the values of \( \alpha \), \( \beta \), and \( c \) for each user type.

<table>
<thead>
<tr>
<th>( j )</th>
<th>( \alpha_i ) (1/hr)</th>
<th>( \beta_i ) (1/hr)</th>
<th>( c_i ) ($/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0.01</td>
<td>5000</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0.04</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 4: Parameters for Two Queue Participant Types

Since we are able to numerically solve for \( \lambda_1 \) and \( \lambda_2 \) for each value of \( \tau \), we can then find how \( c \) behaves as a function of \( \tau \). Figure 1 shows that \( c \) is $50 when the toll is relatively low because the delays are relatively long and the time-sensitive (price-insensitive) operators will have zero demand. Conversely, \( c \) is $5,000 when the toll is relatively high because the price-sensitive (time-insensitive) operators will have zero demand. In the region between these two, \( c \) is increasing.
Figure 1: \( c \) as a function of \( \tau \)

Figure 2 shows that in part of this region, MSC is also increasing with \( \tau \). Figure 2 also shows that there are actually three MSC tolls, which correspond to the three intersections of the Toll and MSC curves.

Figure 2: MSC vs. Toll
Another way of looking at this is using our variable $x(\tau)$. Each time that $x(\tau) = 1$, the toll is equal to MSC. Figure 3 plots $x(\tau)$ and again shows that there are three MSC tolls. It is worth noting that, in general, there must be an odd number of MSC tolls, a fact that we shall interpret shortly. Thus, we have shown through this example that we can certainly not guarantee the uniqueness of a MSC toll and shown that there may be several such tolls. We use this example to understand the significance of multiple MSC tolls.

4.5.3 Significance of Multiple MSC Toll Sets

For convenience, we will refer to the three MSC tolls, in our example, as the lowest, middle, and highest. We recall that MSC tolls are a simple outgrowth of MC pricing, an approach that is meant to maximize welfare by maximizing the difference between the provided benefits of a good or service and the cost of providing it. If the area under the demand curve measures utility and the area under the marginal cost curve measures total cost, than maximizing welfare is the same as maximizing the difference between these two integrals.
Equivalently, we want to make sure that the marginal utility of a unit offered is equal to the marginal cost of offering it and that the second derivative of total utility less total cost, with respect to price, is negative. Setting the price (i.e. marginal utility) equal to marginal cost satisfies the first condition. The second condition is satisfied wherever the derivative of marginal cost with respect to price is less than or equal to one. Often, the second condition is satisfied by virtue of the fact that the marginal cost curve is strictly decreasing with price. However, if the marginal cost were to increase in some region, an intersection may exist with price equals to marginal cost, but a positive second derivative in economic welfare – indicating a minimum in welfare! This is an exact analogy for our case. The toll curve represents the marginal utility while the MSC curve is equivalent to the MC curve. The middle MSC toll, as is seen in Figure 2, is in a region where the slope of the MSC curve rises faster than the toll; the middle toll has a positive second derivative of economic welfare (i.e. total utility less total cost) – indicating that it is a local minimum. At this point, we should also note that even if the second derivative of total utility less total cost is negative, it only guarantees a local maximum, not a global one.

This can be generalized. Where a system has 2n+1 MSC tolls, where n is integer, there will be exactly n+1 local maxima and n local minima. The maxima will occur where the slope of the MSC curve, with respect to toll, is less than one and the minima will correspond to regions where the slope of the MSC curve is greater than one.

This is an important result. We have shown that some of the MSC Tolls, which were thought to be maximizing welfare, are actually minima of welfare. Moreover, the MSC tolls that are maxima, maximize welfare locally, but not necessarily globally. Figure 4 plots welfare as a function of toll in our example and shows that only the highest of the MSC tolls maximizes welfare, while the lowest MSC toll is a local maximum and the middle MSC toll is a local minimum.
The critical lesson from this example is that the focus on MSC tolls is misguided, even in the simplest of queuing systems. Even in a simple steady-state model with continuous, perfectly known demand functions, the quest for the MSC toll seems inappropriate. Moreover, we can have little faith in MSC tolls in more realistic models (i.e. more complex queuing distributions, heterogeneous service time distributions, or time-varying arrivals or departures rates), where we do not know if a MSC toll exists and have no reason to believe that such a toll would be unique. Thus, we conclude our discussion of MSC tolls by noting that MSC tolling at a congested facility, such as an airport, may be an inferior methodology for maximizing welfare.
We noted in Chapter 1 that administrative limits are a commonly proposed solution to airport congestion. Based on these limits, airports issue a limited number of slots per time period. These slots give the recipient the right to perform an operation (typically takeoff or landing) on a daily basis during some narrow band of time. The FAA’s 1969 High Density Rule (HDR) limited the number of scheduled operations per hour at five of the most congested airports. Based on these limits, slots were issued to airlines, with each slot enabling an airline to schedule one daily flight operation at one of these airports during a specified hour. In Chapter 1, we also noted that auctions are the most efficient way of awarding slots. Unlike the other methodologies for awarding slots, auctions award slots to those who value them most, thereby maximizing the utility that the slots provide. However, it is important to remember that auctions are only a distribution mechanism and that the relative merits of auctions versus competing methodologies (e.g. lotteries) speak nothing of the merits of administrative limits in and of themselves.

We start by discussing administrative limits as a methodology for controlling congestion. We continue with an in-depth discussion of using auctions to dole out slots to competing bidders. Our discussion of auctions starts with a brief history, taxonomy, and overview of auctions. We then introduce and discuss the Vickrey auction. Next, we discuss the combinatorial problems of airport slot auctions and the need for the combinatorial auction mechanism. Lastly, we discuss some of the difficulties that persist in using auctions to allocate scarce runway slots.

5.1 Administrative Limits

Administrative limits can be an effective way to reduce airport congestion. If an airport decides on an “acceptable” level of congestion, it can decide how many operations to permit so as not to exceed that level. Even if an airport cannot predict the relationship between the
number of scheduled operations and the congestion level, it can find the right number of operations to permit by a trial-and-error process. The problem with this system is that the airport lacks the capability to determine an "acceptable" level of congestion. It is unclear whether the public finds a 10-minute or a 1-hour delay acceptable or unacceptable?

Another approach to determining administrative limits might be to determine what the capacity of the airport is. That is, if the airport can determine how many operations it can handle per hour, it could then easily determine how many scheduled operations to allow. The problem with this approach is that there is no "magic capacity" that defines the border between an uncongested and a congested airport. The most obvious reason why this is true is that an airport's capacity is typically strongly dependent on weather. For example, airports with parallel runways that are relatively close to each other can conduct simultaneous operations on these runways in good weather, but cannot do so when visibility is poor. Thus, the first question that an airport would need to answer is which capacity to choose: good weather, poor weather, or perhaps something in between the two? This is a question with no clear answers.

Beyond weather conditions, the stochastic nature of airport queues makes it hard to choose a capacity. One component of stochasticity is in the arrivals process to the queue(s) because of the lack of correspondence of actual operations to the schedule. Even if an airport is able to schedule operations down to the minute, there is some unpredictability about when the operations will actually take place. Because some flights are ready early and others are ready late, it is likely that there will be some clumping, a feature which is characteristic of a randomized arrival process (e.g. Poisson). This clumping contributes to congestion even if the airport's service rate is equal to (or exceeds) the arrivals rate to the airport queue(s). Another component of stochasticity is in the service process of the airport because of the less-than-predictable spacing between successive operations. The spacing between successive operations (particularly landings) depends on the size of both the leading and trailing aircraft. Given that an airport typically handles a mix of many types of airplanes and does not know the exact order they will arrive or depart in, this introduces a stochastic component to the time it takes an airport to service an operation.
Thus, administrative limits present a major difficulty. There is no proper methodology that we are aware of for determining the right number of slots to dole out. Choosing capacity based on congestion requires having some acceptable level of congestion, while choosing capacity based on physical parameters is difficult because of the stochastic nature of airports and the effect of weather on operating capacity. Of these two approaches, it seems that the first makes more sense if we could devise a methodology for determining acceptable levels of congestion. This is a topic that we revisit in Chapter 6. In the mean time, we study auctions, which are a useful tool for distributing slots, should we eventually figure out how many slots to distribute.

5.2 Overview of Auctions

Auctions are a price competition either by buyers for the right to purchase a good or service or by sellers for the right to sell a good or service. Where buyers compete, the auction is considered to be a Forward Auction and where sellers compete, it is considered a Reverse Auction. Interestingly, the first known auction was both a Forward and a Reverse Auction. The Greek historian Herodotus reports that, circa 500 BC, the Babylonians would auction off women for the purpose of marriage and that some of the women sold for positive amounts and some for negative amounts (i.e. those women had to pay men to take them). Since the time of Herodotus, auctions have been used to sell or buy a variety of goods and services in many different cultures and societies. We will focus on the Forward Auction, but mention that nearly all the concepts apply in reverse to the Reverse Auction.

Auctions can have many different sets of rules. Public auctions are open to all bidders while Private auctions are only open to a select few. Some auctions have a reserve price, a minimum price at which the good may be bought, while others do not. Some auctions are conducted in a single round while others are conducted in multiple rounds. Some auctions are for a single unit while others are for multiple units of an item.

Perhaps the most basic, distinguishing feature of auctions is whether they are sealed-bid or open. In an open auction, bidders submit their bids publicly by either announcing them or signaling them to the auctioneer. This is the prevalent form of auction in most auction houses and on websites such as Ebay. In contrast, in a sealed-bid auction, each bidder submits a bid to the auctioneer in secrecy. Governments commonly use this type of auction for awarding contracts or rights to use natural resources and business commonly uses this type of auction
for a variety of purposes. This auction affords more privacy to the bidders and also allows the auctioneer to conduct more complex combinatorial auctions, which we shall soon examine in detail.

In open auctions, the highest bidder wins and (almost always) pays a price equal to the highest bid. However, there are several common ways of conducting such an auction. An English auction uses ascending-bids; the auctioneer starts by announcing a low bid and bidders announce successively higher bids until there are no more bids. A Dutch auction uses descending-bids; the auctioneer starts by announcing a high initial price and lowers the price until a bid is made.

Sealed-bid auctions come in two important varieties: first-price and second-price. In a first-price auction, the winning buyer pays the highest bid (i.e. his bid). In contrast, in a second-price auction, the winning buyer pays the second highest bid, or the highest losing bid. These two auction formats, as well as the English and Dutch auctions, are the four most commonly studied auction styles because almost all auctions use one of these four formats or a close variation of one of these formats and because they have interesting properties. We explore some of these properties under varying models of auctions.

The simplest model is that all bidders know the value of the particular good to all other bidders. In such a case, the Revenue Equivalence Theorem predicts that, using any of the four auction types mentioned above, the bidder with the highest valuation will win and will pay a price equivalent to the second-highest valuation. Thus, these four auctions will produce identical, efficient allocations if bidders have perfect information (Bierman and Fernandez 1998).

A slightly more complicated and realistic model assumes that bidders have independent private valuations (IPV) - each bidder knows his valuation of the good, but does not know how others value it and the valuations are independent. In such a case, there are several important results. First, if bidders are risk-neutral, each bidder will bid strictly less than his true valuation if participating in a first-price sealed-bid auction and will bid his true valuation if participating in a second-price sealed-bid auction. Second, the Revenue Equivalence Theorem states that the expected price paid is the same under both of these sealed-bid methods. Third, the Dutch
auction is strategically equivalent\textsuperscript{11} to the first-price sealed-bid auction and the English auction is strategically equivalent to the second-price sealed-bid auction. The overall implication is that, even in this more complex environment, these four methods will, on average, produce the same winners and prices (Bierman and Fernandez 1998).

Lastly, it is notable that an auction is said to be efficient if the bidder who most values the item wins the auction. We already noted that under the assumption of perfect information, any of the four auction types are said to be efficient. An obvious next question is when would an auction be efficient under the IPV assumption. Game theory predicts that if bidders have IPV and the valuations come from a known distribution with continuous density, the four methods mentioned above are all efficient. This is predicated on the assumption that the auction has a "symmetric, pure-strategy equilibrium bidding function that is strictly increasing in private value"\textsuperscript{12} (Bierman and Fernandez 1998, p. 305). While this last result is of theoretical importance, its assumptions are too strong for realistic cases. For example, if a first-price sealed-bid auction is conducted in the IPV environment, buyers will bid strictly below their valuations. However, it seems unlikely that bidders will have symmetric strategies. It is thus possible that bidder A, who values the item slightly more than bidder B, will actually bid below bidder B and that bidder B will then win the auction. Thus, the first-price sealed-bid auction, and by extension the Dutch auction, do not necessarily encourage efficient allocation of the auctioned item. In contrast to these auctions, bidders participating in the English Auction and the second-price sealed-bid will bid their true valuations. Thus, the English and second-price sealed-bid auctions will result in efficient allocations, even if bidders do not have symmetric bidding functions. We hope that this overview explains the basic concepts to the unfamiliar reader and refer those looking for a thorough overview to Klemperer (1999). We now explore the second-price sealed-bid auction, commonly known as the Vickrey auction, in more depth.

\textsuperscript{11} Strategic equivalence means that the auctions are identical from the perspective of game theory, which entails identical choices and payoffs. Strategically equivalent auctions will theoretically produce the same bids, allocations, and prices.

\textsuperscript{12} For the unfamiliar reader, some definitions are in order. Symmetry implies that the strategy is the same for all bidders. Pure strategy denotes a nonrandom strategy for playing the game. In our case, the bidder will bid a certain predictable bid rather than use a random strategy to confuse opposing bidders. Equilibrium implies that such a strategy is optimal for each bidder assuming that the other bidders follow it.
5.3 Vickrey Auctions

Vickrey’s (1961) article entitled *Counterspeculation, Auctions, and Sealed Tenders* is the seminal work in auction theory and was a major factor in Vickrey winning the 1996 Nobel Prize in Economics. Vickrey begins by analyzing the simplest case, an auction of a “single unique indivisible object” (p. 14). Vickrey notes, as we did above, that the first-price sealed-bid auction (and by extension the Dutch auction) is close to Pareto-optimal. However, it can lead to inefficient allocations if “there is much variation in the state of information or the generally expected intensity of desire of the various players for the object, or where the bidders are insufficiently sophisticated to discern the equilibrium point strategy or for some other reason fail to use this strategy” (Vickrey 1961, p. 20). In contrast to this, Vickrey notes that the second-price sealed-bid (and by extension the English auction) encourages the bidders to place bids equal to their true valuations and thus guarantees a Pareto-efficient allocation. While previous research had shown that the English auction is Pareto-optimal, Vickrey was the first to establish the equivalence of the second-price sealed-bid auction and the English Auction and was the first to discover a Pareto-optimal method for sealed-bid auctions. Vickrey notes that the second-price auction, compared to the first-price one, sometimes generates a higher price and sometimes generates a lower price, but that these differences are small and that the second-price method, unlike the first-price method, is always Pareto-optimal. Thus, from the vantage point of buyers and sellers the first-price method and second-price methods may each have advantages and disadvantages, but from the standpoint of trying to maximize societal surplus, the second-price method is preferred.

Vickrey (1961) also notes that the English auction and the second-price sealed-bid auction do not require the same level of gaming as the Dutch auction and the first-price sealed-bid auction. The bidder is relieved of the burden of appraising the value of the good to the other bidders and only needs to appraise his valuation of the good. This elimination of gaming makes the English and second-price sealed-bid auctions more administratively efficient than their counterparts.

Finally, we note that Vickrey (1961) discusses the possibility of an auction for multiple identical items. Vickrey shows that if each bidder is interested in purchasing at most one item, an auction based on the second-price concept is efficient. The only modification is that, in place of the second-highest bid, the final price is the highest losing bid. In other words, if there are m
identical items for sale, the final price should be the \( m + 1 \)^{st} highest price. This discipline encourages all bidders to bid their true valuations and, as a result of this, can guarantee an efficient allocation. However, Vickrey notes that, in the case where a bidder may be interested in more than one unit, the \( m + 1 \)^{st} price mechanism is no longer optimal. The reason is that any buyer bidding on more than one unit will need to consider the possibility that if some of his bids are accepted and some rejected, the price he pays for winning bids could be influenced by the rejected bids. It is therefore in the bidder's interest to understate all bids except the highest bid. Klemperer (1999) reviews some of the recent research on pricing in multi-unit auctions and discusses some of the scenarios in which we can find efficient prices.

The problem that Vickrey (1961) describes is an auction for multiple identical items with bidders who may be interested in more than one unit of the item, but have decreasing marginal value for these items. If the problem of auctioning airport slots was similar to this problem, it would be sufficient to state that the \( m + 1 \)^{st} price mechanism is not efficient, but is nearly so. We would further note that this mechanism is very close to efficient as the number of bidders becomes large and the potential rewards from gaming decrease. Lastly, we would mention the possibility of using non-uniform pricing mechanisms (i.e. winners do not necessarily pay the same amount), which may be similar in spirit to the \( m + 1 \)^{st} price mechanism, but may avoid its pitfalls. Alas, the airport problem is far more complex than the one that Vickrey describes and needs a different approach than the traditional sealed-bid, \( m + 1 \)^{st} price auction.

5.4 Complications of Slot Auctions

The first difference that we note between the case of the airport and the case Vickrey describes is that airlines may have increasing convex valuations of the number of slots awarded. In other words, the total utility that an airline derives can be locally (or globally) convex in the number of slots it receives for some period (or for the day as a whole). This might be the case for an airline running a hub-and-spoke system, where the hub’s network effect makes the marginal utility of a slot increase with the number of slots awarded. The problems here go beyond the gaming that Vickrey describes; airlines may find themselves with no appropriate bidding strategy at all. For example, an airline may value one slot at $100 and two slots at $300
If it places one bid at $100 and another at $200, it could end up with one slot at $200, which is an unacceptable outcome.

There are also combinatorial complications in slot auctions since any auction would involve multiple non-identical items. For example, the auction might include slots for each hour of the day. While slots might award the right to a daily operation between 8 AM and 9 AM, others might award the right to a daily operation between 8 PM and 9 PM. Such an environment poses difficulties. First, an airline’s demand for slots during one period might depend on how many slots are awarded during another period. For example, the value that an airline places on a slot between 8 AM and 9 AM, which the airline intends to use for a landing, might well depend on the availability of a slot between 9 AM and 10 AM, which the airline could use for a subsequent departure. Without this latter slot to depart soon after the arrival, the former slot would be far less valuable. Likewise, there could be interdependencies between slots at different airports. The value of a slot for departure at one airport may strongly depend on a slot for landing at another airport.

Other combinatorial problems arise when an airline has system constraints. One example of such a constraint is a budgetary constraint. Another example of a system constraint would be based on the slot needs of an airline. For instance, if an airline wants to use a slot in the early morning, the airline may desire a slot between 6 AM and 7 AM or between 7 AM and 8 AM, but not both. These problems are serious in as much as they would lead to grossly inefficient allocation of resources if not properly addressed. Fortunately, all of the slot auction’s complications can be addressed through use of a Combinatorial Auction.

5.5 Combinatorial Auctions

Rassenti, Smith, and Bulfin (1982) suggest a combinatorial auction mechanism for airport slot auctions that would allow bidders to bid on packages of slots in addition to bidding on standalone slots. Moreover, this format would allow bidders to impose logical conditions such as, “accept no more than p of the following q packages’ or ‘accept package V only if package W is accepted’” (p. 404). This formulation addresses the variety of problems that we noted above.
First, we note that this mechanism is well suited for cases where bidders have convex valuations of the number of slots awarded. As an example, an airline may find that getting one slot is worth $100, two is worth $300, three is worth $600, and four is worth $1000. With a combinatorial mechanism, the airline can submit four separate bids: $100 for one, $300 for two, $600 for three, and $1000 for four. This along with a constraint stating that the airline wishes to be awarded no more than one of these four bids is equivalent to bidding $100 for 1, $200 more for a second, $300 for a third, and $400 for a fourth. This capability eliminates the need for gaming and promotes efficiency by encouraging truth telling.

It is also well suited for dealing with cases where an airline bids on substitutable slots. For example, an airline may value a slot between 8 AM and 9 AM at $100, a slot between 9 AM and 10 AM at $75, and the combination of these slots at $400. It can place three separate bids: $100 on the earlier slot, $75 on the later slot, and $400 on the combination of the two. If the airline does not want to have more than one slot in any one period, it can also submit a constraint that says “accept the bid on the earlier slot or on the combination of the earlier and later slots” and another constraint that says “accept the bid on the later slot or on the combination of the earlier and later slots”. Such constraints can be built to match the specific desires of the bidder.

Lastly, we note that this type of auction is effective for dealing with system constraints. As an example, an airline participating in such an auction could submit a global budgetary constraint. Another example of a system constraint might relate to the ability to create an economically efficient hub in an airport. For instance, an airline could submit a constraint that says that some subset of its bids is only applicable if it receives at least twenty slots in at least four different periods of the day, where each period is no longer than two hours and the end of one period is no less than two hours before the start of the next period.

These examples show how combinatorial bidding allows bidders to remove much of the guessing involved in slot auctions, which involve multiple, non-identical goods. This type of auction allows the bidder to focus on appraising how much the item(s) for auction are worth instead of expending considerable energy on figuring out how to beat the system and prevent unwanted outcomes. However, two important questions remain. First, is there a pricing mechanism for the combinatorial auction that is equivalent to the \( m + 1 \) price mechanism in
the simple auction? Second, when bidders are interested in purchasing more than one item, will the gaming that is characteristic of the \( m+1 \) price mechanism persist in the combinatorial environment?

5.6 Prices in Combinatorial Auctions

To find a pricing scheme that is analogous to the \( m+1 \) price auction in the single commodity environment, we would like to finding a uniform set of prices where the losing bidders would be no better off winning and paying at those prices and the winning bidders would be no better off had they lost. Finding such a set of price encourages the bidders to bid honestly. While we may on occasion find such a set of prices, Rassenti et al. show that the existence of such a set cannot be guaranteed and that in many cases there is no set of prices that “support the optimal division of packages into accepted and rejected categories” (p. 405). They demonstrate this by solving a discrete project selection (or knapsack) problem with one resource constraint:

\[
\text{Maximize } Z = 5X_1 + 3X_2 + 6X_3 + 5X_4 + 6X_5 + 3X_6 + 4X_7 + 3X_8 + 2X_9 + X_{10} \\
\text{s.t.: } 3X_1 + 2X_2 + 6X_3 + 7X_4 + 9X_5 + 5X_6 + 8X_7 + 8X_8 + 6X_9 + 4X_{10} \leq 24; \quad (45) \\
X_j \in \{0,1\}, \; \forall j = 1,2...10.
\]

If we relax the last constraint so that \( 0 \leq X_j \leq 1 \), we find the solution as:

\[
(Z,X_1,X_2...X_{10}) = (23,1,1,1,1,66,0,0,0,0,0). \quad (46)
\]

Clearly, if we could relax the global constraint, we would choose to increase \( X_5 \), since it is more valuable, per unit of constraint, than \( X_6, X_7, \ldots X_{10} \). We can see from (46) that the critical return rate (i.e. objective function to constraint use) of \( X_5 \) is \( 6/9 \). Thus, the shadow cost of the constraint is \( 2/3 \); an incremental increase of \( \varepsilon \) in the constraint would be worth \( 2/3 \varepsilon \). However, when the constraint is not relaxed the solution is not so simple. The optimal integer solution is:
In this case, \( X_6 \) is chosen even though its critical return rate is \( \frac{3}{5} \), below the critical return rate of \( X_5 \), because choosing \( X_5 \) would violate the resource constraint. This leaves us with the difficult question of what the shadow cost is? Is it the lowest critical return rate of the projects accepted (i.e. \( \frac{3}{5} \)), the highest critical rate of a rejected project (i.e. \( \frac{2}{3} \)), or something else altogether? This forms a perfect analogy for if these were 10 bids on 24 slots, with each bid for some number of slots, we would be choosing bid 6 over bid 5 despite the fact that the bidder who submitted bid 5 has a higher per-slot valuation than the bidder who submitted bid 6. There is no price (per unit) at which the bidder who submitted bid 5 will be happy to have lost and the bidder who submitted bid 6 will be happy to have won. A uniform price mechanism of this type is clearly not possible. However, it is intuitive that as the number of bids and bidders increase, relative to the auction size, the difference between the per-unit highest losing bid and lowest winning bid should become relatively closer and this problem should decrease in severity (Rassenti et al. 1982).

This same problem that we have described clearly extends itself to the types of large integer programs that need to be solved to determine which bids on slots should be accepted so as to maximize the benefit that the slots provide. However, when there is more than one constraint, it is no longer possible to find the per-unit highest losing bid and lowest winning bid since that is only applicable when there is a single resource constraint. To find upper and lower limits of appropriate prices, Rassenti et al. propose solving two pseudo-dual programs to the primal integer program. First, we note their integer-programming formulation, which maximizes the sum of accepted bids subject to resource constraints (e.g. number of slots per hour) and logical constraints submitted by the bidders:

\[
(Z, X_1, X_2 \ldots X_{10}) = (22, 1, 1, 1, 0, 1, 0, 0, 0, 0).
\]

Note that Rassenti et al. suggest that the solution is \((Z, X_1, X_2 \ldots X_{10}) = (21, 1, 1, 1, 0, 0, 0, 0, 0, 0)\), but their solution is erroneous.
\[
\begin{align*}
\text{Maximize} & \quad \sum_j c_j x_j \\
\text{s.t.} & \quad \sum_j a_{ij} x_j \leq b_i \forall i, \\
& \quad \sum_j d_{kj} x_j \leq e_k \forall k, \\
& \quad x_j \in \{0,1\}
\end{align*}
\] (48)

where:

- \(i\) subscripts resource types (slots at different times);
- \(j\) subscripts a package of slots that some airline bid on;
- \(k\) subscripts logical constraints submitted by the airlines;

\[
a_{ij} = \begin{cases} 
1 & \text{if package } j \text{ includes slot } i, \\
0 & \text{otherwise}; 
\end{cases}
\]

\[
d_{kj} = \begin{cases} 
1 & \text{if package } j \text{ is in logical constraint } k \\
0 & \text{otherwise}; 
\end{cases}
\]

- \(b_i\) and \(e_k\) are some positive integers; and
- \(c_i\) is the bid for package \(j\) by some airline (Rassenti et al. 1982, p. 402).

The two quasi duals are then given by:

\[
\begin{align*}
\text{Minimize} & \quad \sum_R y_r \\
\text{s.t.} & \quad \sum_j w_i a_{ij} \leq c_j, \quad \forall j \in A \\
& \quad y_r \geq c_r - \sum_j w_i a_{ir}, \quad \forall r \in R \\
& \quad y_r, w_i \geq 0
\end{align*}
\] (49)

where:

- the optimal solution to P is \(\{x_j^*\}\);
- the set of accepted packages is \(A = \{j | x_j^* = 1\}\);
- the set of rejected packages is \(R = \{r | x_r^* = 0\}\);
- the set of lower bound slot prices to be determined is \(\{w_i^*\}\);
- and the amount by which a rejected bid exceeds the lower bound slot prices (if at all) is \(y_r^*\).
Minimize $\sum_{A} y_j$

s.t. $\sum_{i} v_i a_{ir} \geq c_r \forall r \in R,$

$y_j \geq \sum_{i} v_i a_{ij} - c_j \forall j \in A,$

$y_j, v_i \geq 0$

where:

the set of upper bound slot prices to be determined is $\{v^*_i\}$ and the amount by which an accepted bid is below the upper bound slot prices (if at all is) $y_j$ (Rassenti et al. 1982, p. 405).

Thus, the first dual problem defines a lower bound on the shadow prices, which all accepted bids must match or exceed. The second dual problem defines an upper bound, which all rejected bids must be less than or equal to. Rassenti et al. divide the bids into three categories:

1. Those that exceed the sum of component values in the set $\{v^*_i\}$, which are accepted;

2. Those that are less than the sum of component values in the set $\{w^*_i\}$, which are rejected; and

3. Those that do not exceed the sum of component values in the set $\{v^*_i\}$ and exceed the sum of component values in the set $\{w^*_i\}$. These bids are accepted or rejected on the basis of an efficient solution to the resource utilization problem with integer constraints (Rassenti et al. 1982, p. 406).

While Rassenti et al. (1982) have a solution for how to pick which bids to accept, they note that there is no incentive compatible pricing policy. Since we want to charge bidders no more than they bid and yet want to have uniform pricing, it is easy to see why the best available strategy is to price according to the set $\{w^*_i\}$. However, this pricing strategy provides incentive for gaming because inflated bids help win slots, but do not necessarily increase the price the way that would happen in the $m + 1^{st}$ price auction. In other words, in the $m + 1^{st}$ price auction, the bidder with the $m + 1^{st}$ highest valuation has no incentive to inflate his bid.
because it will lead to paying an excessive amount. In the combinatorial auction, the bidder who would lose if he did not game may have valuation between \( \{w'_i\} \) and \( \{v'_i\} \) and could therefore inflate his bid and possibly win at a price below his valuation. Gaming by purposefully misstating valuation can ultimately lead to inefficient allocations.

If our objective function is to maximize the value that the slots provide (by maximizing the value of the accepted bids), we have shown that we cannot find a uniform price set that differentiates between the winning bids and the losing ones. Thus, in some sense we are not able to find a pricing mechanism for combinatorial auctions that is equivalent to the \( m+1 \)' price auction in the single commodity environment. The combinatorial problem suffers from both the problems of the \( m+1 \)' price auction (because a losing bid can still influence the price paid for a winning one) and its own problems of not being able to find a price set that makes winners happy that they won and losers happy that they lost.

While it would be inappropriate to minimize the theoretical importance of these problems that we have outlined, the problems may not be too severe in practice. Rassenti et al. (1982) suggest that there is good reason to believe that there will be very little gaming in practice. First, they note that as the size of the auction gets bigger (i.e. more slots and more bids), the sets \( \{w'_i\} \) and \( \{v'_i\} \) will be relatively closer. This will decrease the opportunity and incentive for strategic bidding (or gaming). Second, even in a simple single commodity auction, gaming is both difficult and dangerous when the bidder has incomplete information. Those difficulties and dangers should be even greater in the case of a complex combinatorial auction. It is thus likely that bidders will largely focus their energy on determining their own valuations and preparing appropriate bids and will spend relatively little energy on manipulating bids to take advantage of flaws in the auction mechanisms.

Thus, this problem may be more of a theoretical one than a practical one. As a theoretical problem, it poses a challenge for future research. It remains an open question of whether one can devise a mechanism to efficiently allocate several types of goods and/or services that are complimentary in nature. With current auction technologies, a guaranteed efficient allocation is theoretically unattainable.
Chapter 6

THE MARGINAL SOCIAL COST AUCTION

A New Proposal for Dealing with Airport Congestion

We have explored two market-based approaches to dealing with airport congestion. Congestion pricing, which we explored in Chapter 4, would cause airlines to internalize the congestion externality that they create when operating flights at a congested airport. Such internalization should align the airline’s incentives with those of society; the airline will only schedule an operation if the benefits outweigh the costs, including the congestion costs to the rest of society. With this approach, the number of permitted flights will be a function of how high a marginal cost the airlines are willing to bear because airlines will internalize the complete marginal cost of operating a flight.

In Chapter 5, we explored the use of auctions as a method for distributing a limited number of slots. We observed that auctions have the ability to allocate slots in a relatively fair and efficient manner by awarding them to those willing to pay the most for them. Moreover, the combinatorial auction mechanism, with surplus maximization as its objective function, is an effective way for dealing with many of the complex problems associated with a slots auction.

While each of these two approaches offers certain capabilities, we have previously noted that each one has its problems. With regard to congestion pricing, even if the demand curves can be assumed to be continuous functions of cost, there are problems with the Marginal Social Cost (MSC) Toll. First, to establish a toll set that results in usage levels consistent with MSC pricing principles, we would need to have a complete knowledge of the airlines’ demand functions. The normal congestion pricing approach does not have a means for collecting such information. The second problem is that MSC tolls do not maximize societal welfare (surplus) because there can be sub-optimal MSC toll sets, which we established in Chapter 4.
Auctions do not suffer from the problems that MSC tolls do. Unlike MSC Tolls, Auctions are an effective method for revealing the slot valuations because the bidders essentially specify their demand functions when submitting bids. Moreover, a combinatorial auction, unlike MSC tolling, actually enables bidders to detail the substitutability of one slot time for another, something that is essential for successfully implementing MSC tolling. Lastly, the auction does not have the problem of an incorrect objective function, as does the MSC Toll; the auction is merely an allocation mechanism that can be used with a variety of objective functions. However, the traditional slot auction, which bases the number of slots on an administrative decision, is sub-optimal in its "judgmental" approach to picking the number of slots to auction. In other words, auctions are good for doling out slots, but need to be combined with a toll that determines the right number of slots (or the acceptable level of congestion).

Table 5 shows the relative strengths and weaknesses of these two methodologies. What this table makes apparent is that they are complimentary. One’s weakness is the other one’s strength. Choosing one in favor of the other does not satisfy our quest for a solution that can both choose the right capacity based on MSC pricing principles and efficiently allocate that capacity to maximize surplus.

<table>
<thead>
<tr>
<th>Capability</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Auctions</td>
</tr>
<tr>
<td>Estimate Demand Functions</td>
<td>Yes</td>
</tr>
<tr>
<td>Maximize Welfare</td>
<td>Yes</td>
</tr>
<tr>
<td>Choose Optimal Capacity</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 5: The relative advantages and disadvantages of Auctions and MSC Pricing

6.1 A New Objective Function

We can look at the two methodologies that we reviewed above in terms of objectives. Traditional auctions attempt to maximize the surplus from some number of scarce slots; they are a relatively efficient way of allocating scarce resources. MSC pricing attempts to align the user’s cost of using the facility with society’s marginal cost for such use. Neither of these promotes overall welfare by maximizing the total surplus that the facility provides, something that we would like to do through a new objective function. It is worthwhile to note that
maximizing economic welfare, which heretofore we have called welfare, does not necessarily maximize the overall welfare, but that economists have long used the maximization of economic welfare as a goal because it is impossible to measure total welfare and because “the qualitative conclusions about the effect of an economic cause upon economic welfare will hold good also of the effect on total welfare” (Pigou 1932, p. 20). Thus, we would like to devise a methodology that determines which bids to accept and which to reject to maximize the economic welfare that the airport system can provide.

In thinking about a new objective function, let us first assume that there is a congestion-less value that an airport user has for any package of slots. This is the maximum price that the user would be willing to pay for the package of slots assuming that there is no congestion. We further assume that we can quantify a cost per unit of expected time delay for each possible slot-operation combination. With this, we state that the objective is to assign slots in a way that maximizes the congestion-less value of those slots to the operators that use them minus the cost of delays to those operators. Thus, the new objective function requires deciding how many slots to allocate and to whom they are to be allocated in a way that maximizes surplus. To accomplish this goal, we introduce a new methodology: the MSC Auction.

### 6.2 The Marginal Social Cost Auction

The MSC Auction is essentially a hybrid of MSC pricing and the traditional slot auction. It captures the benefits of both of these methods, but does not suffer from the drawbacks of either. The process starts with the airport defining the slots, or the units to be auctioned. Those interested in purchasing these slots would submit bids on any combination of the available slots along with logical constraints on their bids of the type described in Chapter 5. In addition to these bids, each bidder would be required to submit a statement detailing their cost per unit of time delay for each aircraft type for both landings and departures. The airport authority would then decide which bids to accept so as to maximize the expected economic welfare that the airport can provide. Each winning bidder would be required to pay a toll that is no greater than the amount of their bid. To relieve bidders from the complexity of estimating their private congestion costs, the airport would credit (i.e. pay) each winning bidder for the cost of actual congestion. In other words, whereas the bidding and price setting are based on expectations and done on a periodic basis, the airport reimburses the airline for the
actual congestion that they incur. We now describe this process, its advantages, and the challenges associated with it, in greater detail.

6.2.1 Slots

The first decision that an airport would have to make is what time intervals to use for slots. Smaller time intervals lend themselves to greater congestion model accuracy, while larger intervals reduce the complexity of the bidding process. Airports will need to balance the need for model accuracy with the advantages of a simpler, smoother bidding process.

The slots could be auctioned on a periodic basis (e.g. monthly or quarterly) for use on a daily basis for the length of the period. In addition, if airlines want slots that are not on a daily basis, additional slots could be created to meet these needs. Thus, an auction participant could theoretically bid on a slot for the right to depart between 8:00 AM and 8:15 AM on every other Tuesday and Thursday.

There are other variations on slots that could be used in such an auction. An airport could allow bidders to bid on restricted slots that are only good under certain weather conditions. For example, there could be slots that are only for use when Visual Flight Rules (VFR) are in place. Since large airlines routinely cancel and consolidate a significant percentage of their flights when weather conditions are poor, they might place bids with a certain percentage of the slots in the bids for VFR use only, which could translate to substantial savings. For example, an airline bidding on forty landing slots between 11:00 AM and 12:00 PM and forty departure slots between 12:00 PM and 1:00 PM, may wish to bid in a way that only thirty-five of the forty landing and departure slots are for use when Instrument Flight Rules (IFR) are in effect. The five slots for use under VFR are likely to be significantly cheaper for reasons that will soon become apparent.

Other decisions to be made regarding slot definitions include how to categorize the different operation types and aircraft sizes. At one extreme, there would be two separate categories for each aircraft model, one for landing and one for departure. At the other extreme, there would be just one category of slot that is for all runway operations. Neither of these is necessarily right or wrong. The airport will need to determine how many different categories are necessary to properly model the congestion.
6.2.2 Bidding and Cost Estimates

The bidding process is similar to what we described it in Chapter 5. Bidders submit packages of slots, each one with a bid that represents the maximum price that the bidder is willing to pay for that package. The bidders also submit logical constraints that help them deal with some of the uncertainties inherent in the auction process, a process that we also described in Chapter 5.

Bidders would also submit a delay cost estimate to help the airport determine what the cost of delays is. These estimates would be subject to the guidelines and review of the airport authority. Such a review or audit can be of particular importance since there seems to be a strategic advantage to overstating the cost of delay, a matter requiring further investigation. Alternatively, the airport could determine the costs based on manufacturer specification and verifiable parameters such as wages for the flight crew or use standard cost estimates for all airlines.

We recall that in Chapter 2, we discussed whether the airline or the passengers really bear the passenger cost of congestion. It is with regard to these cost estimates and reimbursements that it has practical significance. Theoretically, it would be preferable to reimburse the passengers directly. This would eliminate the cost of delays for passengers and would help the airline recover the loss of passengers and yields that such delays cause. However, there are some problems with reimbursing passengers. The easiest of these problems is how to reimburse them. In today’s day and age, this should be relatively simple. Airlines could serve as the conduit for reimbursing them, by refunding the passengers through their credit cards or through other means. A more serious problem is in estimating the cost of delay. There is likely to be a very large heterogeneity of sensitivities to delay among passengers. At one extreme, a partner in a law firm, might be willing to pay $500 to avoid one hour of delays and at the other extreme, a vacationing minimum-wage worker, might only be willing to pay $5 to avoid the same one hour of delays. This is a significant problem, on both theoretical and practical levels. For these reasons, it may be easier to reimburse the airport rather than passengers; however, this an open-ended question requiring further research.

There could be further complexities in these cost estimates. Such complications could include a nonlinear element to the cost per unit of time delay. For example, airlines may find that the cost per unit time rises with the length of delay and that they wish to specify a nonlinear
function of delay time as their cost estimate. Another example is that an airline may decide that delays for a 10-hour international flight are less costly than for a 4-hour domestic flight, even though the aircraft used are identical. Yet another example is that delays in the beginning of the day might be far more costly to airlines than those at the end of the day. There are clearly several reasons why a simple cost per unit of time delay for each operation type-aircraft type combination may not be specific enough. While this is an open ended question of whether it is necessary to get an exact figure for the cost, we should note that cost accounting is always a tricky business and that cost estimates, as their name implies, are just estimates; airports and airport operators may find a simple cost estimate both administratively efficient and sufficiently accurate.

6.2.3 The Congestion Model

For the purposes of calculating the expected queuing costs, we want a model that can predict information on the expected queuing time as a function of time-of-day. We suggest a model based on Koopman (1970), which we reviewed in Chapter 3. However, Koopman's model is highly unrealistic in assuming that the queue service times are independently and identically distributed. Clearly, when the weather is bad one minute, there is a very high chance that it will be bad in the next minute. As a result, the distributions of service times from minute to minute are highly correlated. We therefore propose a simulation model like Koopman's, but without the assumption that the service times are independently and identically distributed. To accomplish this, we first simulate service time distributions patterns. In other words, using historical weather and airport operations data, we simulate some large number of days with appropriate weather and service patterns for the time period of interest. These simulated days can then be used, one day at a time, to simulate how the service time distribution evolves during the course of a day. This method will offer the advantage of simulating the possibility of having a snowstorm, afternoon thunderstorm, windy day, or any other weather event in as realistic a manner as possible.

For example, let us assume that an airport has two runways and that can both be used under VFR conditions (one for landing and takeoff each), but that only one can be used under IFR conditions. Let us further assume that all slots are good for all weather conditions and that the service time distributions are drawn from a large number of simulated days (as we described). We might model the airport as having two queues, one for landings and one for departures,
when the airport is operating under VFR rules, but having only one queue, shared by both landings and departures and with landings taking precedence, when the airport is operating under IFR rules. For each of the simulated days, we would run a simulation of the type Koopman (1970) describes using the fact that we know exactly when there should be one queue, when there should be two queues, when the queues merge, and when the queues separate. Armed with this information and approximations on the queue arrival rates (by time of day) and service times for the various aircraft-operation type combinations, we can produce statistics such as expected queue length and queuing time, by time of day, for each of these simulated days. With this information, we can also calculate the expected total queuing costs for each of these simulated days. To get statistics for the future period of interest such as expected queue length, expected queue time, or total queuing cost, we can average the statistics from the many simulated days that were studied.

6.2.4 Optimization

We previously stated an objective function of maximizing the economic welfare that an airport produces by maximizing the difference between the value of slots that the airport assigns and the cost of congestion. This requires selecting the best subset of bids from the entire set of submitted bids, a massively complex problem. To illustrate this point, let us consider a highly simplified example of an airport. The airport has one runway, one type of aircraft using the runway, and runways and landings take the same amount of time. In this example, bidders submit single, non-combinatorial bids on slots that are divided on an hourly basis and go from 7:00 AM to 10:00 PM. Since the airport can rank bids for slots in each hour, the only decision that the airport must make is how many slots to permit in each hour. The airport may be able to further simplify this problem by only examining a small subset of solutions. Suppose that the airport is able to limits its decisions to picking a number of slots, between 61 and 70, for each hour of the day. Even in this highly simplified model, there are $10^{15}$ possible solutions that the airport could explore. More realistic models may contain a far greater number of solutions to be considered. Clearly, exhaustive searches are not possible; this type of combinatorial optimization problem requires a smart solution technique since an exhaustive search is usually impossible. While integer optimization problems with analytical objective functions can often be solved with relative ease through techniques such as Branch-and-Bound and decomposition, solving problems without analytical objective functions, such as
our problem, can be a source of significant difficulty. Heuristic techniques need to be developed to solve our difficult problem. This is an area requiring much further research.

6.2.5 Pricing

As with any auction, we must define a rule for how much winning bidders pay. In Chapter 5, we discussed this issue at great length in the context of combinatorial auctions. In that context, where there were a fixed number of slots for auction, we noted that Rassenti et al. (1982) solve two pseudo-dual problems and determine approximate shadow prices that approximately divided the winning bids from the losing bids. Rassenti et al. find that while there is no set that perfectly divides the losing and winning bids, gaming does not offer very large rewards and is fraught with large risks. Thus, if it were possible, solving such a pseudo-dual to find approximate shadow prices would seem to be an adequate solution to our problem. It would offer a convenient, analytical approach, without much gaming. However, due to the non-analytical nature of our objective function, there is no dual problem to speak of. With this in mind, we must seek an alternative pricing solution. We would like to find a set of prices that encourage bidders to submit bids in the amount of their true valuations. Such a set of prices would have the following properties:

- The amount that a winning bidder pays is never more than the amount of the bid.
- The amount that a winning bidder pays is never influenced by the amount of that winning bid.
- The price that a losing bidder would pay if he had won would be greater than or equal to the amount of a losing bid; the losing bidder is better off losing than winning at that set of prices.

Let us consider a MSC pricing approach for a moment. We recall that in this scheme the winning bidder will be paid back for actual private delay costs incurred. MSC pricing principles would dictate that we would charge a toll equal to the expected MC, including both the expected MSC and expected private costs (since the operator will be reimbursed for the actual private delay costs). To what extent does it meet our criteria? The first one is definitely met because if the bid is lower than the expected MC, we would reject it because we could improve our objective function by removing the bid from the set of winning ones. Likewise,
the second one is met because the winner pays the expected MC, a quantity completely independent from his bid. It is the third criterion that is not satisfied. A trivial example illustrates the point. Suppose that the MC of a flight when there is one flight during a certain time period is $10, that the MC of a flight when there are two flights in this period is $20, and that two bidders A and B each submit a bid between $10 and $20 to operate one flight in this period. Solving our objective function will cause us to pick the higher bidder as the winner, but the winner will only pay $10. Clearly, the losing bidder would have been happy to win the bid if he had known that the price is a mere $10; it would have been to the losing bidder’s advantage to overstate the bid and bid as high as $19.99 in an effort to win. We now describe a slight modification to the MC pricing approach that will help diminish, if not eliminate, this problem.

This new pricing approach requires thinking about the origins of MC pricing. MC pricing stems from the idea that if a consumer is unwilling to pay for the marginal cost of consuming a good or service, society would be better off preventing that consumer from consuming it. In this case, we should think of an expanded MC, a MC that includes both the MC of congestion and the MC of preventing some losing bidders from having their bids accepted and enjoying the surplus that would accrue to them in that case. In other words, this price has two components. The first is the expected MC of congestion from using the awarded package of slots. The second is what loss of welfare we have from accepting this bid (completely ignoring the welfare that the winning bidder gets from being awarded the bid). This second component can be calculated as follows. First, calculate the expected surplus of all the accepted bids minus the surplus of the bid of interest. Then rerun the optimization problem optimizing over the set of bids excluding the bid of interest. This second quantity, the surplus that could have been created in lieu of having accepted the bid, is at least as great as the first, the surplus created under the current allocation less the surplus of the bid of interest; the difference between the two quantities is the second component to the price that the winning bidder should pay. With this pricing scheme, the incentive to overstate the bid is greatly reduced, as there is a greater danger of the bidder paying too much when overstating a bid, even if the MC of congestion is less than the amount of the bid. For example, in our case that we mentioned before, it would no longer be to the losing bidder’s advantage to bid $19.99 because the bidder will end up paying $10 plus the difference between the second highest bid and $10 (i.e. will pay
a price equal to the highest losing bid). While not every case is this simple, where there is no advantage to overstating the bid, this example demonstrates the dangers of overstating bids.

6.2.6 Administration

We must ask ourselves who could run such a program and how they would administer it. As the law currently stands, the federal government would need to legislate such a program or grant the F.A.A. the necessary powers to create such a program. To be most effective, the federal government should work with local airport authorities to establish this program with the goal that the airports take an active role in managing the program. There will surely be some issues that are beyond the scope of an individual airport, but the airports are going to be the best equipped to work through the tedious details.

6.3 Making the Marginal Social Cost Auction Viable Through Buy-in

Like any market-based approach to handling congestion, the MSC Auction will be difficult to implement because many of the system stakeholders may oppose it. Some may oppose it because of substantive concerns like the loss of their current property rights, while others may simply oppose it because they favor momentum and are suspicious of new things. To gain the support, or buy-in, of the system stakeholders, we need to be sensitive to their needs and incentives.

With regard to airport runway users, they can be divided into two broad categories: general aviation and commercial aviation. General aviation will mostly be against any market-based measures to alleviate congestion. The ratio of MSC to private cost of congestion for these general aviation operators is typically far higher than that of commercial aviation. In other words, they bear a relatively small portion of the cost of delays that they create in congested airports. They will therefore be unlikely to favor a proposal that forces them to pay for this large, essentially subsidized, MSC. However, general aviation will continue to have access to less congested airports at reasonable rates. The portion of the general aviation community that will be hit the hardest is the very wealthy who prefer to fly their private jets into more convenient airports such as Reagan National, La Guardia, Logan, and O'Hare. In our opinion, we should have little sympathy for making the very wealthy pay for what they use. To counter the general aviation’s lobbying efforts, a strong public relations campaign will need to be undertaken. Such a campaign could focus on how the current system subsidizes the wealthy,
general aviation operators by allowing them to unduly delay the less-wealthy segment of the population that uses commercial aviation. The importance of a campaign cannot be overestimated as the general aviation community has, in recent history, successfully blocked congestion pricing efforts at Boston's Logan Airport. While this community is politically influential and will likely try to stop any market based initiative, they are unlikely to overcome the desires of commercial airlines and the large segment of society who fly on commercial airlines, both of whom are currently bearing the burden of general aviation use of the nation's busiest airports and should be eager to shift the burden back to the general aviation community. Thus, in our opinion, increasing the likelihood of the MSC Auction is more dependent on being both fair and favorable to the airlines and flying public than on satisfying the general aviation community's desire for continued implicit subsidization.

For commercial airlines, a MSC auction will, in general, increase the cost of operating out of a congested airport. One way of lessening the burden of this increased cost would be to use this money to help the airline industry through subsidization. The proceeds of these auctions could be used to reduce the many taxes that airlines currently pay (or charge their customers), finance infrastructure improvements that would benefit the airline industry as a whole, or subsidize the airlines in some other way. Whatever the money is used for, the question of incentive compatibility must be addressed carefully. If airlines (particularly large ones) believe that a certain portion of the money that they pay comes back to them, they will have incentive to overstate bids, knowing full well that it does not cost them what they are paying, but only a fraction of that payment. There are ways to avoid this. For instance, the government could decide to cover the security cost of airports and hope that the revenues from these tolls will cover it, but ultimately accept responsibility for any shortfall. In this manner, airlines do not directly benefit from the amount of money raised and will be unlikely to think of the money as coming back to them through a subsidy. We have noted that without subsidization, the out-of-pocket costs will go up for many commercial airlines; however, the MSC Auction has many benefits for airlines. First and foremost, airlines will be able to reduce expenses such as crew and fuel by spending less time in holding patterns. Second, airlines will be able to increase productivity levels on their fleets and on their airport equipment, further reducing their cost structure. Third, passengers will be more satisfied with the service, possibly leading to increased load factors and/or yields. Lastly, the government should consider subsidizing the
industry to offset some of the increase in out-of-pocket costs. Such a plan, along with a coordinated information campaign to explain its merits to the airlines, may many of the airlines’ support.

We should also consider the impact on airports and the municipalities they serve. Airports are likely to oppose a MSC Auction because it involves significant implementation cost to them and does not really benefit them. Moreover, even if the airport is forced to adopt this scheme, they are unlikely to do a good job at managing it unless there is a performance-based incentive. To avoid opposition or inert implementation, the government could provide a financial incentive to the airports. One smart way of aligning the airports’ incentives with society’s would be for the government to pay the airports an amount equal to the increase in economic welfare that the airport is able to achieve plus reimburse them for the cost of their cost of reimbursing the runway operators for the cost of actual delays. In this way, the profit that the airport would make would be aligned with the increase in welfare that they create. This would not only encourage airports to actively manage such a process, but also encourage airports and their municipalities to make capacity adjustments to increase surplus. Airports will make capacity improvements when, and only when, the costs of such improvements are justified by the potential gain in surplus to airport users. Aligning the airports’ incentives with societal goals is of clear importance as the recent controversy over plans for El Toro Marine Corps Air Station suggests.

6.3.1 Case Study: El Toro Marine Corps Air Station

Orange County uses its John Wayne Airport to service commercial and general aviation. In 1993, when the military decided to shut down the El Toro Marine Corps Air Station, located in Orange County, the local government proposed turning it into a commercial airport. The county’s residents were divided over the proposal and were still split over whether to use it for an airport, parkland, or some other use in 1999, when the base closed. Many arguments have been cited for and against the airport, but the argument essentially boils down to two simple facts. The demand in the area of Los Angeles and Orange counties for airport use is growing and many of the area’s residents believe that more capacity is needed to meet this demand.

---

14 The increase in welfare plus the cost of delays is equivalent to paying them the value of the accepted bids less the welfare that existed at the airport before putting this scheme into place. This latter value may be hard to measure, but could be estimated by economists.
While not necessarily disagreeing with this assertion, the people living near El Toro do not want the burden of living near an airport (i.e. increased pollution and noise from the airplanes flying just a few feet above their houses). Moreover, even for the residents who live a little further out, it may be more desirable to turn the base into golf courses and parks and let the airport be a little further out or, even better, in neighboring Los Angeles County.

On March 5th, 2002, voters in Orange County chose to rezone El Toro for park and park-related use and ruled out the possibility of turning it into an airport. This vote was a large victory for the many, mostly affluent, local residents who had opposed the plans for an airport. However, not everyone was pleased. Many people living in the vicinity of Los Angeles International Airport (LAX) were particularly upset because they believe that ultimately it will lead to more traffic at and eventual expansion of LAX. This is a classic externalities problem. Everyone wants to be able to fly, but few people enjoy living near the airport and having low-flying planes pass over their houses. Let us look at how the MSC Auction that we proposed would address this problem.

First of all, if airports were to receive increased funding from the government in an amount equal to the increase in surplus that they create, reasonable levels of expansion may be more favorable. Expanding an airport would have two simultaneous effects. First, it would increase the number of slots that the airport can award in the auction process. Second, it would decrease the overall congestion. These two effects add up to creation of greater surplus. The county or city that runs that airport would benefit from the new funding. With financial incentive, cities and counties may be more inclined to increase runway capacity. Residents living near El Toro (or LAX) might be willing to see a commercial airport (or additional runway) built if they received something in exchange (e.g. more money for the local school system). There must be greater incentive to having an airport in the neighborhood than there currently is if this country is to meet the growing demand for airport access in major metropolitan areas; the proposal that we have outlined could create just such an incentive.

Of course, the story at El Toro is only one of many such stories. Here, in the Boston area, similar battles have fought, and will continue to be fought, over plans to expand Logan Airport. Many cities might be able to resolve such conflicts if they were able to directly profit from airport expansion.
6.4 Objections to the Marginal Social Cost Auction

While it may be difficult to anticipate every objection to this proposal, we can surely anticipate a few of them. Most of these objections are not specific to the approach that we suggest here, but rather are objections to market-based solutions to airport congestion. We address two such potential objections.

One of the objections to any form of market based approach to congestion management is that it would make it prohibitively expensive for airlines to provide nonstop service from small communities to congested airports and that such service is necessary for encouraging economic activity in these small communities. Thus, small communities have objected to market-based pricing schemes that have been proposed at LGA, commonly citing the critical nature of having nonstop service to the nation’s most important commercial city. These arguments are sound in as much as nonstop service may be important, but are flawed in one key respect. If it is very important for these communities to have nonstop service, they could pay airlines to offer such service and eliminate the financial burden of the MSC Auction for the airline. It makes more sense to have some small community offer direct subsidy if they think it is worthwhile than for all the passengers traveling to and from many other cities to subsidize the service by bearing increased delays as a result of such service. Moreover, it is worth noting that many cities offer subsidies to airlines to serve them and that such a concept is already widespread and should therefore not be objectionable, in concept, to small cities. Lastly, we note that such subsidies are also quite similar to the incentive that many of these cities offer sport franchises for locating in their city and that exempting these cities from a market-based mechanism to manage congestion would be no more logical than the residents of Boston, Massachusetts having to sponsor a sports team in Dayton, Ohio.

Along similar lines, many of the cities with congested airports allow general aviation to use those airports, even where there are less congested alternatives, because they want to encourage commerce. Currently, the access fees to these airports are often so low that there is little, or no, incentive for general aviation to use other airports. Here to, if the cities, business bureaus, or whoever wants to allow the private jets to use these congested airports, does not want the jet operators to face the increase in cost, they should subsidize their operations. However, it would be disingenuous to disguise the subsidization of general aviation through exemptions from congestion pricing and thereby hide the amount of the subsidization from
the general public. Rather, they should offer them direct financial subsidization so that it comes out of a budget that the public is able to monitor.

6.5 Operational Considerations

With the adoption of the MSC Auction, there would certainly be many operational issues that would need to be carefully considered. The intention of this chapter is to merely provide an overall blueprint to the MSC Auction, but not to address every possible issue that would need to be addressed to successfully implement a MSC Auction. Nonetheless, we shall address a few of the operational issues that we think will be quite important to making this scheme practicable.

One of the concerns that we have about the MSC Auction is how to enforce slot times, an issue that becomes particularly difficult if the slot times are very narrow. For instance, if slot times come in 15-minute intervals, how should an airport handle a flight that was scheduled to arrive/depart between 4:45 PM and 5:00 PM and instead requests permission to arrive/depart between 5:00 PM and 5:15 PM? On the one hand, denying all such requests would be unbearable for airlines because they cannot guarantee an on time operation under the best of circumstances. On the other hand, the auction will not work if an airline can buy a slot for between 4:45 PM and 5:00 PM and instead regularly use it to depart between 5:00 PM and 5:15 PM, which is perhaps a more desirable time. However, the airport has several means at its disposal to encourage flying in the allotted time, while allowing for the irregular operations that are sure to take place. For instance, if the airline requests to use its slot in a more desirable time, the airport can charge (in addition to the toll) the difference in MSC between the two periods, removing any incentive to bid on a cheaper slot and use it to fly during a more desirable time. Alternatively, the airport could use priority queuing and give priority to those who are operating on time. In this manner, a flight which operates before or after its scheduled time would essentially be operating on a stand-by basis, which is less desirable, but may be acceptable to those airlines that occasionally operate a flight ahead of or behind schedule.

Another concern that we have about this scheme is that there are operators who are really unsure of their planned operation time. For example, suppose that slots are given in 15-minute intervals and that some private jet service would like to have a slot to operate any time
between 7:00 AM and 8:00 AM because the actual time of operation will depend on when their client, on a given day, wishes to depart. Given the queuing model that we have suggested, it should not be difficult for the airport to estimate such a slot’s expected impact on congestion. The airport can estimate, based on experience, the relative likelihood of operating in each of the four 15-minute intervals and use this to calculate the expected impact on congestion and find the appropriate price to charge for such an operation.

Lastly, we should mention that we are concerned about the ability of general aviation to participate in such a scheme because many general aviation operators do not use an airport with any predictable pattern. For instance, the corporate executive who wishes to fly on the company jet into John F. Kennedy Airport on some afternoon may have planned that trip on the same day and would not have been able to reserve a slot a month or more in advance. How can an airport accommodate such flights under this scheme? There are at least two ways that we think this problem can be handled effectively. The first would be for the airport to forecast general aviation operations at the time of the slot auction, perhaps based on historical trends. With this forecast, the airport could properly optimize and charge prices that would include the expected general aviation operations in the calculations of MSC. Then, the general aviation operations could be charged some sort of MSC toll based on the forecasted MSC at the time that they request permission to conduct the operation. Another way to handle this would be to create a secondary market for slots. Companies that already handle the groundside of general aviation operations at airports would be well suited to buy slots and resell them to general aviation users. These companies would probably be interested in operating such secondary markets if they could use profit-maximizing pricing to recoup their investment and make a profit as well.

We have seen that there are several important complexities to address before this scheme can successfully be implemented. With some mix of creative thinking and trial-and-error, airports should be able to find satisfactory ways of dealing with these issues. While the details are clearly important, the main thing is for the airports to follow the basic principles of welfare maximization and auctions. That is, if airports use auctions to reveal valuations and a combination of optimization and a congestion model to determine how to best maximize welfare, the particulars of how they address tricky situations are of secondary importance.
Chapter 7

SUMMARY, CONTRIBUTIONS, AND FUTURE RESEARCH DIRECTIONS

7.1 Summary and Contributions

This thesis addresses four different research areas dealing with market-based solutions to airport congestion. The first of these areas is congestion models. In Chapter 3, we reviewed six different models of congestion. We compared them and discussed their relative advantages and disadvantages. Koopman (1972) presents a queuing model that is particularly appealing in its ability to predict queuing conditions given a schedule of operations. In Chapter 6, we suggested a new congestion model based on this model. This new model would primarily differ from Koopman's in its treatment of weather. Rather than assume that the conditions at any given time are independent of conditions immediately preceding or succeeding that moment in time, the model we suggest would use a two-staged approach. The first stage would characterize weather through a Monte Carlo simulation and the second stage would predict queuing conditions, using Koopman's model, based on the results of the Monte Carlo simulation.

In Chapter 4, we discussed MSC tolling. Jansson (1998) and others claim that, if demand functions are perfectly known, numerical methods can be used to find a toll set, which results in usage levels with the MSC of an operation type equal to the toll for that operation type. Jansson and others further claim that such a toll set would be optimal. We proved that, under a limited set of assumptions, a MSC toll set exists for a steady state queuing system with operators with varying sensitivities to cost, values of time, and service time distributions; however, we were unable to generalize this proof to a very broad range of models. Importantly, we showed that the MSC toll set is degenerate and that each MSC toll set corresponds to a local minimum or maximum in welfare. In general, only one MSC toll set corresponds to the optimal toll set, the one that globally maximizes welfare. Thus, we found that the MSC toll set is largely a misapplication of marginal cost pricing principles and should not be used as a method for maximizing welfare.
In Chapter 5, we explored slot auctions as a method for fairly and efficiently allocating a limited number of slots to those who value them most. In this context, we reviewed some of the basic results of auction theory as well as some of the particular characteristics of using a combinatorial auction for a slot auction. We found that, while a combinatorial auction will not eliminate all incentives for gaming and is thus not perfectly efficient, it is a good practical tool for allocating scarce slots. However, we found that the traditionally proposed slot auction is sub-optimal because it chooses how many slots to auction using a “judgmental” approach.

In Chapter 6, we introduced a new method for dealing with congested transportation facilities, the Marginal Social Cost Auction. This new method combines a combinatorial auction mechanism; a two-stage congestion model based on Koopman’s (1972), which we described above; and an objective function to maximize the net surplus that the airport can provide through issuing slots. Unlike either MSC pricing or the traditional slot auction, the MSC Auction is able to use information on both demand and congestion functions to both choose how many slots to make available and which bids to accept.

7.2 Future Research Directions

The goal of this thesis is to present a blueprint for the Marginal Social Cost Auction. As a blueprint, the thesis only provides a few of the details of how such an auction could be conducted and what the results of implementing such an auction would be. Further research is required to advance the ideas we have presented here into a detailed proposal.

One area that needs to be developed is the congestion model. The approach we suggested to congestion modeling has many details to be worked out. First, we need to carefully examine the Monte Carlo simulation of weather and airport operating patterns because it is supposed to characterize the variety of weather patterns that occur over some time period, which may be difficult given weather’s chaotic and unpredictable nature. Second, future research should investigate how well the model works when it combines the Monte Carlo simulation with the method that Koopman (1972) proposed. Third, we would like to know how well the model can integrate the many complexities that might need to be modeled in a congestion model (e.g. priority queuing, delayed operations, irregular general aviations operations, etc.).
Another area of research is investigating the properties of large-scale combinatorial slot auctions. We would like to know how the auction that Rassenti et al. (1982) proposed would work for a realistically large-scale problem. An experiment should be designed to test how well airlines that know the valuation of their slots are able to put together meaningful packages of slots to bid on and use logical constraints to maximize their benefit from the auction. This experiment will also need to look at what incentives, if any, bidders have to place bids that are not equal to their true valuations.

The third and largest area of research that needs to be conducted is on unique issues facing the Marginal Social Cost Auction. One challenge of the MSC Auction is developing an optimization routine that can find optimal or near-optimal solutions in a reasonable amount of time. In Chapter 6, we noted that it would be difficult to find such a solution because it involves solving a large combinatorial problem without an analytical objective function, something that is notoriously difficult. Further research is required to determine whether an analytical approximate congestion model can be developed to assist in developing an optimization technique.

Another challenge in the MSC Auction is to determine to what extent the auction encourages truth-telling behavior on the one hand and gaming on the other hand. This is a similar question to the one that we asked about the combinatorial auction combined with administrative limits that Rassenti et al. (1982) proposed, but differs in two key respects. First, the pricing scheme that we have proposed in the MSC Auction is different than the one Rassenti et al. proposed and we therefore cannot infer from the experimental results on gaming behavior that Rassenti et al. reported. Thus, the question of to what extent we can find a fair pricing scheme that discourages gaming in bids is an important area for future experimental research. Second, we noted that the MSC Auction has the possibility for gaming on the cost estimates, something unique to this auction. While we noted that this might not be a severe problem because the cost estimates could be partially or completely controlled by the airport, it is definitely of interest to know whether the auction administrator should or should not believe the cost estimates.

Lastly, we mention that determining what to do with the proceeds is a promising area of research that is of some theoretical importance and very important for practical reasons. The
practical aspect is that smart use of the proceeds could help in getting acceptance of this scheme from airlines and airports, either of which may be able to stop it through lobbying efforts. The theoretical aspect of interest is in designing uses of the proceeds that encourage efficient decision-making. For instance, we would like to design and test subsidies that encourage airlines to only fly when it is efficient for them to do so and airports to add capacity when it is beneficial to society.

In closing, we would like to emphasize the importance of continued research in this field. The basics of auction theory and MSC pricing that we have drawn on in suggesting the MSC Auction were laid out in the 1960s and 1970s, some quarter of a century ago. Since that time, airport congestion has worsened, but there has been little movement toward finding practical, yet theoretically sound solutions that can fix the deteriorating air transportation system. For this to happen, there must be a reinvigoration of research in this area. It is our hope that the introduction of the MSC Auction contributes in this respect.
REFERENCES


