Yield Management for Telecommunication Networks: Defining a New Landscape

by

Salal Humair

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

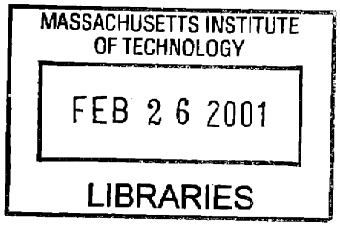
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2001

Author .....
Operations Research Center
January 18, 2001

Certified by .....
Richard C. Larson
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Cynthia Barnhart
Co-Director, Operations Research Center



Yield Management for Telecommunication Networks: Defining a New Landscape

by

Salal Humair

Submitted to the Sloan School of Management
on January 18, 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Can airline Yield Management strategies be used to generate additional revenue from spare capacity in telecom networks? Pundits believe "yes", based on several analogies between the industries such as, for instance, perishable inventory and negligible marginal cost of usage. However, no one has yet described how, one of the chief difficulties being the vastly different nature of airlines products and telecom services.

Motivated to show how Operations Research can play a role in structuring this area, we: (i) argue that telecom Yield Management should be based on 'innovative' services explicitly designed to use only spare capacity, (ii) propose, borrowing from airlines, a framework to simplify related decision modeling, and (iii) demonstrate both our argument and the framework by articulating several 'innovative' telecom services and modeling them to varying degrees of depth.

This thesis focuses only on the decision-making that will be required within a large infrastructure for operating new 'Yield Management' services. For each service, several decision variables can be considered to maximize revenue from available capacity, e.g. pricing, capacity limits and admission control, among others. Incorporating all such decisions in a single model usually leads to complicated formulations. A framework that decouples the decisions from each other to obtain simpler, more insightful models is therefore immensely helpful.

We propose using the airlines modeling framework to separate the decisions involved in the operation of each new service. This framework classifies models into *forecasting, over-booking, seat-inventory control, pricing* and *market segmentation* to reduce the complexity of the system-wide problem. To make this framework useful for telecom, we provide a detailed interpretation of each category in the telecom context.

Finally, the majority of this thesis is the six service ideas that illustrate our argument and the models that demonstrate how the framework might be used. For each service we propose, we discuss possible markets and practical issues. We then formulate a model for one of the decisions resulting from the framework. These models are analyzed to varying depths to demonstrate the operating rules one can discover for revenue maximization.

The contributions of this work are at multiple levels. In addition to our argument and examples of services proposed for telecom Yield Management, it structures the modeling questions in a coherent manner, exploiting more than only the high-level connections between airlines and telecom. Finally, the models themselves are useful and their contributions are at the analytical level. This thesis makes clear several connections between airline and telecom Yield Management that people have found difficult to establish in the past.

Thesis Supervisor: Richard C. Larson
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I place the guilt of this document on the following people, as partial atonement for the injustice that it bears my name but could not have existed without them.

Dick and my committee: Dick of course, as my advisor and mentor, deserves the first mention for his tremendous patience and understanding during my many floundering years, for giving me complete freedom and for un-shackling my academic chains. My committee: Les Servi, John Tsitsiklis and Ismail Chabini. Les, who it was my extreme good luck to have met, who taught me much about research and writing, and for his incomprehensible belief in my abilities. John and Ismail for being extremely supportive and understanding during the development of this thesis, even through the initial unclear and unfocused attempts. Special thanks are also due to Iraj Saniee and Debasis Mitra[1], for helping develop the foundations of three models in this thesis during one fruitful summer.

The wonderful OR center staff: Paulette, Danielle and Laura who were instrumental to my MIT existence – always entertaining my sporadic silly demands and protecting me from being kicked out when I routinely forgot to honor bureaucratic obligations, such as registering each term.

The 1.00 clan who were my real community at MIT: The faculty, Professors Nigel Wilson, Steve Lerman, Jud Harward, George Kocur, Eduardo Kausel and Bob Logcher were always supportive whenever balancing teaching and research became a circus act. Nigel deserves special thanks for his advice during many difficult times and for the wonderful dinners at his house. George showed me again, after I had forgotten, that there was no substitute for hard work and conscientiousness when it comes to teaching. My fellow 1.00 TAs, Terry, Nadine, Li and many others over the years who were my kinsfolk at MIT. And the Civil Engineering staff who dealt with our interminable 1.00-related requests, Cynthia Stewart, Pat Dixon, Stephanie Bowes and Ginny Siggia; they were all marvelous.

My friends: Asad Mehboob Ali, Musa and Raza; Khurram, Kanwal, Adil, Zarmeena, Adil Najam, Huma, Babar and Samia; and Catalina and Greg, my co-sufferers through MIT, who were all smart, caring and funny enough to keep me distracted.

A final tribute remains for the people who matter the most. To arrest their contribution in words would belittle it. During these years of deep contradictions, they were the vital cords of my life. My father, mother and grandmother; my brother and sister; and my two closest friends, Iqbal Ahmed and Qadir Wahid; this thesis I dedicate to them!

[1]Both at Lucent Technologies, Bell Laboratories, NJ.

Contents

Chapter 1

Introduction

Motivated by the success of *Yield Management* (YM), later called *Revenue Management* in airlines and other industries, we ask how similar ideas might apply to telecom networks. It has been widely recognized that airline YM can have a strong impact in telecom (c.f. section 1.6), but no clear case has been demonstrated yet. The chief difficulty here is the stark difference between an airline seat and a telecom service, both in terms of the price sensitivities and the market structure.

The first contribution of this thesis is the argument that telecom YM should be based on innovative services offered explicitly to segment the market and *only* use spare capacity. The reasons motivating this approach are discussed in section 1.3. Our argument, which sounds rather simple in hindsight, once one has seen examples of the services that can be created, is not yet a common practice in telecom.

Offering and operating the services we envisage requires an infrastructure of software agents and other network modifications. Of the many involved issues, the focus of our work is the decision-making that can be pre-programmed into such agents to maximize revenue from the service operations. Because many decisions such as pricing, capacity determination, admission control etc. are relevant, and because incorporating all of them into a single model is usually too complicated, a framework that decouples the decisions to obtain simpler, more insightful models is immensely useful.

The second contribution of this thesis is the modeling framework it proposes for decomposing the decision models. We borrow the YM framework from airlines, of *forecasting, over-booking, seat-inventory control, pricing* and *market segmentation* to separate the decisions involved in the operation of each service. Even though this framework is intended for decision problems related to airlines operations, the abstract versions of the problems are very relevant for telecom. Chapter 2 outlines how this framework can be interpreted in the telecom context. This allows us to use not only the high-level intuition about YM from airlines but also the modeling lessons learnt from their successes.

The final contribution of this thesis, and the majority of the work herein, are the services and models that illustrate how our argument can be realized and our modeling framework can be used. We propose several services, articulate possible modes for their operations, outline their target markets and discuss practical issues. For each service, we then model one of the decisions resulting from the framework. These models are analyzed to differing depths to (i) illustrate the decision rules one might discover for revenue maximizing operation of the services and (ii) demonstrate how these rules parallel many of the airline YM rules. We hope to highlight the connections between airline and telecom YM at this level, even though the telecom services are seemingly far-removed from the airlines context.

This first chapter introduces our argument for telecom YM, which is original, but mainly provides the background and the context for our work. All other chapters present completely original work,

including the translation of the airlines YM framework to telecom, the ideas for services that will facilitate telecom YM, and the model formulations. Chapter 2, which details our modeling framework is particularly useful as it also presents an overview of the services and the models considered in this thesis. Each subsequent chapter addresses a model for a particular service. Several other issues which are not explored in this thesis but will arise in offering YM services are considered only briefly in section 1.3.2, such as marketing, software design, protocol design and information technology.

Ideally, this thesis should excite three groups of people. *Researchers and developers* in telecom who will encounter new challenges in implementation of innovative services for revenue management, *operations researchers* interested in modeling new applications, and the *management* in telecom companies, who might discover a new competitive tool in a fast changing environment. In a less than ideal world, however, our least hope is that the models explored in this paper are interesting in themselves, and relevant to some context even if not immediately to revenue management in telecom.

The rest of this chapter is organized as follows. Section 1.1 is a primer on airline YM practices, the underlying intuition and the perspective it brings to telecom. Section 1.2 presents our motivation, which stems from more than the revenue potential of YM, a perspective on our contributions and the limitations of this work. We present and discuss our argument for using innovative services for telecom YM in section 1.3. This section also places our modeling effort in context, discusses several practical considerations and how a similar approach might apply to other networks. Section 1.4 presents the outline of the thesis chapters. The rest of the chapter contains the background for this work. Section 1.5 is short history of airline YM and section 1.6 presents relevant telecom background, such as the industry trends motivating telecom YM and analogies as well as differences between airlines and telecom. Section 1.7 presents a literature review of airline and telecom YM. Section 1.8 is the summary of this chapter.

1.1 A Yield Management Primer

To understand the motivation, consider YM as invented by the deregulated U.S. airlines in the 1980's. American Airlines claims that intelligent revenue management brings annually an additional $500,000,000 to their bottom line [SLD92]. The concepts brought to the airlines include optimal over-booking policies, nested discount-seat allocation rules for flight cabins, demand forecasting, dynamic adjustments of discount capacities in each flight, multi-hop pricing structures and frequent-flyer programs that minimally interfere with revenue producing passengers, among many others.

Airline YM capitalizes on the simple intuition that one can provide service for as low as the marginal cost when one is certain of excess capacity which will not be utilized. The canonical example is of an empty seat on a flight that yields no revenue after the plane takes off. If one knew for certain that a seat will fly empty otherwise, one would be willing to sell it for as low as the price of a lunch or dinner - the marginal cost of serving an extra customer. Among other factors, the complexity in the process stems from the following: (i) it is not known for certain if a seat will go empty and (ii) since tickets are not negotiated on an individual basis, making seats available at a discount price always increases the chance that people who would otherwise be willing to pay more would probably pay lower prices, thus decreasing overall revenue for the flight. Airlines therefore seek to intelligently control the availability of discount fares on individual flights given demand forecasts, using a combination of the following activities: *forecasting* demand, *over-booking* flights in anticipation of cancellations and no-shows, using sophisticated *seat-inventory control* rules to prevent revenue cannibalization across multiple fare-classes and *pricing* to increase revenue from the network.

This last decade has also seen the application of YM to other industries such as hotel/resort management and car rental agencies, among others. Telecommunications and utility industries are often mentioned as the next targets [All97] [Bru97], and several companies specializing in 'Revenue

Management' are actively engaged in such attempts[1]. Every application of YM to another industry, however, has required considerable tailoring to its specific constraints. Also, the revenue improvements are highly dependent on cultural acceptance of the whole paradigm in the organizations and the willingness of the managers to support the related decision tools. Several difficulties involved in applying YM to telecom are presented in section 1.6.2.

The fundamental viewpoint brought to a telecom network by YM is to view the bandwidth of the network as a perishable commodity which vanishes if not immediately used. Further, the cost of using the network when capacity is available is negligible. Therefore, it makes sense to use this capacity when it is available and attempt to maximize revenue from it when it is scarce. Operationalizing this simple intuition, however, is not obvious. Several complications arise when trying to map a one-dimensional commodity like an airline seat to a continuous commodity like transmission rates and switch utilizations. The different structure of the market for telecom services and the relative magnitude of the revenue involved in servicing a request makes it further unclear if the benefits of airline YM can be realized in telecom. Hence our argument for introducing new services to capitalize on the YM intuition, as described in section 1.3.

One additional note on revenue benefits is that short-term revenue increase is just one aspect of a larger picture. As exemplified by airlines, YM is usually accompanied by a longer-term reduction of investment for capacity increases. This results from the attendant increase in the utilization levels of the resources – such as fuller airplanes, by displacing peak demands to off-peak times. In a competitive market, long-term cost reductions for the providers should ultimately lead to lower costs for the consumers. With our argument for telecom YM, additional benefits to consumers could come from the many new services created to 'yield manage' networks, leading to greater choice and flexibility.

1.2 Motivation, Contributions and Limitations of this Work

The successes of airlines YM were one motivation for this work. Similarly, the contributions to telecom YM are one aspect of its value. Here we discuss our other motivation and related contributions, as well as the limitations of this work.

Our main motivation from an Operations Research (OR) perspective was to: (i) show how OR modeling and analysis can help structure a nebulous domain, and (ii) demonstrate a fertile new modeling area to the OR community, resulting from opportunities to optimally operate new YM services. Airline YM has been one of the great success stories of applied OR in the last few decades and has spurred much interest in the discipline. Telecom YM will likely excite and motivate significant OR modeling work.

One value of this work therefore is in re-iterating the value of OR as more than a narrow discipline restricted to modeling existing operations and processes. Our work demonstrates again how one can use modeling lessons from one domain to "invent" services and operations in another. There is as much room for invention of operations in OR as in other fields, but this requires an imaginative license and an awareness of the real environment in which such services are offered. Examples of such work exist in OR but are relatively infrequent – a purely personal opinion of the author.

The relationship of modeling to reality needs to be understood here in order to place this contribution. As always, reality is far more complicated than desired. The practical complexities of telecom YM, very similar to airline YM, will be daunting – including the implementation of an infrastructure needed to support the operations of new services, their marketing and data collection. Modeling and analysis, even though simplified and only one aspect of the operational infrastructure, can still play a strong supporting role by discovering optimal operating rules for these services. It is important to realize that similar modeling, by approximating actual operations in a highly simplified

[1] A cursory search of the Internet immediately reveals the number of players in the revenue management arena.

manner, has made a strong impact on the profitability of airlines. The contributions of modelers are to discover the first-order factors that can help achieve significant revenue increases from operations of such services. This can only be done once the services are implemented. This is why we can only make the case in this thesis for how telecom YM might evolve according to our argument. The practical implementation might rely on vastly different services and models than those considered in this thesis, using our argument and modeling framework.

As for our second stated motivation, we structure a modeling area for the OR community, and make the case for its use. It is also useful to look at the relationship of YM modeling to other models in telecom and airlines as shown in figure 1-1, to understand this aspect. However, the speculative nature of this work means unequivocal claims are not possible. Detailed contributions expected from the modeling and analysis work will only make sensible reading once the models have been presented. Therefore they are presented at the end of each modeling chapter. As a general comment, the models we analyze not only illustrate the connections to airline YM rules but also reinforce interest in some of the emerging analytical directions, such as, for instance, analysis of multi-server queues with heavy-tailed service times (c.f. chapter 3). Further, not only are these models of interest in telecom but their analyses can perhaps also contribute to the airlines YM literature, a connection which we do not pursue in this thesis.

A last contribution of our argument and modeling framework for telecom YM could be the useful perspective it brings to management of other networks, such as energy or utility networks, which share the perishability of inventory and negligible marginal cost characteristics. We hope that examples of new telecom services and use of the airlines framework as we propose will excite researchers to identify and model new services for other networks, such as transportation or electricity distribution networks (c.f. section 1.3.3).

On the limitations front, the speculative nature of this work makes it hard to predict its practical usefulness. Market acceptance of the services, validity of the models and their effect on actual revenue are all important matters, which cannot be addressed by a single thesis. At this point, models for YM-type services can only serve as proof-of-concept that one can build and operate such services if desired.

A very pertinent example of speculative modeling in the same spirit as ours was Vickrey's [Vic72] bold solution for airline over-booking in the early 70's. Envisioning a future in which airline reservations were a commodity business, Vickrey proposed and developed a remarkably elaborate model for over-booking decisions. His vision was too futuristic for its time and most people dismissed it immediately. Interestingly though, markets are beginning to now evolve on lines proposed and modeled by Vickrey. A similar example was Simon's proposed auctioning procedure [Sim68] for bumping excess customers on over-booked flights. Again, it was debunked when it first appeared, a practice that is now prevalent among almost all carriers.

1.3 Our Argument: Using Innovative Services for Telecom Yield Management

Our argument is simple. Drawing on the trends of new service offerings in the telecom sector (c.f. section 1.6.3), we ask the question, why not offer services that (i) create 'flexible' demand, i.e. which can be managed easily to use *only* spare capacity and/or (ii) use network information advantageously to generate additional revenue.

This exactly parallels the start of airlines YM. The first steps towards airline YM date back to the early 70s, to experiments with products that could effectively segment the market, such as the 21-day advance bookings [Lit72] by BOAC around '72, Super-saver fares by American Airlines around '77 and the now familiar Saturday night stay-overs. Once the existence of market segments for the different products was clearly established, it created the problem of protecting seats across different fares, leading to Yield Management techniques. In a similar spirit, we propose creating new services

that operate at near-zero marginal cost by using only spare capacity and segment the market using network information whenever possible. The vision is that similar to airlines, many such services will find market acceptance, their implementation will spawn problems similar to airline YM problems and the rules for their profitable operation will resemble the airline decision rules, such as optimal over-booking levels, nested seat-allocations, etc.

Our argument is motivated by the difference in the nature of existing telecom services compared to airlines, making a direct translation of airline YM to telecom difficult. Airline YM involves modeling and controlling an accepted existing process for reservations, optimizing it to yield maximum revenue from resources. It is by far unclear if the existing telecom services allow similar flexibility in resource usage. Current network usage is mostly on-demand, making pricing the most obvious option for revenue management – controlling peak usage and discounting in non-peak hours. With new services created explicitly for YM, one obtains far more flexibility in usage of network resources such as spare capacity.

What services might one think of in the spirit of our argument? Consider some of the ideas proposed in this thesis. Further details are presented in chapter 2 where connections to airline YM are also discussed. Chapter 5 models a data-courier service similar to FedEx^{TM}, where the provider uses spare capacity to guarantee content delivery by fixed deadlines. Similarly, chapter 4 presents a decision model for a service where users request the network to place a call before a user-specified latest-start-time rather than require instantaneous service, allowing the provider to schedule these calls using spare capacity. Chapter 7 presents a model for offering service guarantees for Web-server performance to better segment the market, using network information to compute the guarantees.

For any new YM service, a network infrastructure will be needed to support its operation once a case has been made for its implementation (c.f. section 1.3.2). This is likely to involve network agents/software for managing capacity and demand to maximize revenue. Decision-making will be an integral part of these agents and much of it will need to be pre-programmed within them.

It is here that our next contribution lies. Opportunities exist for models, optimization, and construction of decision rules for these agents. This thesis structures and simplifies such modeling using lessons from airline YM. Chapter 2 discusses in detail how the airlines framework can be interpreted in a telecom context to construct useful models.

1.3.1 Yield Management models vs. existing telecom models

To situate the YM modeling effort among existing telecom modeling literature, we briefly characterize the current space of telecom models. This highlights how YM models fill a niche in this space and the difference in perspective brought by YM to telecom modeling.

	Design	**Operational**	**Yield Management**
Telecom	Network design Routing design	Component design (e.g. switches, routers, ...) Performance evaluation (e.g. blocking probs, ...)	Pricing?
Airlines	Network design	Fleet assignment Crew assignment	Forecasting Over-booking Seat-inventory control Pricing

Figure 1-1: A perspective on telecom and airlines Yield Management models.

Extensive modeling and analysis has been done for telecom networks, and a large body of literature exists for every aspect of their design and operation. In spite of the extremely complicated nature of telecom networks – with millions of components such as switches, routers, peripherals, etc. acting on local information, the network behaves as a single, almost intelligent organism. These networks respond to demand and traffic fluctuations to reconfigure themselves, re-route traffic and to do fault checking seamlessly. Much of this behavior relies on a large modeling effort over the past century. A simplified perspective on this body of knowledge is to classify the models according to time-scales. For instance, depending on the abstraction level and the time-scale at which one is interested in system behavior, one can classify telecom models as either *cell/burst-scale*, *call-scale* or *network-scale*.

When the questions of interest are at the component design and analysis level, such as switch buffer dimensioning, switching algorithm design etc. the *cell/burst-scale* is of interest. Here, traffic consists either of discrete entities, the cells, or the instantaneous transmission rates of individual sources [RMVe96]. Cell arrival models range from Poisson and renewal processes to more complicated Markovian models [RMVe96]. Instantaneous rate models can range from short to long range dependent models, such as on-off models, renewal rate processes and others [RMVe96]. Typically, the network-wide effect of component analysis and modeling is difficult to estimate and is explored through simulations.

When the questions of interest are in optimal use of network resources using admission control, routing, etc. one considers system behavior at the *call-scale*. Here one ignores the granularity of cells and the rate fluctuations of individual sources. Traffic is characterized by arrival rates and holding times of calls. Call arrival processes are typically smoother than cell arrival processes and models of interest are Stochastic Knapsack-type [Ros95]. Demand is assumed exogenous, and some state information about the network is used, such as the number of calls on the various routes, but no forecasting or other knowledge of demand is assumed.

Finally, when network-topology design and resource dimensioning is needed, one considers the system at the *network-scale*. Deterministic measures of demand such as peak arrival rates are used and the models of interest are mixed-integer linear/non-linear deterministic/stochastic programs [RMVe96]. At this level, demand is again exogenous, and relatively static decisions need to be made. Extensive work has been done in this and the areas described above. Two good references for such models are [Ros95] and [RMVe96].

In this space, room exists for models which use network traffic information such as daily or hourly patterns to operate services that use only available capacity at any time. Further, such models may also have some limited information about existing demand, such as the amount of unfinished content to be transmitted between locations, and or the number of calls that must be served before some given deadline. With partial information about demand and capacity, such models may more effectively match demand and capacity to generate positive revenue. This is exactly the perspective we hope to demonstrate in our work: that given tactical level information about the network – somewhere between the static level planning and the bit-byte level detail, useful decision models for services can make a substantial impact on revenue by capitalizing on the information.

1.3.2 Practical Issues: marketing, software, protocols, modeling

Implementation of YM services will not be limited to modeling and analysis and will require assault on multiple fronts. We comment on several practical issues without laying claim to this being an exhaustive list. The analogy, not surprisingly, is again from airlines. Airline YM systems were built on an existent information technology infrastructure that could be leveraged by YM models to make intelligent decisions. Benefits from YM directly motivated investments into better infrastructure for obtaining more information, integrating YM models with the reservations systems and developing software – interfaces, back-end systems and decision models.

Similarly, one expects that some basic information and capacity management infrastructure

will be needed in order to control capacity usage across "Yield Managed" traffic and for service-operating software to obtain adequate network information in order to make intelligent decisions. When resulting benefits from such YM services are demonstrated, they might lead to investments in better management infrastructure for these services. Both the nature of decisions that can be made and implementation constraints are so tightly coupled that one cannot be considered without the other.

For instance, if a service requires transporting bulk content between locations, its operating software may consist of distributed agents at select ingress points to the network, making decisions about if to accept content or off-load it to an alternate provider. In this case, each agent will need some notion of available capacity between locations and the amount of bulk content at other locations that need transport. This information may either be obtained directly from other agents, or from a central database tracking summary information for all agents. In either case, one will need protocols for information exchange between the agents and/or the controller. Clearly, if decision-making is embedded in the agents, the models will be restricted by the amount of information that can be reasonably exchanged by the agents. On the other hand, if the agents are "dumb", simply querying the decisions from a central model, the nature of the decision models will be completely different.

Several such issues that will arise in implementing a service idea once its market viability has been established are listed below.

1. *Obtaining network information*: YM services cannot be operated without information about available capacity. Collecting and summarizing network information is therefore essential to any such effort. Unfortunately, in spite of its simple sounding nature, collecting network information is a significant problem for Internet-type networks, see for instance [Pax97], and will need to be addressed.

2. *Allocating capacity*: To ensure that discount or yield-managed traffic interferes minimally with premium or existing traffic on the network, mechanisms must be built. Several possibilities arise here, depending on the implementation. For instance, limits may be set, e.g. on a *time-of-day* basis, on the maximum capacity usable by all discount or "yield-managed" traffic. In this case, protocols and mechanisms will need to be developed to allow reconfiguring the network at periodic intervals, including a central database which contains information on historical traffic patterns.

3. *Developing service-management software*: Implementing controllers or decision makers for operating the services will be an interesting and challenging problem encompassing software, modeling and communication protocols. The decision of a central or distributed control architecture will govern how much information can reasonably be exchanged between distributed agents and will limit the range of suitable decision models. It will also govern the software implementation of the agents and that of the central databases. In any architectural choice, protocols will need to be developed for agents, controllers and databases to interact with each other.

4. *Marketing*: Finally, organizational and marketing issues will need to be addressed for many YM services. Marketing new services may require channels for easy dissemination of service information to the most interested parties. One may also wish to enter into agreements with other network providers for off-loading selected traffic to their networks during periods of overloads, such as bandwidth exchanges for instance.

Fortunately, creating new services for telecom networks is now easier than before. Original telecom systems were massive and integrated, making changes cost-prohibitive. Furthermore, functionality needed for the management of new services was difficult and time consuming to provide. The life cycle for new service development extended from one to two years. Since market opportunities afford far less time – generally six months or so, Java^{TM} architectures such as JINI^{TM} have been developed to facilitate the creation of new services, cutting development time down to the required six months in some instances.

1.3.3 Applications to other networks: energy, utility, etc.

Our argument for using 'innovative' services for telecom YM could find useful application in management of other networks which have fixed capacity, time-varying utilization, and negligible marginal cost of providing service. Examples could include electricity, oil-and-gas pipelines and transportation networks.

Since most such networks suffer from problems similar to telecom networks, namely, that the market and service structure does not directly map to airlines, the approach is to first discover creative uses of the networks, i.e. design and offer services to segment the market and manage demand, to fill in the valleys and shave the peaks. Again, this involves both an understanding of the practical structure of the market these networks cater to and an operational understanding of the networks themselves.

In fact, it might be useful for the reader to indulge in the exercise of imagining new services for these networks to get a sense of the contribution involved in this thesis, and the work entailed in creating practical service ideas. We mention a possible idea below to demonstrate the kind of thinking that might result from such an exercise, for purely illustrative purposes.

Consider an oil-and-gas pipeline company. Demand is strongly seasonal and there is little one can do in terms of pricing to influence demand. Consumers simply do not appreciably increase their gas usage during summer months proportional to discount, even if they bought storage facilities. However, during summer months electricity usage peaks. How are gas and electricity related? Well, some electricity generation plants use oil and/or gas. If one could discount gas in the summer to these plants, perhaps one could keep the networks filled. But discounting usually means that in order to preserve profits, sales volume have to go up significantly. If one is concerned that simply discounting may result in cannibalizing revenue as these plants may have huge storage facilities allowing them to store gas in the summer for the winter, one might consider alternatives. For instance, one way to preserve revenues could be to buy stock in public electricity generation plants and sell gas to them at deep discounts, since marginal costs of transport are nearly zero, thus increasing the profitability of the plants and the overall revenue from the combined infrastructure of the network and the plant investments.

Once one has settled on some of the service ideas thus created, the next job is to demonstrate the benefits that might result, by modeling the services to discover the decisions required for their operation – an exercise we carry out for telecom networks.

1.4 Thesis Outline

We list brief synopses of the chapters that follow. This first chapter provides a broad overview of the context in which this work is situated. Here we also state our argument for telecom YM to be based on innovative services. Chapter 2 is a detailed explanation of our proposed modeling framework. Each subsequent chapter articulates a YM service in detail and models one of the decisions involved in its operation.

Chapter 2 argues the need for using a modeling framework. It translates each component of the airlines YM framework to telecom, circumscribing decisions that might correspond to each framework category. It also introduces briefly the services proposed in this thesis. A summary of the modeling work of later chapters is also presented.

Chapter 3 is the start of the modeling chapters. There we consider a service where the operator attempts to use spare capacity on the backbone to allow bulk transport. We articulate the service and describe its likely markets. We model a capacity determination decision for a single-link. Using exponential random variables, we propose a heuristic policy which performs very close to the optimal and is insensitive to the distributions of the random variables involved.

Chapter 4 outlines two possible services for which customers request to use the network be-

fore a latest-start-time. Here we again consider a capacity determination decision for a single-link model. We outline the complexity of the decision and analyze a simplified policy for several different distributions of the involved random variables.

Chapter 5 proposes a digital courier service where users ask for content to be delivered between locations by fixed deadlines. We discuss aspects of the service's operation and model a decision to determine which content sizes to accept, given the amount of work in the system and the time till the deadline. We analyze in detail a canonical single-link single-deadline model and extend the results to several interesting cases. Many directions for further research are also outlined.

Chapter 6 considers a data-collection related problem for Internet-type networks. It illustrates the problems that might arise for forecasting capacity in telecom. The issue is to decide where to locate a given number of probes on the network to build a picture of the traffic. We formulate an integer program for the problem, show its NP-hardness and outline a greedy heuristic with a provable worst-case performance.

Chapter 7 proposes segmenting the market for Web/Ftp-server hosting using service-level guarantees. We discuss the problem and propose using stochastic facility location theory to compute the service-level guarantees. A few results are included to show the utility of this approach.

Chapter 8 considers a service which offers discounts for long-distance calls at periodic intervals. We discuss the service and consider the pricing decisions. Our argument is to consider revenue cannibalization as the main factor in the discounts. For this we outline a very simple model to make our case.

The level of analysis in chapters 7 and 8 is far less than in the others. Their main purpose is (i) to illustrate our modeling framework of chapter 2 and (ii) to outline interesting research directions.

Chapter 9 is the final summary of the thesis, addressing issues like contributions not mentioned elsewhere and future directions.

1.5 Airline Yield Management: A Brief History

The following is an abridged history of YM to provide a context for the work to follow. Detailed accounts of airline YM history can be found in several references listed in the literature review section 1.7.

The YM practice in airlines can be segmented into three portions, pre-1972, 72-77 and post 77. Airlines were some of the earliest investors in digital computers and automated reservations systems and had access to detailed reservations records for their networks. Leveraging this information to increase revenue from the system was a strong motivating factor in the development of YM tools.

Prior to 1972, most research focused on *forecasting* and controlled *over-bookings*. Forecasting was the problem of predicting probability distributions of customers from among the accepted reservations actually appearing for boarding. Over-booking was the problem of deciding how many reservations to accept for a given flight based on the forecasts. Naturally, given a positive *no-show* probability for reservations, one would accept more requests than the capacity of the plane, the trade-off being the risk of denying a passenger a seat because of over-booking. For a relatively long time, over-booking was a well-kept secret and the airlines refused to acknowledge that there was any such practice. See [Rot85] for a fascinating perspective on the history of over-booking in airlines.

In early 1970s, airlines began offering different fares for the same flights resulting in mixing passengers with different revenues in the same cabin. Among the first such offerings was BOAC's (now British Airways) *early-bird* bookings, selling discount seats to passengers booking at least 21 days in advance of the flight. This offered an opportunity to fill seats that would otherwise be empty, but created the problem of deciding how many seats to protect for later booking customers. If too few seats were protected, the airline ran the risk of denying a reservation request to a later arriving – and higher revenue, customer. If too many were protected, one ran the risk of flying empty seats.

Simple rules, such as protecting fixed percentages of seats on different flights, were not useful because of widely different demand patterns for different flights. This led to investigation of *seat inventory control rules*, marking the beginning of *Yield Management*, now referred to as *Revenue Management*. It also created new forecasting problems, namely the prediction of the arrival processes of various customer classes for reservations.

The final phase of Yield Management was spurred in 1977 by the launch of the American Airlines' *super-saver* fares in April, 1977, shortly before the de-regulation of the U.S. airline industry. A decade-long process was catapulted into spotlight in 1987 by American Airlines announcing $500 million/yr increase in revenue from YM systems alone. Since 1977, YM systems have developed from simple *single-leg flight control*, to *segment control*[2], where multi-hop flights are considered to consist of separate segments with seat-level control applied to each segment, to finally *origin-destination control*. These advances have in turn required more investment in sophisticated information systems and quantitative research. Currently, almost all major carriers in the world and a significant number of smaller airlines have revenue management capabilities. Additionally, it is considered a major competitive tool in price wars.

1.6 Telecom Yield Management: Setting the Stage

This section includes information relevant to YM in telecom, including a brief overview of the telecom industry in the US, the analogies and differences between airlines and telecom pertinent to YM, perspectives from industry regarding the need for telecom YM and a discussion of the possible impacts of YM in telecom.

1.6.1 The telecommunications industry

The Telecommunications Industry is one of the largest sectors of the US economy: revenues for 1993 being in excess of $200 billion with 80% of the revenues coming from the sale of services and not of physical products [Nol96]. Many orders of magnitude larger than airlines – compare, for instance, the largest carrier American Airlines at $17 billion to AT&T with $75 billion in revenues for 1997 (The smaller "Baby Bells" routinely average around $10-15 billion in revenues) – it is also an industry in a major state of flux for the past two decades. Almost totally regulated for the entire duration of its existence, the industry now finds itself in uncharted waters in the wake of the Telecommunications Act of 1996 which is aimed at encouraging competition in all kinds of communications facilities and providing *Universal Service* to the entire nation. Historical parallels with the evolution of the airline industry are remarkable. See [Tra97] for a time-line of the telecom industry evolution in the US. Impetus to change is also being provided by technological developments enabling the convergence of voice, video and data services on the same networks. All of this has resulted in making the industry much more revenue conscious and customer focused than at any point in the past. This has resulted in huge data-mining efforts for analyzing customer behavior from networks operations data. Several real-time network management systems are being marketed to allow companies to manage their networks more efficiently in response to changing demand. The relevancy of yield management to such an environment is difficult to miss. Section 1.6.3 provides some very interesting industry perspectives on the need for application of YM to telecom.

[2]Segment control refers to controlling discount sales for multi-hop flights independently of each other, ignoring the network-wide revenue effect due to passengers using more than one segment.

1.6.2 Analogies and differences between airlines and telecommunications

Analogies

Several analogies between airlines and the telecom industry make it very relevant to consider the application of YM to telecom networks. This section discusses most of them.

Perishable inventory: Like airlines and hotels, one can view the bandwidth/sec in telecom networks as perishable inventory, making it attractive to utilize unused capacity whenever it is available. Bandwidth/sec, however, is just one measure of inventory. Several notions of perishable inventory exist in telecom networks, such as the streaming capacity of a video server, the switching capacity etc.

Large sunk cost and low marginal cost: YM is usually useful when the relative difference between the cost of increasing capacity is large compared to the marginal cost of providing service. For example, short-term capacity such as aircraft seats is fixed. The cost of adding another aircraft to the fleet is huge compared to the cost of providing service when capacity is available - which in airlines might be as low as the cost of a bag of peanuts. It is obvious that the same situation exists for telecom networks, with the marginal cost of providing service negligible when capacity is available.

Varying but predictable demand volume: Demand for various services is strongly temporal with predictable daily and weekly trends.

The competitive structure of the industry: Telecom industry evolution has been tracking the airline post-deregulation history rather closely. See the article "You'll Never Guess Who Wants to Be Your Phone Company ", by Mark R. Bruneau in Telecommunications^R, October 1997, for a fascinating listing of the parallels between airline and telecom industry evolution.

Differences

The perspective brought to telecom networks by YM is attractive, but only the tip of the iceberg, and further, the analogies with airlines are only at a very high level. Several important differences exist between airlines and telecom networks, in terms of resources and the market structure for telecom services. These differences make it difficult to see an immediate translation of airline YM practices to telecom. We contend, however, that the differences are liberating rather than limiting, and that one can do much more in telecom for yield management than in air-travel. The resulting modeling problems are also seen to be richer and more complex than their analogues in airlines. A discussion of important differences between the two industries follows.

Lack of reservations: In airlines, hotels and the transportation industry, there is a *market accepted* existing reservation process which is modeled. In contrast, majority of telecom services are on-demand, with exceptions such as video-conferencing forming a negligible fraction. This makes it difficult to have airline-like decision rules for controlling the use of bandwidth. However, nothing precludes setting up a market for services on the network that could allow some form of a reservation process. This is exactly what we propose in chapter 2.

Pricing and market structure: The pricing structure of telecom services is a legacy of its regulatory history. It is not clear if flexible pricing and/or discounting telecom services similar to airlines will have comparable effects since the elasticity of demand may not be as high. In addition, the existing market structure is subscription-based, meaning discounting services may result in cannibalization of already existing demand – although this is expected to be changing rapidly by introduction of deep-discount dial-around numbers such as MCI's 10-10-321.

Complexity and heterogeneity of the networks: As mechanisms, telecom networks are orders of magnitude more complex than airlines and hotels. With convergence of telecom services and seamless inter-connection of voice-switched, ATM networks and the Internet, it is hard to even characterize the resources needed for servicing particular requests. For instance, a voice call may get routed over

a circuit-switched network, onto the public Internet and then onto an ATM network. How much capacity does such a call use? This is itself a difficult question and makes it hard to design uniform rules for network operation.

Availability of capacity: In airlines, the reservations system accurately tracks network-wide available capacity. In telecom, capacity characterization based on historical patterns is itself a significant problem. Older voice networks could adequately use measures such as average hourly utilization, for example. With Internet-type networks, such a characterization might be too crude to determine the available capacity. It is now an accepted fact that traffic in the Internet is extremely bursty and unpredictable, implying that available capacity fluctuates at very short-time scales [WP98]. Further, the presence of dynamic and local routing in such networks means an adequate picture of the network is not always available [Pax97].

In spite of the above mentioned difficulties, the relevance of yield management to telecom has been widely recognized in the past few years. Several perspectives from industry on the need for YM are collected in section 1.6.3. The challenge in yield managing telecom networks is two-fold: organizational and operational. The organizational challenge is to step out of the rigidity of the telco mindset of basic telephone service offering, and the operational challenge is to offer and intelligently manage services that capitalize on the simple yield management intuition.

1.6.3 Industry trends towards Yield Management

This section highlights current trends towards YM in the telecommunications industry. Finding ways to increase revenues and maintain profitability is a first-order problem for telecom companies, due to increased competition and diminishing margins. Other developments which may reinforce the application of YM are also discussed, such as the emergence of bandwidth exchanges. Some very common opinions in the industry are:

- Telecom companies should capitalize on their existent huge databases of operational and billing data by: (i) building data mining tools and (ii) creating processes for making intelligent business decisions based on that data.
- Carriers and Service Providers should provide more innovative and value-added services to differentiate themselves from the competition.
- Carriers should reduce time-to-market of new services as much as possible by streamlining development processes.
- Carriers should focus on outsourcing/building more flexible network management tools that allow real-time monitoring and resource allocation of their networks.

Such thinking manifests itself in the new service offerings of the last few years, such as call-back numbers, caller-IDs, service-bundling and others. Companies actively seek new services to generate positive revenue streams from their networks. Partly, this is because creating new services is a much easier task than before, with recent developments in architectures based on the Java^{TM} language which dramatically reduce the time-to-market of new services. We comment on these practical issues in section 1.3.2. It is an accepted viewpoint that network operators who cannot segment their market better will lose the competitive battle.

Many of the perspectives mentioned above can easily be seen by glancing through magazines such as Telecommunications^{R}. An excerpt which succinctly summarizes the thinking of the industry regarding YM follows below, from the article "Network Management: A Core Skill for Future Telcos" by Stephen Allott in Telecommunications^{R}, August 1997 (italics added).

> Once carriers have mastered the skills necessary to run their new networks properly, they then have to work out how to profit from them. *Yield management* is the name of the game. In PSTN terms, this meant cheaper off peak calls and little else. The

telcos can learn a lot from the airlines or major oil refineries who have both developed extremely sophisticated yield management systems. Commodity industries such as petrol refining or time-based industries such as transport already understand the capacity yield management, the link with transaction and product pricing, and how the two skills are essential to generating superior returns. Although the public service utility mindset of the PTTs will take time to fade away, it is expected it will become increasingly obvious that yield management will be essential as time goes by.

It is no coincidence that British Airways is the world's most profitable airline and its planes are nearly always full. Filling up a network is not merely a question of how much you charge but also how the tariffs are structured. With PSTN services, time-based billing is the norm. At the other extreme, Internet is flat rate. For ATM, both time-based billing and usage-based billing are being tested. *Hand in hand with tariff structure, systems must capture billing records and 'right fit' customers with appropriate traffic profiles targeted to fill up the spare capacity on the network.*

Finally, once carriers have bought the right network, run it reliably, and filled it up with paying customers, they will then have to compete to attract and retain business. Proof of service levels is the key basis of competition with Web delivery of service information at the forefront. Key innovative service providers such as UUNet, GE Information Services and BT are now able to offer customer views of real time network faults. We expect carriers to offer Web-based united views of service information during 1997.

Other developments that will motivate the practice of YM are liquid bandwidth markets, resulting from the emerging bandwidth exchanges.[3] Such exchanges may allow bandwidth trades to occur at widely differing time scales - from minutes to months. This will give network providers flexibility for many things. For example, networks might be dimensioned to run at very high utilizations, with less profitable traffic classes off-loaded to capacity bought on-the-fly if needed, a classic YM over-booking type of practice. Many such ways will be found to more closely match available capacity to the actual traffic profiles, maximizing yield from the infrastructure.

In summary, there is no doubt that the future of telecommunications services will be market driven. Telcos in particular will need to focus on the individual customer and will have to find ways to segment the market, differentiate their products and manage capacity well to preserve their revenue streams and counter competitive pressures. It is far from clear, however, what that means. Companies like MCI, Sprint, GTE have all invested in large data warehouses to allow them to make better managerial decisions by analyzing vast amounts of customer data. Also, software products are emerging that allow operators to manage their networks in real time in response to fluctuating loads. But a demonstration of these so called "intelligent decision tools" has not yet surfaced.

1.6.4 Possible impacts of Yield Management in telecom

YM has been demonstrated to be an effective competitive tool in airlines. Large carriers such as American Airlines can only compete with smaller carriers during price-wars by matching the prices of their competitors while taking care not to dilute their overall revenue too much to not be able to meet costs. YM systems are the front-line defense for such tactical issues. One expects, and indeed observes, similar and in fact more aggressive situations in telecom. In fact, some economists argue that purely competitive markets are not viable in the case of digital services since the marginal cost of production is so low that the producers can always afford to undercut the competition. See for instance, [Var95].

The following excerpt from the article, "Balancing Infrastructure Against Profitability" by Joao Baptista and Edward Ainsworth in Telecommunications^R, June 1997, summarizes the nature of

[3] See, for instance, the article by Thomas L. Friedman on bandwidth exchanges in the *New York Times*, titled "TheLandgrab.com", January 18, 2000.

competitive tactics in telecom, necessitating protective measures such as YM for incumbent network providers to not dilute their revenue excessively, and for new entrants to sufficiently differentiate their services from the competition.

> A look at the economics of telecoms competition reveals why (infrastructure investments are not a good idea). In any telecoms market, a new entrant's operational efficiency and regulatory advantages will be outweighed by the scale advantages of the incumbent until the entrant reaches a critical level of market share. The exact level of the share threshold varies by country, depending on the market's size and density and the incumbent's efficiency. A competitor's ability to surmount the threshold is affected by its own skill at attracting customers and the number of other new entrants it is competing against.
>
> For a new entrant with a market share below the threshold, the ownership of infrastructure becomes extremely costly, potentially offsetting any shareholder value derived from customers acquired. Infrastructure ownership becomes a millstone around the new entrant's neck, forcing it to pursue marginal customers indiscriminately in an attempt to fill its fixed-cost infrastructure. In practice, the imperative to fill capacity ends up dominating the new entrant's strategy. Management's attention is consumed with waging a price war against the incumbent, rather than seeking differentiation through service innovation or creative customer management. The inevitable result is an industry-wide, margin-sapping cycle of discounts and counter-discounts on 'me-too' services - an outcome that neither new entrants nor incumbents should favor.

Finally, another incentive for telecom providers to initiate YM activities might be the simple goal of 'getting there first'. If useful YM systems for telecom are indeed built, the advantage of being the first to deploy such systems could be significant. The analogy, without surprise, is again from airlines. The yield management process was built and perfected at American Airlines, and was a major strategic advantage leading to their success in the late 80s [SLD92], before widespread acceptance of the potential of these systems. Anecdotal evidence from airlines also suggests that the only reason airlines are currently profitable is because of their YM systems, otherwise they would be running losses.

1.7 Literature Review

Very little literature is available that is directly relevant for telecom YM. We provide a short review of related work in section 1.7.2. But first, we provide a brief review of airline YM literature to give both a sense of the decision problems solved by airlines and the connections we hope to illustrate through our exercise for telecom YM. This review, however, is kept concise for the reason that airline YM models are not directly applicable to telecom YM. This is because analogies between airlines and telecom are only a very high-level as argued in section 1.6.2. We do not attempt to review YM related work in other industries for the same reason.

1.7.1 Airline Yield Management

Survey articles are the corner-stones of literature reviews. Two such articles dealing with revenue management research in the last forty-odd years are by Van Ryzin and McGill [MVR99] and Weatherford and Bodily [WB92]. The first is an excellent review and summarizes the state of the art in revenue management. The second is older and proposes a generalization of the revenue management paradigm for managing any form of perishable inventory. Below we include literature for the problems of *forecasting*, *over-booking*, *seat-inventory control* and *pricing* in airlines. The organization and review below is heavily based on the article mentioned above, by Van Ryzin and McGill [MVR99].

Forecasting

Accurate *forecasting* is extremely important for airlines because of its direct impact on revenues. The earliest models were for forecasting final demand for itineraries, for use in over-booking calculations. This required forecasts for passenger bookings, cancellations and no-shows. Early work in this area investigated the Poisson, Gamma and Negative Binomial models for final demand [BB58]. Several variations for estimating cancellations, no-shows etc. were investigated [Tay68], and later research showed that the normal distribution is usually a good continuous approximation for aggregate demand [Bel87].

Models of arrival processes for bookings are important because of their use in determining rules for seat-inventory control. Much of the above work used arrival models such as the Poisson or the non-homogeneous Poisson processes to estimate final demand. These processes typically under-estimate the variance of the final demand but have the advantage of simplicity in determining seat-inventory control rules. To more realistically model the variance of the final demand, other processes such as the stuttering Poisson [Rot68][Rot71] and the batch Poisson process [BB58] have been proposed.

Two other areas deserving mention are: uncensoring demand data and disaggregate forecasting. The uncensoring problem arises from the fact that airline reservations systems only track accepted reservations, which are affected by the presence of booking limits and capacity constraints. This requires that one be able to build a picture of the total demand from censored data about bookings. Work in this area has included several approaches [Swa90][Lee90][McG95], which we do not describe here. Disaggregate forecasting is desirable because aggregate forecasting may not yield accurate forecasts for less traveled itineraries, and specially rare itineraries. Network-wide, these include an important part of the total demand [Wil92]. Further, disaggregate forecasting has been argued to yield more accurate estimates of aggregate demand [Sa87]. Work on disaggregate forecasting techniques has been done mostly by practitioners [HM83][SLD92][L'H86].

Over-booking

The objective in *over-booking* is to admit the right number of requests such that flights depart as full as possible, given the presence of cancellations and no-shows. The trade-offs involved are between admitting too many request, resulting in excessive denied boardings against not admitting enough and flying empty seats. This problem had been part of airlines practices well before yield management but for a long time, its use was denied by airlines [Rot85]. Another component of the over-booking problem is in determining equitable bumping procedures for over-booked flights [Fal69][Sim68][Sim72][Nag79] for which some of the earliest work was by Vickrey [Vic72].

Early over-booking research used statistical models for predicting single-fare show-ups to compute over-booking limits [BB58][Tay68][RS67][Lit72]. Some work was done on the multiple-fare over-booking problem [Bel87]. All of these models were non-dynamic in the sense that they did not incorporate the passenger cancellation and reservation processes subsequent to the over-booking decision.

Dynamic optimization models have also been investigated for over-booking to maximize revenue from flights sold-out at departure. Such models quickly become complicated and unsuited for implementation. Multiple fare-classes add another level of complexity. The usefulness of dynamic models is in obtaining structural results, indicating the optimality of control-limit type policies [Rot68]. Work has also been done on extensions involving the setting of joint over-booking levels for multiple inventory classes [KVR98] when fares can serve as substitutes as in, say first-class seats serving as substitutes for coach class.

Seat-inventory control

The *seat-inventory control* problem is that of determining how to allocate seats across multiple fare classes. Alternatively stated, given a booking request within a fare class for an itinerary, the problem is whether or not to accept it, given available capacity on all legs and forecasts of future demand for all fare classes. The respective trade-offs for the accept/reject decisions are denial of service to a higher paying customer later in case the flight is full, vs. not being able to sell all seats before the flight takes off. The statement sounds simple, but the actual computation of such a decision can be extremely complicated. The influence of a decision propagates across the entire network by potentially displacing other bookings which will have displacement effects of their own. Similarly, the influence propagates forward in time because displaced bookings may terminate at a later date than the current booking. Also, since most itineraries usually have a return component, there is a downstream effect on capacity.

The earliest seat-inventory control models focused on single flights, starting from Littlewood's rule for two-fare classes [Lit72]. Much work was done on testing the assumptions under which Littlewood's rule is optimal and on empirical testing of its performance [BP73][Ric82][May76]. Belobaba [Bel87] extended Littlewood's rule to multiple fare classes and proposed the Expected Marginal Seat Revenue (EMSR) rule, which is not optimal in general apart from the two-fare case but is very easy to implement and usually gives good results for common demand distributions [McG89][Wol92]. Unfortunately, one can construct distributions for it behaves arbitrarily badly [Rob95] but these are usually not natural. Work has been done on extensions of EMSR to produce better approximations of optimal booking policies [VRM98].

Using dynamic programming approaches, optimal booking limits have been derived under a set of assumptions on the arrival processes (see [MVR99] for a discussion of the assumptions), a typical instance of which is an assumption of low-fare customers booking before the higher fare ones. One typically obtains structural results from such models [McG89][BM93] but there are significant computational limitations to their implementation. Several extensions of the of the single-flight problem have been investigated under relaxed assumptions. Work on these is too extensive to cite here, see [MVR99] for a collection of references.

Network-wide seat-inventory control has become increasingly important with the development of hub-and-spoke networks which result in a large number of passenger itineraries involving connections to different flights. Since single-leg control does not optimize network-wide revenues, airlines have attempted to develop control rules to increase revenue from the network. The first approaches were deterministic and involved solving either min-cost network-flow formulations [GGLM82][Won90] or linear programming approaches [Wol86]. These approaches sometimes produced non-nested allocations, which was counter-intuitive and consequently undesirable for operational level control. To use stochastic models, one needed to make them tractable. This was achieved by clustering the thousands of O-D itineraries into a smaller number of controllable classes, for which several methods were proposed [SP88][Wil88]. Once one has a small number of controllable booking classes network-wide, one needs to decide on control rules. This is currently done computing bid-prices using information from deterministic linear programming/network models [SP88][Wil88][Wil92] to compute dual prices, which are interpreted as the marginal values for incremental seats on different legs in a network. These dual prices, when summed across the legs in an itinerary, are assumed to provide an approximate displacement cost, called the bid-price, and a request is accepted only if its fare is above the bid-price of the itinerary.

Pricing

Airline now view *pricing* as an important part of the revenue management practice. At the strategic and planning level, it has always been important. There is an extensive literature on airline pricing from an economic perspective which addresses issues at an industry level. Tactical or yield management level pricing is now seen as important because the opening and closing of booking

classes for seat-inventory control can be seen as changing the fare structure for customers. Work on this can be found in [GVR97][GVR94], where dynamic pricing problems are treated to determine optimal pricing policies. Little has been published on joint capacity allocation/pricing and market segmentation. Some work can be found in [Bot94].

1.7.2 Telecom Yield Management

There is little or no research under the umbrella of YM in telecommunications, the only paper known to the author that explicitly mentions a *Yield Management* link is by Paschalidis and Tsitsiklis [PT98]. Here the authors investigate congestion-dependent pricing policies for maximizing revenue from circuit-switched calls with exponential arrivals and holding times. The authors show that under stationary demand functions, fixed or static prices (such as *time-of-day* prices) are asymptotically optimal in a number of limiting regimes, such as light and heavy traffic, and a large number of small users. A tentative conclusion is that when demand statistics are slowly varying, time-of-day pricing will often suffice. The behavioral assumption in such a model is that by lowering prices, customers will call more often and demand is infinite. Optimal pricing policies fluctuate at very short time-scales in reaction to congestion, a behavior natural for an algorithm that tracks fast-changing prices but probably not for human reactions to prices. Additionally, in the current market structure consisting of fixed subscriber-ship, one needs to consider the possibility of cannibalizing existing revenue because of discounting.

All other literature has focused on the pricing problem in communication networks at either packet-arrival time scales or pricing flows. For the former, Mackie-Mason and Varian [MMV95] proposed a "smart-market" mechanism in which each packet carries a bid-price and the network only forwards packets with prices above a certain threshold based on congestion. For the latter, Kelly [KMT98][Kel97] proposes rate-based schemes in which flows are charged based on the fraction of network resources used when the network is congested. Clark [Cla97] proposes an expected capacity based scheme where users contract for a given aggregate expected capacity from the network, and packets may get dropped if the user exceeds their capacity profile *and* the network is congested. Charges are based on the expected capacity contract.

The network-wide revenue impact of packet or flow-based schemes is not clear. More seriously, it is not clear if these are appropriate revenue control mechanisms, since users do not typically think in terms of packet flows and use the networks in a task-oriented way, such as send email, retrieve web-information etc. Therefore other schemes have been proposed to investigate how to price the Internet [CSEZ93][WPS97]. Work is relatively scattered and evolving in this domain and the best research tool is often the Internet, with several researchers listing collections of papers at their web-sites[4].

Admission control for loss networks has been used to maximize revenue from multiservice loss networks [Ros95] but here the prices are assumed fixed. Usual models are the form of stochastic knapsacks with class dependent exponential arrival and service times. A good summary of the results can be found in [Ros95].

It is also useful to understand how the modeling work proposed in this thesis relates to existing telecom models – in design, operations etc. Section 1.3.1 discusses this briefly and shows how YM models fill a yet unfilled niche in the space of telecom models.

[4]See for instance, Frank Kelly's site *http://www.statslab.cam.ac.uk/ frank/* and Hal Varian's site *http://www.sims.berkeley.edu/resources/infoecon/*.

1.8 Summary

This chapter has presented the problem of Yield Management for telecommunication networks by drawing motivation from the remarkable success of airlines. We discussed the expected contributions and limitations of this work. Because of difficulties in directly translating airline YM to telecom, we have proposed an argument for designing new services to better exploit demand and capacity to increase revenues from the network. Several aspects of this argument are discussed, including practical issues, its relationship to current telecom modeling and possible application to other networks. The rest of the chapter provided short backgrounds on airline YM and the telecom industry, including high-level analogies between the two and the difficulties in implementing YM in telecom. A short literature review on airline YM has been included but the corresponding review for telecom YM does not include much of significance.

Chapter 2

A Modeling Framework for Telecommunications Yield Management

This chapter ties the thesis together. Here we explain how the airline YM modeling framework of *forecasting*, *over-booking*, *seat-inventory control*, *pricing* and *market segmentation* can be used for modeling operations of new telecom YM services motivated by our argument of section 1.3. We interpret each component of the airline framework in the telecom context and introduce the service ideas modeled in this thesis, each idea serving as a preview of a subsequent chapter.

The outline of the chapter is as follows. Section 2.1 presents the argument for using the airlines modeling framework for telecom. Section 2.2 translates the framework for a telecom context and presents the YM services modeled in this thesis to illustrate the use of the framework components. Section 2.3 presents a summary of the modeling work in this thesis and section 2.4 lists ideas for some other YM services which we do not model in this thesis. Section 2.5 is the summary of the chapter.

2.1 Need for the Airline Yield Management Modeling Framework

To understand our point-of-view and the need for using the airlines framework, consider a specific idea for a YM service which attempts to transport bulk content between locations by fixed deadlines using only spare capacity. Suppose the software architecture for service operation consists of agents distributed across the network that accept content from customers and schedule its transmission. These agents will routinely need to make decisions involving: (i) forecasting available capacity and demand for the service, (ii) determining the pricing levels for bulk shipments, (iii) deciding if an arriving request can be shipped by a given deadline and (iv) deciding which requests to accept in case customers pay differently. A model that incorporates all decisions for the network-wide version of the service is likely to be complicated to formulate and implement. Hence a framework that guides modelers by identifying the common decisions and decomposing the resulting models in a simple manner is extremely valuable.

This motivates using the airlines method of decomposing decision models into the following classes: how to forecast demand and capacity for the service - *forecasting*, how many to accept to fill up available capacity - *over-booking*, which class of traffic to accept in case customers pay differently - *seat-inventory control*, how to price the classes - *pricing*, and how to segment the market for a

service - *market segmentation.*

It is useful to point out that the airlines break-up of the system-wide YM problem into manageable chunks is also a direct result of the complexity of the system-wide YM problem – Smith [SLD92] mentions an estimate that the system-wide problem for American Airlines network could involve as many as 250 million variables. Not only are the problems made tractable by the airlines breakup, many of the rules obtained from YM models are also interpreted intuitively, resulting in insights into system behavior rather than simply giving a numerical solution.

Using the airlines framework, for any given telecom YM service, any or all of the airlines modeling activities of *forecasting*, *over-booking*, *seat-inventory control*, *pricing* and *market segmentation* might make sense. It is also obvious that depending on the service offering, some of the areas might not be relevant. For example, if all bulk shipments were flat-rate or not significantly different in revenue, one may not need seat-inventory-like control to distinguish between customer classes.

Another motivation for using the airlines framework is to highlight the connections between airline YM and telecom YM. These connections are obvious at a high-level – perishable inventory and time-varying capacity usage, but seem to disappear as one tries to imagine an application of airline-like controls in telecom. The problem arises due to lack of existing services in telecom that allow for airline-like decision-making. Even using our proposed argument (c.f section 1.3), ideas for YM motivated telecom services are decoupled from the airline YM products[1] owing to the nature of the telecom services market. Consequently, the services we propose for telecom YM seem far removed from the airline YM products, the only connection being that they capitalize on the fundamental intuition of market segmentation and operation at near-zero marginal cost. However, by decomposing the models using the airlines approach, we re-establish connections to airline YM models at a deeper level. With this approach, many of the rules one obtains for telecom services parallel airline-like rules and the intuitive connections are made explicit. This way, we capitalize on more than the raw intuition of using spare capacity and zero marginal cost, by drawing lessons from the modeling approach of the airlines as well.

To demonstrate how the framework can be used for modeling telecom YM services, we take an eclectic approach. We consider several YM service ideas and for each service, model one aspect of its operation which falls into an area in *forecasting*, *over-booking*, *seat-inventory control*, *pricing* or *market segmentation*[2]. This leads to an interesting exploration of several intriguing YM service ideas rather than demonstrating the entire framework for a single service. We feel that the speculative nature of the services justifies modeling limited aspects of many of these rather than a complete set of models for one, with the hope that some of the service ideas will find their way into practice. Section 2.2 presents in detail how these activities are relevant in a telecom context and the services modeled in this thesis.

A final note concerns the modeling effort in this thesis and its connection to real-world applications (also see section 1.2 for comments on connections between modeling and the real-world). Most models we consider are single-link versions of what will typically end-up being complicated network-wide models. This is a usual first-step. For instance, consider the airlines practices and their evolution. Airlines typically cannot implement optimal decision rules for their networks owing to the complexity of the problem [MVR99]. Even single-leg models are not solved optimally and heuristics such as the Expected Marginal Seat Revenue [Bel87] are extensively used. However, single-leg models are useful for understanding structural properties of the decisions and are insightful for building network-wide heuristics. In a similar spirit, our single-link models may serve as starting points, to gain insight into building network-wide heuristics.

[1] The term *product* in airlines is used to refer to the fare for an itinerary in combination with restrictions, such as 14 day advance bookings, Saturday night stay-over, non-refundability etc., and is the primary mechanism for segmenting the market for an itinerary.

[2] *Forecasting* is a bit peculiar in this regard, as shall be seen later.

2.2 Translating the Framework for Telecom Yield Management

This section, in a sense, circumscribes the decisions for telecom YM services that map into the modeling framework of *forecasting*, *over-booking*, *seat-inventory control*, *pricing* and *market segmentation*. For each category in the framework, we also briefly introduce a service idea for which a related decision is modeled in this thesis. The discussion on the services is brief, with most of the details relegated to the relevant chapters.

All services considered in this section, and indeed the thesis, are assumed to operate at around 0 marginal cost, using only available capacity, for which mechanisms can be easily designed (c.f. section 1.3.2). Also, it is useful to reiterate here that generating ideas for new services is not a formal process. Not only are the services proposed a small subset of possible services but even the ideas described here are only one possible offering mode for the proposed services.

2.2.1 Forecasting

What telecom decisions might fall into the forecasting category? This is usually clear. We comment on the nature of forecasting problems that will arise for many of the YM services. It is anticipated that many of these might be harder than their counterparts in airlines.

Capacity forecasting

An extra dimension involved in telecom is forecasting the physical available capacity for various times of day/week. This problem does not arise in airlines, since the reservations system accurately tracks available capacity at any time, even though one can think of customer cancellations and no-shows introducing uncertainty into the actual available capacity for an aircraft.

The capacity forecasting problem in telecom networks is hard because of several reasons. Even if statistical characterization of available capacities on links is obtained, it is difficult to obtain a measure of the end-to-end available capacity. For instance, the question of how many extra calls can be accommodated between another origin-destination pair is not always answerable using only link data, since the networks employ routing strategies which use local information at time of call setup to route calls.

The problem becomes significantly harder in Internet-type networks for several reasons : the burstiness of traffic, the multiplexing of a large number of sources and the local behavior at the routers, possibly causing several routes to be employed for servicing a request. This makes it very hard to build an accurate picture of network traffic and available capacity at any time. Vern Paxson's thesis [Pax97] addresses the issue of adequate characterization of end-to-end behavior of the Internet in great detail. The tremendous heterogeneity of these networks exacerbates the problem. Intense research activity has recently focused on accurate statistical characterization of Internet traffic at all time scales to understand the operational stability and efficiency of these networks.

The good news is that models for YM services may not require very accurate models of available capacity at all time scales. One may need only a crude estimate of the residual capacity for components. Such estimates might be reasonably obtained from aggregate data such as average link utilizations for various times of the day.

Demand forecasting

Demand forecasts are necessary to appropriately manage capacity, exactly as in airlines. Several demand processes giving rise to requests for telecom services have been relatively well-investigated in literature. For instance, the Poisson process is known to be a good model for long-distance call

arrivals. Similarly, other requests taking place at the human-activity time scale are well modeled as Poisson processes or non-homogeneous Poisson processes with slowly varying arrival rate. It is generally accepted that this is because the Poisson process can be obtained by the superposition of a large number of renewal processes. For other machine initiated arrival processes such as the requests from Web-servers etc, researchers are actively trying to validate useful models.

We make reasonable assumptions for arrival processes whenever needed since revenue for telecom is likely to be much less sensitive than airlines to the exact statistical characterization of the processes, simply because of the magnitude involved per service request. For human-initiated requests, the Poisson process is used. The only other arrival process used in this work is that of request arrivals at a Web-server. For this, we use a batch Poisson process and provide qualitative arguments for its reasonability, but do not perform rigorous statistical analyses of the claims. This is because of our limitation in obtaining realistic data.

Forecasting-related work in this thesis

In light of the above discussion and because almost all services proposed in this thesis are speculative, we do not present any work on demand forecasting. Also, for capacity characterization, we do not find ourselves well-situated to contribute to traffic modeling literature at short time-scales. Work on traffic modeling at this level is fast becoming extensive [Pax97][RMVe96]. Further, useful and practical models require extensive data and may only be marginally more useful for YM-services. We assume that crude measures of available capacity for components such as links, for example, are obtainable.

We do however, consider one problem related to capacity forecasting that has arisen recently for Internet-type networks [WP98]. This is the problem of probe-location. The issue is gathering data for building an accurate picture of traffic on the network. The mechanism is to locate monitoring devices, called probes, on the network to capture information. Since probe locations are fixed and traffic behavior dynamic, one needs to decide the most effective location for these probes to collect the most information in a highly stochastic environment. This problem is mentioned as a newly emerging research problem in [WP98]. We present a first-cut model in chapter 6.

An added advantage from probes might result in the seemingly unrelated area of market segmentation. There is an emerging market 'need' for performance data on ISP networks [Bor98] in addition to the need of network operators for better utilization of existing infrastructure[3]. Corporate customers of ISPs are usually deeply concerned with the service-levels and quality of service they actually receive from the network, and currently use ad-hoc measures for monitoring network performance. Several companies such as Savvis Communications^{TM} and At-Home Service^{TM} make performance guarantees the backbone of their marketing strategies. For an impact of performance monitoring on customer retention, see the article [Bor98] in *Business Communications Review*^{TM} which makes the case that ISPs cannot ignore the impact of such monitoring on sales and customer retention. This is explained in detail in chapter 6.

The only other forecasting related work in this thesis is the presentation of actual data to qualitatively argue that the demand process at Web-server is reasonably modeled as a batch-Poisson process (c.f. chapter 7).

2.2.2 Over-booking

How can over-booking arise in telecom? The essential question here is one of quality of service (QoS) and capacity determination. The main characteristics of the airline over-booking problem are fixed capacity (airline seats) and known reservations, with uncertain behavior. The decision on how many

[3]The difference in operating efficiency between the best and the worst operators in North America and Europe exceeds 30% [Bay96].

to admit is constrained by pre-specified limits on denied boardings. Such a problem could arise in telecom in several forms. For instance, suppose that demand is certain, i.e. customers who show up always want service before some deadline but capacity is stochastically available. Then the decision problem is similar to before, one needs to decide how many to admit such that the possibility of not being able to serve a request is kept within reasonable bounds. Capacity stochasticity in such cases can naturally arise if the resource usage of customers is not pre-determined. One could imagine other services where over-booking might arise. For instance, if customers actually reserve time on the network but are free to not show up, similar to airlines. Again, the decision is exactly as before. In every one of these versions of telecom over-booking, one has some information about existing reservations and the uncertainty is either due to reservations behavior or capacity availability or both.

At the bit/byte level, the over-booking problems mentioned above are somewhat similar to the problem of ensuring an adequate QoS, and are not new to telecommunications literature. They occur in several forms as in, for example, multiplexing of sources at ATM switches, but at a very short time-scale. Using a measure of resource requirement called *effective bandwidths* [Kel96], one admits sources such that their peak transmission rate is greater than the available transmission rate of the resource, but the cell-loss probability for each source is within pre-specified limits. Similarly, the problem of dimensioning a long-distance network to ensure that the fraction of blocked calls is within some QoS bears some resemblance to an over-booking problem without future reservations.

The most important distinctions between these QoS problems and an over-booking problem as we envision for YM are in (i) accounting for the accepted reservations at the time of admission control and (ii) the time-scale. The difference from the cell time-scale QoS problem is obvious since the over-booking problem happens at a call time-scale. But even at the call-scale, it is different from other, more static problems, such as computing blocking probabilities for calls at the dimensioning level, which do not assume partial information about demand, such as reservations. In one sense, if the bit/byte level problem can be considered as an operational problem, and the dimensioning as a strategic decision, the over-booking decision can be considered more as a tactical decision where one has some limited information about future demand and capacity.

Over-booking related work in this thesis

We briefly mention some possible YM services for which we model an over-booking type decision. Detailed discussions follow in relevant chapters. In all services below, the intent is to use only excess capacity for their operation, at near-zero marginal cost.

Network-usage by a Latest Start Time: Two Possible Services

Services that require network usage by a latest start time can arise in more than one context. For instance, an executive or engineer working for an extended period of time in her office or telecommuting from a home office may not care exactly when a phone call or FAX transmission is executed. She might be willing to accept a latest time during which the call or FAX is guaranteed to be placed rather than require immediate service, provided that the fee structure is reduced to reflect her flexibility. Say that the professional's assistant, a software agent, is called Lucy. Lucy 'talks' to a network 'agent' and lets it know the latest time before such a call should commence. In a more compelling example, the customer might be a corporation wishing to obtain network time before a given deadline for its use.

Another service for which usage by a latest start time arises implicitly is when the service provider acts as a content courier, guaranteeing non-preemptive content delivery at a fixed rate to a destination before a customer-specified deadline. With these constraints and the amount of content, a deadline to completion of delivery is the same as specifying a latest time to start of transmission. We discuss this service in detail in chapter 4, including why the constraints of a fixed rate and non-preemptible service might be desirable. Again, the interface for this service could be the same

as Lucy which 'talks' to a network agent.

It is clear that many different decisions arise as part of these services' operation. In the simplest case, if the discount was fixed, the network 'agent' only needs to decide if the additional request can be started before its latest start time with a pre-specified probability, given the number of calls already in session and the accepted reservations. This is an over-booking problem, with the added complexity of the scheduling possibility, which does not arise in airlines. We present a model in chapter 4 which also clarifies other details about the operation of such services. We do not model many other aspects of the services which would also result in interesting models.

What would be a possible market for these services? With a bit of imagination, we can consider more than one. For the first service, consider video-calling as an instance. As a service, it could be expensive enough for a telecommuter to not consider it unless deeply discounted. To the service provider, it costs nothing if capacity is expected to be available and the user would not have bought this service otherwise. Similarly, consider down-loading video titles from libraries on the Internet. One might ask 'Lucy' to connect and down-load titles any time before 6pm for viewing. This would be a case where the user wishes not to swamp her connection to the Internet with these transfers and asks 'Lucy' to down-load content using a separate connection. The market for the second service is very much like the market for another service, which we mention in section 2.2.3.

The revenue potential of such services might be significant. Consider back-of-the-envelope calculations for the first service, for instance. With 80 million customers for a company like AT&T[4], say 1% telecommute occasionally and of these only 10% actually subscriber to a 'Lucy'-type service. If each such subscriber on average uses 'Lucy' only once every two weeks, and each call yields only $2 in revenue, one obtains added revenues of approximately $4 million per year. Similar calculations for the second service can be made, as illustrated in section 2.2.3.

Operating a High-Speed Backbone as a Transport Network

Consider a provider such as MCI^{TM} operating a high-speed backbone with time-varying utilization. Suppose MCI installed web-sites around the backbone, allowing users to upload bulk content to a geographically close web-site, which would then be transmitted at a guaranteed high-rate over the backbone, as soon as it finishes uploading. The service is designed such that transmission only takes place using spare capacity on the backbone, for which mechanisms can be designed[5].

Depending on the details of the service offering, the entire airline framework of pricing, seat-inventory control etc. might be relevant to its operation. We consider only an the over-booking type decision for it in chapter 3, namely, how many to admit given the number in transmission and the number uploading onto a web-site, to maximize revenue from the service.

A variant of such a service exists for private digital network operators such as Vyvx^{TM}. Corporations, for instance, CNN^{TM}, NBC^{TM} or other media companies needing to transport programmed content across geographical locations call into the Vyvx operator to reserve time on the network in fifteen minute increments. They pre-specify the locations that need to be connected. At any time during their requested time window, they effectively have a private virtual network. The carrier in this case, however, has control of two decisions - whether capacity is available for setting up a network within the future time-window and, at the time of setup, what is an optimal topology. For users wishing only to transport content, the service mentioned above might be a possible substitute to renting network capacity.

[4]Source: *Wall Street Journal* (3 Star, Eastern (Princeton, NJ) Edition), 7 April, 1998.
[5]For instance, by setting *time-of-day* limits on the amount of capacity used by such transmissions.

2.2.3 Seat-inventory control

In what contexts could a seat inventory-control problem be relevant for telecom? The abstract version of this problem in airlines is easily stated as a valuation problem. Given two competing uses for a resource, one certain but low-paying (the discount offer) and the other future expected, which is better ? In the telecom context, such a decision naturally arises whenever customers with differing revenues request to use the network at a future time. The exact nature of future usage can be of many forms, as exemplified by the service ideas of section 2.2.2 for which we model only an over-booking type decision. Both these services could easily require a seat-inventory like decision to decide which class of customer of customer to admit if the customers yielded significantly different revenues.

We use another service to illustrate how a seat-inventory control type decision might arise for telecom YM services. The service offers deep-discount courier delivery of digital content, a digital version of FedEx^{TM}. Operation of such a service can be carried out at almost zero marginal cost by designing appropriate protocols.

Seat-inventory control related work in this thesis

Consider a FedEx-type service where customers request delivery of bulk content between locations by a fixed set of network-wide deadlines, such as 5pm, 12 am etc. The service operates exactly analogous to the FedEx company, with the modification that since capacity availability is highly stochastic, the network provider must decide at time of arrival whether to accept the arriving content for transmission on its own network, or ship it to an alternate provider. The objective is maximization of revenue from available capacity while ensuring that no accepted requests miss their deadlines.

For this service, we consider a single-link model in chapter 5 to derive seat-inventory like decision rules, where the objective is maximizing revenue from the system. We extend the results to networks. The evolutionary parallels to airline seat-inventory control methods will be hard to miss.

The motivation for the service proposed above is that many users (e.g., creators of distance education, commercial web-sites that needing regular updates such as UBID^{TM}, CNN^{TM}, search engines such as ALTAVISTA^{TM}, YAHOO^{TM} that need daily replication) view the network as a logistics type distribution system, not dissimilar to FedEx, UPS^{TM} or the USPS^{TM}. From this point of view, there is a package of bits that must be shipped from A to B by time T. Instant or near-instant communication is neither required nor expected. This allows the service provider to place the package of bits onto the network (perhaps in chunks) at different low activity times, thereby minimizing the risk that users having guaranteed high service are turned away or given inferior service. In airline parlance, it minimizes the risk of denied boarding to first class or business class customers. Commercial data courier services already exist for the Internet. For example, the company *e-Parcel*^{TM} is an "Internet Courier" which hosts *Virtual Warehouses* on the Internet to carry any type of file of any size[6]. The speculation that users do not always want instantaneous service is strengthened by considering other services in which network 'delay' is inserted into the service. For example, *SightPath*^{TM}[7] guarantees MPEG quality streaming video off the Internet if the user is willing to wait some amount of time, allowing them to download the content to a buffer close to the customer.

To get a sense of the revenue potential for such a service, consider a company that does not have extensive infrastructure of its own or which needs guaranteed deliveries. To cite a particular context[8], an Internet-auction company A has its auction web-server in city X in the mid-west, whereas its data processing center is in city Y on the west-coast. Every night at around 3 am, after processing

[6] *http://www.e-parcel.com.*

[7] *http://www.sightpath.com.*

[8] From personal conversations with an employee of an existing company, the name not revealed because of confidentiality.

has been completed, data is shipped off from Y to X. It usually takes 2-3 hrs for the transmission to complete. Data has to be in Y and uploaded onto the server at 6am sharp because there are usually 200+ users already logged in and waiting for the quotes. A delay in data arrival can mean the loss of several thousand dollars of revenue. Network problems sometime cause data to arrive late enough such that manual intervention is needed to manage the system[9]. Such a company seems like a primary candidate for a data delivery service if delivery by the deadline can be guaranteed. Even a nominal $5 charge for delivery each day results in additional revenue of $1820/yr. If there were 1000 such customers, one would obtain a $1.8 million/yr added revenue from this service alone.

2.2.4 Pricing

The relevance of pricing for telecom, and indeed any industry, does not need much argumentation. It must be recognized, however, that it is usually governed by competitive pressures and other constraints, which limits the range of options for pricing services and products.

In particular, to understand the possibilities of pricing in telecom, it is instructive to consider the role it plays in airlines YM. It has been argued that dynamic pricing is not the key mechanism leading to the success of YM in airlines. For single flights and most itineraries, prices are published six months or more in advance of departure dates and are relatively constant across the large carriers, since unilateral reduction of prices by a carrier might incite a price-war. The result is relatively static pricing. On the other hand, it is true that there is a natural duality between pricing and seat-inventory control decisions and the opening and closing of booking classes can be viewed as changing the price structure faced by the customers. Also, dynamic pricing is clearly advantageous in pricing custom-itineraries for which fares are not pre-published and pricing fares in a market where bidding mechanisms for airline seats exist – as are evolving over the Internet.

For almost any telecom YM service, setting the prices, whether static or dynamic, will be a relevant problem, but its benefit may be indirect rather than direct as in airlines. The most obvious pricing options may simply not be the most effective. For instance, spot-pricing network usage may not result in the anticipated benefits because of (i) competitive pressures, (ii) revenue sensitivity and (iii) the structure of the market, such as fixed subscriber-ship, which may result in revenue cannibalization.

If one does indeed consider spot-pricing a telecom service, a very important question is the right time scale at which prices should evolve. The benefit of pricing at the cell arrival time-scale or at the call-arrival time scale is not clear. Several schemes proposed in literature (c.f. section 1.7) consider dynamic pricing at cell or flow time-scales. Their overall effect on revenue of an entire flow is relatively uncertain, and network-wide effects are almost impossible to guess or control. Therefore these schemes might more appropriately be called metering mechanisms where the rate charged depends on short time-scale congestion levels in the network. Another model [PT98] investigates pricing at the call-level to conclude that time-of-day pricing might be sufficiently near an optimal dynamic pricing scheme. None of the schemes consider the structure of the market.

If pricing at the cell-level and at the call-arrival level is not beneficial for telecom networks, could a time-scale in the middle work? We consider a service for which pricing decisions need to be made at fixed time intervals and which explicitly considers the possibility of revenue cannibalization.

Pricing-related work in this thesis

Consider discounting the use of the network for long-distance calls when capacity is available and raising the price when it is scarce. The concept itself is not novel but a fundamental issue is the nature of the service. Suppose at fixed intervals of time, given some information about the state of the network and the currently offered prices/products between pairs of locations, one announces

[9] I am not sure what it entails.

to the users new prices/products for a future interval to maximize the revenue from the network. The term "product" refers to a fixed discount price relative to a base tariff for a long distance call between an origin destination pair. Naturally, it is assumed that communication channels of some form exist which allow the users to be notified of the new price/product offerings at every interval[10].

The idea of real time price/product offerings hinges on the belief that some fraction of customers will react to the price offerings and that the benefit in revenue from the changed behavior of these customers will be significant. Clearly, demand has to come, more or less, from a fraction of existing subscribers, and one needs working mechanisms to ensure that their regular usage is not unduly displaced to discount periods – the issue of cannibalization. We address this and other details regarding the viability of such a service in chapter 8.

The service proposed above makes sense, but several important issues can only be settled by a serious market study. These issues relate to the revenue sensitivity of the users and the fact that given the competitive nature of industry currently, the market may simply not accept complicated pricing schemes. Since these issues cannot be addressed in a speculative environment, we leave them for future research, pending the usefulness of the model formulated in this thesis.

To estimate the revenue impact of such an offering, say a company charges \$1/month for subscription to be notified of these price changes. Then consider a company like AT&T. With approximately 80 million customers[11], even if only 1% of the subscribers take up this offer, one immediately obtains \$800,000 per month as the revenue from subscription for the service. In addition, if an increased revenue of \$0.50 per month per subscriber is achieved on average, we have \$400,000 million per month increase. These are annual increases of \$14.4 million, with little added costs. Remember that we are assuming a market penetration of only 1%. Of course, if one loses more than \$1 per person in call revenue, one achieves a net loss. For smaller companies, the effects may not be as dramatic, but are still significant. Sprint, for instance, provides long-distance service to around 15 million homes, and the above calculations result in added revenue of \$2.7 million annually. With these calculations, one needs around 0.6 million people to subscribe to such a service to get \$1 million added revenue.

2.2.5 Market segmentation

Market segmentation is a peculiar activity in the sense that it does not lend itself to formalization like other YM activities. The usual method is to try variations of products and services to gauge the market reaction. It is likely that the situation will be similar for telecom networks.

Again, the airlines examples serve as useful guides for the possibilities in telecom. Market segmentation activities in airlines consist of mechanisms to separate business travelers from leisure travelers. These mechanisms have evolved over the last 30 odd years by repeated experimentation with fare products, starting from from British Airways' *Early-bird bookings* and American's *Super-Saver* fares, maturing to include weekend-stay-overs, multi-hop pricing structures and frequent-flyer mileage programs that minimally interfere with revenue producing customers. The informal nature of the process has produced little published research. Even literature surveys on revenue management do not usually usually include models for such activities. It remains an area where creativity and understanding of the market are required.

One might expect a similar situation for telecom, where market segmentation will arise on multiple fronts. In the YM services context, generating ideas for YM services is itself an informal process and will automatically result in market segmentation, since most services will be created to cater to users with different needs. For instance, the services mentioned in sections 2.2.2 and 2.2.3 target users who do not need the network in real-time. Similarly, dynamic pricing is an attempt to segment the market using prices.

[10]For instance, posting the prices on the web, or using a ticker tape to display the prices in a small window on the user's computer.

[11]Source: *Wall Street Journal* (3 Star, Eastern (Princeton, NJ) Edition), 7 April, 1998.

Table 2.1: Services proposed in this thesis.

Service Ideas for Telecom Yield Management
Network-usage by a latest start time: two possible services
Operating a High-speed Backbone as a Transport Network
The Digital FedEx Service
Probe location
Guaranteeing web server performance
Quasi real-time pricing of long-distance calls

However, other possibilities also arise where one can use modeling to segment the market for existing services. We choose to illustrate the use of market segmentation in the context of an existing telecom service, using network information to create a qualitative difference in the service offering. The general principle is straightforward. Use service guarantees to differentiate customers wherever possible. We present an idea below in which such a concept could prove useful.

Market segmentation-related work in this thesis

Consider web-hosting services which are integral parts of most Internet Service Provider (ISP) offerings. Suppose an ISP includes a guarantee of service of the following form when agreeing to host a customer's site. The fraction F of files transfers with transfer times greater than t will be less than P over every period of length T. We expect to be able to charge a premium for such a guarantee since the contract will also typically include a penalty for not meeting the guarantee. The main issue for the ISP is to understand its network well enough to be able to meet this guarantee reliably with the current traffic patterns on the network. Chapter 7 discusses the main considerations in obtaining such a guarantee using a model for quantifying the delays experienced by files.

Suppose one charged \$50/month on average from a corporate customer for a guarantee such as above. Of course, different customers will have differing levels of guarantees and will in all likelihood be paying different amounts, but for preliminary calculations, some number such as above serves fine. Then one obtains \$600/year as added revenue from one customer. With 100 such customers, this becomes \$60000/year in additional revenue alone. This may seem small but one has to keep in mind that the numbers we have used are relatively small and arbitrary.

2.3 A Summary of the Modeling Work in this Thesis

Table 2.1 lists the services proposed in this thesis for telecom YM. Figure 2-1 illustrates the models attempted for each service, and their situation within our proposed modeling framework. Figure 2-1 also reveals the modeling and analysis opportunities that lie ahead, as we "invent" new services – rows in the figure, fill in the empty slots with new models, and construct increasingly better models for the already filled ones.

2.4 Ideas for Other Services

In this section, we list some other interesting operations and services that might be relevant to telecommunications YM, but for which no work is presented in this thesis.

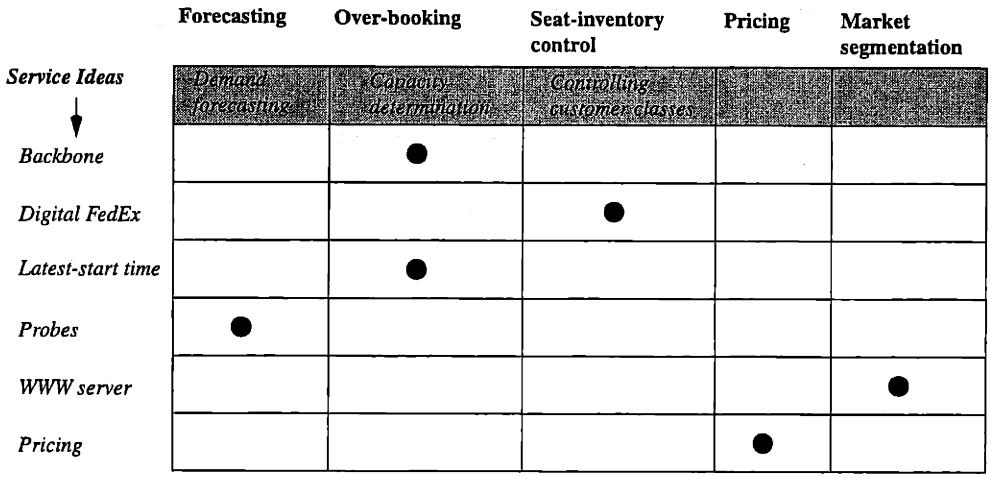


Figure 2-1: A summary of the modeling work in this thesis and its relation to the proposed modeling framework. The •'s are the models attempted.

Variants of "Lucy"

The same infrastructure underlying the *Network-assisted Calls* services of section 2.2.2 might be used to offer other services. For instance, the network provider might use agents like "Lucy" as traffic management mechanisms, posting "spot prices" every T time units, for 'Lucy' to initiate network usage as soon as some algorithm – perhaps from a stochastic dynamic program, observes optimal conditions for action. Even more straightforward might be for the network to post time-of-day prices and for Lucy to act solely on that information. The range of possibilities for implementation is large. For instance, instead of only user-side agents like 'Lucy', the network could operate a population of its own agents which have limited information about the network, and have them interact with 'Lucy' to maximize net revenues from the entire network.

Operational network design in a 'Yield Management' environment

Once a network is in place, one may wish to determine optimal routing and transmission policies that handle discount or 'yield managed' traffic differently from premium traffic. Revenue management policies, coupled with time of day demand variations - may suggest routing and multi-casting schemes that appear to the eye to be far from optimal yet in fact are optimal. Again, in the airlines, think of routings for low fare customers: sometimes it is better to route a person from A to B on three separate flight segments instead of a non-stop flight, increasing the number of hops from one to three. Yet such a policy minimizes net congestion on the network, where congestion is appropriately defined across the ensemble of customers.

Strategic network-design in a 'Yield Management' environment

One may wish to capitalize on the effects of YM on network traffic to minimize capacity costs. In the absence of revenue management policies that aim at smoothing demands for network services, one must design the network to respond to (unmanaged) peak loads. This requires a larger network capacity – hence larger capital investment – than would otherwise be required. This fact is a key reason why electric utilities pay customers to find ways to save electricity, especially during peak

times. The utilities do not want to invest capital toward creating a power generation and distribution network that is fully utilized only at (infrequent) peak demand times; the investment simply does not pay for itself. It is cheaper to pay customers to find ways not to demand their product! Such effects might deserve modeling when designing networks. Particularly in an environment where capacity can be leased on short notice, one may be able to minimize the rental costs of additional capacity.

Service-level agreements

At the strategic level, by linking Internet economics and Internet traffic via service level agreements, contract design, and efficient market clearing mechanisms for bandwidth, one may be able to maximize yield from capacity usage. The central decision to be made here for a large network-owner is essentially to determine a system of pricing with associated service level guarantees; for example, given traffic patterns on a network, how many options contracts on capacity can be priced for premium or guaranteed services.

Examples of such market clearing mechanisms may be: (a) a system of capacity options, where users are able to purchase the "option" to use capacity for a fixed period in the future with associated service level guarantees. These contracts may be medium to long range time-scales of use (say a few months), and also allow for interesting hedging opportunities; (b) spot market pricing of current capacity to clear the capacity on a shorter time-scale of use (say a day or a week); (c) capacity partitioning, a hybrid of the above, where a network owner makes the decision to use part of its capacity for spot-market sales and part for options type contracts.

Modeling congestion cost for an "Internet Service Provider" (ISP)

Customer response to congestion levels is obviously important in several different ways, but the seemingly simple question of how increase in congestion levels – appropriately defined – affect consumer behavior has not yet been addressed in telecom literature. Research, instead, has focused on other aspects of telecommunications economics such as, for instance: finding socially optimal pricing schemes for usage of digital networks, see MacKie-Mason [MMV95] or Kelly [KMT98], allocation of resources between users to maximize unspecified user utilities and telecommunications economics policy such as Bailey and McKnight [MB97] (add inter-providers' settlements literature).

There is room for new thinking in this arena and an idea is to borrow the framework of random utility theory, or more precisely discrete choice theory, which has proved extremely valuable in other related contexts such as transportation demand modeling – see, for instance, Lerman and Ben-Akiva [BAL94]. In particular, discrete choice modeling has the attractive feature that it provides a verifiable medium for testing if our analysis and predictions of consumer behavior are reasonable. This framework can be used to operationalize our intuition about customer behavior and may result in unforeseen benefits.

2.5 Summary

In this chapter, we proposed using the airlines YM modeling framework of *forecasting*, *over-booking*, *seat-inventory control*, *pricing* and *market segmentation* for modeling telecom YM services. We discussed possible telecom versions of each component of the framework. We briefly described the new service ideas articulated in this thesis and the decision problems that will be modeled in later chapters. We also included a summary of the modeling work and its relation to our modeling framework. Finally, we mentioned some ideas for other telecom services and operations that could be relevant to a YM effort in telecom.

Chapter 3

Overbooking: Operating a High-Speed Backbone as a Transport Network

This chapter demonstrates how over-booking type decision models might arise for telecom YM. The vehicle is a YM service designed to utilize available capacity on a high-speed backbone (c.f. section 2.2.2 for the context). We first introduce and discuss the service, including some practical aspects of its operation. We then model a decision problem related to capacity determination similar to the over-booking decision in airlines. Specifically, a single-link model is considered for determining the optimal number of requests to admit on the backbone, given existing traffic. We analyze the performance of a simple and attractive class of policies for revenue maximization. Several interesting properties of these policies are established and a fast algorithm is presented for obtaining them. Their performance is compared to the optimal policies and the model is extended in several directions.

The chapter is organized as follows. Section 3.1 introduces and discusses the service. Section 3.2 presents a single-link model and its discussion. Section 3.3 proposes the use of linear admission policies for revenue maximization and analyzes them in detail, presenting an analytical solution and investigating their sensitivity to service time distributions. In section 3.4, we explore an alternate formulation to compare the performance of these linear policies to truly optimal policies. Section 3.5 considers extensions of the model. Section 3.7 is the summary and lists the contributions of this chapter.

3.1 The service

Consider a provider offering spare capacity on a backbone for content transport. Users wishing to transmit bulk content between locations can upload it to a web-site. Once uploaded, it is automatically transmitted to a destination address over the backbone at a guaranteed high fixed-rate. Incentives for the customers to use the service will be (i) a deep discount since renting network capacity for content transmission is expensive, and (ii) performance, since web-sites will typically be close to the end-users resulting in better performance than the public Internet. The motivation for the provider is generation of revenue from available capacity, as long as existing traffic is not adversely impacted. Several practical issues, both technical and marketing, must be considered to ensure that the service offering indeed capitalizes on the motivation stated above. We comment on several of these in section 3.1.1.

In this setting, we explore how to maximize revenue from the service, using an over-booking type decision as follows. When a user accesses the web-site and 'clicks' a button to upload content, the

web-server must decide, based on the number of requests under transmission on the backbone and the number of requests being uploaded onto the web-site, if it can guarantee a high-speed connection to this user as soon as she finishes uploading the content. Further, assume that the operation of the service is as simple as possible. Specifically, this means that (i) the user is not required to enter any information about the nature and size of the content, and (ii) the server keeps no state information such as the amount of content already transmitted for the requests in service. Then the server has two options at each arrival, it can either redirect an upload request to an alternate provider's web-site at some cost, or let the upload proceed. If when the request is uploaded, it cannot obtain a guaranteed connection on the backbone, we assume it is re-directed to an alternate network instantaneously at a higher cost. This entire process is assumed to be transparent to the user, i.e. the shunting of the content to an alternate provider and the shipping of the content over an alternate backbone are all invisible to the user, for business reasons.

3.1.1 Practical Issues

The *yield management* motivation for the service is (i) to either use spare capacity in the network and/or (ii) to operate a network at near full capacity, by not designing it for peak demands (c.f. section 1.3). In both cases, the excess demand would need to be deflected to alternate networks. This could be an arrangement with an alternate provider, or perhaps, instantaneous acquisition of more bandwidth whenever demand exceeds capacity. The second option could be exercised through electronic *bandwidth-exchanges*, such as the ones being developed[1]. Such arrangements are increasingly likely in the emerging telecom market where bandwidth liquidity is already significant.

Examples of providers for such a service might be companies such as MCI^{TM} which operate a high-speed Internet backbone. One can envisage MCI using residual capacity in its network for such a service, perhaps by setting capacity limits at daily or hourly time scales. Customers for this service could include corporations such as CNN^{TM} and NBC^{TM}, for example, which have typically used arrangements with proprietary networks such as Vyvx^{TM} in the past to transport programmed content.

The idea for the service is intriguing, but an actual offering will have to consider many practical issues, such as the infrastructure for managing the service. For instance, with our suggested architecture of geographically distributed web-sites as 'gateways' to the backbone, location of the sites will need to be considered carefully. The objective would be to locate the sites so users can find a site close to them easily, making content download fast. Several interesting modeling questions arise here, even though we focus only on an overbooking type decision in this thesis. For instance, (i) determining the closest/fastest upload site for a request arriving from a location, (ii) determining the number and placement of these upload sites to minimize the mean upload time across the ensemble of customers[2] and (iii) determining the storage space required at the sites to handle demand.

The nature of the service offering and its operation is also subject to several possibilities. For instance, even though we only consider one available transmission rate in this paper, it is more likely that one would offer a range of available rates to cater to customer preferences and to better segment the market. Further, the operation of the web-server itself is subject to many possibilities. One may wish for the server to maintain state information such as the amount of content already transmitted on the backbone, the size of the content the customer is uploading, etc. This information may then be used to make better operating decisions. For instance, if one has information on the distributions of the residual times for content transmissions under way, one can consider using the model of chapter 4 instead of the model in this chapter.

[1] See, for instance, the article by Thomas L. Friedman on bandwidth exchanges in the *New York Times*, titled "TheLandgrab.com", January 18, 2000.

[2] Stochastic facility location models might be particularly useful for this problem.

3.2 A Single-link Model

We model the decision problem presented in section 3.1 to determine a policy for redirecting upload requests to an alternate provider. The model, more or less, captures most of the essential characteristics involved in a revenue maximizing decision, and focuses only on network bandwidth as the constraining resource.

3.2.1 The model

Poisson arrivals with rate λ occur at a single link with C identical circuits. Each arrival can be admitted to the system or redirected at cost c_e to an *alternate provider*. If admitted, it takes the customer a random amount of time to finish uploading content, distributed exponentially with mean $1/\mu_1$. When uploaded, the customer occupies a free circuit on the link, if one is available, for an exponential duration with mean $1/\mu_2$. If no circuit is available when a customer finishes uploading, the request is routed to an *alternate link* at cost $c_i > c_e$, at which point it is considered lost to the system. The assumption $c_i > c_e$ is necessary for the problem to be well-formed, for if $c_i \leq c_e$, admitting everyone would be optimal.

Call the fraction of arrivals that are redirected at arrival *external blocking* B_e, and the fraction of *admitted* requests that find all C circuits occupied, B_i, *internal blocking*. Let $i = 0, \ldots, \infty$, be the number of customers uploading and $j = 0, \ldots, C$, be the number of customers transmitting, both in steady state[3]. Note that the subscript i in c_i stands for *internal blocking*, and does not refer to the state variable i. We preferred this abuse of notation to the burden of remembering a more obscure subscript.

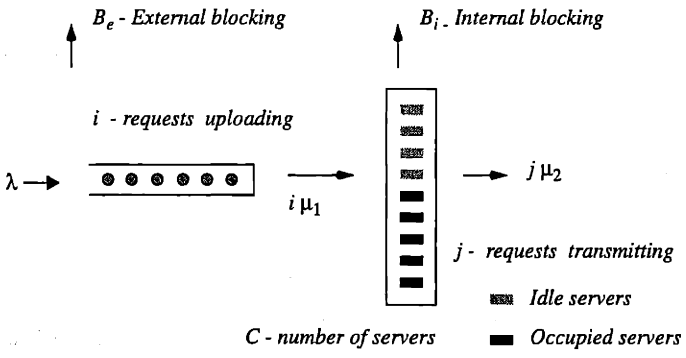


Figure 3-1: The model.

3.2.2 Remarks

1. The exponentiality of the random variables in the model is attractive, making it tractable. However, there is no reason to believe that service times will be memoryless, and a more realistic model would have general distributions for service times. The Markov description of system with general service times, however, becomes too complicated to be of much initial

[3]It is relatively obvious that steady state exists since an uncontrolled system is simply an $M/M/\infty$ queue feeding a loss system.

value. Such systems are useful for getting numerical results, but usually do not yield immediate qualitative insight. Our hope is to gain insight into the behavior of the system using exponential distributions, to help in analysis of more general models.

2. The assumption of memoryless arrivals is reasonable since requests generated at human activity time-scales are usually well-represented by a Poisson process. Much evidence is available in literature supporting this assumption, some of which can be found in [PF94].

3. A priori, one expects $\mu_1 \leq \mu_2$ since the backbone will usually be much faster than the connection to the web-site. We will not, however, impose this restriction in the analysis unless absolutely necessary.

4. The most troublesome aspect of the model is the assumption of independent service times at successive stages for every request. This is clearly not the case, since the time spent uploading by a request should be strongly correlated to the transmission time over the link. We expect, for example, a correlation of 1 if there is no independent disturbance in the available transmission rate over the link. Unfortunately, a Markov description again becomes very complicated for a model with correlated service times. We comment on this problem in detail in section 3.3.4.

5. Instead of a detailed literature review, we briefly comment on related models. The space of Markovian models is extremely rich and varied. Further, Markovian models often have deeper connections to each other than appear at first sight. All of which makes it difficult to say that ours is the first treatment of such a model.

   For instance, several models involve multiple stages, but most consider systems where requests overflowing a first link get a chance at a second link [Kel91]. A related two-stage *retrial* model is discussed in detail in [Wol89], where over-flows from a first loss system are considered as arrivals to a second loss system and the model is extended to an arbitrary number of stages. Admission policies are not considered for this model and analysis is attempted under FCFS operating policies. Common questions in over-flow and retrial models involve performance measures such as fractions of requests lost, etc. Our search has not so far revealed a model directly related – that considers optimization of revenue in a system where requests can get blocked at a second link. Research on over-flow models remains huge but relatively scattered (c.f. [Wol89], chapter 7-14).

   In terms of telecommunications literature, the above model does not appear in mainstream literature since it does not naturally correspond to the mechanics of common services such as real-time telephony etc. It may arise implicitly in modeling some ancillary activity but we have not so far found a report. Over-flow and re-trial models, on the other hand, appear extensively in telecommunications literature.

3.2.3 Objective function

To maximize revenue using an optimal redirection policy at arrival times, it is equivalent to focus on minimization of system cost, defined as the cost of internal and external blocking since the total revenue arriving to the system is constant whenever revenue for each arrival is independent identically distributed – for instance, when revenue per call is fixed or depends only on the length of transmission of a request.

A redirection policy π is a function mapping each state (i, j) to $\{0, 1\}$ for all $i = 0, \ldots, \infty$, and $j = 0, \ldots, C$. $\pi(i, j) = 1$ if an arriving request is accepted when the system is in state (i, j), otherwise $\pi(i, j) = 0$. By convention, we let $\pi(i, j) = 0$ if state (i, j) cannot occur with positive probability under π.

Under any given redirection policy π, the system behaves as a Markov process with a single

recurrent class[4], and the long-term time average cost can be written as follows:

$$z^\pi = \lambda c_e B_e^\pi + \lambda c_i (1 - B_e^\pi) B_i^\pi.$$

We can reduce the above expression to a convenient form expressing z^π in terms of p_{ij}^π, the steady state probabilities of the Markov process under the policy π. Call $\mathcal{S}^\pi$ the set of positive probability states under π, and recall that under π, B_i^π is the fraction of *admitted* requests that are blocked because all C servers are busy. B_i^π is then written as:

$$B_i^\pi = \frac{\mu_1}{\lambda} \frac{\sum_{i=0}^{\infty} i\, p_{iC}^\pi}{1 - B_e^\pi},$$

Then the objective function becomes:

$$z^\pi = \lambda c_e \sum_{\substack{(i,j)\in\mathcal{S}_\pi: \\ \pi(i,j)=0}} p_{ij}^\pi + \mu_1 c_i \sum_{(i,C)\in\mathcal{S}_\pi} i p_{iC}^\pi.$$

The first summation involves states where arrivals are redirected to an alternate system and the second summation involves states where C circuits are full. Because of the clutter in notation, we will drop the superscript π from the algebra unless needed, since the policy under consideration will usually be obvious.

3.2.4 Comments on optimal policies

Under any policy π, one obtains a Markov process with the following transition rates. Figure 3-2 shows these transitions for a linear policy of the form, $\pi(i,j) = 1$ iff $i + j \leq \theta - 1$, for a particular $\theta = 1, \ldots, \infty$.

$$j < C: \begin{cases} (i,j) & \to (i+1,j), & \pi(i,j)\lambda, \\ & \to (i-1,j+1), & i\mu_1, \\ & \to (i,j-1), & j\mu_2. \end{cases}$$

$$j = C: \begin{cases} (i,C) & \to (i+1,j), & \pi(i,j)\lambda, \\ & \to (i-1,j), & i\mu_1, \\ & \to (i,j-1), & j\mu_2. \end{cases}$$

Given policy π, one can write the global balance equations in terms of the above transition rates and solve them numerically to evaluate z^π. This leaves, however, the question of determining the optimal policy. The policy space is large, and a simple search might require evaluating a combinatorial number of policies, each evaluation requiring the solution of a possibly large linear system.

To get a rough sense of the problem, note that even if i is restricted to be always less than some integer N, then with C servers, there are $C \times N$ positive probability states for the Markov process. With the decision is admit, not admit in each state, the total number of possible policies are $2^{C \times N}$, clearly very large. This necessitates the need to either find an algorithm that that prunes the search space significantly, or start the search with a policy already close to the optimal. Of course many of these policies are non-sensical, for instance, the ones that do not assign $\pi(0,0) = 1$, in which case $p_{00} = 1$, or policies involving transient classes.

[4]State (0,0) and all states that are reachable from it.

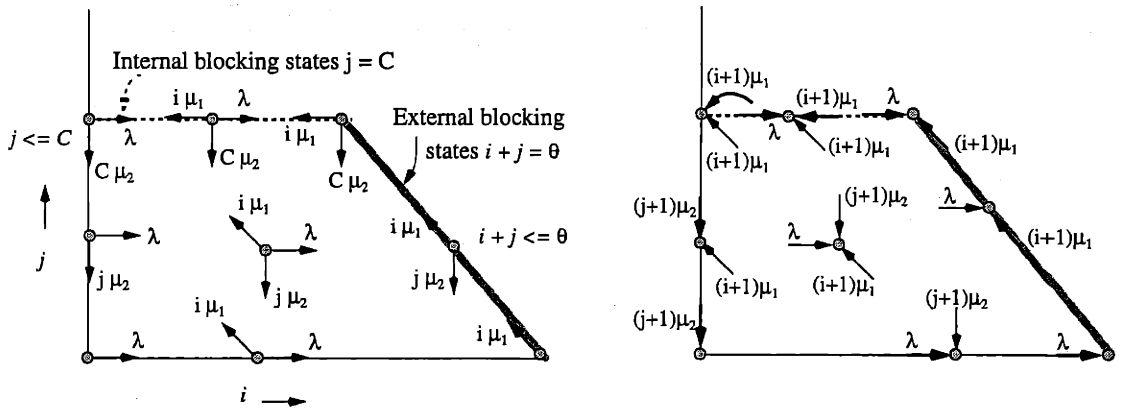


Figure 3-2: The state space under a linear admission policy of the form, admit *iff* $i + j \leq \theta - 1$. Inward and outward transition rates are shown separately.

Some simple policies, however, admit of relatively easier analysis. For instance, under a linear redirection policy, the state space is given by figure 3-2, and one can easily write down the global balance equations. This motivates the analysis of section 3.3 where we try to obtain a policy optimal within the class of linear policies.

3.3 Linear Redirection Policies

Motivated by the difficulty of computing optimal policies, we analyze the performance of linear redirection policies, as defined below. Other motivations are the simple form of a linear policy, which makes implementation easy, and to obtain qualitative insights into the main parameters affecting system cost.

Call a policy linear if it admits arriving requests only when $i+\beta j \leq \theta-1$ and $j \leq C$, where β and $\theta \geq 1$, are real constants. For any such given linear policy, characterized by (β, θ), the objective value $z(\beta, \theta)$ can be computed by solving the global balance equations numerically to obtain $p_{ij}(\beta, \theta)$, the steady-state probabilities.

A strategy might then be to search for an optimal linear policy over the space $\beta \geq 0, \theta \geq 1$. The difficulty of such a strategy is the search, since each evaluation requires the solution of a system of linear equations – likely to be large for anything but the most trivial systems. Further, it is unclear how to determine the direction of the search at each iteration. Fortunately, the class of linear policies with $\beta = 1$ permit of an analytical solution, and are computationally found to be nearly optimal within the class of all linear policies.

The central results obtained are:

1. Linear policies with $\beta = 1$, i.e. admit only if $i + j \leq \theta - 1$, permit a closed-form solution for the steady-state probabilities, resulting in a simple algorithm for obtaining an optimal policy within this class. Further, the closed-form allows for analysis of these policies, giving some qualitative insights into the system.

2. Even within the restricted class of linear policies with $\beta = 1$, one obtains significant improve-

ments over a naive policy – admit-everyone blocking at the link if necessary – by as much as 5%-30%, unless either the system is very lightly loaded or the ratio c_i/c_e is very low.

3. Numerical investigations indicate that the performance of the optimal $\beta = 1$ policy is very close to a policy optimal within the class of all linear policies.

4. Numerical investigations reveal the performance of the optimal $\beta = 1$ policies is within 1-5% of the true optimal policy, (c.f. section 3.4).

5. And most importantly, numerical investigations reveal that the steady state probabilities of the system under a $\beta = 1$ policy are *insensitive to the distribution and the correlation structure of the service times in the successive stages.* This property allows for real investigations into the behavior of more realistic models.

3.3.1 Linear policies: $i + j \leq \theta - 1$

Steady-state probabilities for policies which admit iff $i + j \leq \theta - 1$ can be obtained in closed-form. Call the state space for this policy, $\mathcal{S}(\theta)$. Then $\mathcal{S}(\theta) = \{(i,j) : i + j \leq \theta, j \leq C, i \geq 0, j \geq 0\}$. Letting $p_{ij} = 0$ whenever $(i,j) \notin \mathcal{S}(\theta)$, the global balance equations can be written as follows for all $(i,j) \in \mathcal{S}(\theta)$.

$$\begin{aligned} j < C: \quad & p_{ij}(\lambda + i\mu_1 + j\mu_2) = p_{i-1,j}\lambda + p_{i,j+1}(j+1)\mu_2 + p_{i+1,j-1}(i+1)\mu_1, \\ j = C: \quad & p_{ij}(\lambda + i\mu_1 + C\mu_2) = p_{i-1,C}\lambda + p_{i+1,C}(i+1)\mu_1 + p_{i+1,C-1}(i+1)\mu_1. \end{aligned}$$

With $\rho_1 = \lambda/\mu_1, \rho_2 = \lambda/\mu_2$, the solution to these equations is given by:

$$p_{ij}(\theta) = \frac{1}{G(\theta)} \frac{\rho_1^i \rho_2^j}{i!\ j!}, \quad (i,j) \in \mathcal{S}(\theta), \quad \text{where} \quad G(\theta) = \sum_{(i,j)\in\mathcal{S}(\theta)} \frac{\rho_1^i \rho_2^j}{i!\ j!}.$$

And the objective function can be written as below, using $z(\theta)$ to denote the objective value.

$$z(\theta) = \frac{\lambda}{G(\theta)} \left[c_e \sum_{\substack{(i,j)\in\mathcal{S}(\theta): \\ i+j=\theta}} \frac{\rho_1^i \rho_2^j}{i!\ j!} + c_i \frac{\rho_2^C}{C!} \sum_{i=0}^{\theta-C-1} \frac{\rho_1^i}{i!} \right], \quad \theta = 1, 2, \ldots, \infty. \tag{1}$$

For convenience, we have written the expression in the above form, but the second summation only exists for values of θ with $\theta \geq C + 1$.

The product-form obtained for the steady-state probabilities is interesting and deserves some thought, since the same product-form is not obtained for linear policies with $\beta \neq 1$. One possible connection is to the known Erlang network. For instance, steady-state probabilities for our linear policy are exactly the same as that of the well-known Erlang loss-network model in telecommunications [Ros95], with the topology shown in figure 3-3. This motivates the question if there is a deeper connection between linear policies in our model and some appropriate topology of a product-form loss network. The hope being to discover insights from the substantial analysis available for the Erlang model [Ros95]. We do not pursue this direction in this thesis.

3.3.2 Properties

Expression (1) can now be investigated both analytically and numerically to establish some qualitative properties of the objective function $z(\theta)$.

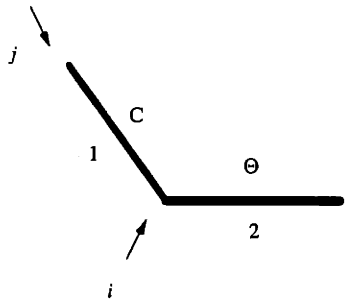


Figure 3-3: An Erlang loss network having the same steady-state probabilities as the $i + j \leq \theta - 1$ policy. Links 1, 2 have capacities C, θ respectively. Calls of type i, j arrive with rate λ each with mean holding times $1/\mu_1$ and $1/\mu_2$ respectively. Calls of type j require a circuit on links 1 and 2 simultaneously while calls of type i only require a circuit on link 2.

Properties which are not very surprising and can be established with some algebra include the following. The algebra is documented in section 3.6.

1. $z(\theta)$ is monotone decreasing from $\theta = 1, \ldots, C$. An intuitively obvious fact since with $\theta < C$, we are blocking requests unnecessarily even when the probability of internal blocking is 0.

2. $z(\infty)$, the cost of *admit-everyone* policy, converges to the cost of Erlang blocking at the link – as if the link of size C was operated under Poisson arrivals with rate λ and mean service time $1/\mu_2$.

3. $B_e(\theta)$, the external blocking, is monotone decreasing in θ.

4. The cost of internal blocking is monotone increasing in θ.

Other properties of the optimal policy θ^* and the optimal cost $z(\theta^*)$ for which there is numerical evidence are demonstrated in representative graphs of $z(\theta)$ against θ shown in figure 3-4. The figures suggest that the objective function either has a unique optimum or monotonically decreases. In the comments below, factors affecting the optimal policy θ^* refer to its sensitivity as long as $\theta^* < \infty$, i.e. whenever *admit-everyone* is not optimal.

1. The *optimal cost* $z(\theta^*)$ seems very sensitive to the arrival rate λ, slightly less so to c_e, c_i but quite insensitive to μ_1. It is also moderately sensitive to μ_2 for which a plot is not shown. This is seen by looking at all the plots.

2. c_e and c_i seem to affect the values of the *cost* $z(\theta)$ for $\theta < C$, and $\theta \geq C$, respectively, which is not very surprising. The behavior is as shown by the bottom-left plot. The surprising fact is that *policy* itself, θ^*, seems not to be very sensitive on the specific ratio c_e/c_i whenever a unique θ^* exists as indicated by this and the top-right plot.

3. The *optimal policy* θ^* does not seem very sensitive to the arrival rate λ, all other parameters being the same, as indicated from the top-left plot.

4. Even for lightly-loaded systems, characterized by moderate values of λ/μ_2, one obtains a unique $\theta^* < \infty$ if the ratio c_i/c_e is large enough, as indicated by the top-right plot. However, when c_i/c_e is moderate, the optimal policy is likely to be to accept everyone, i.e. $\theta^* = \infty$. This is also intuitively obvious, since if the cost of internal blocking is large, it should limit the number of people accepted when the link is full.

5. Finally, the bottom-right plot indicates that θ^* seems most sensitive to the parameter μ_1, with lower values of μ_1 increasing θ^* substantially. The same is true for μ_2, plots for which are not shown. μ_1 and μ_2 therefore seem to be the critical factors controlling the optimal policy.

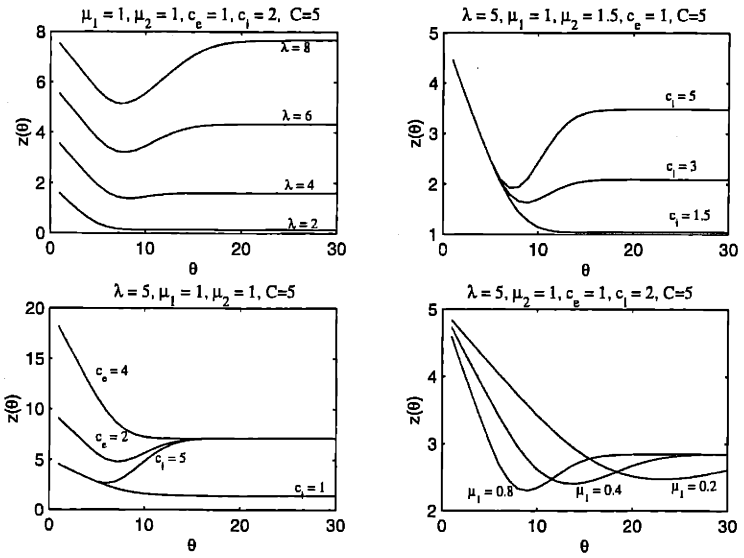


Figure 3-4: Representative plots of $z(\theta)$ vs. θ for the $i+j \leq \theta$ policy.

Finally, some other properties which we think are true but do not have formal proofs for include the following:

1. A characterization of the optimal θ in terms of the parameters of the model, including conditions on the loads ρ_1 and ρ_2 when such a θ exists.

2. A proof that the optimal θ is unique whenever admit-everyone is not optimal.

3.3.3 A fast algorithm for obtaining the optimal $i+j \leq \theta-1$ policy

Finding the optimal $\beta = 1$ policy involves searching over integers $\theta = 1, \ldots, \infty$. Computing the objective function $z(\theta)$ involves summations over the space $i+j \leq \theta-1$ to evaluate the normalization constants $G(\theta)$. A much more efficient algorithm can be obtained by recognizing the recursive structure of $G(\theta)$.

First note that the value of the objective function is strictly decreasing over $\theta = 1, \ldots, C$, as mentioned in section 3.3.2. We can therefore restrict our search to $\theta = C, \ldots, \infty$. This search can be organized very efficiently by re-writing expression (1) in the following form. One can then write a simple algorithm that takes no more than $O(C)$ operations to evaluate $z(\theta+1)$ given $z(\theta)$, instead of $O(C\theta)$ operations that would be required in a brute-force evaluation.

$$z(\theta) = \frac{\lambda}{G(\theta)} \left[c_e \frac{\rho_1^\theta}{\theta!} \sum_{j=0}^{C} \binom{\theta}{j} \left(\frac{\rho_2}{\rho_1}\right)^j + c_i \frac{\rho_2^C}{C!} \sum_{i=0}^{\theta-C-1} \frac{\rho_1^i}{i!} \right], \quad \theta = C, \ldots, \infty. \tag{2}$$

Consider the algorithm below for evaluating $z(\theta)$ for $\theta = C, \ldots, \infty$, which takes advantage of the following recursions involved in (2):

$$\binom{\theta}{j} = \frac{\theta}{\theta-j}\binom{\theta-1}{j}, \quad \text{and}$$

$$G(\theta) = G(\theta-1) + \frac{\rho_1^\theta}{\theta!} \sum_{j=0}^{C} \binom{\theta}{j} \left(\frac{\rho_2}{\rho_1}\right)^j.$$

An Algorithm for Fast Evaluation of $z(\theta)$, $\theta = C, \ldots, \infty$

Input :

$$\lambda, \mu_1, \mu_2, c_e, c_i, C$$

Initialization :

$$\theta := C$$
$$a_j := \binom{\theta}{j}\left(\frac{\rho_2}{\rho_1}\right)^j, \qquad j = 0, \ldots, C$$
$$G := \sum_{(i,j)\in S(\theta)} \frac{\rho_1^i \rho_2^j}{i!\, j!}$$
$$b := 0$$
$$z(\theta) := c_e \sum_{j=0}^{C} a_j$$

Main Loop :

for $\theta = C+1, \ldots, \infty$
  $A := 0$
  for $j = 0, \ldots, C$
    $a_j(\theta) := \frac{\theta}{\theta - j}\, a_j(\theta)$
    $A := A + a_j$
  end for;
  $G := G + A\frac{\rho_1^\theta}{\theta!}$
  $b := b + \frac{\rho_1^{\theta-C-1}}{(\theta-C-1)!}$
  $z(\theta) := \frac{\lambda}{G}\left[c_e A \; + \; c_i b \frac{\rho_2^C}{C!}\right]$
end for;

Output :

$$z(\theta), \qquad \theta = C, \ldots, \infty$$

3.3.4 General service time distributions and correlations

We address here the exponentiality and independence of the service times, assumed in the model of section 3.2. These are clearly unrealistic assumptions since transmission times for the same request must be strongly correlated at subsequent stages. We provide numerical evidence below that the performance of the $i + j \leq \theta - 1$ policies is independent of the distribution of the service times and their correlation structure, depending only on their mean. This makes such policies highly attractive for implementation and analytical investigations. We think a formal proof for this property would be very exciting to obtain.

The representative results below were obtained by simulating the system under the $i + j \leq \theta - 1$ policy with differing distributions of service times but the same means. Percent errors for simulated costs are computed against the analytically obtained cost $z(\theta^*)$ for the following service time distributions: independent exponential in both stages, exponential in the first and perfectly correlated (i.e. second stage time is deterministic and the same as the sampled time in the first stage), independent Erlang in both stages and independent Pareto on both stages. It is seen that the error is less than 1% in all cases, which is more or less simulation noise, as shown by the comparison of the simulated costs for the independent exponentials case against the exact cost $z(\theta^*)$.

Input	μ_1	μ_2	c_e	c_i	C
	1	1	1	2	5

	θ^*	$z(\theta^*)$	Sim cost Exp-Exp	% diff	Sim cost Exp-Det corr	% diff
λ=2	10	0.146	0.14756	1.07	0.14763	0.05
λ=4	8	1.380	1.3796	-0.05	1.3806	0.07
λ=6	8	3.199	3.2042	0.13	3.1941	-0.31
λ=8	8	5.151	5.1431	-0.16	5.1471	0.08

	θ^*	$z(\theta^*)$	Sim cost Erlang-Erlang	% diff	Sim cost Pareto-Pareto	% diff
λ=2	10	0.146	0.14824	0.46	0.14619	-0.92
λ=4	8	1.380	1.3741	-0.40	1.3756	-0.29
λ=6	8	3.199	3.1923	-0.37	3.1928	-0.35
λ=8	8	5.151	5.1461	0.06	5.131	-0.23

3.3.5 Optimal linear policies $i + \beta j \leq \theta$

A natural question to consider is the performance of the optimal policy θ^*, when $\beta = 1$, compared to an optimal linear policy with an arbitrary value of β. Unfortunately, the same product-form solution no longer exists for arbitrary β policies, because of the structure of the state-space, as shown in figure 3-5 for several linear policies. We therefore resort to numerical investigations to determine how to improve a policy with $\beta = 1$.

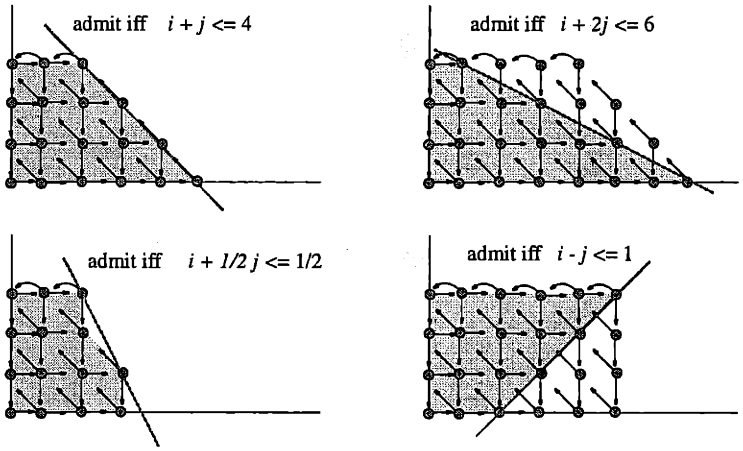


Figure 3-5: Representative state space for some linear policies. The sets of positive probability states are shown for each policy. The shaded regions are states to which one in-bound transition is an admission.

Experiments with several linear policies as shown in figure 3-6 indicate that no significant improvement is obtained by arbitrary β policies over the optimal $\beta = 1$ policy. The results were obtained by numerically solving the global balance equations for certain sets of policies and comparing the resulting $z(\beta, \theta)$ values to the values of $z(1, \theta)$. The slight differences may be attributable to numerical imprecision in the solution of the global balance equations.

We speculate that the θ^* policy captures the major trade-offs within the class of linear policies. For $\beta < 1$, we always get a worse optimal policy, and for $\beta > 1$, the benefit in admitting more external arrivals is offset by the cost of internal blocking, as indicated in figure 3-5.

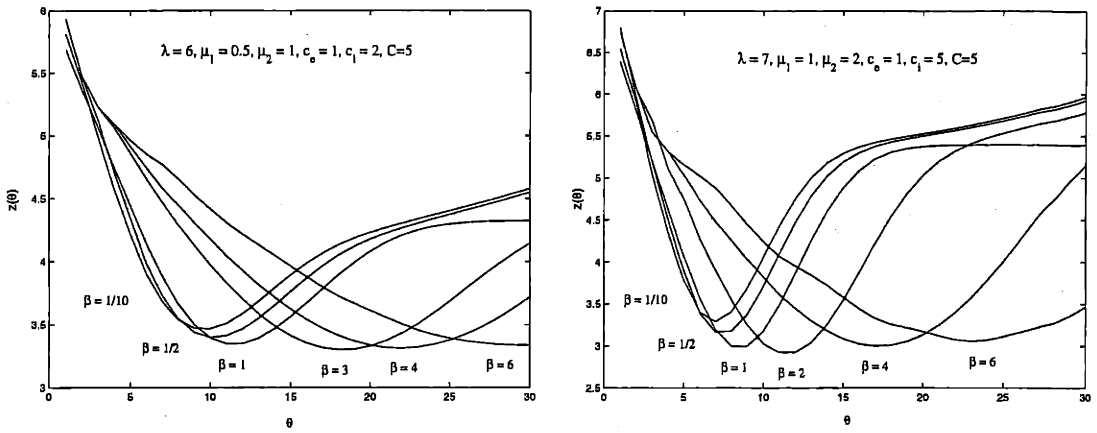


Figure 3-6: Comparison of linear policies for two sets of model parameters. The optimal value $z(\theta)$ does not decrease significantly from $\beta = 1$ as β increases.

3.4 Optimal policies

Here we compare the performance of a θ^* policy to that of the true optimal policy. To obtain optimal policies, we view our decision model as a countable state, continuous time, average reward dynamic programming problem. Since under any policy π, the Markov process obtained has bounded rates, standard dynamic programming theory applies (see [Ber95] for instance) asserting that there must be a stationary optimal policy. Note that in principle, our state space is infinite. However, there is either an optimal policy with a finite state space or *admit-everyone* is optimal. In the latter case, the θ^* policy is also optimal and there is no need to search for an optimal policy. Therefore, only when a finite θ^* policy is obtained do we need to consider its sub-optimality.

We reformulate the problem as that of minimizing average cost over an infinite horizon. We discretize time into intervals of length δ and call $J^*\delta$ the optimal average cost per stage to write Bellman's equation as follows:

$$j \neq C : J^*\delta + h_{ij} = \min_{\pi_{ij}\in\{0,1\}} \{ \; (1-\pi_{ij})\lambda\delta c_e \; + \; (1 - i\mu_1\delta - j\mu_2\delta - \pi_{ij}\lambda\delta)\; h_{ij} \; + \; i\mu_1\delta \; h_{i-1,j+1} \; + \; j\mu_2\delta \; h_{i,j-1} + \pi_{ij}\lambda\delta \; h_{i+1,j} \; \},$$

$$j = C : J^*\delta + h_{ij} = \min_{\pi_{ij}\in\{0,1\}} \{ \; (1-\pi_{ij})\lambda\delta c_e \; + \; (1 - i\mu_1\delta - j\mu_2\delta - \pi_{ij}\lambda\delta)\; h_{ij} \; + \; i\mu_1\delta c_i \; + \; i\mu_1\delta \; h_{i-1,j} \; + \; j\mu_2\delta \; h_{i,j-1} + \pi_{ij}\lambda\delta \; h_{i+1,j} \; \}.$$

Here π_{ij} is the admit/not-admit decision in state (i,j) and h_{ij} the relative cost in state (i,j). We will consider the set of above equations for states with $i \leq N, j \leq C$, where N is some chosen finite number, perhaps $2\theta^*$. Then under the condition $h_{00} = 0$, Bellman's equation has a unique solution in the unknowns J^* and h_{ij}. Given a solution to Bellman's equation, the optimal policy is obtained by choosing a π_{ij} which minimizes the right-hand side. As an aside, the solution to Bellman's equation has the interpretation that h_{ij} is the relative reward in state (i,j). In particular, if a policy that attains the minimum in Bellman's equation on the rhs is followed starting from state (i,j) and starting from state (m,n), the *difference* in *total* rewards over the infinite horizon is $(h_{ij} - h_{mn})/\delta$.

		θ^*	$z(\theta^*)$	J^*	$(z(\theta^*) - J^*)/J^*$	Iterations
$\lambda = 10$	$C = 10$	24	1.3007	1.2993	0.1078	1
$\lambda = 20$	$C = 10$	21	10.2661	10.0836	1.8099	2
$\lambda = 20$	$C = 15$	34	4.6274	4.5882	0.8554	2
$\lambda = 50$	$C = 20$	43	27.847	27.3546	1.8001	3

Table 3.1: Performance comparison of the θ^* policy with the optimal policy for several values of λ and C. All other parameters were held constant for these runs, with $\mu_1 = 1, \mu_2 = 1.5, c_e = 1$ and $c_i = 2$. The last column lists iterations of the policy-iteration algorithm before convergence, when started with a θ^* policy.

With the above formulation, one can obtain an optimal policy using any of the classical dynamic programming algorithms, but the computational complexity increases with the size of the state space, limiting the feasibility of a brute-force approach. We use the policy-iteration algorithm which, when started from an arbitrary policy, obtains an improving sequence of policies in every iteration to converge to the optimal policy. In order to guarantee that the policy iteration algorithm terminates finitely, one needs to restrict it to a finite state space, hence the need for N. Further, the relative rewards h_{ij} need to be bounded. Both are not a problem when the state space is truncated.

We present computational comparisons of the optimal policy found using the policy iteration algorithm, when started from a θ^* policy. Table 3.1 lists the results and figures 3-7 and 3-8 show the state space for an optimal policy vs. that of the corresponding θ^* policy. Remarks are noted below.

1. The policy iteration algorithm takes no more than 2-3 iterations to find the optimal policy when started from a θ^* policy.

2. The performance of the θ^* policy is very close to the optimal, within $1-2\%$ for all cases tested.

3. The computational complexity of obtaining a θ^* policy is significantly lower than that of obtaining an optimal policy[5]. Obtaining $z(\theta)$ in every iteration of the algorithm of section 3.3.3, when finding a θ^* policy, is an $O(C)$ operation, whereas every iteration of the policy iteration algorithm might require $O(C^3N^3), N \geq C$ operations using standard Gaussian elimination, a speed that makes policy iteration infeasible for any but the smallest values of C. To emphasize the point, with $C = 20$, the difference in magnitude is greater than 20^5. In a Matlab^{TM} implementation, for instance, for the case of figure 3-8 with $N = 60$, the policy iteration algorithm took 25 minutes to complete three iterations, whereas the θ^* policy was obtained in less than 10 secs. Further, the storage requirement of the policy iteration algorithm will typically be $O(CM)$, unlike the storage requirement of the θ^* policy, only $O(C)$.

3.5 Interesting directions

We comment briefly on some interesting directions without pursuing them further. For several extensions of interest to the basic model of section 3.2, one might be able to identify product-form policies and compare their performance to that of an optimal admission policy.

For instance, consider the model of section 3.2 with the modification that an external Poisson stream of arrivals at rate γ (say class 2) is allowed to the link or the second stage directly, skipping the first stage. Call the arrivals from the original stream at stage 1 class-1. If service times for all

[5]One may, of course, try and find other more efficient algorithms for obtaining an optimal policy.

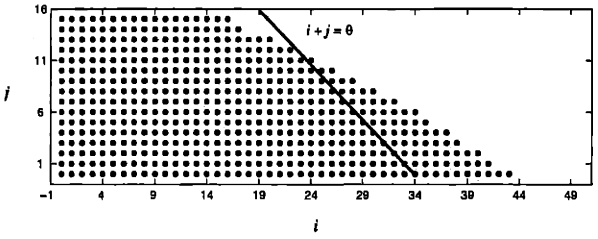


Figure 3-7: The set of positive probability states for an optimal policy vs. the θ^* policy for the case $\lambda = 20, C = 15, \theta^* = 34$, table 3.1.

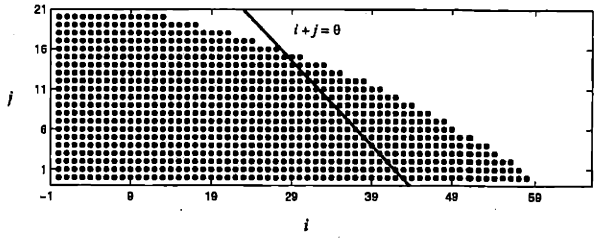


Figure 3-8: The positive probability states for the optimal policy for the case $\lambda = 50, C = 20, \theta^* = 43$, table 3.1.

requests using the link remains exponential with rate μ_2 regardless of class, a policy that admits class-1 arrivals into the first stage and class-2 at the second stage[6] only when $i + j \leq \theta - 1$ retains the product-form solution of section 3.3.1, where now $\rho_2 = (\lambda + \gamma)/\mu_2$. One can then investigate exactly as we did, the performance of this policy compared to alternative policies.

Similarly, the approach of finding a product-form policy might be usefully extended to multiple stages, fixed routing networks and/or multiple classes of arrivals and service times.

3.6 Proofs of properties: the $i + j \leq \theta - 1$ policy

The algebra for many of the properties of a linear policy with $\beta = 1$ is transcribed in this section.

First we note the easy fact that in an Erlang loss system with given capacity C and load ρ, the blocking probability $B(C)$ is monotone decreasing in the capacity C. The Erlang loss formula is:

$$B(C) = \frac{\rho^C/C!}{\sum_{i=0}^{C} \rho^i/i!}.$$

This can be re-written as:

$$B(C) = \frac{1}{\sum_{i=0}^{C} \frac{C!}{i!}\rho^{i-C}}, \quad \text{and}$$

$$B(C+1) = \frac{1}{\frac{(C+1)!}{\rho^{C+1}} + \sum_{i=0}^{C} \frac{(C+1)!}{(i+1)!}\rho^{i-C}}.$$

Now comparing the denominators, we see that $B(C) > B(C+1)$.

Remark 3.6.1. $z(\theta)$ is monotone decreasing in $\theta = 1, \ldots, C$.

Note that:

$$\sum_{i+j=\theta} \frac{\rho_1^i \rho_2^j}{i!j!} = \sum_{j=0}^{\theta} \frac{\rho_1^{\theta-j}\rho_2^j}{(\theta-j)!j!} = \frac{\rho_1^\theta}{\theta!} \sum_{j=0}^{\theta} \frac{\rho_2}{\rho_1} \binom{\theta}{j} = \frac{(\rho_1+\rho_2)^\theta}{\theta!}, \quad \text{for } \theta = 0, \ldots, C.$$

Therefore $z(\theta)$ for $\theta = 1, \ldots, C$ can be written as:

[6] Class 1 arrivals are blocked after the first stage only if the second stage is full.

$$z(\theta) = \frac{\sum_{i+j=\theta} \frac{\rho_1^i \rho_2^j}{i!j!}}{\sum_{i+j\leq\theta} \frac{\rho_1^i \rho_2^j}{i!j!}} = \frac{\frac{(\rho_1+\rho_2)^\theta}{\theta!}}{\sum_{i=0}^{\theta} \frac{(\rho_1+\rho_2)^i}{i!}}.$$

The expression on the right is now the same as the Erlang loss formula with load $\rho_1 + \rho_2$ and capacity θ, and is therefore decreasing in θ.

Remark 3.6.2. $z(\infty)$ converges, and the cost of admit-everyone policy is the same as c_i times the cost of blocking at the link with Poisson arrival rate λ as if there was no first stage.

To get $z(\infty)$, the cost of an admit-everyone policy and simultaneously show that it converges to a limit, note the following for the sub-expressions involved in $z(\theta)$, from (1):

$$\lim_{\theta\to\infty} \sum_{\substack{(i,j)\in\mathcal{S}(\theta):\\ i+j=\theta}} \frac{\rho_1^i \rho_2^j}{i!\ j!} = 0,$$

$$\lim_{\theta\to\infty} \frac{\rho_2^C}{C!} \sum_{i=0}^{\theta-C-1} \frac{\rho_1^i}{i!} = e^{\rho_1} \frac{\rho_2^C}{C!},$$

$$\lim_{\theta\to\infty} G(\theta) = e^{\rho_1} \sum_{j=0}^{C} \frac{\rho_2^j}{j!}.$$

Now it is obvious that $z(\theta)$ converges and that the cost of admitting everyone is the same as the cost of Erlang blocking at the link, as shown below:

$$z(\infty) = c_i \frac{\rho_2^C/C!}{\sum_{j=0}^{C} \rho_2^j/j!}.$$

To show the remaining two facts related to the monotonicity of $B_e(\theta)$ and $B_i(\theta)$, we require some messy algebra.

Remark 3.6.3. External blocking is monotone decreasing in θ.

We know this is the case when $\theta \leq C$. When $\theta \geq C$, we will show that for every $j = 0, \ldots, C$, $p_{\theta-j,j}(\theta)$, the probability of the external blocking state $(\theta - j, j)$ under policy $i + j \leq \theta - 1$, is monotone decreasing in θ. Since $B_e(\theta) = \sum_{j=0}^{C} p_{\theta-j,j}(\theta)$, we will get the desired result. This proceeds relatively similar to the monotonicity manipulations for the Erlang loss formula. For any given J between 0 and C, we have:

$$p_{\theta-J,J}(\theta) = \frac{\frac{\rho_1^{\theta-J}\rho_2^J}{(\theta-J)!J!}}{\sum_{\substack{i+j\leq\theta\\ j\leq C}} \frac{\rho_1^i \rho_2^j}{i!j!}} = \frac{1}{\sum_{k=0}^{\theta} \sum_{\substack{i+j=k\\ j\leq C}} \frac{(\theta-J)!J!}{i!j!} \rho_1^{i-\theta+J} \rho_2^{j-J}}.$$

For the same J, we have $p_{\theta+1-J,J}(\theta+1)$ as the corresponding blocking state under a $\theta+1$ policy, and a similar manipulation as above results in:

$$p_{\theta+1-J,J}(\theta+1) = \frac{\frac{\rho_1^{\theta+1-J}\rho_2^J}{(\theta+1-J)!J!}}{\sum_{\substack{i+j\le\theta+1\\ j\le C}} \frac{\rho_1^i\rho_2^j}{i!j!}}$$

$$= \frac{1}{\sum_{k=1}^{\theta+1}\sum_{\substack{i+j=k\\ j\le C}} \frac{(\theta+1-J)!J!}{i!j!}\rho_1^{i-\theta-1+J}\rho_2^{j-J} + \sum_{\substack{i+j=0\\ j\le C}} \frac{(\theta+1-J)!J!}{i!j!}\rho_1^{i-\theta-1+J}\rho_2^{j-J}}$$

$$= \frac{1}{\sum_{k=0}^{\theta}\sum_{\substack{i+j=k\\ j\le C}} \frac{(\theta+1-J)!J!}{(i+1)!j!}\rho_1^{i-\theta+J}\rho_2^{j-J} + \frac{(\theta+1-J)!J!}{\rho_1^{\theta+1-J}\rho_2^J}}.$$

Now comparing the denominators of the expressions for $p_{\theta-J,J}(\theta)$ and $p_{\theta+1-J,J}(\theta+1)$, we see that $p_{\theta-J,J}(\theta) > p_{\theta+1J,J}(\theta+1)$, which establishes $B_e(\theta) > B_e(\theta+1)$.

Remark 3.6.4. Internal blocking cost is monotone increasing in θ.

From the expression for $z(\theta)$, it suffices to show that $1/G(\theta)\sum_{i=0}^{\theta-C-1}\rho_1^i/i!$ is monotone increasing in θ, which requires us to show that for every θ,

$$\big(G(\theta+1) - G(\theta)\big)\sum_{i=0}^{\theta-C-1}\frac{\rho_1^i}{i!} \le G(\theta)\frac{\rho_1^{\theta-C}}{(\theta-C)!} \qquad (a).$$

Consider first the following inequality which will be used repeatedly, for $C \le k \le \theta$:

$$\sum_{j=0}^{C}\frac{\rho_1^{k-j}\rho_2^j}{(k-j)!j!} \ge \frac{(\theta-C)!}{\rho_1^{\theta-k}(k-C)!}\sum_{j=0}^{C}\frac{\rho_1^{\theta-j}\rho_2^j}{(\theta-j)!j!}$$

To illustrate why this is true, consider $k = \theta - 2$, for which we have:

$$\sum_{j=0}^{C}\frac{\rho_1^{\theta-2-j}\rho_2^j}{(\theta-2-j)!j!} = \frac{(\theta-C)(\theta-C-1)}{\rho_1^2}\sum_{j=0}^{C}\frac{\rho_1^{\theta-j}}{(\theta-C)(\theta-C-1)\ (\theta-j-2)!}\frac{\rho_2^j}{j!}$$

$$\ge \frac{(\theta-C)(\theta-C-1)}{\rho_1^2}\sum_{j=0}^{C}\frac{\rho_1^{\theta-j}}{(\theta-j)!}\frac{\rho_2^j}{j!},$$

since $j \le C$ for all terms in the summation. Now consider the rhs of expression (a), using the above inequality:

$$G(\theta)\frac{\rho_1^{\theta-C}}{(\theta-C)!} \ge \frac{\rho_1^{\theta-C}}{(\theta-C)!}\Big(\sum_{k=C}^{\theta}\frac{(\theta-C)!}{\rho_1^{\theta-k}(k-C)!}\Big)\Big(\sum_{j=0}^{C}\frac{\rho_1^{\theta-j}\rho_2^j}{(\theta-j)!j!}\Big)$$

$$= \Big(\sum_{i=0}^{\theta-C}\frac{\rho_1^i}{i!}\Big)\Big(\sum_{j=0}^{C}\frac{\rho_1^{\theta-j}\rho_2^j}{(\theta-j)!j!}\Big).$$

Whereas the lhs can be written as:

$$(G(\theta+1)-G(\theta))\sum_{i=0}^{\theta-C-1}\frac{\rho_1^i}{i!} = \left(\sum_{j=0}^{C}\frac{\rho_1^{\theta+1-j}\rho_2^j}{(\theta+1-j)!j!}\right)\sum_{i=0}^{\theta-C-1}\frac{\rho_1^i}{i!}$$
$$\leq \left(\sum_{j=0}^{C}\frac{\rho_1^{\theta-j}\rho_2^j}{(\theta-j)!j!}\right)\sum_{i=1}^{\theta-C}\frac{\rho_1^i}{i!}.$$

Comparing the inequalities for the rhs and the lhs, we get the result.

3.7 Summary

This chapter demonstrates how over-booking type models might arise in telecom YM by considering a specific service idea. We proposed an idea for a service, and formulated a simple single-link model to investigate a capacity determination decision during its operation. Specifically, the model considered how many requests to allow in the system given a number of request already being served. We proposed the use of a simple class of policies which admit of an analytical solution, are within 1-2% of the optimal, and can be obtained *much* faster than true optimal policies. Further, these policies have the property of being insensitive to the service time distributions and correlation structure, making them highly attractive for analysis, as well as implementation. We then extended the model in several directions to indicate how a similar line of reasoning might be useful in other cases.

3.7.1 Contributions

Contributions from this chapter include the following:

1. The service idea which is straightforward enough to be no-brainer, if it can be implemented intelligently. It has simple and intuitive appeal and we suspect some variation of it is likely to appear in the market in the near future.

2. In the modeling arena, we view our contribution as the starting point for analysis of larger-scale network models involving variants of the capacity determination question. Even for the single-link case, the unexpectedly good performance of the θ^* policy and its insensitivity remain surprising, and motivate questions about deeper connections of these policies to multi-class queuing networks [Kel79] and product-form loss models [Kel91]. It further seems that optimal policies may have a simpler characterization than the one we have discovered, given the closeness of their state space to that of the θ^* policy (c.f. figure 3-8). Many generalizations of the single-link model are of interest, and we leave these for later research.

3. Finally, the connection with airline over-booking we alluded to earlier in section 3.1 can be seen explicitly by looking at an optimal linear policy as described in section 3.3. Revenue maximizing policies may allow arrivals to be admitted even when the link is full, in order to more fully utilize the system, similar to airline over-booking. As a comment, this behavior arises because we expect μ_2 to be larger than μ_1 in the model of section 3.2, therefore the departure process from the link with rate $C\mu_2$ is likely to be faster than the the uploading process with rate $i\mu_1$. The probability of a circuit becoming free before a download finishes is therefore likely to be significant, and it makes sense to allow uploads even when the link is full.

Chapter 4

Overbooking: Network-usage by a Latest Start Time - Two Possible Services

We demonstrate how over-booking type decisions could arise in telecom YM, in the context of two discount services where users indicate a *latest start time* (LST) before which service must commence, allowing the provider to schedule their requests using spare capacity (c.f. section 2.2.2 for the context). The LSTs may be explicitly specified or may arise implicitly due to the nature of the service, as explained in section 4.1. We outline the service ideas and model the problem of determining the optimal number of requests to accept to maximize revenue from available capacity, given existing requests and their LSTs.

Section 4.1 presents the services and discusses several possibilities for their operation. Section 4.2 presents a single-link model for determining an acceptance policy. Section 4.3 investigates a simple acceptance policy in detail. Several cases are considered separately, including where the service times are exponential, generally distributed and heavy-tailed random variables. Several directions for future research are also outlined. The summary and contributions of this chapter are presented in section 4.5.

4.1 The services

Imagine a service where a software agent 'Lucy' on a user's PC 'talks' to a network agent to request that a call commence anytime before a latest start time (LST) and the network agent places the call whenever capacity is available before the LST. This capitalizes on the notion that not everyone needs to use the network in real-time and many customers might be willing to trade-off on-demand service for a discount. For the network provider, the cost of providing service is negligible if capacity is available. Several practical choices must be made when designing such an offering, for instance, does 'Lucy' indicate the length of the session to the agent? Does every user demand a different rate etc.? We comment on these and other practical matters in section 4.1.1.

Another context in which a service with LSTs occurs is when the provider acts as content courier. Arriving customers request delivery of bulk content by a deadline, at a fixed rate and with non-preemptible transmissions. This gives the service provider flexibility to time the start of transmission when capacity is available. The constraints of fixed rate and non-preemptive service could be a practical requirement, arising for instance, from security concerns from the customers[1].

[1] Allowing arbitrary number of connection attempts to external computers can easily result in a breach of security.

In this context, the LSTs arise implicitly since given the amount of content and the fixed rate, non-preemption constraint, the network needs to start transmission by a LST if the deadline is to be satisfied. Again, exact parameters of the service must be designed to make it practical.

It is clear that the entire airline YM framework could be relevant in modeling different aspects of such services (c.f. section 2.1 for the perspective). For instance, decisions of interest about the services could involve: (i) *pricing*, to determine prices that maximize revenue from the offering, (ii) *seat-inventory control*, to accept the optimal mix of customers to fill available capacity when customers pay differently, and (iii) *forecasting*, to determine available capacity and demand. Finally, agents also need to use a *scheduling rule* to determine the order in which existing requests should be served.

We focus only on *overbooking*, determining the optimal number of customers to accept to maximize revenue, given available capacity and unserved requests. Because the decision for both services is the same, we talk henceforth only about the 'Lucy' service. The agent determines at request arrival time if to accept the arriving request or to refer it to an alternate system at a cost, based on the number of requests in the system, their LSTs and the number of calls in service. This cost can also be interpreted as the cost of using 'real-capacity' instead of 'available capacity'. We assume no information is required of the users in advance, such as the length of their sessions. Further, each user uses some constant bandwidth during the length of her session, which could be an actual circuit or a measure such as *effective bandwidth* of the source. Also, instead of revenue maximization, one could also consider a service-level criteria where the objective is to ensure a QoS - typically specified as the probability that an accepted request cannot be served before its LST. In this chapter, we examine a single-link model to determine such acceptance decisions.

4.1.1 Practical Issues

Several issues need to be addressed before the services can be made operational. First, one needs to decide the exact mode of the offerings, although we focus only on one possibility here. For instance, one may allow users to request time-windows with *earliest start times* in addition to latest start times. One may decide to make available different transmission rates to users. The users might also be given the option to 'cancel' their requests if they change their minds and may be asked to indicate the approximate length of usage in advance.

Providers of such services can include network operators with predictable utilization patterns and a network that allows for capacity management, such as setting time-of-day limits, etc. With a bit of modeling work, one can perhaps also apply it to networks like the Internet which do not have sophisticated capacity management mechanisms, if a measure of capacity usage can be obtained for accepted calls, such as effective bandwidths for instance. Some mechanism to ensure that only spare capacity is used by the agents needs to be constructed. The simplest schemes could involve setting limits on the amount of capacity used by discount calls, which needs to be forecast and set from usage patterns. Alternatively, one could implement a sophisticated control where each network agent monitors usage of the network by premium traffic, and schedules calls based on its own estimates of the available capacity. Also, in case demand can indeed be deflected to other networks for lack of capacity, one would need to seek arrangements with other network providers for capacity management.

Other design decisions relate to the architecture for the service management software. One envisages a few, or possibly only one, network agents located around the network to which 'Lucy' talks. Whether these agents are distributed or centralized will guide decision models. For instance, models that have perfect information on accepted calls at the various agents and their respective LSTs will naturally be different from local decision models at each agent, which may assume only probabilistic information about existing demand at other agents. All such possibilities will raise extremely interesting modeling problems.

4.2 A Single-link Model

Consider a C-server facility with identical servers. Requests arrive to the facility as a Poisson process of rate λ. Associated with each arriving request is a latest start time (LST) X from the time of arrival, before which it must start service. X are assumed to be i.i.d random variables, and we will consider versions of the model when the value of X is specified upon arrival, or available only in distributional form at the decision epoch. Once a call is started, it occupies a server for S units of time, S being i.i.d random variables not known at start of service. Service is not preemptible.

We consider an acceptance policy which admits an arriving job only if the probability of starting it before its LST is greater than a pre-specified QoS. The problem is to determine this probability, with knowledge of either the exact value of the LST X or with only distributional information about it. Section 4.2.1 discusses other information available for the decision and some modeling issues. Section 4.3 presents and discusses a simple acceptance policy motivated by the complexities outlined in section 4.2.1.

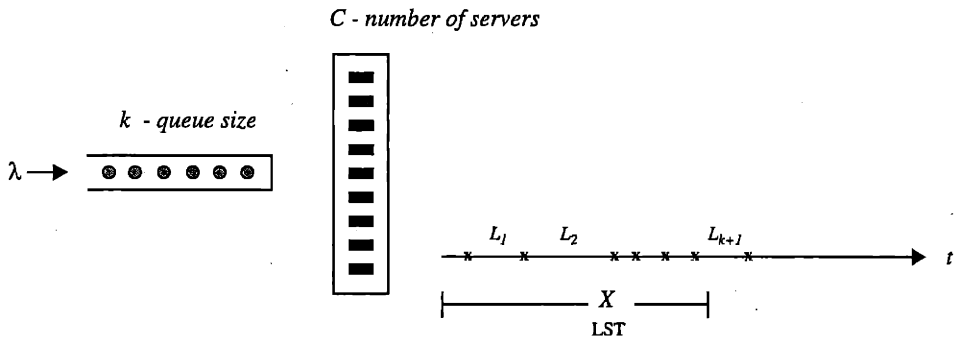


Figure 4-1: A single link model.

4.2.1 Remarks

1. The probability of starting service for an arriving job before its LST clearly depends on all the jobs in the system, their remaining service times, their LSTs *and the scheduling discipline* – which may necessitate taking into account future arriving jobs such as, for instance, under *earliest LST first scheduling.* Not only could such information requirements be too stringent to maintain, using all such information to determine exactly the probability of missing the LST is itself a hard problem. We therefore look for simpler admission control policies which abstract away the scheduler as a black box and assume only distributional information about the service times of jobs in queue and in service. Details follow in section 4.3 where the reasonability of such a policy is also addressed.

2. One needs to specify what happens to accepted calls that miss their LST. We assume that (i) the specified QoS is small enough that this is a rare occurrence and (ii) that a call *must* be serviced even if it misses its deadline, perhaps with an associated penalty.

3. The case of having knowledge of the LST X only in distributional form before making the acceptance decision models situations where users can specify the LST in probabilistic terms, or when the available rate for bulk transmissions is stochastic.

4. Our QoS based acceptance criterion makes this decision problem exactly the same as an airline overbooking problem. The objective in both our problem and the airline overbooking problem is to maximize the number of requests admitted given available capacity, realized demand and stochastic demand behavior. See section 2.2.2 for more details on the connections. However, the complexity in our problem stems from the presence of scheduling, making an optimal policy much harder to obtain.

4.2.2 Literature review

Departure processes from queues are of interest in several applications and have been studied extensively, but results usually focus on the unconditional steady-state departure times, unlike the departure times we consider, which are conditioned on an observed queue size. Therefore we mention only a few papers in this area to serve as useful leads, instead of a true literature survey.

Even for unconditional steady-state departure processes, the exact description is hard to obtain except for the simplest queues. Consequently, much literature focuses on obtaining bounds for the inter-departure time distributions. For instance, Whitt [Whi84a], [Whi88] and [Whi84b] deal with approximations for departure processes in single-server queues, and with light and heavy-traffic approximations for inter-departure time distributions. Several references related to departure processes are cited in these papers. Daley [Dal76] is an earlier work that attempts to address mathematical aspects of output processes of $G/G/s/N$ queues.

4.3 Analysis: A Simple Acceptance Policy

The complexity of incorporating all information in an acceptance decision, as outlined in section 4.2.1, motivates the search for simple admission control policies. Specifically, suppose that scheduling is non-idling FCFS and $e^{-\delta}$, the maximum possible probability of missing the LST for a call, is very small. Then, if we only consider distributional information about X (LST), distributional information about the lengths of calls yet to be started, and distributional information about the remaining service times of calls already started, we can state the admission control problem in a simple form as follows.

For any arrival at time τ that finds k jobs in queue, call $L_i, i = 1, \ldots, k+C$ the inter-departure time between the $i-1$st and ith departure, assuming departure 0 occurs at time τ. Then admit iff:

$$\boxed{\Pr\left\{ \sum_{i=1}^{k+1} L_i > X \right\} \leq e^{-\delta}.} \tag{1}$$

In such an admission control policy, the probability on the lhs is a function of k, X and the distribution of L_i's, the inter-departure times. It needs to be computed on-line at every arrival if either the value of X is specified or the L_i's are non-stationary.

If the L_i's are assumed stationary at every arrival and x, the value of X is known at each arrival, an alternate specification of the policy is $k^*(x)$, the largest queue size an arrival with deadline x must find, for (1) to be satisfied. This follows by noting that the departure times $k, k+1, \ldots$, must be stochastically ordered under a FCFS non-idling policy, with the $k+1$st departure time stochastically larger than the kth departure time. Therefore there must be a largest $k^*(x)$ for every x, which satisfies (1). It is further clear that $k^*(x)$ must be monotone increasing in x.

With stationary L_i's, a further simplification results when only the distribution of X is known at arrival. Here, the policy is characterized by a single number k^* and can be computed off-line. To see this, note that since X are i.i.d, the probability on the lhs is only a function of k, the queue size at arrival. In this case, the admission policy is characterized by the integer k^* defined below, such

that one accepts a job iff the number in the queue $k \leq k^*$.

$$k^* = \arg\max_k \Pr\left\{ \sum_{i=1}^{k+1} L_i > X \right\}$$
$$\text{s.t.} \quad \Pr\left\{ \sum_{i=1}^{k+1} L_i > X \right\} \leq e^{-\delta}.$$

In the remainder of this section, we will consider the evaluation of both the general policy (1) and the simplified k^* policy, assuming the L_i's are stationary.

4.3.1 Remarks

- Note that the L_i's represent the earliest k departures from the system and are not i.i.d, which is the real complication in determining the probability on the lhs of (1).
- The assumption that arrivals find stationary residuals is usually reasonable because of the PASTA[2] property. Stationarity of the residuals implies that the L_i's are stationary, but not i.i.d. as mentioned above.
- Under the FCFS non-idling scheduling rule, the k^* policy when X is not known is not unreasonable, counter to a possible first reaction. In effect, the only assumption it makes in addition to the scheduling rule is the stationarity of the residual service times.
- To understand the effect of information on our proposed acceptance policy, under the FCFS non-idling assumptions, consider the following different cases of the amount of information known at decision time.

  1. If x, the value of X is known at arrival and one knows *exactly* the service times of all jobs in the system and the residual times of the jobs in service, the decision is straightforward, since one can algorithmically determine the $k+1$st departure time, and compare it with the value x to make the acceptance decision.
  2. If X is only available in distributional form, but one has perfect information about all requests in the system as mentioned above, we can still use an algorithm to determine $\sum_{i=1}^{k+1} L_i$ and the admission control reduces to evaluating the CDF of X, $G(x)$ as shown below.

$$\Pr\left\{ \sum_{i=1}^{k+1} L_i > X \right\} = G\left(\sum_{i=1}^{k+1} L_i \right) \leq e^{-\delta}$$

  3. When x, the value of X is known at arrival but the L_i's are random, one needs the CDF of $\sum_{i=1}^{k+1} L_i$, the $k+1$st departure time to make the decision. This is likely to be the most complicated case since the departure-time distributions from multi-server queues are usually difficult to characterize.
  4. Finally, a k^* policy results when both X and L_i's are random. In this case, one expects a more conservative decision than all of the cases above since the randomness of the L_i's and X increases the variance of the outcome.

  These cases help outline the trade-offs in using a k^* policy vs. keeping more state information for making the decision. Since the first two are straightforward, we focus on the last two cases in the remainder of this section.

[2] *Poisson Arrivals See Time Averages.*

4.3.2 Exponential LST X and exponential service times

This base case is of interest because the distribution of the $(k+1)$st departure-time is known explicitly and because it serves as a useful benchmark for comparing approximation schemes for more general distributions. Assume that X is exponential with mean $1/\mu_d$ and the service times at each server are exponential with mean $1/\mu_s$. Then k^* is easily obtainable by solving for the largest k using the following expression, where the following is only valid for $k \geq 1$:

$$\Pr\left\{ \sum_{i=1}^{k+1} L_i > X \right\} = 1 - \left[\frac{C\mu_s}{C\mu_s + \mu_d} \right]^{k+1} \leq e^{-\delta}. \tag{2}$$

This obtains

$$k^* = \left\lfloor \frac{\ln(1 - e^{-\delta})}{\ln(C\mu_s/(C\mu_s + \mu_d))} - 1 \right\rfloor.$$

4.3.3 Generally distributed LST X and exponential service times

A simple expression for the probability of missing the deadline is unlikely to exist in this more general case, since

$$\begin{aligned} \int_{x=0}^{\infty} \Pr\left\{ \sum_{i=1}^{k+1} L_i > X \mid X = x \right\} dG(x) &= \int_{x=0}^{\infty} \int_{y=x}^{\infty} \frac{(C\mu_s)^{k+1} y^k e^{-C\mu_s y}}{k!} dy \; dG(x) \\ &= \sum_{i=0}^{k} \left[\frac{1}{(k-i)!} \int_{x=0}^{\infty} e^{-C\mu_s x} \, (C\mu_s x)^{k-i} \, dG(x) \right] \leq e^{-\delta}. \end{aligned} \tag{3}$$

Still, it is always possible to numerically evaluate the expression above for any value of k and therefore to determine k^*, the maximum k such that the constraint $\leq e^{-\delta}$ is satisfied. This computation is not too difficult and can be organized recursively, with each $k = 2, \ldots, \infty$, requiring the evaluation of only one extra integral, starting from $k = 1$ for which two integrals are needed, as shown below:

$$\begin{aligned} k = 1: \quad \Pr\Big\{ \sum_{i=1}^{k+1} L_i > X \Big\} &= \int_{x=0}^{\infty} e^{-C\mu_s x} \, C\mu_s x \, dG(x) + \int_{x=0}^{\infty} e^{-C\mu_s x} \, dG(x), \\ k \geq 2: \quad \Pr\Big\{ \sum_{i=1}^{k+1} L_i > X \Big\} &= \Pr\Big\{ \sum_{i=1}^{k} L_i > X \Big\} + \frac{1}{k!} \int_{x=0}^{\infty} e^{-C\mu_s x} \, (C\mu_s x)^k \, dG(x). \end{aligned}$$

This is also a good point to pause and experiment with approximations for the probability on the lhs of (1), since when the service times are not exponential, approximations might be the only realistic method for computing an admission decision. We consider the Chernoff-approximation below, to bound the rhs of (3) above and determine an approximate k^*. The expression for the Chernoff-bound involves a single integral as shown below. Comments on the quality of the approximation follow in 4.3.3. Connections and insights from these experiments will appear in later sections.

$$\begin{aligned}\int_{x=0}^{\infty} \Pr\left\{ \sum_{i=1}^{k+1} L_i > X \mid X = x \right\} dG(x) &\leq \int_{x=0}^{\frac{k+1}{C\mu_s}} dG(x) + \int_{x=\frac{k+1}{C\mu_s}}^{\infty} e^{-(r^* x - \ln \phi(r^*))} dG(x) \\ &= G\left(\frac{k+1}{C\mu_s}\right) + \left(\frac{e}{k+1}\right)^{k+1} \int_{\frac{k+1}{C\mu_s}}^{\infty} e^{-C\mu_s x} (C\mu_s x)^{k+1} dG(x) \\ &\leq e^{-\delta}. \end{aligned} \tag{4}$$

Getting k^* from the above expression again involves solving the integral numerically for $k = 1, 2, \ldots, \infty$, until it exceeds $e^{-\delta}$. Some remarks are noted below.

Remarks

1. Although obtaining the exact k^* from (3) and the approximate k^* from (4) is the same amount of work, the Chernoff-bound will often yield a quicker estimate of the probability of missing a deadline than the computation of the actual probability, involving a single integral compared to the summation of $k+1$ integrals for the actual probability.

2. To evaluate the quality of the Chernoff-bound (4), we investigate the case when both the deadlines and the service times are exponential. In this case, the exact probability expression is given by (2), and (4) resolves into the following expression for the Chernoff-bound:

$$1 - e^{-(k+1)\frac{\mu_d}{C\mu_s}} \left[1 - \frac{\mu_d}{C\mu_s + \mu_d} \sum_{i=0}^{k+1} \frac{(k+1)!}{(k+1-i)!(k+1)^i} \left(\frac{C\mu_s}{C\mu_s + \mu_d} \right)^i \right].$$

   Figure 4-3 plots the actual probability and the Chernoff-bound, and figure 4-2 plots the resulting k^*'s. A priori, the Chernoff-bound is expected to be good in the range where x, the value of X exceeds the mean $(k+1)/C\mu_s$ significantly. This follows from the fact that for exponential service times, the $k+1$st departure is the sum of $k+1$ i.i.d random variables, and for sums of a large number of i.i.d random variables, the Chernoff-bound is known to be asymptotically exact, in the sense of the Cramer-Chernoff theorem (c.f. [RMVe96], page 382 for instance). On the other hand, since (4) involves integrating over the entire range of X, the bound is expected to perform worse than it might otherwise for large x.

   This is reflected in figure 4-3. The Chernoff-bound tracks the actual probability well for all k, *but with almost a fixed difference*, as long as the ratio μ_s/μ_d does not get too small. Part of the difference is clearly seen in the factor $G\big((k+1)/(C\mu_s)\big)$ in (4), since the Chernoff bound obtains only for $X > (k+1)/(C\mu_s)$. The influence on the approximate k^* is also significant, which might be off by as much as $5-10$, as shown in figure 4-2.

3. To further understand the effect of the factor $G\big((k+1)/(C\mu_s)\big)$ in approximation (4), which arises from the randomness of the LST X, we experiment with deterministic X. When X is deterministic with value x, the exact probability is obtained as

$$\Pr\Big\{ \sum_{i=1}^{k+1} L_i > x \Big\} = e^{-C\mu_s x} \sum_{i=0}^{k} \frac{(C\mu_s x)^{k-i}}{(k-i)!},$$

   while the Chernoff bound gives

$$\Pr\Big\{ \sum_{i=1}^{k+1} L_i > x \Big\} \leq \begin{cases} 1, & x \leq \frac{k+1}{C\mu} \\ e^{-C\mu_s x} \left(\frac{eC\mu x}{k+1} \right)^{k+1}, & x > \frac{k+1}{C\mu}. \end{cases}$$

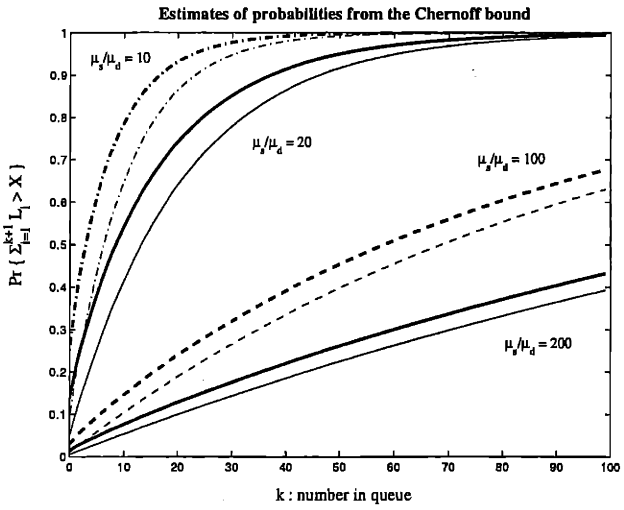


Figure 4-2: The probability estimates from the Chernoff-bound for missing the deadline, plotted against the exact value when service times and deadlines are both exponential. The Chernoff-bound curves are above the exact probability curves.

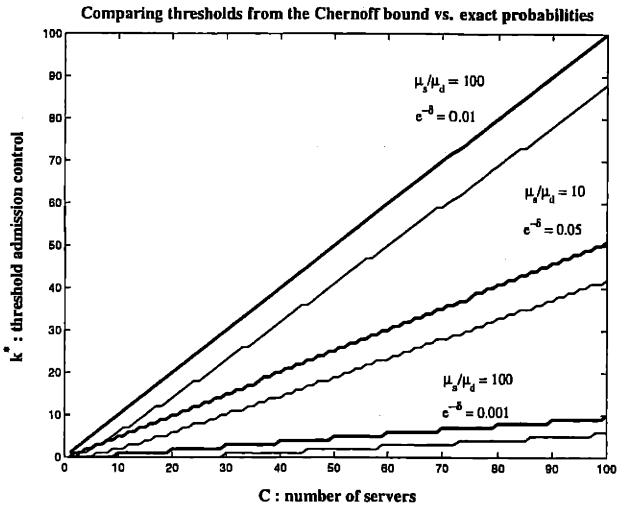


Figure 4-3: Comparing k^* obtained from the Chernoff-bound against that obtained from the exact probability when service times and deadlines are both exponential. The exact k^* are shown with thicker lines.

Figure 4-4 shows the Chernoff-bound obtained and the actual complementary CDFs, giving a sense of the range of x for which the bound might be a reasonable estimate. It is clear that the bound is a good estimate only for large values of x, specially when $x \geq 3(k+1)/(C\mu_s)$. For lower values, it could seriously over-estimate the probability, despite being exponential in x. However, recall our condition that the probability of missing the deadline $e^{-\delta}$ be very small. From figure 4-4, it is clear that when $e^{-\delta} \leq 0.01$ for instance, the difference in the largest deadline admitted using the Chernoff-bound vs. the exact probability is less than 1, and hence the Chernoff-bound performs very well in the range we are interested in. We leave further discussion of the bound for later sections.

4.3.4 Deterministic X and generally distributed service times

The simple integral equation (3) can be used only when service times are exponential. When service times are more general, an easy description of the departure process is not available, motivating approximations for the probability of missing a deadline. We propose two approximations in this section for the case when the value of X is known at the time of decision. Note that since in this case, one actually needs to obtain the probability at every arrival, an additional requirement is that the approximations be easily computable.

The approximations rely on the crucial assumption that at every arrival, the residual-life (service-time) random variables at each server are *independent identically distributed* with the stationary residual-life distribution associated with the service-time random variable. This assumption is motivated by the PASTA[3] property of queuing systems (see, for instance, [Wol89]). We then distinguish two cases with separate approximations for each.

[3] *Poisson Arrivals See Time Averages.*

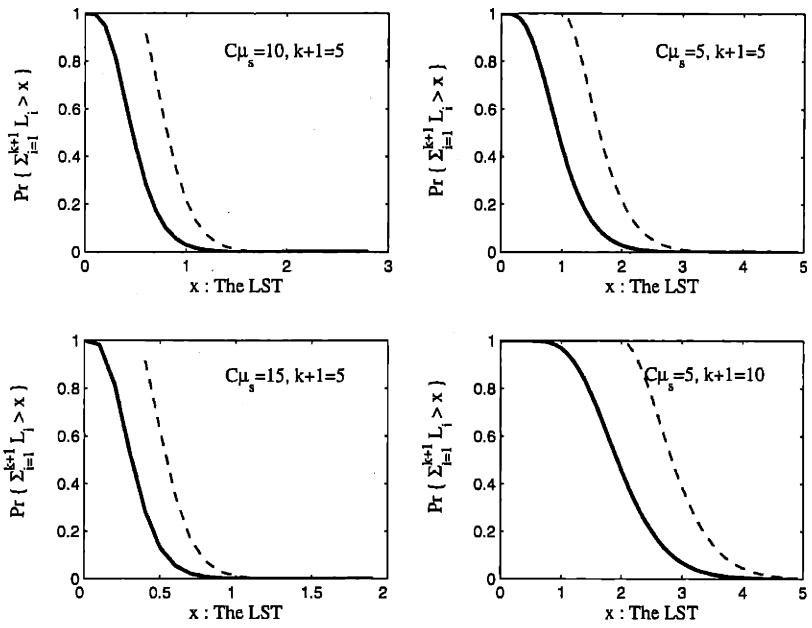


Figure 4-4: Chernoff-bound obtained for deterministic X and exponential service times, for several values of $C\mu_s$ and k. The actual complementary CDFs are shown in thick lines.

When the mean of the residuals exists

Calling R the stationary residual life at a server, and S_i, $i = 1, \ldots, k$, the service time random variables for the k people in the queue at arrival, we propose the following approximation for the probability.

$$\Pr\left\{ \sum_{i=1}^{k+1} L_i > x \right\} \approx \Pr\left\{ \sum_{i=1}^{k} S_i + R > Cx \right\} \leq e^{-\delta}. \tag{4}$$

This speeded single-server approximation can be shown to hold rigorously (see section 4.4 for the proof) under the extra assumption that in addition to the i.i.d residuals seen at arrival, the residual-lives for the jobs in service *at the time this arrival will enter service* are also i.i.d stationary residuals. Such an assumption has been made before in literature [NR78] for deriving approximations for the mean waiting time in an $M/G/c$ queue. Comments on approximation (4) follow.

1. Our approximation is the exact probability if either $C = 1$ or if the service times are exponential.

2. The additional assumption of finding stationary residuals *at start of service* is expected to be reasonable as k, the number in the queue gets larger. We argue this by noting that if a request finds stationary independent residuals at arrival, then by the time she starts service, C delayed renewal processes have been in process for a while. Certainly, as $k \to \infty$, these processes should become independent of each other by the $k+1$st departure time, leading to our assumption. This argument would not hold when k is small, and specially when $k \leq C$.

3. We test the quality of approximation (4) for service times with Gamma distributions. Results are displayed in figure 4-5. The speeded-server approximation was obtained via direct convolution. The actual complementary CDF was obtained via simulation, assuming stationary residuals at arrival times. It is also prudent to keep in mind the magnitude of numerical errors that might exist in the results. In tests including exponential service times and Gamma service

times when $C = 1$, the worst-case percentage errors between the simulation and the convolved probabilities were between 0.1-3%, with most below 0.7%.

The approximation seems to do remarkably well at all values of x. Figure 4-5 also confirms our argument above that the quality of the approximation should get better as k increases. In fact, one sees again that for large values of k, if $e^{-\delta} \leq 0.01$, the difference in the largest deadline admitted using the approximation vs. the exact probability is quite small.

One curious aspect of the curves is that the speeded-server over-estimates the actual probability when x is approximately less than the mean of the $k+1$st departure time, while it under-estimates the probability for larger x. A possible explanation is in terms of the residual R in (4). As x gets larger, we expect the residuals in the actual probability to have less effect on the $k+1$st departure time, a fact not reflected in the approximation, where a single residual always stays. Given that the residual service times have an increasing failure rate (IFR) for gamma distributions, it would explain the discrepancy in the curves.

Finally, we hope that the quality of the approximations extends to mixtures of Gamma distributions. Since mixtures of Gamma distributions have the property of being dense in the family of distributions, a well known result (see [Kel79] for example), this could make the approximation significantly more powerful. A possibly useful direction for future research.

4. There remains the question of evaluating (4). A possibility is to actually convolve the S_i's and R to determine the actual probability. Although not difficult numerically, this is too slow for real-time control, and therefore one might consider quickly computable bounds for (4). We comment on some possible directions below and their associated complexity.

   - When the moment-generating function of the service-time random variable is available, one can attempt to obtain the Chernoff-bound for (4) (the derivation is straightforward, but is listed in section 4.4.2 for easy reference):

$$\begin{aligned} \Pr\left\{ \sum_{i=1}^{k} S_i + R_1 > Cx \right\} &\leq \frac{1}{r^* \mathbb{E}(S)} \phi_S^k(r^*) [\phi_S(r^*) - 1] e^{-r^* Cx} \\ &= \frac{1}{\mathbb{E}(S)} e^{-r^* Cx \, + \, k \ln \phi_S(r^*) \, + \, \ln[\phi_S(r^*) - 1] \, - \, \ln r^*}. \end{aligned} \tag{5}$$

     Where $\phi_S(r)$ is the mgf of the service time random variable S and r^* satisfies:

$$k \frac{\phi_S'(r)}{\phi_S(r)} + \frac{\phi_S'(r)}{\phi_S(r) - 1} - \frac{1}{r} = Cx.$$

     In general, we can go no further in simplifying (5) until we have an explicit form of the mgf $\phi_S(r)$ to manipulate. Unfortunately, the expression above does not usually lend itself to an easy solution for r^*.

   - Possibly useful techniques for bounding (4) might be similar to those used in obtaining bounds for the steady state waiting-time in $G/G/1$ or $M/G/1$ queues, such as the Kingman-bound (c.f. [Gal96] for instance) for example. Useful techniques might also be learnt from bounds for $G/G/c$ queues reported in literature (see, for instance, [Whi84a], [Whi88] and [Whi84b]).

When the mean of the residuals does not exist

For certain distributions of interest, such as Pareto for example, the stationary residual distribution does not have a mean. In this case, the speeded-server approximation is not expected to be good. When $k+1 \leq C$, one alternative is to bound the probability of $(k+1)$st departure from the system by time x, by the probability of the departure of the *first* $k+1$ *of the* C *customers in service* seen by an

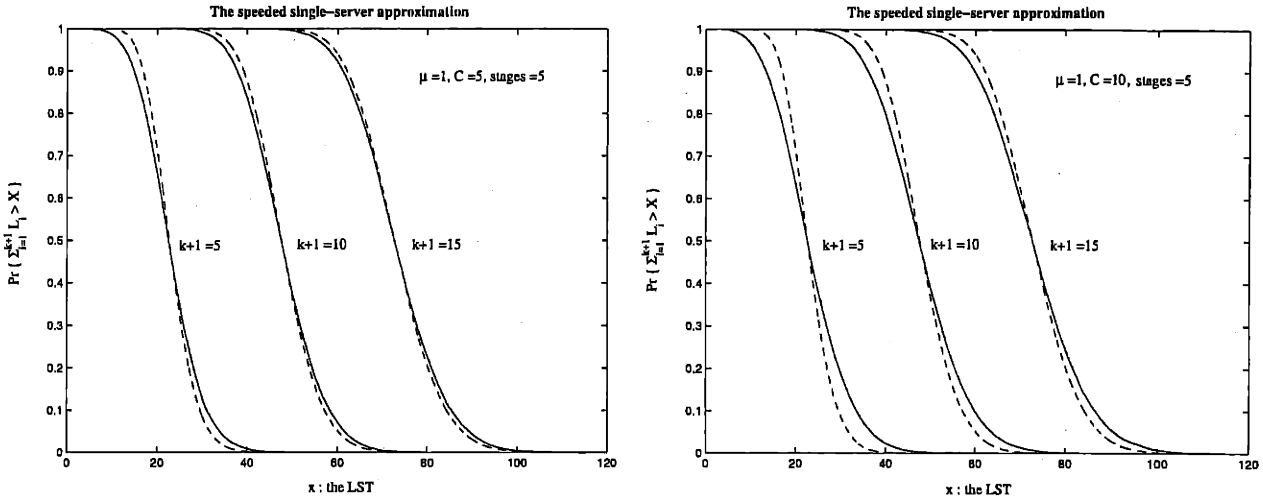


Figure 4-5: The complementary CDF for the speeded server approximation, plotted against the exact complementary CDF obtained via simulation, for service times having Gamma distributions. The Gamma parameter is called *stages* above and the speeded-server curves are dashed.

arrival. Under the assumption of i.i.d residuals, this probability is obtained from the order-statistics as follows, where $H(x)$ is the CDF of the residual life.

$$\Pr\left\{ \sum_{i=1}^{k+1} L_i > x \right\} \leq$$

$$\Pr\left\{ (k+1)st \text{ smallest of } (R_1, \ldots, R_C) > x \right\} = \sum_{j=0}^{k} \binom{C}{j} H^j(x) \, [\, 1 - H(x) \,]^{C-j}.$$

The order-statistics approximation is too conservative when the mean of the residuals exists and is less than or equal to the the mean of the service times. For exponential service times, this is easily seen as the inter-departure times for the order-statistics approximation are exponential with rates $C\mu_s, (C-1)\mu_s, \ldots$, whereas the actual inter-departure times from the system are exponential with rate $C\mu_s$. Order-statistics therefore under-estimate the probability of exceeding a given x. One might extrapolate from this to guess that if the residuals have an increasing failure rate (IFR), order-statistics will significantly under-estimate the actual probability. For decreasing failure rate (DFR) residuals, as in the case of Pareto distribution, one expects the order-statistics to perform very much better.

4.3.5 Directions for future research

Several interesting research directions are revealed by the analysis of our simple acceptance policy. We note some of them below.

1. For our proposed policy, we have demonstrated the quality of the speeded single-server approximation for general service times, but the computation of this approximation itself remains an open issue, convolution not being a practical alternative. Quickly computable bounds might be more easily obtainable for the approximation than the actual probability, and could be very useful.

2. A generalization that would make the model significantly richer is to allow multiple customer classes with class-dependent service time distributions. In this case, the acceptance decision

would need to incorporate information about the position and class of the customers in the queue and in service.

3. The most interesting generalization of our proposed acceptance policy would be to a network, where the departure process is significantly more complicated. However, limit-type results might exist here. The case of fixed routing networks with exponential service times might be the easiest, allowing extension of the ideas of sections 4.3.3 and 4.3.4.

4. Finally, the restriction of an arbitrary policy is unwarranted, since the actual scheduling rule would attempt to utilize capacity as efficiently as possible. Because of the difficulty of incorporating arbitrary scheduling rules into an admission decision, it might make sense to investigate via simulation how the probability of missing the deadline changes if the scheduler follows a rule different than FCFS – *earliest LST first* for example, even though the admission controller assumes FCFS.

4.4 Derivations

4.4.1 The speeded single-server approximation

We derive the approximation of section 4.3.4 under the assumptions:

- The residual lives of the jobs in service found by an arrival are i.i.d random variables having the stationary residual-life distribution of the service time.
- The residual lives of the jobs in service at the time this newly arriving job will enter service are also i.i.d random variables having the stationary residual-life distribution of the service time.

Then when an arrival finds k jobs in the queue and C in service, let

V_a: The total unfinished work seen at the instant of arrival.
R_j: The residual work at server j found at arrival.
E_j: The residual work at server j at the instant this arrival enters service.
S_i: The service time of the ith customer in the queue.
$\sum_{i=1}^{k+1} L_i$: The departure time of the $k+1$st departure from the system relative to the time of arrival - note this is also the instant that the new arrival will enter service.

By definition $V_a = \sum_{j=1}^{C} R_j + \sum_{i=1}^{k} S_i$. Also, under FCFS non-idling scheduling policies, it is true that

$$V_a = C\sum_{i=1}^{k+1} L_i + \sum_{j=1}^{C-1} E_j.$$

Using the two identities for V_a, it is immediate that

$$\sum_{i=1}^{k+1} L_i = \frac{C-1}{C}\sum_{j=1}^{C-1}\frac{R_j - E_j}{C-1} + \frac{R_C}{C} + \sum_{i=1}^{k}\frac{S_i}{C}.$$

Now if we believe the two assumptions, then E_j and R_j have the same distribution for all j and are independent of each other. Therefore, if they have a mean, $\sum_{j=1}^{C-1}(R_j - E_j)/(C-1)$ should go

to zero relatively quickly as C increases, by the strong law of large numbers, and the approximation below follows.

$$\Pr\left\{ \sum_{i=1}^{k+1} L_i > x \right\} \approx \Pr\left\{ \sum_{i=1}^{k} S_i + R_C > Cx \right\}.$$

4.4.2 Chernoff-bound for the speeded single-server approximation

The Chernoff-bound for expression (4) of section 4.3.4 follows from the independence of the S_i's and R_C. For the service-time random variable S with distribution function $G(x)$ and moment-generating function $\phi_S(s)$, the stationary excess-life R has distribution, density and moment-generating functions as shown below giving expression (5) of section 4.3.4 for the Chernoff-bound.

$$F_R(r) = \frac{1}{\mathbb{E}(S)} \int_0^r [1 - G(u)]du$$

$$f_R(r) = \frac{1}{\mathbb{E}(S)}[1 - G(r)]$$

$$\phi_R(s) = \frac{\mathbb{E}(e^{sS} - 1)}{s\mathbb{E}(S)} = \frac{1}{\mathbb{E}(S)} \frac{\phi_S(s) - 1}{s}$$

4.5 Summary

This chapter proposed two services for telecom YM that ask users to specify *latest start times* (LSTs) by which start-of-service must occur. To demonstrate overbooking-type decisions in the YM framework of chapter 2, we explored when an arriving request could be served before its LST given the number of requests and information about the system.

Using a single-link model, we outlined the complexity of the underlying decision and analyzed a simple acceptance policy which bases its decision on the probability that the arriving request can be served before its deadline. Several cases were analyzed, including when the service times are exponential or generally distributed. Several approximations for the probability of missing a deadline were proposed and numerically investigated. We outlined several directions for future research.

4.5.1 Contributions

This chapter contributes the following.

1. Our exercise of articulating services for telecom YM is useful in that it demonstrates how to capitalize on the YM intuition. We hope that it guides and interests researchers and telecom operators in formulating innovative ways to manage capacity using new services.

2. Our model of section 4.3 finds applications in several domains, not only telecom YM. For instance, the admission decision in queues with deadlines spans telecommunication traffic, real-time systems, processor scheduling and services management, to name a few.

   Further, the line of analysis we pursue focuses attention on the departure processes from queues. This is an area of interest, but most results in literature relate to the unconditional steady-state distribution of the inter-departure intervals and the associated results are asymptotic (c.f. the literature review section 4.2.2 for details). We focus instead on the departure process conditioned on the queue size seen at arrival, an approach that has not been dealt with widely.

Our model motivates studies of departure processes from networks as well, conditioned on partial network information.

The derivation of the speeded single-server approximation in section 4.4 is also, to our knowledge, new, in spite of the fact that the approximation itself finds use often in literature (c.f. [Wol89] for instance). It is surprisingly simple and seems to do quite well numerically, as demonstrated in section 4.3.4.

3. Finally, the overbooking connection with airlines YM is transparent, as the acceptance decision we model is directly motivated by the overbooking decision in airlines. Airline over-booking decisions also attempt to obtain the probability of denial of service when admitting an arriving customer [SLD92], but do not have the added complexity of a scheduling rule to contend with, as in our case. Our modeling exercise also reveals how network information might be leveraged to squeeze the maximum out of capacity.

Chapter 5

Seat-Inventory Control: The Digital FedEx Service

This chapter demonstrates how seat-inventory control type models could arise in telecom YM. The context used is that of a digital courier service which utilizes spare network capacity to deliver bulk content (c.f. section 2.2.3 for the context). The service offering is similar to physical courier services such as FedEx^{TM} or DHL^{TM}. We present and discuss the service first and then a model for one associated decision problem, determining content sizes to accept given available capacity. We analyze a canonical model to determine the optimal policy, and then extend the results to several interesting cases.

The chapter organization is as follows. Section 5.1 discusses the details of the service offering. Section 5.2 presents the formulation and analysis of the canonical model, with the actual proofs relegated to section 5.5. Section 5.3 extends the insights from this model to several more interesting cases. Section 5.4 generalizes the structure of the optimal policy to fixed-routing networks and multiple deadlines. The summary and contributions from this modeling work are presented in section 5.6.

5.1 The service

"The digital FedEx service" refers to an offering where customers request delivery of bulk content between locations by a fixed set of deadlines, such as 5pm, 12 am, etc. Imagine a web-based interface, where users enter the destination and source address, the file name and choose from a set of fixed deadlines by which delivery must be completed to the destination. Motivation for customers to use this service could be: (i) deep discount over comparable real-time transmissions, (ii) removing the need to rent expensive and under-utilized network capacity for bulk transmissions, or (iii) offloading non real-time traffic from corporate networks to reduce capacity investments. For providers, the motivation is obvious; generating revenue from spare capacity whenever possible, as long as existing traffic is not unduly impacted. To ensure minimal impact on existing traffic as well as guarantee a service level such that content deliveries occur before their deadlines, one needs to limit the amount of accepted content to available capacity. In practice, however, rejecting customers might not be a desirable option. A better option might be to offload excess content to other providers, with whom arrangements must be sought. Of the many interesting problems raised by the operation of such a service, we model only an airline seat-inventory control problem, namely deciding which customers to retain to maximize revenue from the system. Other problems which we do not consider and several practical issues are discussed in section 5.1.1.

We model the decision of an agent at the web-server that determines whether to accept an

arriving content request with a given size and deadline for a particular destination, or to off-load it to an alternate provider at some cost, using revenue maximization as the criterion. The need for a sophisticated decision model arises because of the uncertainty in capacity availability and in the sizes of future arriving jobs, for which both the times and sizes are unknown. Agents naturally need to have some forecast of the available capacity between the locations, and the demand, to make the decision, and will obviously have access to the untransmitted content at various locations along with their respective deadlines. We assume that an agent transmits continuously at the available rate as long as there is remaining content in the system, policing itself to not flood the network. We comment on how this can be achieved and other factors in section 5.1.1.

5.1.1 Practical Issues

"The digital FedEx service" is classic YM use of spare capacity to generate positive revenue from the network. The constraint of using only residual capacity means the cost of such transmissions is almost zero[1], allowing revenue maximization. Examples of the possible market for such a service are mentioned in section 2.2.3. In this section, we mention the several interesting problems that could arise in developing an operating infrastructure for this service, which we do not address in this thesis.

The architecture for service management could consist of several web-sites located strategically on the network so most bulk shippers can find a close site easily. The placement of these sites will be an issue of interest. Whether the sites manage their decisions centrally or base it only on local information will need to be decided before any decision models can be formulated. We assume perfect information exchange is possible between agents, i.e. an agent can quickly query the amount of untransmitted content at other sites.

Also, questions about the capacity management mechanism by such a service will need addressing. Two possibilities, for instance, are: (i) each agent bounds its maximum transmission rate by a time-of-day available rate downloaded from a central database each day, and manages short time-scale congestion within the bounded rate in a TCP-style[2], which is tantamount to setting aside capacity, or (ii) no capacity is specifically set aside for the service, but the transmission protocol used by the agents is an extremely non-aggressive form of TCP protocol, which backs off transmission as soon as it detects congestion in the network and ramps up its rates very slowly as congestion eases. In either case, some notion of available capacity will need to be forecasted and summarized for the agents to make their decisions at content arrival times. Limits on the total amount of content that can be accepted for each deadline to obey service levels will likely have to come from a capacity model, much like the overbooking-type decision models for airlines. The output of such a model might then be used to decide which classes of content to admit.

The nature of transmissions, direct or intermediate, will need decision. In practice, the network provider might want to use perhaps push-and-store policies to move content to intermediate web-sites closer to their destination instead of direct transmissions to congested locations. This and whether agents can download content immediately from a source or at any time of their choice will dictate the storage requirements for intermediate housing of content. In this chapter, we assume direct transmissions. Further, we assume for simplicity that the agents can transmit content whenever they want, to whichever destination they choose, without constraint. In practice, there might be limitations on keeping an open connection to a destination address, for security reasons for instance.

[1] Easily implementable in practice using one of several forms of rate policing.

[2] TCP stands for *Transmission Control Protocol* and is the predominant protocol used for transmission over the Internet. Its congestion control algorithm is *exponential decrease, linear increase* in case of congestion.

5.1.2 Outline of the model and results in this chapter

The canonical model of this paper is a single-link with fixed available transmission rate, a single class of arriving customers all paying the same regardless of job-size and a single common deadline to completion of transmission for all requests. The optimal admission control policy for this model is shown to be threshold-type with the thresholds depending on the amount of untransmitted content in the system. Several structural results are obtained when the file-size distributions are general. Computational experiments are carried out and the structure of the thresholds is explored in more detail for *concave* file-size distributions. Further, we propose and compare the performance of a heuristic admission policy to the optimal and *greedy* policies.

Insights from the above model easily extend in several directions. In particular, we show the existence of threshold-type optimal policies for the cases listed below.

- For the single-link, single customer class case:
  - When the revenue and cost per request is time varying and the file size distributions are non-stationary.
  - When capacity is a deterministic function of time, or when capacity is stochastic with the distribution of the remaining cumulative transport capacity available.
- When multiple classes of customers arrive, arrivals pay a class-dependent revenue independent of job-size. Here multiple nested thresholds are shown to exist.
- Networks with fixed routing and deterministic available capacity.
- Multiple deadlines on a single-link with the "Earliest Deadline First" scheduling discipline. Here a connection is drawn to fixed routing networks.

5.2 A canonical model

We present a dynamic programming formulation for the following model. A single link has fixed available rate R in some appropriate units (say Mb/s). Requests arise at one end of the link for delivery of content to the other end on or before a fixed deadline D (in some absolute time-units) common to all requests. The arrival process is Poisson with rate λ and the sizes of jobs Y are random variables with known distributions, independent of the arrival process and the amount of work in the system. Each arriving job pays a pre-agreed revenue the structure of which is described later.

The decisions of interest are :

- At the time of arrival, if to accept content of *known* size for delivery by D, or ship it to an alternate system at a cost, given the amount of work already in the system, with the objective of maximizing revenue from the remaining transport capacity.
- How to schedule transmissions among the accepted jobs to maximize the number of jobs that complete transmission by the deadline, where it is natural to only consider non-idling pre-emptible policies for scheduling since the service provider will always have control of the content after it has been accepted.

We will restrict our attention to admitting content only when it is feasible to complete its transmission before the deadline under a non-idling scheduling policy. An equivalent statement is that content is to be admitted only if it does not exceed the total remaining transport capacity of the link, till the deadline. Figure 5-1 shows a sample path of work evolution for the model under some given acceptance policy.

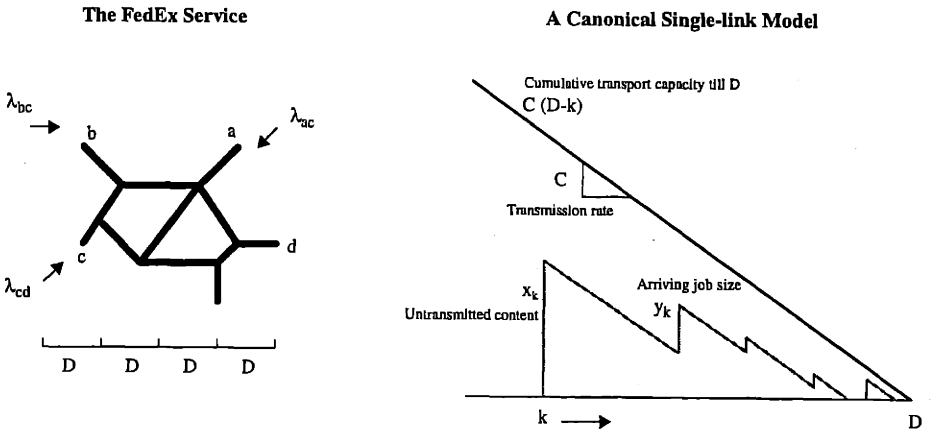


Figure 5-1: The FedEx service and a canonical model.

5.2.1 A Dynamic Programming formulation

The following is a fairly general dynamic programming formulation for determining the admission decision. For purposes of initial analysis, much of this problem will be restricted to the case of stationary job sizes and revenue.

Let $k = 0, \ldots, D$, index time in discrete units with length of each interval δ. Let R be the available transmission rate and D be the common deadline for completion of delivery at the other end. Also, call:

x_k: The amount of un-transmitted content in the system at beginning of the kth interval.

y_k: The amount of arriving work waiting to be admitted/rejected at the beginning of period k.

$u_k(x_k, y_k) \in \{0, 1\}$: Reject/accept decision for the content y_k, defining $u_k(x_k, y_k) = 0$ whenever $y_k = 0$ or $x_k + y_k > D - k$. We will use only u_k for brevity.

w_k: Random size of content arriving during interval $[k, k+1)$ which will await decision in period $k+1$. The distribution of w_k is assumed to be known for all k and work is assumed to arrive in quanta of size $R\delta, 2R\delta, \ldots, \infty$.

$r_k(y_k)$: Revenue for content of size y_k in period k, defining $r_k(0) = 0$ for all k.

$c_k(y_k)$: Cost of off-loading content of size y_k in period k to an alternate system, $c_k(0) = 0$ for all k.

Assume WLOG that $R = 1$, then the state of the system (x_k, y_k), at any period $k = 0, \ldots, D-1$ evolves as follows:

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} \max(0, x_k + u_k y_k - 1) \\ w_k \end{bmatrix}.$$

Note that the state space is integral and when $0 \leq x_0 \leq D$, the definition of the decisions $u_k(x_k, y_k)$ ensures that the state space is the set of all integer-valued vectors in the non-negative orthant $x_k, y_k \geq 0$ satisfying $x_k + y_k \leq D - k$ at every k.

The DP recursion for the admission control problem can now be written as follows (here $J_k(x_k, y_k)$ is the expectation of the optimal "additional" revenue at stage k):

$$J_D\begin{bmatrix}x_D\\ y_D\end{bmatrix} = 0,$$

$$J_k\begin{bmatrix}x_k\\ y_k\end{bmatrix} = \max_{u_k\in\{0,1\}} u_k r_k(y_k) - (1-u_k)\, c_k(y_k) + \mathop{\mathbb{E}}_{w_k} J_{k+1}\begin{bmatrix}\max(0, x_k + u_k y_k - 1)\\ w_k\end{bmatrix},$$
$$k = 0, \ldots, D-1.$$

We can write the above DP in an alternate form by defining the optimal "expected additional revenue" given x_k units of work in the system as:

$$\phi_k(x_k) = \mathop{\mathbb{E}}_{w_k} J_k\begin{bmatrix}x_k\\ w_k\end{bmatrix}, \qquad k = 0, \ldots, D-1.$$

For stationary file sizes with Y representing the file-size random variable at each stage, and fixed revenue and cost per job, i.e. $r_k(y_k) = \bar{r}$ and $c_k(y_k) = c$, the DP recursion in terms of $\phi_k(x_k)$ is noted below. $r = \bar{r} + c$ and $p_k(x_k) = \Pr(1 \le Y \le D-k-x_k)$ for $0 \le x_k \le D-k-1$, the probability that an arrival occurs and the arriving file size is less than or equal to the residual capacity. We define $p_k(D-k) = 0$ for notational consistency.

$$\begin{aligned}
&\phi_D(x_D) \equiv 0, \quad \text{and for } k = 0, \ldots, D-1,\\
&\phi_k(x_k) = -c\Pr\{Y > 0\} +\\
&\qquad \big(1 - p_k(x_k)\big)\ \phi_{k+1}\big(\max(0, x_k - 1)\big) +\\
&\qquad p_k(x_k) \mathop{\mathbb{E}}_{Y|1\le Y\le D-k-x_k} \max\big\{\ r + \phi_{k+1}\big(\max(0, x_k + Y - 1)\big),\\
&\qquad\qquad \phi_{k+1}\big(\max(0, x_k - 1)\big)\ \big\}.
\end{aligned} \tag{5.1}$$

In the following sections, we note remarks on this model and present results on the optimal admission control policy. We also note, in case it is helpful for the reader, that $\Pr\{Y > 0\} = \lambda\delta$ and that conditioned on $Y \ge 1$, $\phi_{k+1}\big(\max(0, x_k + Y - 1)\big)$ is just another way of writing $\phi_{k+1}(x_k + Y - 1)$.

5.2.2 Remarks

1. Note that the model implies that inter-arrival times are memoryless and the arriving file sizes are i.i.d random variables.

2. Since the state space is integral, the functions $\phi_k(x_k)$ are only defined for non-negative integer x_k, with $\phi_k(D-k) = -c\lambda(D-k)$ representing the cost of a full system at stage k, since all arrivals in subsequent stages must be blocked.

3. A trivial threshold for admission is $D-k-x_k$, the feasibility condition. We will show, however, that non-trivial thresholds for admission exist for almost all states and stages.

4. A fixed revenue/job is not as artificial as it may appear at first sight, for two reasons: (i) it is easily relaxed once initial results have been obtained, and (ii) the revenue structure for a FedEx-type service will probably be similar to postal service, with fixed charge for a range of content-size. In fact, if revenue/job is a function of size, it will have to be significantly sub-linear, since proportional revenue cannot be feasible owing to the large variability of content sizes. Sub-linear functions are well-approximated with step-wise functions, or fixed revenues for ranges of content sizes. For this reason, when considering multiple customer classes, we will model each class paying a constant class-dependent revenue.

5. Finally, we note that there is no real need to formulate the problem with a discrete state space. In understanding the behavior of the model, sometimes discrete and sometimes continuous versions of the model lend themselves to easier analysis. For instance, results for *general* file-size distributions are obtained for discrete state space, with computational experiments, which reveal the behavior for the discrete system. In contrast, results in section 5.2.4 for *concave* file-size distributions are obtained for the version in which the state space and arriving work are continuous variables.

5.2.3 General results

This section lists some results obtained for the above model, without any assumptions on the distribution of the file-sizes, except that they take values $0, \ldots, \infty$ with positive probabilities. Proofs of the propositions, called *remarks* below, are provided in section 5.5.

We present first some properties of the optimal "expected additional revenue" functions $\phi_k(x_k)$, and then properties of the optimal admission control policy.

Remark 5.2.1 (Monotonicity in the remaining work). For every k, the optimal "expected additional revenue" function $\phi_k(x_k), x_k = 0, \ldots, D-k$, is monotone non-increasing in x_k. This fact is intuitive as $\phi_k(x_k)$ represents the optimal "additional" expected revenue which can only be lower if one has more unfinished work in the system. Two proofs are given in section 5.5, one by a simple argument and another by induction.

Remark 5.2.2 (Monotonicity in time). For every x, $\phi_k(x) \geq \phi_{k+1}(x)$. Again, intuitive as one can only get more revenue if one has more stages to go and the same amount of work in system.

Remark 5.2.3 (Incremental revenue in both work and time bounded by r). For any x, k, (i) $\phi_k(x) \leq r + \phi_k(x+1)$ and (ii) $\phi_k(x) \leq r + \phi_{k+1}(x)$. This formalizes the intuition that one cannot get incremental revenue greater than r with one extra unit of capacity, in either time or space. The proof requires induction.

A graph of the optimal expected "additional" revenue function for a Pareto file-size distribution is shown in figure 5-2. Geometric file-sizes also exhibit similar behavior. The graph was obtained by running the DP outlined in section 5.2. The results show $\phi_k(x_k)$ to be concave[3]. Unfortunately, the following counter-example illustrates that this is not always the case for arbitrary file-size distributions.

Example 5.2.1 (Counter-example to concavity of $\phi_k(x_k)$). Observe the behavior of $\phi_k(x_k)$ as depicted in figure 5-3 for discretized Gaussian-like random variables. Specifically, with mean μ and variance σ^2 for a Gaussian random variable, the plot is for file sizes having

$$\Pr(Y = k) = \begin{cases} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} e^{(x-\mu)^2/2\sigma^2} dx, & k = 0, \\ \int_{k-1}^{k} \frac{1}{\sqrt{2\pi}\sigma} e^{(x-\mu)^2/2\sigma^2} dx, & k = 1, \ldots, \infty. \end{cases}$$

Note, however, that the behavior of the $\phi_k(x_k)$'s in figure 5-3 is transient, and asymptotically, the functions become more regular, as shown in figure 5-4. We do not have an analytical result for the convergence and asymptotic regularity of the $\phi_k(x_k)$'s, although the convergence of $\phi_k(x_k)/(D-k)$ is shown below.

[3]Using the analogue of concavity for functions with integral domains.

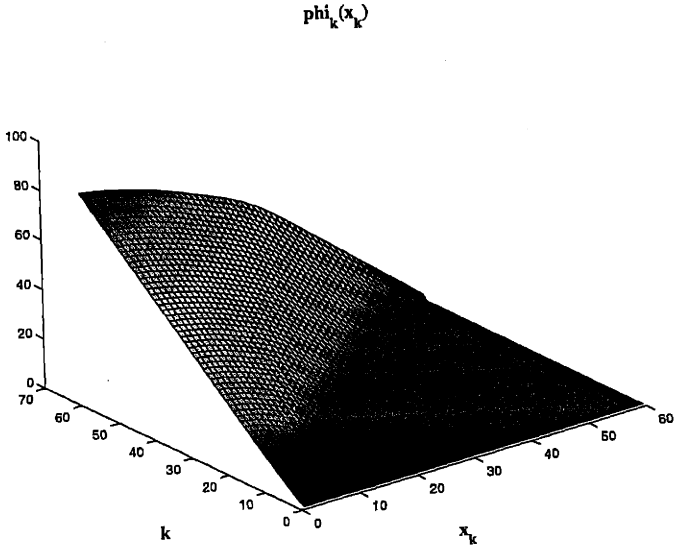


Figure 5-2: A plot of $\phi_k(x_k)$ vs. k and x_k for Pareto file size distributions with the Pareto parameter $a = 1.25, \lambda = 0.9, c = 0$ and $\tilde{r} = 2$. The deadline D is at 0 and k runs backwards.

Remark 5.2.4 (Convergence of the time-average additional revenue). For any given x,

$$\lim_{(D-k)\to-\infty} \frac{\phi_k(x)}{D-k} = \frac{\phi_{k+1}(x)}{D-k-1} = \frac{\phi_k(x+1)}{D-k},$$

i.e. the slope of the functions $\phi_k(x_k)$ becomes constant asymptotically, and therefore one expects linear growth far away from the deadline D *and* the capacity constraint $D-k$. This behavior is exhibited by the plots of figures 5-2 and 5-4.

With the above propositions, we can establish the existence of a non-trivial threshold admission policy which is optimal for the model. Some structural properties of the thresholds are also derived.

Remark 5.2.5 (Optimality of a non-trivial threshold policy). For any x_k, $0 \le x_k \le D-k$, the optimal admission control policy is characterized by a threshold job-size $y_k^*(x_k) \le D-k-x_k$ such that one admits a job of size y_k in period k iff $y_k \le y_k^*(x_k)$. Follows from the monotonicity of $\phi_k(z)$. The formal proof is provided in section 5.5. The intuition is as follows. For any $x_k \ge 1$, consider $r + \phi_{k+1}(x_k + y_k - 1)$ and $\phi_{k+1}(x_k - 1)$ as functions of y_k. Then the picture in figure 5-5 is true. This immediately reveals that as long as $\phi_{k+1}(x_k - 1) > r$, there is a $y_k^*(x_k)$ such that one accepts a job iff $y_k \le y_k^*(x_k)$. The picture is similar for $x_k = 0$.

Remark 5.2.6. $y_k^*(x_k) \le y_k^*(x_k+1)+1$ for all k and $0 \le x_k \le D-k-1$. The threshold can increase by at most 1 with a unit reduction in the amount of unfinished work. This is useful when finding conditions for concavity of $\phi_k(x_k)$'s.

The above two propositions are exhibited in the plots of optimal thresholds obtained from computational experiments with Pareto file-size distributions, figures 5-6 and 5-7. The plots hint that the thresholds might possess some sort of monotonic behavior. Therefore the next two propositions establish conditions for monotonicity of $y_k^*(x_k)$ in x_k. However, computational experiments with arbitrary file-size distributions (not transcribed here) reveal that the thresholds are not unconditionally monotone, therefore necessary and sufficient conditions for monotonicity are obtained.

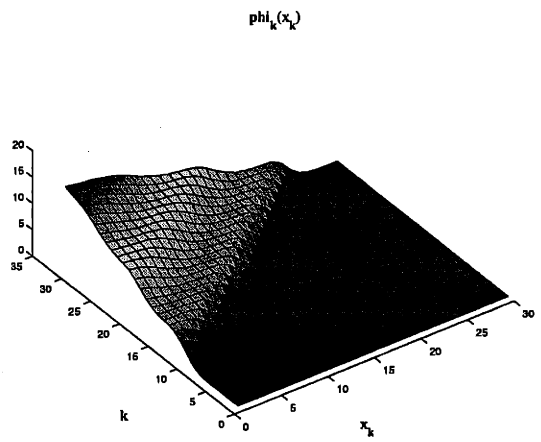


Figure 5-3: $\phi_k(x_k)$ plotted for Gaussian file-size distributions, with $\mu = 10, \sigma = 2, \lambda = 0.9, \tilde{r} = 5, c = 0$.

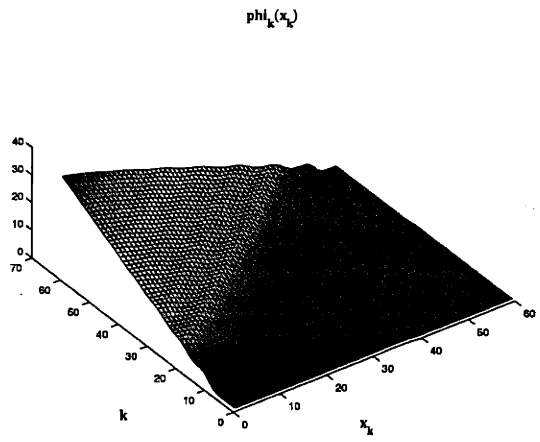


Figure 5-4: Showing transient vs. asymptotic behavior for the same data.

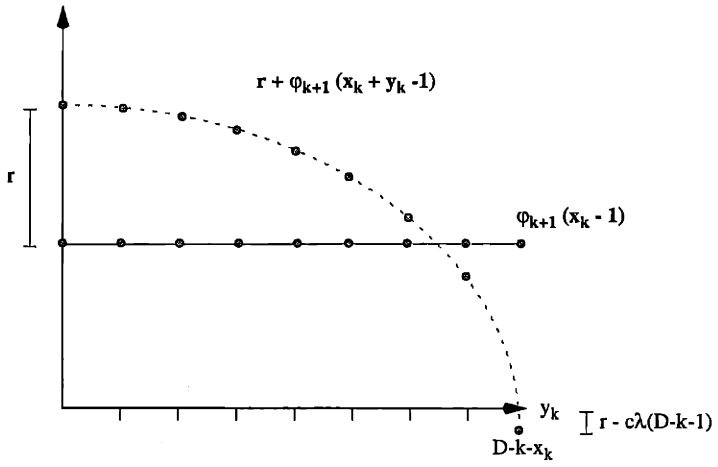


Figure 5-5: A graphic illustration of the proof for existence of optimal thresholds.

Remark 5.2.7 (Sufficient condition for monotonicity of the thresholds). If for every k, $\phi_k(x)$ are concave in $x, 0 \leq x \leq D - k$, then $y_k^*(x_k) \geq y_k^*(x_k + 1)$ for all x_k. This explains the results from computational experiments with Pareto and Geometric file size distributions. We also note that if $\phi_k(x)$ are convex, the direction of monotonicity is reversed, although it is difficult to think of the practical interpretation for why these functions would be convex.

Remark 5.2.8 (Necessary condition for monotonicity of the thresholds). If for all $0 \leq x_k \leq D-k-1$, $y_k^*(x_k) \geq y_k^*(x_k + 1)$, then

$$\phi_{k+1}(x-1) - \phi_{k+1}(x) < \phi_{k+1}(x + y_k^*(x) - 1) - \phi_{k+1}(x + y_k^*(x) + 1).$$

Note that the necessary condition is a relaxation of concavity.

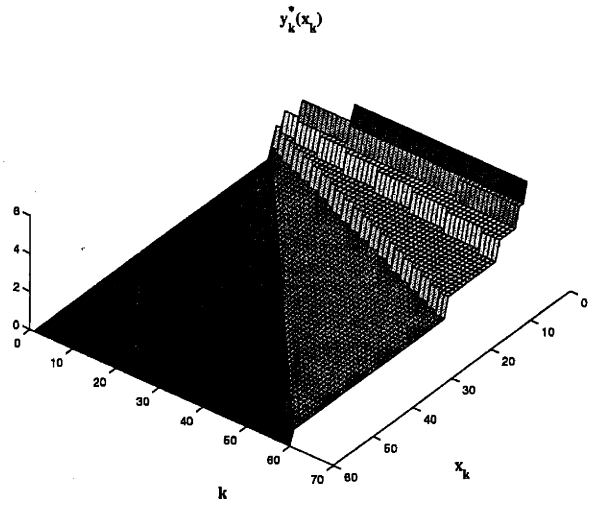


Figure 5-6: The optimal thresholds $y_k^*(x_k)$ plotted for Pareto file size distributions with the Pareto parameter $a = 1.25, \lambda = 0.9, c = 0$ and $\bar{r} = 2$. The deadline D is at 0.

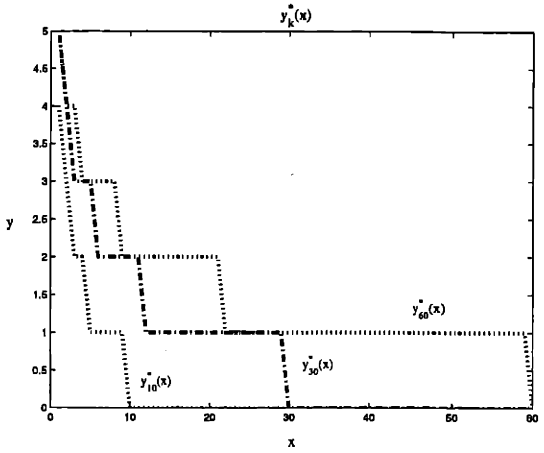


Figure 5-7: Plots of $y_k^*(x_k)$ vs. x_k for some values of k. File sizes are Pareto with $a = 1.25, \lambda = 0.9, c = 0$ and $\bar{r} = 2$. The *admission regions* are below the lines for each k.

5.2.4 Results for concave file-size distributions

The structural results of section 5.2.3 are useful but do not indicate how to compute the thresholds. Since computational evidence suggests the thresholds are sufficiently regular for the Pareto distribution (figure 5-6), we discuss the case of *concave* file-size distributions [4] to further understand their properties. The intention is to explain the experiments with Pareto distributions and guide the search for heuristic admission policies – addressed in section 5.2.5.

We conjecture the form of the optimal policy when file-size densities are convex and establish the convexity of the thresholds based on the assumption that in this case, the optimal additional revenue functions are concave for all stages. Extensive computational testing suggests that this is indeed the case for Pareto and Exponential file-size distributions, but we do not provide a formal proof of this latter property.

The convexity of the thresholds established below is for the *continuous* version of the model with the decision stages still discrete but x_k and Y now continuous variables. We assume that the support of Y is $[0, \infty)$. This leaves formulation (5.1) unchanged, except that $p_k(x_k) = \Pr(0 < Y \leq D-k-x_k)$ for $0 \leq x_k < D-k$ and we define $p_k(D-k) = 0$. The strict inequalities will not worry us too much in the analysis below. Now it is easy to see that continuity of x_k and Y does not affect the optimal policy. In effect, for continuous file-size distributions over $[0, \infty)$:

Remark 5.2.9. Continuity of the optimal additional revenue functions $\phi_k(x_k)$ follows easily, and they remain monotone non-increasing in x_k over $0 \leq x_k \leq D-k$. The proof follows exactly the argument for the discrete case, as outlined in proposition 5.5.1.

Remark 5.2.10. The optimal admission policy remains a threshold $y_k^*(x_k)$, similar to proposition 5.5.5. It is also useful to remember that here $y_k^*(x_k)$ solves

$$\begin{aligned} x_k \geq 1: &\qquad r + \phi_{k+1}(x_k + y - 1) = \phi_{k+1}(x_k - 1), \\ x_k < 1: &\qquad r + \phi_{k+1}\big(\max(0, x_k + y - 1)\big) = \phi_{k+1}(0). \end{aligned}$$

[4] With convex densities.

for every x_k with $\phi_{k+1}(x_k - 1) \geq r - c\lambda(D - k - 1)$. And for larger x_k, $y_k^*(x_k) = D - k - x_k$.

Then the DP recursion can be written as follows, where G is the distribution function of Y:

$$\begin{aligned} \phi_D(x_D) \equiv &\quad 0, \quad \text{and for } k = 0, \ldots, D-1, \\ \phi_k(x_k) = &\quad -c\lambda\delta \;+\; \phi_{k+1}\big(\max(0, x_k - 1)\big) \;+ \\ &\int_{y=0}^{y_k^*(x_k)} \big[r \;+\; \phi_{k+1}\big(\max(0, x_k + y - 1)\big) \;-\; \phi_{k+1}\big(\max(0, x_k - 1)\big)\big]\, dG_Y(y). \end{aligned} \tag{5.2}$$

Now suppose that the file-size distributions are *strictly concave and differentiable* over $[0, \infty)$. This covers the cases of Pareto and Exponential distributions [5]. In this case, we believe $\phi_k(x_k)$ are strictly concave and differentiable for every k, based on computational experiments. This allows us to obtain proposition 5.2.12. First, however, we need the following lemma.

Lemma 5.2.11. *Let $f(x)$ be a monotone decreasing, strictly concave and differentiable function of x, defined over some interval $[a, b]$. Let $h(x) = \frac{df^{-1}}{dx}[f(x) - c]$ for some constant $c \geq 0$. Then $0 \leq h(x) \leq 1$, and $h(x)$ is monotone increasing in x, for every $c \geq 0$.*

Proof. It is obvious that f^{-1} exists, from the monotone decrease and strict concavity of f. Also, differentiability of f implies differentiability of f^{-1}. Now obviously $f^{-1}[f(x)] = x$ and therefore, for $c = 0$, the lemma holds since

$$h(x) = \quad \frac{df^{-1}}{df}[f(x)]\frac{df}{dx}(x) = 1.$$

For $c > 0$, first re-write the above expression as

$$\frac{df^{-1}}{df}[f(x)] = \frac{1}{\frac{df}{dx}(x)}.$$

Then notice that $\frac{df^{-1}}{df}[f(x) - c]$ is just displacing the function $\frac{df^{-1}}{df}[f(x)]$ right by c. Then if $\frac{df^{-1}}{df}[f(x)] = \frac{1}{\frac{df}{dx}(x)} = \frac{1}{z(x)}$, we can write $\frac{df^{-1}}{df}[f(x) - c] = \frac{1}{z(x)-c}$. This implies

$$h(x) = \frac{df^{-1}}{df}[f(x)]\frac{df}{dx}(x) = \frac{z(x)}{z(x) - c} = \frac{1}{1 - \frac{c}{z(x)}}.$$

Now notice that $z(x) = \frac{df}{dx}(x) < 0$ and monotone decreasing in x, from the strict concavity of f. Therefore $\frac{1}{1-c/z(x)}$ is > 0, less than 1 and monotone increasing in x.

□

Proposition 5.2.12 (Convexity of the thresholds). *If $\phi_{k+1}(x)$ is strictly concave and differentiable over $0 \leq x \leq D - k - 1$, the following are true.*

i. *$y_k^*(x)$ is monotone decreasing in x.*

ii. *$y_k^*(x)$ is differentiable over $0 \leq x \leq b_k$ and over $b_k \leq x \leq D - k$, for some constant $b_k \geq 0$. At b_k, it is continuous.*

[5] For Pareto distributions, there is the small matter of the location parameter, which cannot be zero, but the results here can be easily extended to cover this technicality.

iii. $\frac{dy_k^*}{dx}(0) > -1$, *$y_k^*(x)$ is convex in x over $0 \le x \le b_k$ and linearly decreasing over $b_k < x \le D-k$.*

Proof. Part (i) requires only monotonicity and concavity of ϕ_{k+1}. Assume $x \ge 1$, since the case $x < 1$ requires little added effort. Then from remark 5.2.10, $r = \phi_{k+1}(x-1) - \phi_{k+1}\big(x-1+y_k^*(x)\big)$. Therefore for any $z > 0$, $r < \phi_{k+1}(x+z-1) - \phi_{k+1}\big(x+z-1+y_k^*(x)\big)$, i.e. we must have $y_k^*(x+z) < y_k^*(x)$.

For (ii) and (iii), note that strict concavity, differentiability and monotonicity of ϕ_{k+1} brings invertibility. And in fact, the inverse $\phi_{k+1}^{-1}(z)$ is also strictly concave, differentiable, and monotone decreasing over $-c\lambda(D-k-1) \le z \le \phi_k(0)$. Now let $b_k = \phi_{k+1}^{-1}\big(r - c\lambda(D-k-1)\big)$. From remark 5.2.10, first note that we can write:

$$\begin{aligned} 0 \le x < 1: \quad & y_k^*(x) = 1 - x + \phi_{k+1}^{-1}\big[\phi_{k+1}(0) - r\big], \\ 1 \le x \le b_k: \quad & y_k^*(x) = 1 - x + \phi_{k+1}^{-1}\big[\phi_{k+1}(x-1) - r\big], \\ b_k \le x \le D-k: \quad & y_k^*(x) = D - k - x. \end{aligned}$$

Now (ii) is immediate, given differentiability of $\phi_{k+1}^{-1}(z)$, and (iii) can be seen by differentiating the above expressions. $\frac{dy_k^*}{dx}(x) = -1$ over the ranges $0 \le x < 1$ and $b_k \le x \le D-k$, while for $1 \le x \le b_k$, we have:

$$\frac{dy_k^*}{dx}(x) = -1 + \frac{d\phi_{k+1}^{-1}}{dx}\big[\phi_{k+1}(x-1) - r\big], \qquad 1 \le x \le b_k.$$

Lemma 5.2.11 then gives (iii).

□

Note that the property above explains the behavior of the thresholds seen in figure 5-6 for Pareto distributions. Part (iii) above is the analogue of remark 5.2.6 and its interpretation is that the thresholds increase much more slowly than the decrease in the remaining content in the system. A picture illustrates this behavior.

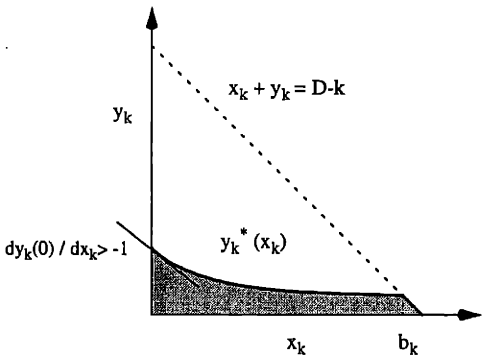


Figure 5-8: Illustrating convexity of the thresholds $y_k^*(x_k)$ for concave file-size distributions. The admission region is shaded.

Part of the behavior of the optimal thresholds for the Pareto case is explained by the above property, but we still lack a complete characterization and an easily implementable policy. Figure 5-6 suggests that the form of an optimal policy for the Pareto[6] case, in the discrete version of the model, is characterized by a set of scalars $\alpha_i, i = 1, \ldots, \bar{y}$, where $\alpha_i > \alpha_{i-1}$ and $\alpha_0 = 1$, such that:

[6]Geometric distributions also have similar behavior.

$$y_k^*(x_k) = i, \qquad \text{for } \alpha_i(D-k) \leq x_k < \alpha_{i-1}(D-k).$$

Where we define $\bar{y} = \lim_{(D-k)\to-\infty} y_k^*(0)$, a limit that seems to exist, from computations. Figure 5-9 illustrates the policy.

We state the above based on strong computational evidence but do not have a proof. However, if one believes that the thresholds indeed have the form as above, an easy computational procedure to determine the optimal policy is to compute the set of scalars α_i off-line, based on a few hundred stages of the DP[7]. The policy is then implemented by simply storing the numbers α_i and performing a single computation at each arrival.

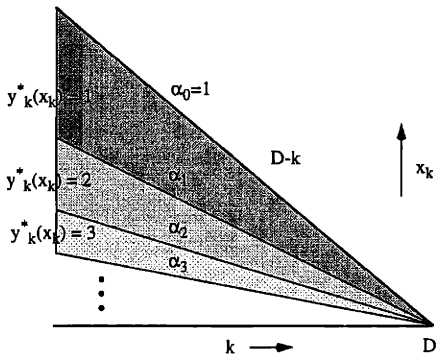


Figure 5-9: Illustrating the conjectured form of the optimal policy for concave file-size distributions.

5.2.5 A heuristic policy: the airline "Expected Marginal Revenue" connection

Two reasons motivate investigating heuristic admission policies: (i) we do not have a formal characterization of the optimal policy, and (ii) when file-size distributions are not concave, the behavior of the thresholds is expected to be more complicated. Even for concave distributions, obtaining optimal thresholds is highly unlikely when the model is generalized to multiple deadlines and networks (c.f. section 5.4). To develop our intuition and understanding of the model, we search for a heuristic policy in the Pareto case and compare its performance to that of the optimal and *greedy* admission policies. The heuristic proposed here is directly motivated by the popular EMSR [Bel87] heuristic for airlines seat inventory control.

We let the discrete version of the model for the Pareto case guide our intuition. Here, if we believe the characterization of the thresholds of section 5.2.4, we expect a nesting structure, where at every stage k, the optimal policy *protects* some capacity for jobs of size 1. Similarly, it protects some amount of capacity for all jobs of size 1 and 2, and so on. One line of reasoning this suggests is as follows:

1. View demand at any stage k as segregated into independent Poisson processes $\lambda_i, i = 1, \ldots, \infty$, with $\lambda_i = \lambda \Pr\{Y = i\}$, i.e. we have independent Poisson arrivals for jobs of size i. If X_{ik} is the total demand for size i jobs that will arrive in the remaining stages till the deadline, then X_{ik} is a Poisson random variable with the Poisson parameter $\lambda_i(D-k)$. Further, since each arrival brings a constant benefit r, we can view the benefit per unit of bandwidth from a customer of size i as r/i.

[7] Around 100-150 stages seem to characterize almost all the cases we tested.

2. Now suppose we held the view that we have $D-k$ amount of capacity to sell in stage k to all customers arriving after that stage, and demand will be realized immediately in the next interval, with arrivals for larger jobs occurring before arrivals for shorter jobs. Then we can look for the minimum units of bandwidth to protect from class $i+1$ and higher sized jobs. Formally, we are looking for protection levels $p_{ik}, i = 1, \ldots, \infty$, such that a job of size $i+1$ is admitted in state x_k only if the remaining capacity $D-k-x_k > p_{ik}$.

   If we ignore knapsack effects and consider only class 1 and 2 jobs, we can use the following analogue of Littlewood's rule [Lit72] from airline seat-inventory control. Let $p_{1k} = D-k$, and keep reducing p_{1k} as long as

$$\frac{r}{2} \geq r \Pr\{X_{1k} > p_{1k}\}.$$

   The interpretation is in terms of the marginal revenue per unit of bandwidth from jobs of size 1 *ignoring knapsack effects*. The revenue per unit of bandwidth from a size 2 job is $r/2$. The rule therefore models a decision where a job of size 2 has arrived, and its immediate benefit is to be compared to the future expected benefit from arrivals of class 1, if all demand for size 1 jobs were to be realized *immediately* in the next interval, *after* the demand for job 2 was realized.

   For jobs of size 3, following the airlines analogy, the EMSR rule gives $p_{2k} = p^1_{2k} + p^2_{2k}$, where p^1_{2k} and p^1_{2k} are individual protection levels given by:

$$\frac{r}{3} \geq r \Pr\{X_{1k} > p^1_{2k}\} \qquad \text{and} \qquad \frac{r}{3} \geq \frac{r}{2} \Pr\{X_{2k} > p^2_{2k}\}.$$

   This process can be continued until the protection levels obtained exceed available capacity.

This completes the description of the heuristic. Below are comments on some of its properties and assumptions. Figure 5-10 illustrates the thresholds obtained by this heuristic for Pareto distributions.

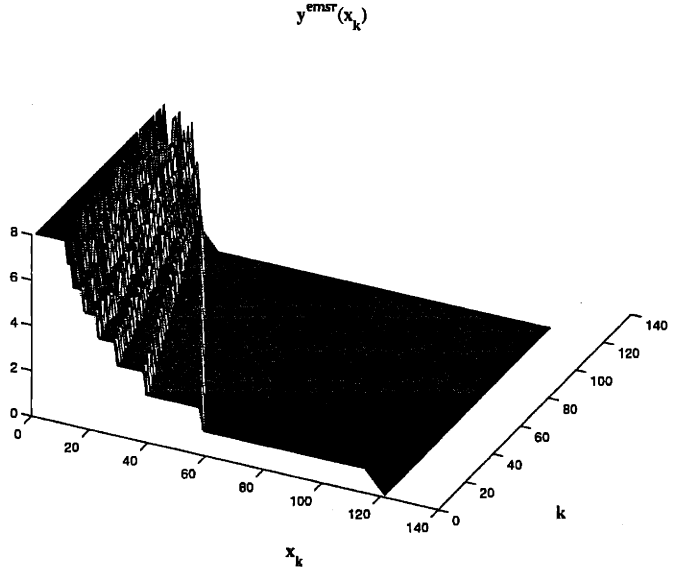


Figure 5-10: Thresholds obtained by the EMSR heuristic for Pareto file size distributions with the Pareto parameter $a = 1.25, \lambda = 0.9, c = 0$ and $\bar{r} = 2$. The deadline D is at 120.

1. Assumptions in proposing this heuristic are many, the most important being static control, realization of all demand in the next interval and the ordering of the demand realizations. One also ignores the time dependency of the available capacity, which decreases at a constant rate.

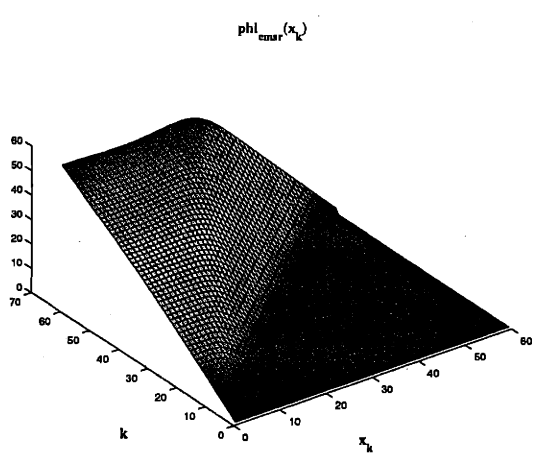


Figure 5-11: The revenue function obtained with the EMSR heuristic for Pareto files-size with $a = 1.25, \lambda = 0.9, c = 0$ and $\tilde{r} = 2$. The deadline D is at 0.

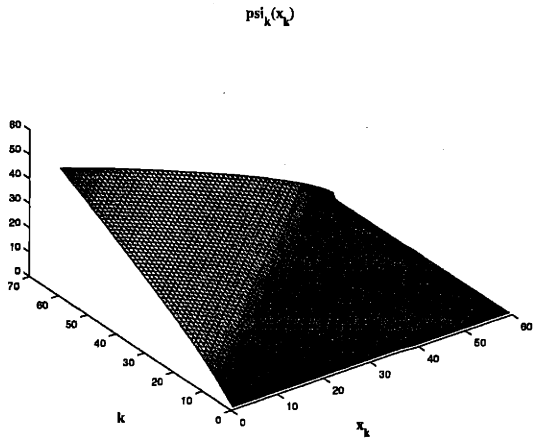


Figure 5-12: The revenue function obtained under a greedy policy for Pareto files-size with $a = 1.25, \lambda = 0.9, c = 0$ and $\tilde{r} = 2$.

2. The non-optimality of the protection levels as we describe is known even in the airlines literature [BM93], where an exact expression for the optimal protection levels is also derived under certain assumptions. The exact expression, however, is more involved and at least in the airlines, case, not significantly more beneficial than the EMSR heuristic [BM93].

3. The net effect of the assumptions and the non-optimality of the heuristic protection levels is reflected in figure 5-10, where the heuristic under-estimates the protection levels for each job class, the optimal thresholds displayed in figure 5-6. We do not attempt to assign the reasons to the many assumptions, but comment that several refinements are possible – a promising direction for future research.

4. The issue of computation needs attention. As described, the heuristic involves obtaining protection levels for each stage k, a procedure that, although straightforward, is still cumbersome. However, it seems that the protection levels are approximately linear in k evidenced by figure 5-10 and only the slopes are over-estimated. If we believe this linearity, then we can obtain protection levels for some stage k, compute the scalars $\beta_i = \frac{p_{ik}}{D-k}$ and implement this policy, instead of computing the optimal α_i's by a DP recursion (c.f. section 5.2.4).

Performance comparison of the heuristic with the optimal policy and the *greedy* policy follow. The performance of the *greedy* policy is obtained by the following recursion, with $\psi_k(x_k)$ denoting the *greedy* expected additional revenue:

$$\begin{aligned}
&\psi_D(x_D) \equiv 0, \quad \text{and for } k = 0, \ldots, D-1, \\
&\psi_k(x_k) = -c\Pr(Y > 0) + \psi_{k+1}\big(\max(0, x_k - 1)\big) + \\
&\qquad \sum_{y=1}^{D-k-x_k} \Pr(Y = y)\big[r + \psi_{k+1}(x_k + y - 1) - \psi_{k+1}\big(\max(0, x_k - 1)\big)\big].
\end{aligned}$$

Figures 5-11 through 5-14 summarize some computational experiments. Figures 5-13 and 5-14 are particularly informative, displaying the revenue function as function of x_0 and k separately for all the policies. We comment on the results but leave the refinement of the heuristic for future research.

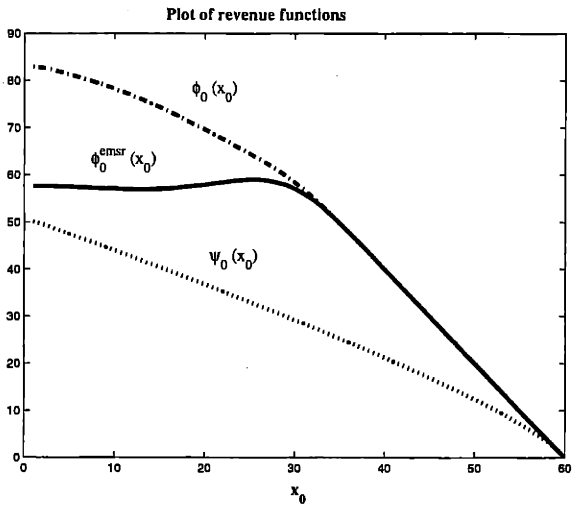


Figure 5-13: The revenue function obtained for different policies plotted as a function of x_0, for Pareto distribution with $a = 1.25, \lambda = 0.9, c = 0$ and $\tilde{r} = 2$. The deadline D is at 0.

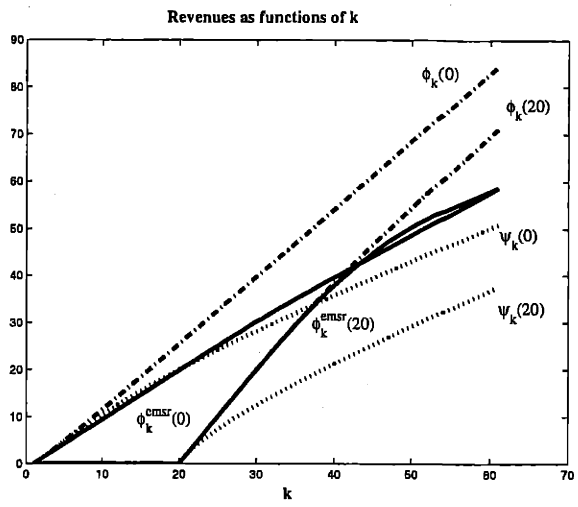


Figure 5-14: The revenue function obtained for different policies plotted as a function of k, for Pareto distribution with $a = 1.25, \lambda = 0.9, c = 0$ and $\tilde{r} = 2$. The deadline D is at 0.

1. From figure 5-13, the EMSR heuristic remains exactly optimal for $30 \leq x_0 \leq 60$, but deteriorates significantly in the range $0 \leq x_0 \leq 30$. This behavior arises since the thresholds for the optimal and EMSR policy match exactly in the range $30 \leq x_0 \leq 60$, shown in figure 5-10. In the range $0 \leq x_0 \leq 30$, the EMSR thresholds over-estimate optimal thresholds, and the performance decreases steadily. A tentative conclusion is that the EMSR heuristic is a good approximation when the system carries a medium to heavy load.

2. From figure 5-14, all policies seem to have linear growth in k, for fixed values of x. Again, owing to the nature of the thresholds obtained for EMSR, one sees that the approximation is better when workload is higher, although the differences between the greedy and the EMSR, and between the EMSR and the optimal policies, grow more or less linearly in k.

3. One sees from figure 5-10 that *knapsack* effects become more important closer to the deadline, an expected behavior. However, we speculate that the revenue impact of knapsack effects is second order compared to the effects of under-estimation of the protection-levels at stages far away from the deadline. A possibility for improving the performance of the EMSR heuristic might therefore be to limit the maximum size of the job that can be accepted in any stage, and scale up the protection levels, for lower size jobs proportionally. We do not pursue this further.

5.3 Extensions

This section presents several interesting extensions of the basic model of section 5.2. The principal result here is the existence of an optimal threshold policy for a variety of more interesting cases.

5.3.1 Time-dependent revenue, rejection cost, file size distributions

A natural extension is the case where revenue, cost and file sizes are non-stationary. Specifically, with revenue for stage k being $\tilde{r}_k$, the cost of rejection c_k, and the arriving file sizes independent random variables Y_k with distributions dependent on k, one obtains *exactly* the DP recursion (5.1) with c, r and Y now subscripted by k, i.e. $c = c_k$, $r = r_k = \tilde{r}_k + c_k$ and $Y = Y_k$. $p_x(x_k)$ likewise

involves Y_k. We assume $c_k < r_k$ for every k, for otherwise it would be optimal to reject everyone in stage k.

With the formulation unchanged, the optimal acceptance policy must still be a work-dependent threshold. To see this, note that $\phi_k(x_k)$ is still monotone decreasing in x_k, since the proof of proposition 5.5.1 does not depend on r, c or the distribution of Y. Likewise, the proof of existence of the threshold policy, proposition 5.5.5, depends only on the monotone decrease of the $\phi_k(x_k)$'s and the file-size random variables being continuous. Following the same argument as proposition 5.5.5, it is easy to see that for $x_k \geq 1$, if $\phi_{k+1}(x_k - 1) \leq r - \lambda \sum_{i=k+1}^{D} c_i$, one accepts a job if it fits, otherwise the optimal policy consists of a non-trivial threshold $y_k^*(x_k) < D - k - x_k$. The situation is depicted in figure 5-15.

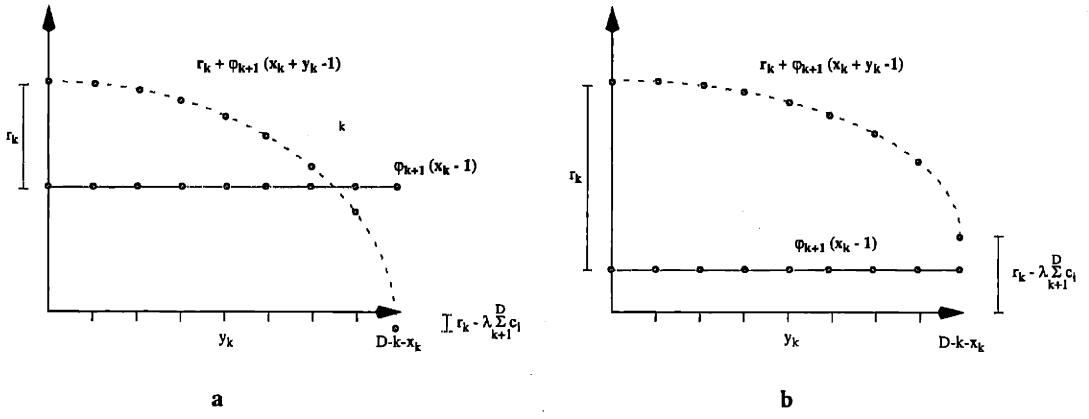


Figure 5-15: The proof for existence of optimal thresholds for non-stationary revenue, cost and file sizes. Figure **a** illustrates the case $y_k^*(x_k) < D - k - x_k$ and figure **b** illustrates when $y_k^*(x_k) = D - k - x_k$.

5.3.2 Deterministic time-varying and/or stochastic transmission rate

The case of a given time-varying deterministic capacity function is easy. Suppose the non-negative integer-valued function $R_k\delta$ represents the available transmission rate for stage $k = 0, \ldots, D$. Calling $\mathcal{R}_k = \delta \sum_{i=k}^{D} R_i$, the formulation (5.1) remains correct with $x_k - 1$ replaced by $x_k - R_k\delta$ and $D-k-x_k$ replaced by $\mathcal{R}_k - x_k$. Now the proof of proposition 5.5.1 for monotonicity of the $\phi_k(x_k)$'s still falls through and we see exactly as in section 5.3.1 that the optimal policy must still be a work-dependent threshold. Finally, combining this and the reasoning of section 5.3.1, we see that the structure of the optimal policy holds for time-varying capacity, revenue, cost and file-size distributions.

The case of stochastic available capacity requires a re-formulation but the existence of optimal thresholds can still be shown. We emphasize that we consider this case more or less academic, since both characterizing the available capacity and computing the thresholds is expected to be a hard problem. Suspending the issue of actually computing the involved quantities for the time being, we proceed as follows.

Suppose we have available the complete description of a stochastic process $\{R_k\delta\}_{k=0,\ldots,D-1}$ describing available transmission rate for each stage k. Assume $R_k \geq 0$ for all k, and is independent of the decisions and the arrival process. Call the associated *cumulative* remaining capacity process $\{\mathcal{R}_k\}_{k=0,\ldots,D-1}$ where $\mathcal{R}_k = \delta \sum_{i=k}^{D} R_k$.

We assume a feasibility constraint of the form $\Pr\{y_k > \mathcal{R}_k - x_k | R_0, R_1, \ldots, R_{k-1}\} \leq e^{-\beta}$, for some given $\beta > 0$. This ensures that the only acceptable content sizes y_k in stage k are those

which can be delivered by the deadline with probability greater than $1 - e^{-\beta}$. Now if $G(.)$ is the distribution function of $\mathcal{R}_k - x_k | R_0, R_1, \ldots, R_{k-1}$, we can see from figure 5-16 that the above constraint translates to a largest size job $\bar{y}_k(x_k)$ that can be accepted in stage k, depending on the work in the system and the history of the transmission rates, i.e. $\bar{y}_k = \max_y G_{\mathcal{R}_k - x_k | R_0, R_1, \ldots, R_{k-1}}(y)$ such that $G_{\mathcal{R}_k - x_k | R_0, R_1, \ldots, R_{k-1}}(y) \leq e^{-\beta}$. The DP (5.1) can now be written as follows, defining $p_k(x_k) = \Pr(1 \leq Y \leq \bar{y}_k(x_k))$.

$$
\begin{aligned}
&\phi_D(x_D) \equiv 0, \quad \text{and for } k = 0, \ldots, D-1, \\
&\phi_k(x_k) = -c\Pr\{Y > 0\} + \\
&\qquad \big(1 - p_k(x_k)\big) \mathop{\mathbb{E}}_{R_k} \phi_{k+1}\big(\max(0, x_k - R_k\delta)\big) + \\
&\qquad p_k(x_k) \mathop{\mathbb{E}}_{Y | 1 \leq Y \leq \bar{y}_k(x_k)} \max\Big\{ \quad r + \mathop{\mathbb{E}}_{R_k} \phi_{k+1}\big(\max(0, x_k + Y - R_k\delta)\big), \\
&\qquad\qquad\qquad\qquad \mathop{\mathbb{E}}_{R_k} \phi_{k+1}\big(\max(0, x_k - R_k\delta)\big) \Big\}.
\end{aligned}
$$

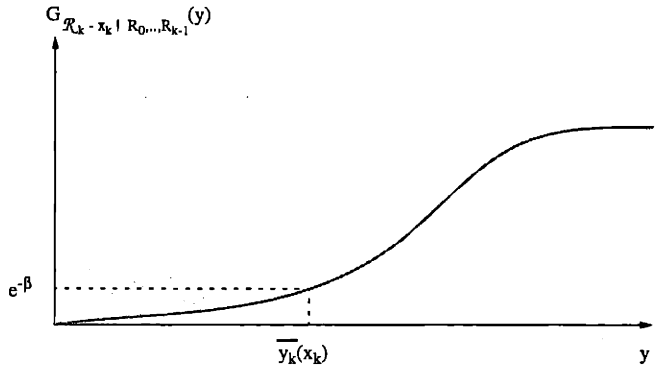


Figure 5-16: The stochastic feasibility constraint for admission.

Now following the proof of proposition 5.5.1, we can see that $\phi_k(x_k)$ is still monotone decreasing as long as the R_k's are independent of the decision process. Then the optimal admission policy must still consist of a threshold job-size at each stage and state, since proposition 5.5.5 depends only on the monotonicity of the $\phi_k(x_k)$'s.

5.3.3 Multiple customer classes

Here the model is the same as section 5.2 with the only difference being that multiple classes of customers arrive according to independent Poisson processes. Every arrival has an associated class-dependent revenue irrespective of size. The most reasonable model here would be to assume that files are divided into size ranges with a fixed revenue associated with each range, larger files having a higher revenue than smaller ones. The existence of multiple nested thresholds can easily be shown, with the thresholds increasing in the revenue of the job.

Specifically, denote the arrival rate λ_i and let $y_i, i = 1, \ldots, n$ represent the arriving file size for customer class i in stage k. Note that we drop the subscript k from the y_i's for notational convenience, assuming that the Y_i's are discretized versions of i.i.d *continuous* random variables in each stage k. Let $\tilde{r}_i, c_i$ denote the revenue and rejection cost for customer class i and assume both $r_1 \geq r_2 \geq \ldots \geq r_N$ and $c_1 \geq c_2 \geq \ldots \geq c_N$. Now since at each stage, at most one arrival can occur, the following DP formulation results, where $\mathbb{I}_{\{y_i > 0\}}$ is the indicator function of the event $y_i > 0$. w_i are the random arriving file sizes for class i in stage k which will await decision in stage $k+1$.

$$J_D \begin{bmatrix} x_D \\ y_1 \\ \vdots \\ y_N \end{bmatrix} = 0, \qquad \forall x_D, y_1, \ldots, y_N,$$

and for $k = 0, \ldots, D-1$,

$$J_k \begin{bmatrix} x_k \\ y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{cases} \max \Big\{ \sum_{i=1}^N \tilde{r}_i \mathbb{I}_{\{y_i>0\}} + \mathop{\mathbb{E}}_{w_1,\ldots,w_N} J_{k+1} \begin{bmatrix} x_k + \sum_{i=1}^N y_i - 1 \\ w_1 \\ \vdots \\ w_N \end{bmatrix}, \\ \qquad - \sum_{i=1}^N c_i \mathbb{I}_{\{y_i>0\}} + \mathop{\mathbb{E}}_{w_1,\ldots,w_N} J_{k+1} \begin{bmatrix} \max(0, x_k - 1) \\ w_1 \\ \vdots \\ w_N \end{bmatrix} \Big\}, \quad 1 \le \sum_{i=1}^N y_i \le D - k - x_k, \\ - \sum_{i=1}^N c_i \mathbb{I}_{\{y_i>D-k-x_k\}} + \mathop{\mathbb{E}}_{w_1,\ldots,w_N} J_{k+1} \begin{bmatrix} \max(0, x_k - 1) \\ w_1 \\ \vdots \\ w_N \end{bmatrix}, \quad \text{o/w.} \end{cases}$$

By defining

$$\phi_k(x_k) = \mathop{\mathbb{E}}_{w_1,\ldots,w_N} J_{k+1} \begin{bmatrix} x_k \\ w_1 \\ \vdots \\ w_N \end{bmatrix},$$

we can now write the following formulation, where $p_k(x_k) = \Pr(1 \le \sum_{i=1}^N Y_i \le D - k - x_k)$ and $r_i = \tilde{r}_i + c_i$.

$\phi_D(x_D) \equiv 0$, and for $k = 0, \ldots, D-1$,

$$\begin{aligned} \phi_k(x_k) = &- \delta \sum_{i=1}^N \lambda_i c_i + \\ &\big(1 - p_k(x_k)\big)\ \phi_{k+1}\big(\max(0, x_k - 1)\big) + \\ &p_k(x_k) \mathop{\mathbb{E}}_{Y_1,\ldots,Y_N | 1 \le \sum_{i=1}^N Y_i \le D-k-x_k} \max \Big\{ \sum_{i=1}^N r_i \mathbb{I}_{\{Y_i>0\}} + \phi_{k+1}\big(\max(0, x_k + \sum_{i=1}^N Y_i - 1)\big), \\ &\qquad \phi_{k+1}\big(\max(0, x_k - 1)\big)\ \Big\}. \end{aligned}$$

Now we have the following results which establish the existence of nested admission thresholds

for each class i, depending on the ordering of the r_i's.

Proposition 5.3.1. *For every k, the optimal "expected additional revenue" function $\phi_k(x_k), x_k = 0, \ldots, D-k$, is monotone non-increasing in x_k.*

Proof. Note that an admissible policy $\pi = \{\mu_k, \mu_{k+1}, \cdots, \mu_{D-1}\}$ is a sequence of functions μ_k such that

$$\mu_k \begin{bmatrix} x_k \\ y_1 \\ \vdots \\ y_N \end{bmatrix} \in \begin{cases} \{0,1\}, & 1 \leq \sum_{i=1}^N y_k \leq D-k-x_k, \\ \{0\} & \text{o/w.} \end{cases}$$

Then starting from state x_k, consider the following policy. Let $x_k \to x_k + 1$ and follow the policy optimal for starting from state $x_k + 1$. This clearly produces a feasible sequence of admissions with additional revenue $\phi_k(x_k + 1)$. The optimal policy starting from state x_k must do at least as well, therefore $\phi_k(x_k) \geq \phi_k(x_k + 1)$.

□

Theorem 5.3.2. *If the file size random variables Y_i can take values $0, \ldots, \infty$, the optimal admission policy consists of a threshold job size $y^*_{ik}(x_k)$ for each class i in stage k, such that one admits a job of size y_i iff $y_i \leq y^*_{ik}(x_k)$. Further, $y^*_{ik}(x_k) \geq y^*_{i+1,k}(x_k), i = 0, \ldots, N-1$, if $r_i \geq r_{i+1}$ for $i = 1, \ldots, N-1$.*

Proof. Consider $x_k \geq 1$, since $x_k = 0$ follows exactly the same reasoning. A job of class i, with size $y_i \geq 1$ is admitted in stage k iff $1 \leq y_i \leq D-k-x_k$ and

$$r_i + \phi_{k+1}(x_k + y_i - 1) \geq \phi_{k+1}(x_k - 1).$$

Now we have

$$y_i = 1: \quad r_i + \phi_{k+1}(x_k + y_i - 1) = r_i + \phi_{k+1}(x_k),$$

$$y_i = D-k-x_k: \quad r_i + \phi_{k+1}(x_k + y_i - 1) = r_i - (D-k-1)\sum_{i=1}^N \lambda_i c_i.$$

Also, from elementary reasoning, for all values of x_k, $\phi_{k+1}(x_k - 1) \geq -(D-k-1)\sum_{i=1}^N \lambda_i c_i$. Using proposition 5.3.1, if $r_i - (D-k-1)\sum_{i=1}^N \lambda_i c_i < \phi_{k+1}(x_k - 1) < r_i + \phi_{k+1}(x_k)$, there must be a $0 < y^*_{ik}(x_k) < D-k-x_k$ such that for $y_i \leq y^*_{ik}(x_k)$, $r_i + \phi_{k+1}(x_k + y_i - 1) > \phi_{k+1}(x_k - 1)$ and for $y_i \leq y^*_{ik}(x_k)$, the converse is true.

The only other possibilities are: $r_i + \phi_{k+1}(x_k) \leq \phi_{k+1}(x_k - 1)$, in which case $y^*_{ik}(x_k) = 0$, i.e. no jobs from class i can be admitted, or $\phi_{k+1}(x_k - 1) \leq r_i - (D-k-1)\sum_{i=1}^N \lambda_i c_i$, in which case $y^*_{ik}(x_k) = D-k-x_k$, i.e. a job of class i is admitted if it fits. In any case, we have the existence of a threshold size $y^*_{ik}(x_k) \leq D-k-x_k$ for admission of class i.

To see that the thresholds are nested, note that if $r_i \geq r_j$, then for all $1 \leq y \leq D-k-x_k$,

$$r_i + \phi_{k+1}(x_k + y - 1) \geq r_j + \phi_{k+1}(x_k + y - 1).$$

Then the definition of $y^*_{jk}(x_k)$ implies the following, and $y^*_{ik}(x_k) \geq y^*_{jk}(x_k)$ follows from the monotonicity of ϕ_{k+1}. Figure 5-17 illustrates this proof.

$$r_i + \phi_{k+1}(x_k + y^*_{jk}(x_k) - 1) \geq r_j + \phi_{k+1}(x_k + y^*_{jk}(x_k) - 1) > \phi_{k+1}(x_k - 1).$$

□

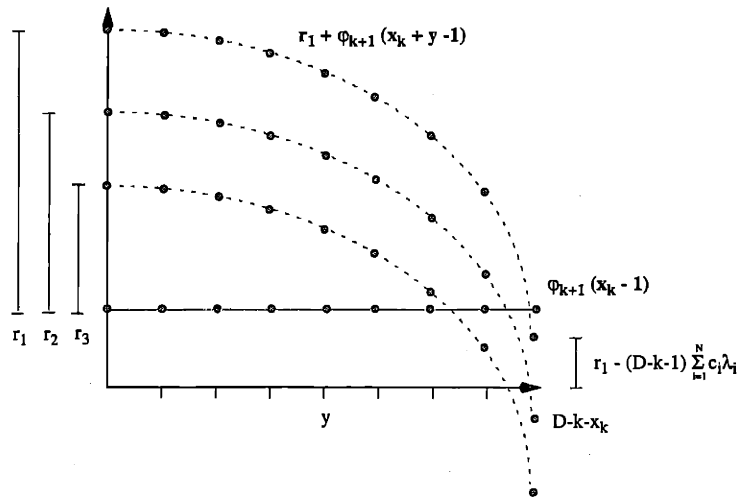


Figure 5-17: An illustration of the proof for existence of nested optimal thresholds.

5.4 Networks and multiple deadlines

We consider the case of a network with fixed routing and a single deadline, and the case of multiple deadlines on a single link. The principal results in this section are that the monotonicity of the optimal additional revenue is a robust property and the optimal policy is of threshold-type for these more complicated cases, even though the actual thresholds are unlikely to be easily computable.

In the case of multiple We demonstrate this, we choose to analyze the case of only two deadlines and a simple network instance with only one deadline and fixed routing, since the essential characteristics of the line of reasoning are fully revealed by these cases. The generalization to the case of a full network is then easy.

5.4.1 Networks with fixed routing

Let a network consist of $j = 1, \ldots, L$, links, with available rate R_j for link j. Let $r = 1, \ldots, \mathcal{R}$, index routes where each route r is a subset of the links. Requests arrive for content delivery over route r as a Poisson process with rate λ_r, and every such arrival requests delivery of a random amount of content Y_r over the route, before a deadline D *common to all requests network-wide.* For each r, Y_r are i.i.d and independent of the arrival process and state of the network. Arrival for route r, if accepted, results in revenue $\tilde{r}_r$, otherwise a cost c_r. The assumption $\tilde{r}_r > c_r$ is natural.

We discretize time as before into intervals of length δ, with $k = 0, \ldots, D$, indexing decision stages, and assume Y_r are appropriately quantized, i.e. arriving in quanta of $\min_j R_j\delta$. Now call x_{rk} the content for route r untransmitted at stage k. Let $a_{jr} = 1$ if link $j \in r$, and $a_{jr} = 0$ otherwise.

The following vector notation can now be defined. Let $\mathbf{x}_k = (x_{1k}, \ldots, x_{\mathcal{R}k})^T$ be the remaining content vector and $\mathbf{y}_k = (y_{1k}, \ldots, y_{\mathcal{R}k})^T$ the *arrived* content vector requiring admission decision in stage k. $\mathbf{Y} = (Y_1, \ldots, Y_{\mathcal{R}})^T$ is the random content vector arriving in every interval, which awaits decision till the next stage, i.e. $\mathbf{y}_k$ is the realization of $\mathbf{Y}$ in stage k. Note that Poisson arrival processes implies that $\mathbf{y}_k$ cannot have more than one non-zero component with positive probability. Finally, let $\mathbf{A} = [a_{jr}]$ of dimension $L \times \mathcal{R}$ be the incidence matrix of routes over links, $\mathbf{R}$ be the

available rate vector indexed by links j and $\tilde{\mathbf{r}}, \mathbf{c}$, be the revenue and cost vectors indexed by routes r.

Now requiring delivery of all admitted content before the deadline is imposing the constraint $\mathbf{A}\mathbf{x}_k \leq (D-k)\mathbf{R}$ for all k. Therefore the feasibility constraint for admitting content vector $\mathbf{y}_k$ is

$$\mathbf{A}\mathbf{y}_k \leq (D-k)\mathbf{R} - \mathbf{A}\mathbf{x}_k.$$

Before writing a DP formulation for the admission decision, we need to address the added complication of the scheduling policy, since the content transmitted for route r in any interval depends on the scheduling policy, which must account for admitted content for all routes in the network, the vector $\mathbf{x}_k$. The key simplifying assumption we make is that in interval k, z_{rk}, the rate allocated to route r is a deterministic quantity that is exogenous input to the admission controller. This models a situation where the admission controller has no knowledge of how the scheduler determines the rates z_{rk}, and assumes that these rates are not affected by its admission decisions. Vector $\mathbf{z}_k$ is then defined as $\mathbf{z}_k = (z_{1k}, \ldots, z_{\mathcal{R}k})^T$. We assume that z_{rk} are integral.

We write two versions of the DP as before since it makes explicit the exact decision model and makes for easier understanding. Below we let $\mathbb{I}(\mathbf{y}_k)$ be the normalized vector $\mathbf{y}_k$, i.e. $\mathbb{I}(\mathbf{y}_k) = 1/(\sum_{r=1}^{\mathcal{R}} y_r)\mathbf{y}_k$. The notation $(\mathbf{x}_k - \mathbf{z}_k)^+$ refers to the non-negative part of the vector $\mathbf{x}_k - \mathbf{z}_k$.

$$J_D \begin{bmatrix} \mathbf{x}_D \\ \mathbf{y}_D \end{bmatrix} = 0, \qquad \forall \mathbf{x}_D, \mathbf{y}_D,$$

and for $k = 0, \ldots, D-1$,

$$J_k \begin{bmatrix} \mathbf{x}_k \\ \mathbf{y}_k \end{bmatrix} = \begin{cases} \max \left\{ \tilde{\mathbf{r}}^T \mathbb{I}(\mathbf{y}_k) + \mathop{\mathbb{E}}_{\mathbf{Y}} J_{k+1} \begin{bmatrix} (\mathbf{x}_k + \mathbf{y}_k - \mathbf{z}_k)^+ \\ \mathbf{Y} \end{bmatrix}, -\mathbf{c}^T \mathbb{I}(\mathbf{y}_k) + \mathop{\mathbb{E}}_{\mathbf{Y}} J_{k+1} \begin{bmatrix} (\mathbf{x}_k - \mathbf{z}_k)^+ \\ \mathbf{Y} \end{bmatrix} \right\}, & \\ & \mathbf{A}\mathbf{y}_k \leq (D-k)\mathbf{R} - \mathbf{A}\mathbf{x}_k, \\ & \mathbf{y}_k^T \mathbf{e} \geq 1, \\ -\mathbf{c}^T \mathbb{I}(\mathbf{y}_k) + \mathop{\mathbb{E}}_{\mathbf{Y}} J_{k+1} \begin{bmatrix} (\mathbf{x}_k - \mathbf{z}_k)^+ \\ \mathbf{Y} \end{bmatrix}, & \text{o/w.} \end{cases}$$

As before, we can re-write the formulation above in a form where the state consists only of the vector $\mathbf{x}_k$. To do this, define the following quantities, $\mathbf{e}_r$ being the rth unit vector.

$$p_{rk}(\mathbf{x}_k) = \Pr\left(\mathbf{Y} \geq \mathbf{e}_r, \ \mathbf{A}\mathbf{Y} \leq (D-k)\mathbf{R} - \mathbf{A}\mathbf{x}_k\right), \qquad \phi_k(\mathbf{x}_k) = \mathop{\mathbb{E}}_{\mathbf{Y}} J_k \begin{bmatrix} \mathbf{x}_k \\ \mathbf{Y} \end{bmatrix}.$$

Now we can let $\mathbf{p}_k(\mathbf{x}_k) = \left(p_{1k}(\mathbf{x}_k), \ldots, p_{\mathcal{R}k}(\mathbf{x}_k)\right)^T$ and write the DP as follows. Here $r_r = \tilde{r}_r + c_r$, $\mathbf{e}$ is a vector of 1's, and $\mathbf{A}_r$ is the rth column of $\mathbf{A}$.

$$\phi_D(\mathbf{x}_D) \equiv 0, \quad \text{and for } k = 0, \ldots, D-1,$$
$$\phi_k(\mathbf{x}_k) = -\mathbf{c}^T\mathbf{p}_k(\mathbf{x}_k) \; + \; \left(1 - \mathbf{e}^T\mathbf{p}_k(\mathbf{x}_k)\right) \phi_{k+1}\left((\mathbf{x}_k - \mathbf{z}_k)^+ \right) +$$
$$\sum_{r=1}^{\mathcal{R}} p_{rk}(\mathbf{x}_k) \underset{Y_r|1\le Y_r,\; Y_r\mathbf{A}_r \le (D-k)R - \mathbf{A}\mathbf{x}_k}{\mathbb{E}} \max\Big\{ \; r_r \; + \; \phi_{k+1}\left((\mathbf{x}_k + Y_r\mathbf{e}_r - \mathbf{z}_k)^+ \right),$$
$$\phi_{k+1}\left((\mathbf{x}_k - \mathbf{z}_k)^+ \right) \Big\}.$$

We can now extend the results of the previous sections.

Lemma 5.4.1. *For every stage k and state $\mathbf{x}_k$, the optimal "expected additional revenue" function is monotone decreasing in $\mathbf{x}_k$, i.e. $\phi_k(\mathbf{x}_k) \ge \phi_k(\mathbf{x}_k + n\mathbf{e}_r), r = 1, \ldots, \mathcal{R}$, for all $n \ge 1$ satisfying $\mathbf{A}(\mathbf{x}_k + n\mathbf{e}_r) \le (D-k)\mathbf{R}$.*

Proof. A relatively straightforward extension of the idea of the single-link proof. Notice that the set of vectors $\mathbf{y}$ for which an admission is feasible in state $\mathbf{x}_k + \mathbf{e}_r$ at stage k is

$$\mathbf{A}y \le (D-k)R - \mathbf{A}(\mathbf{x}_k + \mathbf{e}_r) \le (D-k)R - \mathbf{A}\mathbf{x}_k.$$

Therefore if some $\mathbf{y}$ is admissible in state $\mathbf{x}_k + \mathbf{e}_r$, it is also admissible in state $\mathbf{x}_k$. Then starting from state $(\mathbf{x}_k, \mathbf{y}_k)$, let $\mathbf{x}_k \to \mathbf{x}_k + \mathbf{e}_r$ and follow the policy optimal for starting from state $\mathbf{x}_k + \mathbf{e}_r$. The cost of this policy is $J_k(\mathbf{x}_k + \mathbf{e}_r, \mathbf{y}_k)$. Since the optimal policy must do at least as well, we have

$$J_k\begin{bmatrix}\mathbf{x}_k \\ \mathbf{y}_k\end{bmatrix} \ge J_k\begin{bmatrix}\mathbf{x}_k + \mathbf{e}_r \\ \mathbf{y}_k\end{bmatrix}, \qquad \forall \mathbf{y}_k.$$

Expectation over $\mathbf{y}_k$ gives

$$\phi_k(\mathbf{x}_k) \ge \phi_k(\mathbf{x}_k + \mathbf{e}_r).$$

□

Theorem 5.4.2. *If for each r, the file size random variables Y_r can take values $0, \ldots, \infty$, then the optimal admission policy consists of a work-dependent threshold job size $y^*_{rk}(\mathbf{x}_k)$ for each route r and stage k, such that one admits a job of size y_r iff $y_r \le y^*_{rk}(\mathbf{x}_k)$.*

Proof. Consider the admission decision for a request of size y_r for route r. The job is admitted only when

$$r_r \; + \; \phi_{k+1}\left((\mathbf{x}_k + y_r\mathbf{e}_r - \mathbf{z}_k)^+ \right) \; \ge \; \phi_{k+1}\left((\mathbf{x}_k - \mathbf{z}_k)^+ \right).$$

Now letting $\bar{y}_r(\mathbf{x}_k) = \min_j (D-k)R_j - \sum_{r=1}^{\mathcal{R}} a_{jr}x_r$, we can see that there are only three possibilities because of lemma 5.4.1.

$$r_r + \; \phi_{k+1}\left((\mathbf{x}_k + \mathbf{e}_r - \mathbf{z}_k)^+ \right) \; < \; \phi_{k+1}\left((\mathbf{x}_k - \mathbf{z}_k)^+ \right) \quad \Rightarrow \text{No job can be admitted,}$$
$$r_r + \; \phi_{k+1}\left((\mathbf{x}_k + \bar{y}_r(\mathbf{x}_k)\mathbf{e}_r - \mathbf{z}_k)^+ \right) \; \ge \; \phi_{k+1}\left((\mathbf{x}_k - \mathbf{z}_k)^+ \right) \quad \Rightarrow \text{Admit a job if it fits.}$$

Otherwise there must be a $y^*_{rk}(\mathbf{x}_k) > 0$ and $y^*_{rk}(\mathbf{x}_k) \le \bar{y}_r(\mathbf{x}_k)$, such that it is optimal to accept a job of size y_r for all values $y_r \le y^*_{rk}(\mathbf{x}_k)$, and to reject otherwise.

□

5.4.2 Multiple deadlines

The case of multiple deadlines on a single link is very similar to a network with fixed routing, with a polyhedral region characterizing the set of feasible decisions. The only modifications are that the terminal cost function is not 0 for deadlines before the last, and we let the scheduling policy depend explicitly on the state of the system. Results obtained are similar to before. The optimal revenue function is found to be monotone decreasing, and the optimal policy remains of threshold-type. Instead of handling an arbitrary number of deadlines, we formulate and analyze an instance with only two consecutive deadlines. The lines of reasoning are fully revealed by this case.

Consider two *equally spaced deadlines* on a single-link, with the same available transmission rate for both deadlines. The model maintains information about undelivered content for both deadlines. As before, we require that all admitted content be transmitted completely before its associated deadline. Further, we assume that the scheduling policy is *non-idling* and *infinitely pre-emptible*, where transmission for the second deadline is pre-empted if work is accepted for the first deadline, and resumes only when there is no remaining work for the first deadline. Below is a a discretized formulation of the problem. We do not provide details of the notation, since it is a natural extension of earlier notation.

Let $i = 1, 2$, index deadlines with 1 being the earlier deadline. Arrivals for the two deadlines occur as independent Poisson arrival processes with rates λ_i and arrivals for deadline 1 cease to occur after the deadline expires. Call x_{ik} the untransmitted content and y_{ik} the size of the job for which an admission decision is needed in stage k, with $\tilde{r}_i, c_i$ the revenue and rejection cost for a job associated with deadline i.

Now if $D \leq k \leq 2D$, the problem involves only remaining content x_{2k} and arriving content y_{2k} for the second deadline, the model reduces to our canonical model of section 5.2, and all results for the optimal policy apply, allowing us to write:

$$J_D \begin{bmatrix} x_{1D} \\ x_{2D} \\ y_{1D} \\ y_{2D} \end{bmatrix} = \tilde{J}_0 \begin{bmatrix} x_{2D} \\ y_{2D} \end{bmatrix} \qquad \forall x_{1D}, x_{2D}, y_{1D}, y_{2D},$$

where $\tilde{J}_0$ is the optimal cost function obtained from the canonical model of section 5.2. We therefore consider only $0 \leq k \leq D$ in this section. The DP formulation can now be written as follows, noting that at most one of y_{1k} and y_{2k} can be non-zero in any given stage. $\mathbb{I}_S$ is the indicator function of set S and Y_i are the random arriving file sizes in stage k which will await decision in stage $k+1$. We define $\bar{y}_{2k}(x_{1k}, x_{2k}) = 2D - k - x_{1k} - x_{2k}$ and $\bar{y}_{1k}(x_{1k}, x_{2k}) = \min(\bar{y}_{2k},\ D - k - x_{1k})$,

$$J_D \begin{bmatrix} x_{1D} \\ x_{2D} \\ y_{1D} \\ y_{2D} \end{bmatrix} = \tilde{J}_0 \begin{bmatrix} x_{Dk} \\ y_{Dk} \end{bmatrix}, \qquad \forall x_{1D}, x_{2D}, y_{1D}, y_{2D},$$

and for $k = 0, \ldots, D-1$,

$$J_k \begin{bmatrix} x_{1k} \\ x_{2k} \\ y_{1k} \\ y_{2k} \end{bmatrix} = \begin{cases} \max \Bigg\{ \sum_{i=1}^{2} \tilde{r}_i \mathbb{I}_{\{y_{ik}>0\}} + \mathop{\mathbb{E}}_{Y_1,Y_2} J_{k+1} \begin{bmatrix} \max(0, x_{1k} + y_{1k} - 1) \\ \max(0, x_{2k} + y_{2k} - \mathbb{I}_{\{x_{1k}+y_{1k}=0\}}) \\ Y_1 \\ Y_2 \end{bmatrix}, \\ \quad - \sum_{i=1}^{2} c_i \mathbb{I}_{\{y_{ik}>0\}} + \mathop{\mathbb{E}}_{Y_1,Y_2} J_{k+1} \begin{bmatrix} \max(0, x_{1k} - 1) \\ \max(0, x_{2k} - \mathbb{I}_{\{x_{1k}=0\}}) \\ Y_1 \\ Y_2 \end{bmatrix} \Bigg\}, \\ \qquad\qquad 1 \le y_{1k} \le \bar{y}_{1k}(x_{1k}, x_{2k}), \\ \qquad\qquad 1 \le y_{2k} \le \bar{y}_{2k}(x_{1k}, x_{2k}), \\ - c_1 \mathbb{I}_{\{y_{1k}>\bar{y}_{1k}(x_{1k},x_{2k})\}} - c_2 \mathbb{I}_{\{y_{2k}>\bar{y}_{2k}(x_{1k},x_{2k})\}} + \mathop{\mathbb{E}}_{Y_1,Y_2} J_{k+1} \begin{bmatrix} \max(0, x_{1k} - 1) \\ \max(0, x_{2k} - \mathbb{I}_{\{x_{1k}=0\}}) \\ Y_1 \\ Y_2 \end{bmatrix}, \\ \qquad\qquad \text{o/w.} \end{cases}$$

We now define:

$$p_1 \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} = \Pr(1 \le Y_1 \le \bar{y}_{1k}(x_{1k}, x_{2k})), \qquad p_2 \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} = \Pr(1 \le Y_2 \le \bar{y}_{2k}(x_{1k}, x_{2k})),$$

$$\phi_k \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} = \mathop{\mathbb{E}}_{Y_1,Y_2} J_k \begin{bmatrix} x_{1k} \\ x_{2k} \\ Y_1 \\ Y_2 \end{bmatrix}.$$

The DP is then re-formulated as follows, where $r_i = \tilde{r}_i + c_i$ and $\tilde{\phi}_D$ is the optimal additional revenue function in stage D, for the remaining single-deadline problem.

$$\phi_D \begin{bmatrix} x_{1D} \\ x_{2D} \end{bmatrix} \equiv \tilde{\phi}_0(x_{2D}) \quad \text{and for } k = 0, \ldots, D-1,$$

$$\phi_k \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} = -c_1 \Pr(Y_1 > 0) - c_2 \Pr(Y_2 > 0) +$$

$$\left(1 - p_1 \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} - p_2 \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix}\right) \phi_{k+1} \begin{bmatrix} \max(0, x_{1k} - 1) \\ \max(0, x_{2k} - \mathbb{I}_{\{x_{1k}=0\}}) \end{bmatrix} +$$

$$p_1 \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix}$$
$$\mathop{\mathbb{E}}_{Y_1 | 1 \le Y_1 \le \bar{y}_{1k}(x_{1k}, x_{2k})} \max \left\{ r_1 + \phi_{k+1} \begin{bmatrix} \max(0, x_{1k} + Y_1 - 1) \\ \max(0, x_{2k}) \end{bmatrix}, \ \phi_{k+1} \begin{bmatrix} \max(0, x_{1k} - 1) \\ \max(0, x_{2k} - \mathbb{I}_{\{x_{1k}=0\}}) \end{bmatrix} \right\} +$$

$$p_2 \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix}$$
$$\mathop{\mathbb{E}}_{Y_2 | 1 \le Y_2 \le \bar{y}_{2k}(x_{1k}, x_{2k})} \max \left\{ r_2 + \phi_{k+1} \begin{bmatrix} \max(0, x_{1k} - 1) \\ \max(0, x_{2k} + Y_2 - \mathbb{I}_{\{x_{1k}=0\}}) \end{bmatrix}, \ \phi_{k+1} \begin{bmatrix} \max(0, x_{1k} - 1) \\ \max(0, x_{2k} - \mathbb{I}_{\{x_{1k}=0\}}) \end{bmatrix} \right\}.$$

Now the following analogues of earlier results can be shown.

Lemma 5.4.3. *For every k, the optimal "expected additional revenue" function is monotone decreasing in x_{1k} and x_{2k}, i.e. for all x_{1k}, x_{2k} satisfying $x_{1k} + x_{2k} \le 2D - k - 1$, $x_{1k} \le D - k - 1$,*

$$\phi_k \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} \ge \phi_k \begin{bmatrix} x_{1k} + 1 \\ x_{2k} \end{bmatrix},$$
$$\phi_k \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} \ge \phi_k \begin{bmatrix} x_{1k} \\ x_{2k} + 1 \end{bmatrix}.$$

Proof. As before, here the definition of an optimal policy starting from any state $(x_{1k}, x_{2k}, y_{1k}, y_{2k})$ is that it is optimal among all *admissible* policies $\pi = \{\mu_k, \mu_{k+1}, \ldots, \mu_{D-1}\}$ consisting of sequence of functions μ_k such that

$$\mu_k \begin{bmatrix} x_{1k} \\ x_{2k} \\ y_{1k} \\ y_{2k} \end{bmatrix} \in \begin{cases} \{0,1\}, & 1 \le y_{1k} \le \min(2D - k - x_{1k} - x_{2k}, D - k - x_{1k}) \\ & 1 \le y_{2k} \le 2D - k - x_{1k} - x_{2k}, \\ \{0\} & \text{o/w.} \end{cases}$$

Then the following is an obvious consequence.

$$1 \le y_{1k} \le \min\left(2D-k-(x_{1k}+1)-x_{2k}, D-k-(x_{1k}+1)\right) < \min(2D-k-x_{1k}-x_{2k}, D-k-x_{1k})$$
$$1 \le y_{2k} \le 2D-k-(x_{1k}+1)-x_{2k} < 2D-k-x_{1k}-x_{2k},$$

and

$$1 \le y_{1k} \le \min\left(2D-k-x_{1k}-(x_{2k}+1), D-k-x_{1k}\right) \le \min(2D-k-x_{1k}-x_{2k}, D-k-x_{1k})$$
$$1 \le y_{2k} \le 2D-k-x_{1k}-(x_{2k}+1) < 2D-k-x_{1k}-x_{2k}.$$

Therefore if μ_k is admissible for state $(x_{1k}+1, x_{2k}, y_{1k}, y_{2k})$ or for state $(x_{1k}, x_{2k}+1, y_{1k}, y_{2k})$, it is admissible for state $(x_{1k}, x_{2k}, y_{1k}, y_{2k})$.

Now starting from any state $(x_{1k}, x_{2k}, y_{1k}, y_{2k})$, let $x_{1k} \to x_{1k}+1$, and follow the policy optimal for starting from $x_{1k}+1$, always possible owing to the definition of admissible functions, defined above. The cost of this policy is $J_k(x_{1k}+1, x_{2k}, y_{1k}, y_{2k})$. Since the optimal policy must be at least as good, we can take expectations over y_{1k}, y_{2k} to conclude

$$\phi_k \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} \ge \phi_k \begin{bmatrix} x_{1k}+1 \\ x_{2k} \end{bmatrix}.$$

Follow exactly the same reasoning to see that

$$\phi_k \begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} \ge \phi_k \begin{bmatrix} x_{1k} \\ x_{2k}+1 \end{bmatrix}.$$

□

Theorem 5.4.4. *If the file size random variables Y_i can take values $0, \ldots, \infty$, the optimal admission policy consists of a work-dependent threshold job size $y^*_{ik}(x_{1k}, x_{2k})$ for each deadline $i = 1, 2$, such that one admits a job of size y_{ik} iff $y_{ik} \le y^*_{ik}(x_{1k}, x_{2k})$.*

Proof. Consider the admission decision for a job for deadline 1. A job of size y_{1k} is admitted iff $1 \le y_{1k} \le \min(2D-k-x_{1k}-x_{2k}, D-k-x_{1k}) = \bar{y}_{1k}(x_{1k}, x_{2k})$ and

$$r_1 + \phi_{k+1}\begin{bmatrix} x_{1k}+y_{1k}-1 \\ x_{2k} \end{bmatrix} \ge \phi_{k+1}\begin{bmatrix} \max(0, x_{1k}-1) \\ x_{2k}-\mathbb{I}_{\{x_{1k}=0\}} \end{bmatrix},$$

Now there are only three possibilities as outlined below, using the monotonicity of ϕ_{k+1} from proposition 5.4.3.

$$r_1 + \phi_{k+1}\begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix} \le \phi_{k+1}\begin{bmatrix} \max(0, x_{1k}-1) \\ x_{2k}-\mathbb{I}_{\{x_{1k}=0\}} \end{bmatrix}, \quad \Rightarrow y^*_{1k}(x_{1k}, x_{2k}) = 0,$$
$$r_1 + \phi_{k+1}\begin{bmatrix} x_{1k}+\bar{y}_{1k}(x_{1k}, x_{2k})-1 \\ x_{2k} \end{bmatrix} \ge \phi_{k+1}\begin{bmatrix} \max(0, x_{1k}-1) \\ x_{2k}-\mathbb{I}_{\{x_{1k}=0\}} \end{bmatrix}, \quad \Rightarrow y^*_{1k}(x_{1k}, x_{2k}) = \bar{y}_{1k}(x_{1k}, x_{2k}).$$

Otherwise, if

$$r_1 + \phi_{k+1}\begin{bmatrix} x_{1k}+\bar{y}_{1k}(x_{1k}, x_{2k})-1 \\ x_{2k} \end{bmatrix} < \phi_{k+1}\begin{bmatrix} \max(0, x_{1k}-1) \\ x_{2k}-\mathbb{I}_{\{x_{1k}=0\}} \end{bmatrix} < r_1 + \phi_{k+1}\begin{bmatrix} x_{1k} \\ x_{2k} \end{bmatrix},$$

we must have a $y^*_{1k}(x_{1k}, x_{2k})$ with $0 < y^*_{1k}(x_{1k}, x_{2k}) < \bar{y}_{1k}(x_{1k}, x_{2k})$, such that for all $y_{1k} \leq y^*_{1k}(x_{1k}, x_{2k})$, the job is admitted, and for all greater y_{1k}, it is rejected.

Following exactly similar reasoning, using the monotonicity of ϕ_{k+1} from proposition 5.4.3 we can prove the existence of $y^*_{2k}(x_{1k}, x_{2k})$.

□

5.4.3 Remarks

This section lists some remarks about the network and multiple deadline formulations.

1. Section 5.4.2 suggests that a multiple deadline problem can always be viewed as a fixed routing network problem with a terminal cost function different from 0. This is because the polyhedral region corresponding to the set of feasible decisions during any inter-deadline interval corresponds to a network topology with a single deadline.

2. The above remark and sections 5.4.1 and 5.4.2 suggest that a proof for optimality of threshold-type policies should be available for fixed routing networks with multiple deadlines.

3. It further seems that the only properties that were essential to proving the optimality of threshold policies were that the files sizes take values $0, \ldots, \infty$ and that the scheduling rule be either fixed or independent of the admission decisions. This leads to the speculation that the results above can be extended to the case when capacity is either time-varying or stochastic, as long as it is independent of the admission decisions.

4. There seems little hope of being actually able to compute the optimal thresholds. The more interesting question raised is thus of finding heuristic rules for approximating the behavior of the thresholds. We have attempted this for the single-link case, in section 5.2.5, but the case of a network remains open.

5. It also seems easily provable that multiple classes of customers should have nested thresholds on a route-by-route basis, i.e. if multiple classes of customers paying differently request to use the same route, one should obtain separate thresholds for each class, with the threshold for the highest paying customer the highest, and likewise ordered for all classes. This can be seen by following the proof for existence of nested thresholds for the single link case, in section 5.3.3.

6. Finally, for the multiple deadline case, we wonder if the existence of the threshold-type policy can be extended to arbitrarily small inter-deadline intervals, to obtain an existence result for the case when deadlines are continuous random variables. We do not pursue this direction in this work.

5.5 Proofs for the canonical model

This section collects proofs for the remarks listed in section 5.2.3.

Proposition 5.5.1 (Monotonicity in remaining work). *For every k, the optimal "expected additional revenue" function $\phi_k(x_k), x_k = 0, \ldots, D-k$, is monotone non-increasing in x_k.*

Proof. Two alternative proofs are noted here. The first is more compact while the second is more mechanical and uses induction on the functions $\phi_k(z)$.

The first proof relies on the viewing a finite horizon dynamic program as maximization of the additional expected revenue over all admissible policies (c.f. [Ber95]), starting from a given initial state. Here, we define the optimal additional revenue function as

$$J_k\begin{bmatrix}x_k\\y_k\end{bmatrix} = \max_{\pi\in\Pi} J_k^{\pi}\begin{bmatrix}x_k\\y_k\end{bmatrix},$$

where Π is the set of all *admissible* policies, and an *admissible* policy $\pi = \{\mu_k, \mu_{k+1}, \dots, \mu_{D-1}\}$ is a sequence of functions μ_k that map states (x_k, y_k) into sets $U_k(x_k, y_k)$ as defined below. $J_k^{\pi}(x_k, y_k)$ is the expected additional revenue starting from state (x_k, y_k) and following policy π.

$$U_k(x_k, y_k) = \begin{cases}\{0,1\}, & y_k \geq 1 \text{ and } x_k + y_k \leq D - k,\\ \{0\} & \text{o/w.}\end{cases}$$

Now notice that the definition of admissibility implies that a policy π admissible when starting from state $(x_k + 1, y_k)$ is also admissible when starting from state (x_k, y_k).

Then given state (x_k, y_k), let $x_k \to x_k + 1$, and follow the policy optimal for starting from state $(x_k + 1, y_k)$. Clearly this produces a feasible set of decisions for every sequence of arrivals and has cost $J_k(x_k + 1, y_k)$. An optimal policy must do at least as well starting from (x_k, y_k) and since J_k's are the optimal costs, for any $x_k \leq D - k - 1$,

$$J_k\begin{bmatrix}x_k\\y_k\end{bmatrix} \geq J_k\begin{bmatrix}x_k+1\\y_k\end{bmatrix}, \quad \forall y_k.$$

Taking expectation over y_k yields

$$\phi_k(x_k) \geq \phi_k(x_k + 1).$$

□

An alternative proof using induction

We use induction on the monotonicity of the functions $\phi_k(z)$. The proposition is trivially true for $\phi_D(z)$ since it is the zero function. Now suppose that for stage $k+1$, $\phi_{k+1}(z) \geq \phi_{k+1}(z+1), z = 0, \dots, D - k - 1$.

Rewrite the expressions for $\phi_k(0)$ and $\phi_k(1)$ as follows:

$$\phi_k(0) = -c\lambda\delta \;+\; \big(1 - p_k(0)\big)\;\phi_{k+1}(0) \;+\; \sum_{y=1}^{D-k} \max\big\{r + \phi_{k+1}(y-1),\; \phi_{k+1}(0)\big\}\Pr(Y = y),$$

$$\phi_k(1) = -c\lambda\delta \;+\; \big(1 - p_k(1)\big)\;\phi_{k+1}(0) \;+\; \sum_{y=1}^{D-k-1} \max\big\{r + \phi_{k+1}(y),\; \phi_{k+1}(0)\big\}\Pr(Y = y),$$

simply writing $p_k(0)\,\mathbb{E}_{Y|1\leq Y\leq D-k} \max\big\{r + \phi_{k+1}(Y-1),\; \phi_{k+1}(0)\big\}$ as $\sum_{y=1}^{D-k} \max\big\{r + \phi_{k+1}(y-1),\; \phi_{k+1}(0)\big\}\Pr(Y = y)$. Then,

$$\begin{aligned}&\phi_k(0) - \phi_k(1) =\\ &\quad \phi_{k+1}(0)\big[p_k(1) - p_k(0)\big] +\\ &\quad \sum_{y=1}^{D-k-1} \Pr(Y = y)\Big(\max\big\{r + \phi_{k+1}(y-1),\; \phi_{k+1}(0)\big\} - \max\big\{r + \phi_{k+1}(y),\; \phi_{k+1}(0)\big\}\Big) +\\ &\quad \Pr(Y = D - k)\max\big\{r + \phi_{k+1}(D-k-1),\; \phi_{k+1}(0)\big\}.\end{aligned}$$

Noting that $p_k(1) - p_k(0) = -\Pr(Y = D-k)$ and $\phi_{k+1}(D-k-1) = -c\lambda(D-k-1)$, we rewrite:

$$\begin{aligned}
&\phi_k(0) - \phi_k(1) = \\
&\quad \Pr(Y = D-k)\Big(\max\{r - c\lambda(D-k-1),\ \phi_{k+1}(0)\} - \phi_{k+1}(0) \Big) + \\
&\quad \sum_{y=1}^{D-k-1} \Pr(Y = y)\Big(\max\{r + \phi_{k+1}(y-1),\ \phi_{k+1}(0)\} - \max\{r + \phi_{k+1}(y),\ \phi_{k+1}(0)\} \Big) \\
&\quad \geq 0.
\end{aligned}$$

The inequality on the last line follows since each term weighted by the probabilities is ≥ 0, either by monotonicity of $\phi_{k+1}(z)$ or by a simple elementary argument.

Similarly, we can write for $z \geq 1$,

$$\phi_k(z) = -c\lambda\delta \ + \ \big(1 - p_k(z)\big)\ \phi_{k+1}(z-1) \ + \sum_{y=1}^{D-k-z} \max\{r + \phi_{k+1}(z+y-1),\ \phi_{k+1}(z-1)\} \Pr(Y = y),$$

$$\begin{aligned}
\phi_k(z+1) &= -c\lambda\delta \ + \ \big(1 - p_k(z+1)\big)\ \phi_{k+1}(z) \ + \sum_{y=1}^{D-k-z-1} \max\{r + \phi_{k+1}(z+y),\ \phi_{k+1}(z)\} \Pr(Y = y) \\
&\leq \ -c\lambda\delta \ + \ \big(1 - p_k(z+1)\big)\ \phi_{k+1}(z-1) + \sum_{y=1}^{D-k-z-1} \max\{r + \phi_{k+1}(z+y),\ \phi_{k+1}(z-1)\} \Pr(Y = y).
\end{aligned}$$

The last inequality for $\phi_k(z+1)$ follows from the inductive hypothesis. Now exactly as before, we see that $\phi_k(z) - \phi_k(z+1) \geq 0$.

□

Proposition 5.5.2 (Monotonicity in time). *For every x, $\phi_k(x) \geq \phi_{k+1}(x)$.*

Proof. For $x = 0$,

$$\begin{aligned}
\phi_k(0) &= -c\lambda\delta \ + \ \big(1 - p_k(0)\big)\ \phi_{k+1}(0) \ + \sum_{y=1}^{D-k} \max\{r + \phi_{k+1}(y-1),\ \phi_{k+1}(0)\} \Pr(Y = y), \\
&\geq \big(1 - p_k(0)\big)\ \phi_{k+1}(0) \ + \sum_{y=1}^{D-k} \phi_{k+1}(0) \Pr(Y = y) \\
&= \ \phi_{k+1}(0).
\end{aligned}$$

For $x \geq 1$,

$$\begin{aligned}
\phi_k(z) &= -c\lambda\delta \;+\; \big(1-p_k(z)\big)\;\phi_{k+1}(z-1) \;+\; \sum_{y=1}^{D-k-z} \max\big\{r+\phi_{k+1}(z+y-1),\;\phi_{k+1}(z-1)\big\}\Pr(Y=y),\\
&\geq \big(1-p_k(z)\big)\;\phi_{k+1}(z) \;+\; \sum_{y=1}^{D-k-z} \phi_{k+1}(z)\Pr(Y=y),\\
&= \phi_{k+1}(z),
\end{aligned}$$

where the inequality on the second line above follows from proposition 5.5.1, $\phi_{k+1}(z-1) \geq \phi_{k+1}(z)$.

□

Proposition 5.5.3 (Incremental revenue in both work and time bounded by r). *For any x, k,* (i) $\phi_k(x) \leq r + \phi_k(x+1)$, *and* (ii) $\phi_k(x) \leq r + \phi_{k+1}(x)$.

Proof. Consider (i), for $x = 0$, we immediately have

$$\begin{aligned}
&\phi_k(0) - \phi_k(1)\\
&= \phi_{k+1}(0)\big[p_k(1) - p_k(0)\big] \;+\; \Pr(Y = D-k)\max\big\{r+\phi_{k+1}(D-k-1),\;\phi_{k+1}(0)\big\} \;+\\
&\qquad \sum_{y=1}^{D-k-1} \Pr(Y=y)\Big(\max\big\{r+\phi_{k+1}(y-1),\;\phi_{k+1}(0)\big\} - \max\big\{r+\phi_{k+1}(y),\;\phi_{k+1}(0)\big\}\Big)\\
&\leq \Pr(Y = D-k)\Big(\max\big\{r+\phi_{k+1}(D-k-1),\;\phi_{k+1}(0)\big\} - \phi_{k+1}(0)\Big) \;+\; \sum_{y=1}^{D-k-1} r\;\Pr(Y=y)\\
&\leq \sum_{y=1}^{D-k} r\;\Pr(Y=y) \;\leq\; r.
\end{aligned}$$

The first inequality above follows from the facts: $p_k(1) - p_k(0) = -\Pr(Y = D-k)$ and $\phi_{k+1}(D-k-1) = -c\lambda(D-k-1)$, and that for every $y \geq 1$,

$$\begin{aligned}
&\max\big\{r+\phi_{k+1}(y),\;\phi_{k+1}(0)\big\} \quad \geq \phi_{k+1}(0) \qquad \text{and}\\
&\max\big\{r+\phi_{k+1}(y-1),\;\phi_{k+1}(0)\big\} \leq r+\phi_{k+1}(0) \quad \text{using proposition 5.5.1,}\\
\Rightarrow &\max\big\{r+\phi_{k+1}(y-1),\;\phi_{k+1}(0)\big\} - \max\big\{r+\phi_{k+1}(y),\;\phi_{k+1}(0)\big\} \;\leq\; r.
\end{aligned}$$

When $x \geq 1$, we use induction on the functions ϕ_k. (i) is trivially true for $\phi_D(x)$ since it is the 0 function. Now assume that $\phi_{k+1}(x) \leq r + \phi_{k+1}(x+1)$ for all $x \geq 0$, then noting that $1 - p_k(x+1) = 1 - p_k(x) + \Pr(Y = D-k-x)$, we can write,

$$\phi_k(x) - \phi_k(x+1) =$$
$$\big(\phi_{k+1}(x-1) \;-\; \phi_{k+1}(x)\big)\big[1 - p_k(x)\big] \;+$$
$$\sum_{y=1}^{D-k-x-1} \Pr(Y=y)\Big(\max\big\{r + \phi_{k+1}(x+y-1),\; \phi_{k+1}(x-1)\big\} \;-$$
$$\max\big\{r + \phi_{k+1}(x+y),\; \phi_{k+1}(x)\big\}\Big) +$$
$$\Pr(Y = D-k-x)\Big(\max\big\{r + \phi_{k+1}(D-k-1),\; \phi_{k+1}(x-1)\big\} - \phi_{k+1}(x)\Big).$$

Note each term weighted by the probabilities is $\leq r$ by the inductive hypothesis – using an argument similar to before for the middle term, i.e. for $y \geq 1$,

$$\begin{aligned}
&\max\big\{r + \phi_{k+1}(x+y),\; \phi_{k+1}(x)\big\} && \geq \phi_{k+1}(x) \\
&\max\big\{r + \phi_{k+1}(x+y-1),\; \phi_{k+1}(x-1)\big\} && \leq \max\big\{r + \phi_{k+1}(x), \phi_{k+1}(x-1)\big\} \\
\Rightarrow &\max\big\{r + \phi_{k+1}(x+y-1),\; \phi_{k+1}(x-1)\big\} - \max\big\{r + \phi_{k+1}(x+y),\; \phi_{k+1}(x)\big\} && \leq\; r
\end{aligned}$$

This proves (i).

Now consider (ii) $\phi_k(x) \leq r + \phi_{k+1}(x)$. When $x = 0$, the following is true for $y \geq 1$, using $r + \phi_{k+1}(y-1) \leq r + \phi_{k+1}(0)$ by proposition 5.5.1.

$$\begin{aligned}
\phi_k(0) &= -c\lambda\delta \;+\; \big(1 - p_k(0)\big)\;\phi_{k+1}(0) \;+ \sum_{y=1}^{D-k} \max\big\{r + \phi_{k+1}(y-1),\; \phi_{k+1}(0)\big\}\Pr(Y=y), \\
&\leq \big(1 - p_k(0)\big)\;\big(r + \phi_{k+1}(0)\big) \;+ \sum_{y=1}^{D-k} \big(r + \phi_{k+1}(0)\big)\Pr(Y=y) \\
&= \; r + \phi_{k+1}(0).
\end{aligned}$$

When $x \geq 1$, we use (i) for $\phi_{k+1}(x-1) \leq r + \phi_{k+1}(x)$ and proposition 5.5.1 for $r + \phi_{k+1}(x+y-1) \leq r + \phi_{k+1}(x)$ to show the following:

$$\begin{aligned}
\phi_k(x) &= -c\lambda\delta \;+\; \big(1 - p_k(x)\big)\;\phi_{k+1}(x-1) \;+ \sum_{y=1}^{D-k-x} \max\big\{r + \phi_{k+1}(x+y-1),\; \phi_{k+1}(x-1)\big\}\Pr(Y=y), \\
&\leq \big(1 - p_k(x)\big)\;\big(r + \phi_{k+1}(x)\big) \;+ \sum_{y=1}^{D-k-x} \big(r + \phi_{k+1}(x)\big)\Pr(Y=y), \\
&= r + \phi_{k+1}(x),
\end{aligned}$$

□

Proposition 5.5.4 (Convergence of the time-average additional revenue). *For any given x,*

$$\lim_{(D-k)\to-\infty} \frac{\phi_k(x)}{D-k} = \frac{\phi_{k+1}(x)}{D-k-1} = \frac{\phi_k(x+1)}{D-k},$$

Proof. To see the first equality, note from propositions 5.5.2 and 5.5.3 that $\phi_{k+1}(x) \le \phi_k(x) \le r + \phi_{k+1}(x)$, which can be written as

$$\frac{\phi_{k+1}(x)}{D-k-1}\frac{D-k-1}{D-k} \le \frac{\phi_k(x)}{D-k} \le \frac{r}{D-k} + \frac{\phi_{k+1}(x)}{D-k-1}\frac{D-k-1}{D-k}.$$

Letting $(D-k) \to -\infty$, we get the desired result.

Similarly, to see the second equality, we can write $\phi_k(x+1) \le \phi_k(x) \le r + \phi_k(x+1)$ from propositions 5.5.1 and 5.5.3. This gives us

$$\frac{\phi_k(x+1)}{D-k} \le \frac{\phi_k(x)}{D-k} \le \frac{r}{D-k} + \frac{\phi_k(x+1)}{D-k}.$$

The result follows by letting $(D-k) \to -\infty$.

□

Proposition 5.5.5 (Optimality of a non-trivial threshold policy). *For any x_k, $0 \le x_k \le D-k$, the optimal admission control policy is characterized by a threshold job-size $y_k^*(x_k) \le D-k-x_k$ such that one admits a job of size y_k in period k iff $y_k \le y_k^*(x_k)$.*

Proof. From the DP recursion (5.1), one admits a job of size y_k, $1 \le y_k \le D-k-x_k$ in stage k iff

$$r + \phi_{k+1}\big(\max(0, x_k + y_k - 1)\big) \ge \phi_{k+1}\big(\max(0, x_k - 1)\big).$$

This condition, written separately for $x_k = 0$ and $x_k \ge 1$ is:

$$\begin{aligned} x_k = 0: &\quad r + \phi_{k+1}(y_k - 1) \ge \phi_{k+1}(0), \\ x_k \ge 1: &\quad r + \phi_{k+1}(x_k + y_k - 1) \ge \phi_{k+1}(x_k - 1). \end{aligned}$$

View the condition for a fixed $x_k \ge 1$ as a function of y_k and consider the case $x_k \ge 1$ since the same argument holds for $x_k = 0$.

At $y_k = 1$, $r + \phi_{k+1}(x_k + y_k - 1) = r + \phi_{k+1}(x_k) \ge \phi_{k+1}(x_k - 1)$ from proposition 5.5.3, part (i). At $y_k = D-k-x_k$, we have $r + \phi_{k+1}(x_k + y_k - 1) = r - c\lambda(D-k-1)$ since $\phi_{k+1}(D-k-1) = -c\lambda(D-k-1)$. In addition, from proposition 5.5.1, $\phi_{k+1}(x_k + y_k - 1)$ is monotone non-increasing in y_k.

It follows that if $\phi_{k+1}(x_k - 1) \le r - c\lambda(D-k-1)$, one accepts a job if it fits. Otherwise $\exists\, y_k^*(x_k) < D-k-x_k$ which is the largest y_k such that $r + \phi_{k+1}(x_k + y_k - 1) > \phi_{k+1}(x_k - 1)$. From monotonicity of ϕ_{k+1}, for all $y_k(x_k) \le y_k^*(x_k)$, the optimal decision is *accept* while for $y_k(x_k) > y_k^*(x_k)$ the optimal decision is *reject*. Combining the two cases, the optimal policy consists of a threshold job-size $y_k^*(x_k) \le D-k-x_k$ such that jobs are accepted *iff* $y_k(x_k) \le y_k^*(x_k) \le D-k-x_k$.

□

Proposition 5.5.6. $y_k^*(x_k) \le y_k^*(x_k+1) + 1$ *for all k and $0 \le x_k \le D-k-1$.*

Proof. We consider $1 \leq x_k \leq D-k-1$. The same argument is valid for $x_k = 0$. By definition of the thresholds $y_k^*(x_k)$, we have

$$\begin{aligned} r + \phi_{k+1}(x_k + y_k^*(x_k) - 1) &\geq \phi_{k+1}(x_k - 1), \\ r + \phi_{k+1}(x_k + y_k^*(x_k)) &< \phi_{k+1}(x_k - 1). \end{aligned} \tag{2}$$

Similarly, for $x_k + 1$,

$$\begin{aligned} r + \phi_{k+1}(x_k + y_k^*(x_k + 1)) &\geq \phi_{k+1}(x_k), \\ r + \phi_{k+1}(x_k + y_k^*(x_k + 1) + 1) &< \phi_{k+1}(x_k). \end{aligned} \tag{3}$$

Now suppose $y_k^*(x_k) > y_k^*(x_k + 1) + 1$, which is equivalent to $y_k^*(x_k) - 1 \geq y_k^*(x_k + 1) + 1$, then we must have, from proposition 5.5.1,

$$r + \phi_{k+1}(x_k + y_k^*(x_k) - 1) \leq r + \phi_{k+1}(x_k + y_k^*(x_k + 1) + 1),$$

but then the definitions of $y_k^*(x_k)$ and $y_k^*(x_k + 1)$ imply the following,

$$\phi_{k+1}(x_k - 1) \leq r + \phi_{k+1}(x_k + y_k^*(x_k) - 1) \leq r + \phi_{k+1}(x_k + y_k^*(x_k + 1) + 1) < \phi_{k+1}(x_k),$$

which is a contradiction to proposition 5.5.1.

□

Proposition 5.5.7 (Sufficient condition for monotonicity of the thresholds). *If for every k, $\phi_k(x)$ are concave in $x, 0 \leq x \leq D-k$, then $y_k^*(x_k) \geq y_k^*(x_k + 1)$ for all x_k.*

Proof. First note that the proposition is true for $x = 0$ regardless of the structure of $\phi_k(x)$, and in fact $y_k^*(0) = 1 + y_k^*(1)$. This follows from the definition of $y_k^*(1)$, which is

$$\begin{aligned} r + \phi_{k+1}(y_k^*(1)) &\geq \phi_{k+1}(0), \\ r + \phi_{k+1}(y_k^*(1) + 1) &< \phi_{k+1}(0). \end{aligned}$$

Which can be written as

$$\begin{aligned} r + \phi_{k+1}\big((y_k^*(1) + 1) - 1 \big) &\geq \phi_{k+1}(0), \\ r + \phi_{k+1}\big((y_k^*(1) + 2) - 1 \big) &< \phi_{k+1}(0), \end{aligned}$$

and therefore $y_k^*(1) + 1$ is the largest y satisfying $r + \phi_{k+1}(y - 1) \geq \phi_{k+1}(0)$, which happens to be the definition of $y_k^*(0)$.

For $x \geq 1$, we require conditions on $\phi_k(x)$. Note that the definition of concavity for ϕ_k is that for any $x \geq 1$,

$$\phi_k(x - 1) - \phi_k(x) \leq \phi_k(x + y - 1) - \phi_k(x + y), \text{ for all } y \geq 0.$$

This clearly implies the following for any $x \geq 1$.

$$\phi_{k+1}(x - 1) - \phi_{k+1}(x) \leq \phi_{k+1}(x + y_k^*(x) + 1 - 1) - \phi_{k+1}(x + y_k^*(x) + 2 - 1),$$

which can be re-written as

$$\phi_{k+1}(x) - r - \phi_{k+1}(x+1+y_k^*(x)+1-1) \geq \phi_{k+1}(x-1) - r - \phi_{k+1}(x+y_k^*(x)+1) > 0.$$

The strict inequality on the second line following from the definition of $y_k^*(x)$. Now it is clear that $1+y_k^*(x)$ does not satisfy $r+\phi_{k+1}(x+1+y-1) \geq \phi_{k+1}(x)$ and therefore $y_k^*(x+1) < 1+y^*(x)$ using the monotonicity of ϕ_k's from proposition 5.5.1, or equivalently $y_k^*(x+1) \leq y^*(x)$.

□

Proposition 5.5.8 (Necessary condition for monotonicity of the thresholds). *If for all* $0 \leq x_k \leq D-k-1$, $y_k^*(x_k) \geq y_k^*(x_k+1)$, *then*

$$\phi_{k+1}(x-1) - \phi_{k+1}(x) < \phi_{k+1}(x+y_k^*(x)-1) - \phi_{k+1}(x+y_k^*(x)+1).$$

(which is a relaxation of concavity).

Proof. Given $y_k^*(x_k) \geq y_k^*(x_k+1)$ for some $x \leq D-k-1$, we also have from proposition 5.5.6 that $y_k^*(x_k) \leq y_k^*(x_k+1)+1$. The two together imply that either $y_k^*(x_k) = y_k^*(x_k+1)$ or $y_k^*(x_k) = y_k^*(x_k+1)+1$.

Suppose $y_k^*(x_k) = y_k^*(x_k+1)$, then the definitions (2) and (3) imply that

$$\phi_{k+1}(x-1) - \phi_{k+1}(x) < \phi_{k+1}(x+y_k^*(x)-1) - \phi_{k+1}(x+y_k^*(x)+1),$$

Now suppose that $y_k^*(x_k) = y_k^*(x_k+1)+1$, then (2) and (3) yield

$$\begin{aligned}\phi_{k+1}(x-1) - \phi_{k+1}(x) &< \phi_{k+1}(x+y_k^*(x)-1) - \phi_{k+1}(x+y_k^*(x)) \\ &\leq \phi_{k+1}(x+y_k^*(x)-1) - \phi_{k+1}(x+y_k^*(x)+1),\end{aligned}$$

the last inequality following from proposition 5.5.1.

□

5.6 Summary

This chapter proposed a FedEx-like service offering for digital networks to generate positive revenue from spare capacity. We modeled the admission decision for determining the optimal job sizes to accept, to maximize revenue.

We analyzed in detail a single deadline, single link model with a single customer class, to discover properties of the optimal admission policy and the associated optimal revenue. Analysis was carried out for general file-size distributions and *concave* file size distributions. An easily implemented heuristic policy was proposed and tested in computational experiments.

Insights from this simple model were extended to several successively more complex cases. For instance, we showed existence of the thresholds for the stochastic capacity case and the existence of nested thresholds for multiple customer classes with class-dependent revenues.

The model was further extended to fixed-routing networks and multiple deadlines. Existence of optimal threshold policies was shown for both cases. We used connections from these analyses to conjecture that the optimal admission policy was of threshold type for any problem with a polyhedral feasible admission region. This covered the case of an arbitrary network with fixed routing and multiple deadlines.

5.6.1 Contributions

We think of this chapter having contributed the following.

1. The articulation of the service idea itself – an often ignored contribution. The digital-FedEx service idea has *both* marketing and modeling appeal. On the marketing side, it relates well to consumers acceptance of the mechanisms of physical courier deliveries. On the modeling side, it cleverly lumps together content for each deadline, minimizing the need for complicated state information for each consumer request, such as deadlines, content sizes etc. The decision problem is made simpler – though not trivial, and a cleaner formulation results. We note, however, that nothing essential in lost in the service offering, since one can let the difference between the deadlines be as small as desired. In practice, we doubt anyone will care between having a choice of deadlines which is arbitrary vs. deadlines spaced 5 minutes apart.

2. In terms of modeling and analysis, our contribution might be viewed as that of the two-fare seat-inventory control model to airlines [Lit72], or the single-link Erlang model [Kel91] to tele-traffic. Our single-link canonical model of section 5.2 has revealed promising directions for more complicated cases. Our proposed heuristic (c.f. section 5.2.5) seems a promising foundation for constructing practical operating rules. Similarly, our network model could play a role similar to fixed-routing networks in tele-traffic literature [Kel91]. Also, the results for networks and multiple deadlines indicate several directions for future research, some of which are summarized in section 5.4.3.

3. Finally, a goal was to illustrate the connection to the models used in airlines for seat inventory control (c.f. section 2.2.3 for the motivation). The threshold admission policies optimal for our model are directly analogous to the threshold admission decisions for multiple customer classes in airlines, and depend only on the remaining work in the system. Our thresholds are similarly difficult to compute and motivate the study of EMSR-like [Bel87] heuristics (c.f. section 5.2.5). The existence of multiple nested thresholds for admitting jobs from different classes again bear a strong resemblance to the airlines results [SLD92].

Chapter 6

Forecasting: Locating Probes for Determining Traffic Patterns

Integral to any 'Yield Management' effort in telecom is the ability to forecast available capacity and demand for services. We have so far assumed in earlier chapters the availability of such data, adequately summarized for the problem at hand. In many situations, such data might be available from existing mechanisms. However, for networks like the Internet, obtaining adequate information and summarizing it is itself a burgeoning research area. Even though we do not focus on forecasting in this thesis (c.f. section 2.2.1), this chapter serves to demonstrate the nature of the problems that arise and the room for modeling and optimization in this domain.

A fundamental issue for IP[1] networks is to forecast available capacity and traffic patterns on the network. These networks require novel data collection mechanisms such as *probes* to track traffic because of several reasons mentioned in section 6.1.1. The decision of interest here is the optimal location of a given number of probes to build as complete a picture of network traffic as possible. For this we formulate an integer program which results in an NP-hard formulation. We propose a heuristic for its solution and bound the performance of the heuristic. Several interesting research directions are outlined at the end.

In section 6.1, we discuss the context in which the probe location problem arises and some considerations relevant to modeling. Section 6.2 presents the integer programming formulation, the assumptions behind the formulation and a survey of related modeling literature. Section 6.3 describes and discusses a greedy heuristic for obtaining a solution and section 6.4 provides its analysis. In section 6.5 we address the complexity of the heuristic and outline two variants which differ in the manner of obtaining the solution to a sub-problem within the greedy heuristic. A proof of NP-hardness of the sub-problem also proves the NP-hardness of our integer program. Section 6.6 provides the summary and contributions from this chapter.

6.1 The Probe Location Problem

Several networking research projects are working on deploying sets of 'probes' on networks – potentially hundreds of them, to try to capture both orchestrated and independent snapshots of traffic on IP networks [WP98]. Such "Network Tomography" is expected to enable Internet Service Providers (ISPs) to better manage traffic across their links and exchange points. Further, it is anticipated that performance monitoring will be a strong marketing and sales strategy for many ISPs in the future (c.f. section 6.1.1).

[1] *Internet Protocol.*

The problem arises because of a desire to characterize capacity usage and traffic dynamics on the network, to better utilize existing resources and because of a strong market demand for performance monitoring (c.f. section 6.1.1). Capacity usage as indicated by link and switch utilizations is usually not sufficient for traffic profiling and network capacity management. In recent years, many end-to-end studies of the Internet have been carried out which try to characterize the performance of 'paths' in the Internet, usually in terms of performance metrics like the round-trip times, packet loss rates and routing behavior, see for instance [Pax97]. Unfortunately, building a network-wide view of traffic dynamics is not possible with such strategies. Hence the need for alternate mechanisms such as probes.

Probes are small workstation-size devices that can be attached to a limited number of links at routers/switches in a network, allowing them to capture copies of all packets transmitted through these links. This information can then be analyzed off-line for any desired purpose. The location of the probes is a question of interest as one attempts to build as complete a picture of network usage patterns as possible with a limited number of probes. It is important to get the location somewhat right at the first attempt for relocation of probes is logistically and organizationally inconvenient. Further, the location policy affects the total number of probes needed to monitor a network of a given size, with obvious cost implications. The issues involved in locating probes on a network are discussed in section 6.1.2, along-with the relevant modeling considerations.

6.1.1 The context

The probe location problem is part of a larger effort to understand traffic patterns and behavior of IP networks - a.k.a the Internet. In recent years, intense effort has been dedicated by researchers towards monitoring and understanding IP networks [Pax97, TMW97]. The problem arises because we do not adequately understand how the local behavior of thousands of components such as routers, switches and traffic sources translates to the macroscopic behavior of the network.

The Internet is peculiar as compared to traditional telephony networks because the network does not carry state information. Because of its tremendous autonomy, heterogeneity and the paradigm of best-effort service, an integrated monitoring framework was never envisioned or developed for the Internet. As a result, appropriate mechanisms for traffic measurement do not exist. From the limited information available at the routers, it is difficult to build a picture of traffic flows over the network and the causes of congestion that arise thereof, necessitating the need for measurement mechanisms and devices such as probes that can perform the function.

Strategies employed for Internet measurement studies can be crudely classified into *end-to-end* or *flow-based*, depending on whether they try to obtain information between pairs of locations only or if they obtain information about the flow passing some given point in the network. An excellent example of an end-to-end study is the doctoral work of Vern Paxson [Pax97] in which several aspects of network behavior such as end-to-end delay and its variations, routing behavior and bandwidth availability between pairs of sites are studied. The finding of this study on routing is particularly relevant to our modeling. It is reported that for most pairs of communicating nodes, routes are usually dominated by a single route, and that lifetimes of routes can range from seconds to many days, with most lasting for days. Probes naturally fall into the second category of measurement strategies[2]. Network-wide deployment of probes is a more recent phenomenon as corporations and large companies attempt to lower the cost of maintaining large network infrastructures and achieve higher utilizations[3].

There is also a strong market 'need' for performance data on ISP networks [Bor98] in addition to the need of network operators for better utilization of existing infrastructure[4]. Corporate customers

[2]For a background on the importance of flow-measurements and their relevance, see the home page of CAIDA - *Co-ooperative Association for Internet Data Analysis, http://www.caida.org/.*

[3]There is evidence that utilizations for corporate networks might be as low as 5%.

[4]The difference in operating efficiency between the best and the worst operators in North America and Europe

of ISPs are usually deeply concerned with the service-levels and quality of service they actually receive from the network, and currently use ad-hoc measures for monitoring network performance.

Several companies such as Savvis Communications^{TM} and At Home service^{TM} make performance guarantees the backbone of their marketing strategies. For an impact of performance monitoring on customer retention, see the article [Bor98] in *Business Communications Review* which makes the case that ISPs cannot ignore the impact of performance monitoring on sales and customer retention.

6.1.2 Modeling considerations

It is pertinent to point out two things: (i) probe location is just one of the problems raised in the monitoring and performance analysis of the Internet, and (ii) the objective for locating probes is usually vaguely phrased, as in, for instance, given a number of probes for monitoring traffic, where should they be located on the network ? We find it necessary to mention this because questions about our objective function will inevitably arise when the model is presented. The objective function explored in this paper is a result of our modeling license. This is natural since the problem as posed is not very meaningful.

We assume that our objective is to build as complete a picture as possible of important traffic patterns in the network. What precisely constitutes important traffic patterns and what is meant by a complete picture will become clear later. In fact, it will be seen that the objective function of our formulation is more general than the preceding statement. Constraints are usually that probes need to be located at a router/switch and there is a limit on the maximum number of links each probe can monitor simultaneously.[5]

An additional consideration for modeling is the nature of the probe deployment process which itself admits of various possibilities. For instance, one may decide to ship all probes simultaneously to locations with attendant instructions on the links to which they must be attached. Alternatively, one may wish to deploy them sequentially; installing a few probes, getting additional information about the traffic patterns from them and deploying the next set. In our particular study, the logistical inconvenience and organizational challenges involved in sequential location were considerable enough dictating the choice of a single shot deployment. This is reflected in the proposed integer programming model.

There is also the question of whether *a priori* information is available about the network that can aid the probe location decision. For instance, data about link utilization levels is usually not very difficult to get from the network. We assume any such information to be summarized in the probability of traffic between a node pair taking certain routes through the network.

6.1.3 Outline of the model and results in this chapter

We formulate an integer programming model for probe location that tries to maximize the amount of flow captured by a given number of probes, subject to a constraint on the number of links each probe can monitor. The data required for the model are the network topology and the probabilities of traffic between pairs of locations taking a given route through the network[6]. The model is found to be NP-hard, motivating the use of a 'greedy' heuristic to obtain good solutions or alternatively, to provide a good initial lower-bound for an exact solution method such as branch and bound. We analyze two variants of the greedy algorithm which trade-off efficiency against worst-case performance.

When the number of links each probe can monitor is relatively small[7], the greedy algorithm runs

exceeds 30% [Bay96].

[5]Usually a low number, around 4-6 links.

[6]This information is usually easily obtainable from the network using standard utilities such as *ping* or *trace-route*.

[7]More precisely, when the number of links L is a fixed small number, then $\binom{N}{L}$ is obviously a low order polynomial in N, where N and L are the number of nodes in the network and the number of links each probe can monitor

in time polynomial in the number of nodes in the network. We analyze the heuristic to bound its worst case performance to within 37% of the optimal. More specifically, we show that our problem is actually an instance of maximizing a sub-modular set function subject to a cardinality constraint and appeal to a result by Nemhauser, Wolsey and Fisher [NW78] for sub-modular function optimization to obtain the bound.

When the number of links that can be monitored by probes is relatively large, we propose another variant of the greedy heuristic which greedily selects the links to monitor at each node. Here, however, the bound on the worst-case performance no longer applies.

6.2 An Integer-Programming Model for Probe Location

This section presents an integer programming formulation of the probe location problem. The data for the model are (i) the topology of the network, (ii) the probability of pairs of nodes communicating with each other (or alternately the relative importance attached to traffic between pairs of nodes), (iii) the probability of traffic between each pair of nodes taking a particular path through the network, (iv) the number of probes to be located, (v) the candidate locations for the probes and (vi) the maximum number of links a probe can monitor at any given location. Probes are assumed to be indistinguishable from each other.

6.2.1 Notation

Specifically, let $G = (V, A)$ be a graph with bi-directional edges representing the network topology with $|V| = N$ and nodes indexed by i.

Let k index origin-destination (O-D) pairs in the graph and $\mathcal{R}_k$ represent the set of routes between O-D pair k. Every route is an acyclic ordered sequence of nodes or alternately, links in the network and belongs to exactly one O-D pair. Call the probability that node pair k communicates P_k, and the conditional probability (given that they communicate) that traffic between k takes route $r \in \mathcal{R}_k$, $P_{r|k}$.

Denote by M the total number of probes to be located on the network and let L be the maximum number of links any probe can monitor. Probes are assumed to be indistinguishable from each other and can only be located at nodes.

In this model, to maximize the expected number of O-D pairs that are being monitored by the probes, one proceeds as follows. The probability of O-D pair k being monitored is the sum of probabilities $P_k.P_{r|k}$ across routes $r \in \mathcal{R}_k$ that have at least one link being monitored by a probe. But note that we cannot double count the probabilities, i.e, if more than one link in route r is being monitored, we can only count the probability once. Preventing this double counting is the major reason for the complexity of the ensuing model formulation .

6.2.2 Formulation

The following formulation attempts to maximize the expected number of O-D pairs that can be monitored by probe location and assignment (which links to monitor). Call $\delta(i)$ the set of outgoing arcs incident on node i, $\delta(i) = \{(i,j) : j \in V, j \neq i\}$, and let $x_i \in \mathbb{Z}^+$ be the number of probes located at node i. Call the probability of traffic between k taking route $r \in \mathcal{R}_k$ as $P_r = P_k.P_{r|k}$. Let $\mathcal{R} = \cup_k \{r \in R_k : P_r > 0\}$ be the set of positive probability routes in the network.

Let

---

respectively.

$$y_{ij} = \begin{cases} 1, & \text{if probe located at node } i \text{ assigned to monitor link } (i,j) \in \delta(i), \\ 0, & \text{otherwise,} \end{cases}$$

and

$$z_r = \begin{cases} 1, & \text{if route } r \text{ monitored,} \\ 0, & \text{otherwise.} \end{cases} \qquad r \in \mathcal{R}$$

Then our formulation is:

$$\begin{aligned} P: \quad \max \quad & \sum_{r \in \mathcal{R}} P_r z_r \\ \text{s.t.} \quad & \sum_{i=1}^{N} x_i \leq M, \\ & \sum_{(i,j) \in \delta(i)} y_{ij} \leq L x_i, \quad i = 1, \ldots, N, \\ & \sum_{(i,j) \in r} \left(y_{ij} + y_{ji} \right) \geq z_r, \quad \forall r \in \mathcal{R}, \\ & z_r \in \{0,1\}, \quad \forall r \in \mathcal{R}, \\ & y_{ij} \in \{0,1\}, \quad \forall (i,j) \in A, \\ & x_i \in \mathbb{Z}^+, \qquad i = 1, \ldots, N. \end{aligned}$$

The first and second set of constraints are obvious, limiting the total number of probes to be located on the network and the maximum number of links that can be monitored at any node. The third set of constraints is necessary for preventing double counting of the probabilities by setting $z_r = 1$ in an optimal solution if at least one link on route r is monitored by some probe and forcing it to be zero if no link is monitored. The objective function is then the expected number of O-D pairs that are monitored by a given probe location and assignment. Some observations on the model follow.

6.2.3 Remarks

Input data for the model

Data for the above program requires the specification of the probabilities P_r for each route $r \in \mathcal{R}$. In any general network, the number of paths is usually exponential unless the network is pathologically sparse. If we required therefore, the specification of a positive number for all paths, finding a solution to P would be nearly hopeless. The way data is assumed to be constructed for the above model is therefore as follows.

For the important O-D pairs in the network, guess the probability that they communicate with each other (if it is certain that they do, set this probability to 1). Conditioned on the event that an O-D pair communicates (has positive traffic), assign a probability distribution to a small set of routes between the two nodes. Usually, this is not an impossible task since some routing information is available for the network. In the absence of any information, assign equal probabilities to all routes in a likely set of routes $\mathcal{R}_k$ for O-D pair k. We assume henceforth that the likely number of routes between any k is reasonably small so that the resulting input is polynomial in instance size rather than exponential.

Prioritizing traffic instead of maximizing monitored O-D pairs

We comment on the generality of the proposed objective function. Note that it is not necessary in general for the numbers P_r to be probabilities. They could in fact be arbitrary non-negative weights assigned to routes to reflect the importance of monitoring traffic on that route. Given that each route corresponds to at most one O-D pair, this allows us to rank the importance of sites on the network, usually a very important practical consideration since it allows the exclusion of certain nodes from consideration altogether[8].

Routing

The most significant assumption in the formulation of P is that the probability of traffic between pairs of nodes taking a particular route through the network can be determined and remains static over time. Determination of these probabilities is not too significant a task. In fact, one can select the important node pairs and monitor them for some length of time using utilities such as *ping* or *trace-route* to collect routing data. To justify that the probabilities remain static, we appeal to reported results in a study of end-to-end Internet dynamics [Pax97] where it is found that Internet traffic between pairs of nodes usually traverses a single route that dominates most of the time. Even when traffic oscillates between a small set of routes during a period of interest, say a day or a week, we can view our probabilities as the fraction of time traffic spends on one route and can interpret the objective function as the fraction of traffic intercepted over a given time horizon.

Unlimited probe memory

One final assumption in the formulation is that probes have unlimited memory. In practice, probes are small computers with limited memory and can therefore only capture packets for short durations[9] depending on the volume of traffic flowing through the links being monitored. It is clear that our model ignores this limitation which begs the question of a reasonable interpretation of our objective function. We propose the interpretation that our objective locates the probes so they are in a position to intercept the maximum amount of relevant flow whenever desired for any given duration. In future formulations, one may wish to explicitly account for the capacity constraint.

6.2.4 Literature review

Literature related to data measurement and performance monitoring on the Internet is rapidly growing, with several independent and concerted initiatives underway, the most notable of which is the *Cooperative Association for Internet Data Analysis* (CAIDA). The class of problems needing to be addressed is highly varied and broad, relating not only to measurement, but also to analysis of data and visualization [Cla99] at several network layers, as articulated in Willinger [WP98], for instance. However, we have found no reports yet on optimization models for locating measurement devices.

Optimization literature similarly reports no models for the probe location problem on the Internet. In a different context, specifically transportation, the problem of locating facilities on the network to maximize intercepted flow without double counting was first formulated by Berman, Larson and Fouska (BLF) [BLF92] where a facility located at a node could intercept all the flow passing through the node. The problem was shown to be similar to another problem in location theory called the Maximum Covering Location Problem [MZH83], which is NP-hard. A greedy heuristic was proposed in the same paper in view of the problem complexity and a bound on the performance of the heuristic was derived. In a later paper [BBL95], a sharper tight bound was provided on

[8] One may not, for instance, want to monitor sites with low traffic volumes or remote sites that are not of value.
[9] Usually of the order of minutes.

the performance of the same greedy algorithm and the model was generalized in several directions, where facilities were allowed to intercept flow within a pre-specified radius from their location. So far, however, an extension to the case where facilities can monitor only a given number of links is not available. In fact, in the original formulation in [BLF92], links are irrelevant to the formulation as long as the paths (routes) are specified as sequences of nodes.

At first glance, it seems that our formulation P should be at least as hard as the formulation in BLF. Informally, this can be seen by noting that P has many more variables corresponding to link assignments and that by setting $L = N-1$, we can make the second set of constraints redundant and get the formulation of BLF. A more insightful comment is that given a probe located at a node, the problem of which links it should monitor is exactly the same as the BLF formulation for a restricted network topology. We show this problem to be NP-hard in section 6.5 by transformation from the NODE COVER problem.

6.2.5 The solution approach

The NP-completeness of our problem motivates the use of a greedy heuristic for obtaining quick approximate solutions. We propose a heuristic in section 6.3 which is similar to the one proposed in BLF [BLF92] with appropriate modifications to include the constraint on the maximum number of links that can be monitored. The analysis of the complexity and worst-case performance of our heuristic is complicated by these modifications and is addressed in section 6.4. Specifically, our greedy heuristic relies on solving a sub-problem of assigning a single probe to L links at each node to intercept the maximum flow. This is itself an NP-hard problem for even one node. We therefore outline two possible ways of solving this sub-problem depending on whether L is a relatively small or large number and analyze both cases to bound their worst-case performance.

6.3 A Greedy Heuristic

The basic idea behind the heuristic is to assign the probes sequentially to subsets of links at nodes with cardinality $\leq L$ which intercept most of the un-intercepted flow (probabilities are interpreted as flows) that remains in the network. Every time a subset is selected, the links in the set and the set of flows intercepted by it is removed from the network. This continues until M subsets have been selected or until no more flow remains in the network. This idea is formalized below.

6.3.1 Notation

Recall that $\delta(i), i = 1, \ldots, N$, is the set of outgoing arcs at node i. Let $E_i = \{F \subseteq \delta(i) : |F| \leq L\}$ be the family of all subsets of $\leq L$ links at node i and let $E = \cup_{i=1}^{N} E_i$ be the set of all possible assignments for the probes in the given network.

6.3.2 The algorithm

Initialization: Set $t = 1$, $\tilde{\mathcal{R}} = \mathcal{R}$, $\tilde{E}_i = E_i, i = 1, \ldots, N$. t is the iteration index, $\tilde{\mathcal{R}}$ is set of routes with positive flows remaining in the graph and $\tilde{E}_i$ is the family of subsets (of size $\leq L$) of the remaining links at node i.

Step 1: For each node $i = 1, \ldots, N$, compute

$$V_i = \max_{\tilde{E}_i} \sum_{r \in \tilde{\mathcal{R}}: r \cap \tilde{E}_i \neq \emptyset} P_r$$

and set

$$S_i = \arg\max_{\tilde{E}_i} \sum_{r \in \tilde{\mathcal{R}}: r \cap \tilde{E}_i \neq \emptyset} P_r$$

breaking any ties in the choice of S_i arbitrarily.

This step computes the value V_i of a node as the maximal remaining flow that can be intercepted at that node by monitoring any subset of links of size $\leq L$ and stores this optimal subset as S_i.

Step 2: Let $j = \arg\max_{i=1,\ldots,N} V_i$ and assign the tth probe to monitor the set of links S_j. In other words, assign the probe to monitor the links with maximal interception of flow across all nodes.

Step 3: Set

$$\tilde{\mathcal{R}} := \tilde{\mathcal{R}} \backslash \{r \in \tilde{\mathcal{R}} : r \cap S_j \neq \emptyset\},$$
$$\tilde{E}_i := \tilde{E}_i \backslash \{F \cap S_j \neq \emptyset : F \in \tilde{E}_i\}, \quad i = 1, \ldots, N.$$

Remove all intercepted flows from the graph. Remove all subsets of links that have at least one link being monitored from each node.

Step 4: If $t = M$ or if $\tilde{\mathcal{R}} = \emptyset$, stop. Otherwise set $t := t + 1$ and return to step 1.

6.3.3 Remarks

The greedy heuristic clearly produces a feasible solution, but the issues of its complexity and worst case performance need further scrutiny.

An outstanding issue is the ability to compute the values V_i and finding the sets S_i at each iteration. We note that this is itself a combinatorial problem. For instance, in a complete graph, there are $\binom{N-1}{L}$ possible subsets of cardinality L at each node in the first iteration, and more if we consider lower cardinality subsets. Consequently, obtaining V_i and S_i is not a trivial task. One may be tempted to look for an optimal algorithm for solving the sub-problem at each node by noting that the graph at each node has a very simple hub structure and that flows through that graph can only occupy at most two links. Unfortunately, we show in section 6.5 that this observation is not beneficial and the sub-problem itself is in fact NP-hard.

This motivates the use of two different variants of the greedy heuristic which differ only in their manner of solving the sub-problem in *Step 1*. We address this in more detail in section 6.5. At this point, it is useful for presentation purposes to ignore complexity issues and assume that we can obtain the optimal solution to the sub-problem in each iteration. This helps us present the main analysis of the greedy heuristic in section 6.4. We return to the sub-problem in section 6.5.

6.4 Analysis of the Greedy Heuristic: Bounds on Performance

We show in this section that our problem is an instance of maximizing a sub-modular function subject to a cardinality constraint by appropriately picking a ground set from our graph and defining a sub-modular set function. In this setting, our greedy algorithm is the same as a generic greedy algorithm proposed for sub-modular function optimization by Nemhauser, Wolsey and Fisher [NW78] for which they prove that the worst case performance is always within 37% of the optimal and that this bound is tight – i.e. there are instances in which greedy performs at least as bad as 37% of the optimal. Two points are worth noting, however: (i) this is the worst case bound and the greedy solution might be much closer to the optimum, and (ii) the tightness of the bound for the general sub-modular optimization problem does not imply its tightness for our problem.

The analysis below proceeds along lines similar to the development in [BBL95]. We first provide the background on sub-modular functions, the sub-modular function optimization problem and the generic greedy algorithm for it solution. We then show that our proposed heuristic fits this framework by reformulating our integer program as a sub-modular function optimization problem.

6.4.1 Sub-modular functions and Sub-modular function optimization

Given a finite set N, let f be a real valued function defined on subsets of N. f is *non-decreasing* if $f(S) \leq f(T)$ for $S \subseteq T \subseteq N$. f is *sub-modular* if $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$ for $S, T \subseteq N$. A more detailed introduction to sub-modular functions can be found in [NW88].

A sub-modular function optimization problem subject to a cardinality constraint is formulated as:

$$\bar{P} : \max_{S \subseteq N, |S| \leq M} f(S)$$

where $f(S)$ is sub-modular. The case when $f(S)$ is non-decreasing has been studied by Nemhauser et. al. in [NWF78, NW78]. Obtaining an exact solution to the problem is known to be NP-hard. Consequently, the following generic greedy algorithm is proposed for obtaining 'good' solutions, and the worst case performance of the greedy heuristic is bounded by the subsequent theorem. In the description below, R_G is the set picked by the greedy algorithm and Z_G is the value $f(R_G)$ of this set.

The Generic Greedy Algorithm

Input: $f(S), M, N$

Initialization: $R^0 := \emptyset$, $t := 1$

Main Loop: For $t = 1, \ldots, M$.

$j_t := \arg\max_{j \in N \setminus R^{t-1}} f(R^{t-1} \cup \{j\})$

$R^t := R^{t-1} \cup \{j_t\}$

Output: $R_G = R^M$

$Z_G = f(R^M)$

The following theorem is proved by Nemhauser, Wolsey and Fisher [NWF78].

Theorem 6.4.1 (Nemhauser, Wolsey and Fisher). *The value Z_G returned by the greedy algorithm when applied to the problem*

$$\bar{P} : \max_{S \subseteq N, |S| \leq M} f(S),$$

for $f(S)$ non-decreasing and sub-modular satisfies

$$\frac{Z_G}{Z_*} \geq 1 - (1 - \frac{1}{M})^M \geq 1 - \frac{1}{e} \cong 0.63,$$

where Z_* is the optimal value for the problem.

The theorem ensures that the greedy heuristic is optimal when $M = 1$ and is within 37% of the optimal for any other value of M. In addition, the bound above is tight in that there are instances where $Z_G = Z_*[1 - (1 - 1/M)^M]$.

An observation on the generic greedy algorithm as related to our problem: note that it requires the computation of a function $f(S)$ for any given subset of N and the ability to find the maximum over the remaining elements in N in the main loop. Therefore if either of the steps has exponential running time, the algorithm is itself exponential. We avoid the first issue by requiring the input to be polynomial thereby ensuring that $f(S)$ can be evaluated in polynomial time. The second issue is the subject of section 6.5.

6.4.2 Reformulating the integer program

We will show in this section how to reformulate our original problem P as a sub-modular function optimization problem to which the bound described above can be applied.

Recall from section 6.3.1 that E_i are the families of sets with $\leq L$ links at each node i and $E = \cup_{i=1}^{N} E_i$ is the set of all possible assignments for the probes in the network. E is clearly finite. With a slight abuse of notation, let $N = \{1 \ldots, |E|\}$ be our ground set for the new formulation and index the sets in E by $n \in N$ arbitrarily, with E_n being the nth set in E. For any set $S \subseteq N$, let $E(S) = \cup_{n \in S} E_n$.

Now reformulate P as:

$$\bar{P} : \max_{S \subseteq N, |S| \leq M} f(S),$$

with

$$f(S) = \sum_{r \in \mathcal{R}: r \cap E(S) \neq \emptyset} P_r.$$

In other words, for any set S of links monitored by probes, $f(S)$ is the sum of flows intercepted by these links without double counting flows that traverse more than one link. We show below that $f(S)$ is sub-modular. The proof is similar to the proof in [BBL95] applied to our formulation.

Proposition 6.4.2. *If $P_r \geq 0$ for all $r \in \mathcal{R}$, then $f(S)$ is non-decreasing and sub-modular.*

Proof. If $S \subseteq T$, then $E(S) \subseteq E(T)$, implying that $f(S) \leq f(T)$ when $P_r \geq 0$, so $f(S)$ is non-decreasing.

For sub-modularity, it is sufficient to show that for *all* $S \subseteq T$ and $k \notin T$, $f(T \cup \{k\}) - f(T) \leq f(S \cup \{k\}) - f(S)$. To see this, for any $S \subseteq T$, let $\mathcal{R}_T = \{r \in \mathcal{R} : r \cap E(T) = \emptyset\}$. $\mathcal{R}_T$ is the set of routes not covered by assigning probes to the set T. Now since $S \subseteq T$ implies $E(S) \subset E(T)$, we have $\mathcal{R}_T \subseteq \mathcal{R}_S$. Then,

$$\begin{aligned} f(T \cup \{k\}) - f(T) &= \sum_{r \in \mathcal{R}: r \cap E(T \cup \{k\}) \neq \emptyset} P_r \; - \sum_{r \in \mathcal{R}: r \cap E(T) \neq \emptyset} P_r \\ &= \sum_{r \in \mathcal{R}: r \cap E(T) = \emptyset, r \cap E(k) \neq \emptyset} P_r = \sum_{r \in \mathcal{R}_T: r \cap E(k) \neq \emptyset} P_r \\ &\leq \sum_{r \in \mathcal{R}_S: r \cap E(k) \neq \emptyset} P_r \\ &= f(S \cup \{k\}) - f(S), \end{aligned}$$

showing that $f(S)$ is sub-modular. □ □

6.5 Complexity

We address here in detail the complexity of the proposed heuristic of section 6.3.2 and the solution of the sub-problem in *Step 2* of the heuristic. We consider cases when the number of links that can be monitored by each probe is small and large respectively, and propose selecting links greedily at each node to obtain a solution to the sub-problem in the latter case. The need for such a heuristic solution arises due to the NP-hardness of the sub-problem itself, which we establish using a transformation from the NODE COVER problem. We also show the sub-optimality of this heuristic by a simple example. For this modified scheme, the bound of section 6.4 on worst-case performance of the greedy algorithm no longer holds.

6.5.1 Complexity of the greedy algorithm

Assume for now that the sub-problem is NP-hard. We show this later. We ask, when is the optimal solution to the sub-problem difficult to obtain ? Since $\binom{N-1}{L}$ is a loose upper-bound[10] on the number of possible assignments at each node, when the number of links relative to the number of nodes in the network is small, the computational burden is not too great. In fact, the number of possible subsets of links at each node is bounded by $(N/L)^L$ and we can simply enumerate the possibilities for each node. This bounds the complexity of the greedy heuristic by $O((N/L)^L NM)$, where L is a number as low as perhaps $3-5$. Note that this is a very loose upper-bound, since we're assuming both that the graph is complete at each iteration, which is simply not true, and that there are at least L links at each node that have positive flow. The performance of this heuristic is then bounded by Theorem 6.4.1.

When L (and N) gets to be significantly large relative to N, $(N/L)^L$ becomes prohibitively large, and enumeration no longer remains feasible for solving the sub-problem. In this case, we propose greedily selecting L (or less if needed) links at each node and using the sets obtained in this manner as the S_i's. This bounds the complexity of the greedy heuristic by $O(LNM)$. Note however, that now the Nemhauser-Wolsey performance bound $1-(1-1/M)^M$ no longer applies since the analysis of the heuristic in section 6.4 relies crucially on being able to find the *optimal* S_i's at each iteration.

6.5.2 Complexity of the sub-problem - proof of NP-hardness

We show that our sub-problem is NP-hard by demonstrating that every instance of a NODE COVER problem can be solved in polynomial time if we have a polynomial time algorithm for our sub-problem.

NODE COVER is one of the fundamental NP-complete problems and can be found in any text on computational complexity [GJ79]. The problem statement is: given graph $G=(V,A)$ and a positive integer $K \leq |V|$, is there a subset $S \subseteq V$ with $|S| \leq K$ such that for each edge $\{i,j\} \in A$, at least one of i,j, belongs to S?

Given any instance of NODE COVER, create a hub-and-spoke graph $\bar{G}$ with the same number of links as the number of nodes in NODE COVER, numbering the links from $1,\ldots,|V|$ arbitrarily. For each edge $\{i,j\}$ in G, assign a unit flow to the links corresponding to nodes i and j in $\bar{G}$. It is clear that this construction is polynomial. Note that the total flow in $\bar{G}$ is $|A|$. Now solve the problem of locating K facilities to maximize intercepted flow for $\bar{G}$. Report YES to node cover if the optimal solution is $|A|$, NO otherwise. It is clear that $\bar{G}$ is an instance of our sub-problem. Therefore a polynomial time algorithm for this problem cannot exist unless $P=NP$.

[10] We consider only maximum cardinality subsets of size $\leq L$ at each node since it is clear that omitting a link with positive flow is never optimal and including a link with zero flow is never worse, allowing us to eliminate from enumeration all subsets of lower cardinality.

Note that this also shows that the general problem in BLF [BLF92] is NP-hard since the sub-problem is really just their general problem on a restricted network.

6.5.3 Counter-example to the optimality of a greedy solution to the sub-problem

To see that selecting the links greedily at node i does not yield the optimal S_i's in the greedy heuristic of section 6.3.2, consider the following example.[11] We assume that the greedy heuristic picks the link with the maximum flow, removes the flows from all other links, picks the next maximal flow link and so on until no more flow remains or L links have been chosen. Suppose a node has three links labeled $1, 2, 3$ incident in it as shown in figure 6-1 with the following flows:

$$\begin{aligned}\text{Link } 1 &: f_1 = 10, f_2 = 2,\\ \text{Link } 2 &: f_1 = 10, f_3 = 3,\\ \text{Link } 3 &: f_3 = 3, f_4 = 1.\end{aligned}$$

Then our greedy heuristic would pick links 2 and 1 (in that order) with a resulting value of 15 whereas the optimal choice is links $1, 3$ with a value of 16.

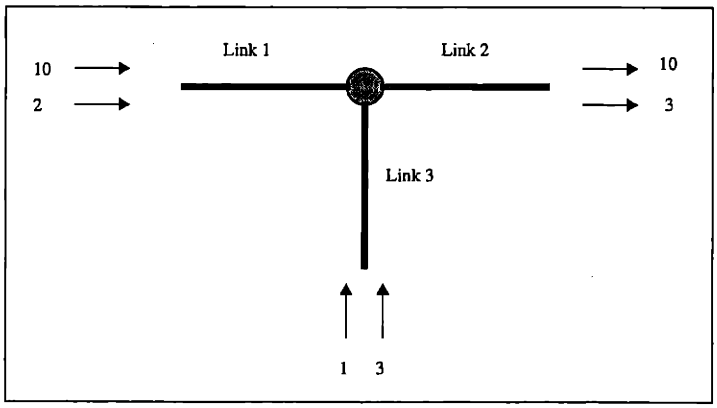


Figure 6-1: A counter-example to the optimality of greedily selecting links at each node.

6.6 Summary

This chapter highlighted the potential for modeling and analysis in data collection and forecasting for Internet-type networks, by focusing on the particular problem of locating probes on IP networks. The decision problem of finding the optimal location for a given number of probes and determining the best links to monitor at each location was formulated as an integer program. We proposed a greedy heuristic for its solution and bounded its worst-case performance by transforming it to a sub-modular function optimization problem. We proved that a sub-problem in one step of the greedy heuristic was itself NP-hard, also showing the NP-hardness of our integer program. We therefore distinguished separate cases where each probe could monitor a small vs. a large number of links, and proposed a variant of the greedy heuristic for the latter case, where one selects the links greedily at each node. The results obtained extended earlier results on discretionary facility location.

[11] This particular example is courtesy of Les Servi at GTE Laboratories, Waltham, MA.

6.6.1 Contributions

Our contributions in this chapter are:

- To our knowledge, this is the first attempt at modeling the probe location problem for the Internet. This is partly because large scale network monitoring is an emerging activity for the Internet and partly because the operational problems usually outweigh optimization problems, as mentioned in section 6.1.1.

- In the modeling context, a similar model for locating facilities to maximize intercepted flow has been formulated and studied before by Berman, Larson and Fouska [BLF92] and a bound on a similar greedy heuristic has been obtained in [BBL95], when a facility located at a node can intercept all flow through that node. The bound in [BBL95] is tight and is within 37% of the optimal.

  We extend the model, the greedy heuristic and its analysis to the case where the number of links monitored by each facility is fixed, to obtain the following main results:

  1. When the number of links that can be monitored by each probe is small, the greedy heuristic runs in polynomial time and its worst case performance is still bounded to within 37% of the optimal. Slightly surprising since our problem is much more complex requiring the determination of probe locations and assignments to links simultaneously.
  2. When the number of links is relatively large, we propose an alternate scheme which selects links greedily at each node, but for which the worst-case bound does not apply.
  3. In addition, as part of the discussion motivating the use of two variants of the greedy heuristic, we provide a short proof of NP-hardness of our problem for a very simple class of networks that gives some insight into the hardness of our general problem (cf. section 6.5.2).

Chapter 7

Market Segmentation: Guaranteeing WWW/FTP Server Performance

What sort of modeling might be relevant and useful for market segmentation activities in telecom YM? This chapter presents an example where models can help use network information to differentiate a service and increase revenues. We outline some formulations and indicate directions that are likely to yield useful results.

The service considered is web-hosting, offered by almost all Internet Service Provider (ISPs). The question of interest is if this service can be differentiated to increase revenues using service-level guarantees. The differentiation attempted here might be considered qualitative, not unlike the practices employed by airlines, such as 14-day advance bookings, Saturday might stay-overs, etc. (c.f. section 2.2.5 for the context.) Computing service-level guarantees for this purpose, however, is a non-trivial task (c.f. section 7.1.2). We borrow the framework of stochastic facility location theory [MF90] to model the problems, which allows us to compute certain service-levels as a function of network traffic patterns and the server location. Some results are immediately obtainable from this framework, such as the optimal location of an FTP server to minimize mean session delays. The models we formulate and analyze in this chapter simplify many of the complexities in real web-server operations, providing the basic building blocks for further investigations of these systems, the behavior of which is largely not well-understood.

The chapter outline is as follows. Section 7.1 discusses some possible service-level guarantees and several practical aspects of the server location problem, including the difficulty in its modeling. Section 7.2 presents the models of the network, FTP and WWW servers. We discuss optimizing the location of FTP servers in section 7.3 and that of WWW servers in section 7.4. Section 7.6 lists the summary and contributions from this chapter.

7.1 The Service Offering

Briefly, a web-hosting service is when a corporation contracts with an ISP to manage its web-servers and all associated details. These services are already differentiated by the obvious physical parameters, such as server type (Windows NT or UNIX), server speed, connection speed to the Internet, monthly traffic volume, backup frequencies, technical support[1], etc. etc. We consider

[1] A cursory search of the Internet using "Web-hosting services" immediately reveals the plethora of options accompanying these services.

another option, namely some service-level guarantee, for differentiating the service. These guarantees could result not only in added-revenue from the differentiation but also an extra competitive edge for an ISP. Further, the cost of computing this guarantee is essentially zero from the network provider's point-of-view as long as adequate network traffic data is available. However, making sure that the offered guarantees are reliably determined as a function of traffic levels is a non-trivial task, and different models may yield drastically different answers. We mention several difficulties in this regard in sections 7.1.1 and 7.1.2.

In this setting, consider the following possibilities for a WWW or an FTP server service guarantees: (i) maximum transfer time to a random user shall be below T with probability P or that (iii) the mean transfer time across all users shall be bounded by M. Since questions of computing useful service-level guarantees can quickly become very complex (c.f. section 7.1.1), in this chapter we focus on the simplest question. Specifically, where should a single FTP or WWW server be placed on the network to maximize a given service-level objective? The answer naturally depends on the objective being optimized. We formulate models where the objective is minimization of the mean delay experienced by a random request, or of the expected file transfer time across all random requests. Computing such service guarantees require knowledge of the traffic patterns on the network, the expected demand patterns on the server and some characteristics of the content, such as the typical sizes of files on the server. We borrow a framework from stochastic facility location theory to incorporate these factors in location models in sections 7.2 through 7.4. The remainder of this section discusses related contexts in which the server location problems arise, the difficulties in modeling such problems and the usefulness of stochastic facility location theory in answering them.

7.1.1 Practical Issues

To understand possible impacts of service differentiation, consider the web-hosting market. Forrester Research estimates from 1998 projected that revenues from hosting complex web-sites will reach $8 billion by 2002 while revenues for hosting simple web-sites may reach $1 billion in 2002 [Cau99]. Almost all ISPs offer web-hosting services, including every major player such as IBM^{TM}, GTE^{TM}, MCI^{TM}, Sprint^{TM} and AT&T^{TM} etc. Competition is therefore severe. It is expected to become more intense as companies move towards hosting business applications on the Internet, an already emerging trend as companies like Microsoft^{TM} position themselves strategically – for instance, Microsoft's $200 million investment in Qwest Communications International in December 1998. ISPs find that customers are increasingly looking to replicate content and are asking for service-level agreements to ensure their sites are both available and responsive, and for pushing content closer to the end users [Cau99]. To further understand the relevance of service guarantees for businesses, see [Bul97, Hal99] for instance. Articles routinely cite that the most important factors for corporations in choosing carriers are performance and customer care [For98] and that many businesses would gladly pay more for a better level of service [Kau99].

In addition to service-level agreements, optimal server-location problems arises in other contexts. For instance, when a provider such as IBM^{TM} decides to host complex web-sites for events such as the Olympics^{TM}, the U.S. Open^{TM} or the Grammy^{TM} awards, an obvious first question is the location of these sites. Similarly, when locating large sites such as Yahoo^{TM}, Travelocity^{TM} etc., with millions of subscribers worldwide, there is major incentive to consider the location decision carefully, since even a small fraction of subscribers accessing the site simultaneously can substantially affect network traffic. Aside from an ISP context, server location and caching remains an important issue in general for the Internet. The importance of this question for the Internet is widely recognized. See for instance, the WWW consortium's "Replication and Caching Position Statement"[2].

In complete generality, a server location decision and the computation of a service-level guarantee is likely to be quite complicated. Commercial web-sites usually involve complex hardware in the form of clustering environments, co-operating servers and caches, all of which we abstract away

[2] *http://www.w3.org/Propagation/Activity.html.*

as a single server. Moreover, hosting a web-site is increasingly likely to involve sets of network caches strategically located to push content to the end-users. The location of these caches and the replenishment strategies for refreshing content at these sites makes obtaining service-level guarantees hard. Even appropriately simplified models, which abstract away much of the detail can be quite complicated to analyze, as mentioned below.

7.1.2 Difficulties in modeling server location

A preliminary literature search has revealed little on analytical modeling of server location for the Internet. On related issues such as dynamic selection of already existing servers, the primary mode of investigation seems to be trace-driven simulations [CC96, CC95, GS95]. This is to be expected since we are only now beginning to understand traffic models for Internet services such as FTP, Telnet etc. [PF94].

An open question is whether location is even a first-order factor in web-server performance, given the "physics" of the Internet and the nature of web-traffic. Large-scale and unpredictable variation, often termed as "flash crowding", is inherent in Internet traffic. This makes it unclear if useful service-level guarantees can even be offered since there might always be enough likelihood of their violation. Indeed, several other factors could easily dominate the performance of web-servers.

The chief modeling difficulty in answering the above question is that the natural formulations, which are queuing-theoretic, require significant departure from the assumptions of classical queuing theory. This results in a lack of analytical machinery available for answering relevant questions. For instance, new understanding of the network indicates that arrival processes for many Internet services are usually not Poisson [PF94], that service times may be heavy-tailed [CB96, BC98, PKC96] and that the network traffic is self-similar and long-range dependent [LTWW94]. Consequently, few results can be directly imported from known queuing models. One either finds attempts to use classical queuing models, with the results inapplicable, or attempts to incorporate more realistic assumptions, with few analytical results. This has not so far led to practically useful models for Internet operations, to our knowledge.

Stochastic facility location theory might be a useful framework for studying location problems in the Internet. Facility location has been usefully studied in other domains, starting from the seminal work of Hakimi [Hak64b] for deterministic facility location to the location of mobile units in stochastic environments [BCL$^+$90]. A good collection of such models and results is [MF90]. We seek to apply some of this vast modeling literature to Internet server location whenever appropriate. We will, in particular, find this framework very useful in formulating our models in section 7.2. Our chief difficulty will then be in analyzing our models when the classical assumptions do not hold.

7.2 Basic Models

This section presents the basic models of the network, an FTP server and a WWW server. Modeling assumptions and their motivation is argued in detail.

7.2.1 Network model

Let $G = (V, A)$, an undirected graph, represent our communication network. Index the nodes by $i = 1, \ldots, N$, and denote the links $\{i, j\} \in A$. We specify in this section a common notation for a demand process and end-to-end delay metric for this graph.

Requests for service arise at the nodes of the graph according to some arrival process with mean rate λ, with the relative frequency of requests from node i being $w_i \geq 0$, $\sum_{i=1}^{N} w_i = 1$. The long term mean rate of requests arising from node i is then λw_i. We have so far left unspecified what it

means for a request to arise from a node. This is intentional since it depends on the type of server (FTP or WWW) being modeled and will be addressed later.

Network delay metric

Call $D(i,j)$ the round-trip time (RTT) for a single packet to travel from node i to j and back to i. $D(i,j)$'s are assumed to be random variables and will constitute our basic measure of network delay.

The use of RTTs as the network delay metric is motivated by the behavior of the TCP transmission protocol used by FTP and web-servers. TCP re-transmission depends on the sender receiving acknowledgments from the receiver, making RTTs the main determinant of file-transfer times. It is reasonable, therefore, to consider file-transfer times to be roughly proportional to the $D(i,j)$'s. On a practical note, measuring RTT's is relatively easy using a utility such as *ping*, for instance.

It is also useful to understand the nature of variations in the RTTs. The randomness of $D(i,j)$'s is attributable to at least four distinct factors:

1. The routine random fluctuations in network packet traffic that is caused by the burstiness and multiplexing of a large number of sources. We understand from various studies on Internet traffic that this type of randomness is very bursty, implying wide fluctuations in round-trip times (RTTs) for packets spaced even minutes apart, and that this burstiness persists over several time scales. The seminal work in this area is that of Leland et. al. [LTWW94].
2. The effect of network routing and other operational policies. Such effects contribute to the randomness of the RTTs but occur at much longer time-scales than the changes in traffic volume [Pax97].
3. The changes in the average value of the *volume* of traffic during hourly, weekly and possibly seasonal cycles (end of year reports for instance). Such changes have been observed in various studies to be clear and predictable over daily and weekly periods [TMW97].
4. Accidents such as router or link outages, or other events causing non-routine randomness.

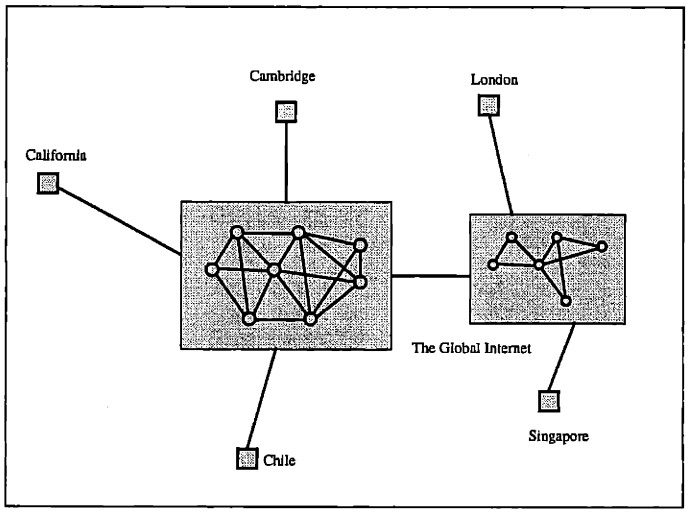


Figure 7-1: The Internet.

With server location decisions being of the tactical type, i.e. ones that make sense over reasonably long periods of time – of the order of hours or days, it is variations of the third kind that most interest us.

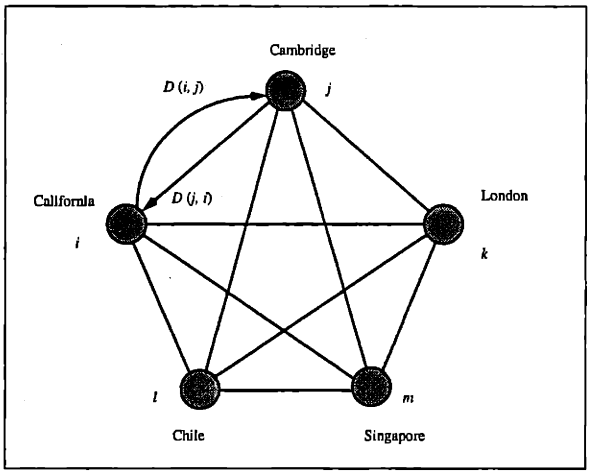


Figure 7-2: The Model.

On a further note, in order to model the effects of predictable volume changes, one can assume that the network has a state which is characterized by the distribution of RTTs between all pairs of nodes. The network might then be assumed to transit between states at fixed intervals as a Markov chain, each state differing from another by the difference in distribution of the RTT between at least one pair of nodes. For the moment however, to not confuse issues, we restrict our models to the case where the network remains in one state. Alternatively stated, this means that the RTTs have given distributions that do not change. This can be interpreted as modeling some fixed interval of time during which traffic volume does not change appreciably.

7.2.2 FTP server model

The model for an FTP server is as follows: requests for FTP sessions (connections) arrive at the server according to a Poisson process, an assumption which has been tested and found reasonable for session arrivals [PF94]. From section 7.2.1, the rate of the process is λ and the arrival rate from node i is λw_i. The FTP server can allow a maximum of C concurrent sessions. An arriving request, if it finds C sessions in progress, is blocked and is considered lost to the system, otherwise it is accepted and occupies a single port[3] for the duration of the session, whether or not it is transmitting data.

Several characteristics of FTP sessions over the Internet have been investigated [PF94]. Typical sessions have been found to have on-off characteristics where data transmission only takes place during on-periods. The on-periods themselves have heavy-tailed durations. Individual transmissions during on-periods are extremely bursty, corresponding to transfers of individual or groups of files.

With such mechanics, it is reasonable to expect the session duration to be some non-decreasing function of the network delay (and the number of users in the system), because of the file transfers. A first approximation could be linear, the sum of a random *uncontrollable* default length of a session, and *the network induced delays* in file transfers. Formally, conditional on a request from node i being accepted when the server is located at node x, the duration of its session is the random variable $S(x|i)$ described as $S(x|i) = F_i + \beta D(x,i)$, where F_i is the default session length of a user from node i, $D(x,i)$ is the RTT from node x to i and $\beta \geq 0$ is some chosen scalar. It immediately follows that if the random variables involved have finite means, the mean duration of a session from node i is $\bar{S}(x|i) = \bar{F}_i + \beta \bar{D}(x,i)$.

[3]Usually the FTP session control port.

7.2.3 WWW server model

We model WWW servers as multi-server queues with FCFS discipline, Batch-Poisson arrival processes and heavy-tailed distributions for service times. Brief explanation of these modeling assumptions follows below. The number of 'queue-servers' in the model is the maximum number of TCP connections the WWW server can have open simultaneously. Stochasticity in service times arises from network-induced delays in file-transfers. An infinite buffer is assumed for queuing requests — a reasonable assumption in practice considering the queuing capacity compared to the total number of available TCP connections.

Individual arrivals

Arrivals in the model are identified with GET[4] requests made from the WWW server. These are requests for *individual files* (html pages, images etc.) as seen by the WWW server. A GET request is not the same as a client request for a URL[5]. In short, a single request for a url usually initiates multiple GET requests for images and other files with embedded links from the requested url. GET requests may be human *or* machine initiated and may not be identified easily with human actions. The size of content requested by each GET is assumed to be random variable with a known distribution. Each GET request is assumed to be handled separately by the WWW server and is allocated a separate TCP connection[6] for the duration of the transmission.

The motivation for considering GET requests as the unit of arrival instead of using a behavioral model for human generated requests is that server logs permit easy access to data associated with each GET request (time of arrival, amount of content, referring page, client etc.). One the other hand, inferring the behavior pattern of the process originating these GET's is more troublesome. Several studies have attempted to relate the behavior of a population of users to the resulting traffic from WWW servers [BC98], with the resulting models often involved and not suited to optimization.

The arrival and service processes

We discuss below why the batch-Poisson process is a reasonable model for arrivals of GET requests at a WWW server, and why the service times should be modeled as heavy tailed random variables.

For realistic modeling, the arrival and service processes must be chosen to agree with empirical observations. It has been observed that (i) file-sizes associated with GET requests exhibit extremely high variations and (ii) the resulting output traffic from WWW servers is self-similar in nature [CB96]. Incorporating (i) is obvious, i.e. select heavy-tailed service times. It turns out that in fact, (i) gives rise to (ii) automatically even though there are several ways of producing self-similar processes. For instance, studies such as [PKC96] report that selecting heavy-tailed file-size distributions may be sufficient to produce the kind of self-similarity desired and that this self-similarity is relatively invariant to the distribution of the inter-arrival times [PKC96]. It is further reported that the retransmission rate and packet losses are not unduly affected by increasing self-similarity [PKC96]. This has positive implications for using RTTs as the measure of network delay, independent of the details of the server.

The insensitivity of the server-output process to the arrival process motivates approximating GET requests by a simple stochastic process. For instance, a non-stationary Poisson process or perhaps a Poisson process with bulk arrivals may suffice, while retaining the property of self-similar output using heavy tailed file size distributions. This is actually not far from the case. For instance, the figures below illustrate data from the logs of an actual WWW server. Figure 7-3 shows a sample path

[4] *Terminology used by most WWW servers.*

[5] *Universal Resource Locator.*

[6] Note that this models accurately the HTTP 0.9/1.0 protocol where separate TCP connections are used for each file, compared to HTTP 1.1, which allows the use of a single TCP connection for multiple files.

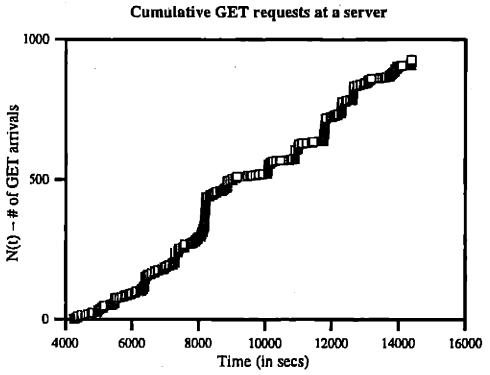


Figure 7-3: An empirical sample-path of cumulative 'GET' arrivals – from an actual WWW server.

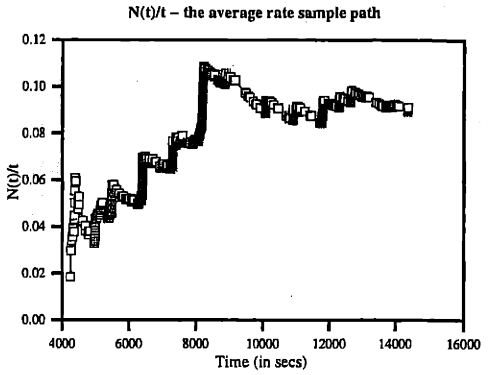


Figure 7-4: An empirical sample-path of the time-average arrival rate of GET requests.

of cumulative GET requests arriving within a time window of approximately four hours. Figure 7-4 shows the time-average arrival rate from the same data set. Qualitatively, we can see that in figure 7-3, the slope of the arrivals – or the arrival rate, is almost constant between the very steep climbs. This confirms the speculation that arrivals may be well-approximated by a batch-Poisson process. Further reinforcement is provided by figure 7-4 in which the time-average arrival rate almost seems to be converging to a value as the time horizon gets longer. This, as is well-known, is a 'renewal process' property. Certainly, when the batch sizes are finite, one expects the time-average arrival rate of a batch-arrival Poisson process to converge to a value.

7.3 Optimizing FTP Server/s Location

In this section, we formulate the optimal location problem for FTP servers to minimize mean session length. For single servers, results are immediately available from stochastic facility location theory. For multiple FTP servers, we formulate the problem, informally discuss its complexity and outline some promising directions for its analysis. The network delay metric used is the RTT between pairs of locations, as discussed in section 7.2.1.

7.3.1 Minimizing mean delay and a penalty cost

When the objective is minimizing the sum of expected session length and a penalty for rejection, we can directly apply a result from Berman et. al. [BCL+90]. On a modeling note, we assume here that the "penalty for rejection" is expressed in units equivalent to the mean session length, i.e. it is in terms of units of time, converted if necessary, from other units such as dollars.

Consider a single FTP server located at node x, then the cost $z(x)$ per unit time of this location when the system is in steady state is:

$$z(x) = [1 - B(x)]\bar{S}(x) + B(x)\bar{Q},$$

where $B(x)$ is the stationary probability of an arriving request being blocked, $\bar{S}(x)$ is the mean session length of an arriving request, given by $\bar{S}(x) = \sum_{i=1}^{N} w_i \bar{F}_i + \beta \sum_{i=1}^{N} w_i \bar{D}(x,i)$ and $\bar{Q} \geq \sum_{i=1}^{N} w_i \bar{F}_i = \bar{F}$ is the cost per rejected customer. Note that $\bar{Q} \geq \bar{F}$ is necessary to make the problem meaningful, for if it would cost less to reject a request than serve one, one would reject all requests.

Rewriting $z(x)$ as $\bar{F} + [1 - B(x)]\bar{D}(x) + (\bar{Q} - \bar{F})B(x)$ where $\bar{D}(x) = \beta \sum_{i=1}^{N} w_i \bar{D}(x, i)$, the single FTP server location problem is seen to be exactly the same as a problem studied in stochastic facility location, called the *p-server single facility loss median* (p-SFLM) problem in Berman et al. [BCL$^+$90]. Some results are then immediate. In particular, the optimal location for the p-SFLM is the 1-median, defined as the location that minimizes the mean system delay $\bar{S}(x)$. Part of the development in [BCL$^+$90] is presented below with appropriate notational changes.

First, we drop $\bar{F}$ from $z(x)$ since it does not depend on x and write $z(x)$ as $z(x) = [1-B(x)]\bar{D}(x)+ QB(x)$, where $Q = \bar{Q} - \bar{F} \geq 0$. Then, an expression for $B(x)$ is obtained by noting that the FTP server model is exactly an Erlang loss system [GH85, Kle75], with

$$B(x) = \frac{\rho(x)^C / C!}{\sum_{k=0}^{C} \rho(x)^k / k!},$$

where $\rho(x) = \lambda \bar{S}(x) = \lambda \bar{F} + \lambda \beta \sum_{i=1}^{N} w_i \bar{D}(x, i)$.

Now the single FTP server optimization problem is stated as:

$$P_1 : \min_{x \in V} z(x).$$

The following theorem is proven in Berman et al. [BCL$^+$90]. We adapt the proof in section 7.5 for our problem.

Theorem 7.3.1. *The optimal location for P_1 coincides with the 1-median location when $Q \geq 0$.*

The above result may seem a bit peculiar because there are no restrictions on the penalty cost Q other than non-negativity. Some intuition into this behavior might be gained by reasoning about the case when $C = 1$. With a single channel, consider the expected system cost by viewing the system to be incurring a unit cost only when busy plus a possible penalty cost. Then the expected cost at a random instant is 0 conditioned on the system not being busy and is $1 + \lambda Q$ when it is, which makes the expected system cost $B(x)[1 + \lambda Q]$ which is monotonically increasing in $\bar{D}(x)$. Thus the expected system cost is minimized by minimizing $\bar{D}(x)$.

For a multi-channel system, i.e. $C > 1$, the above reasoning extends as follows. The expected system cost per unit time when the system is not idle, is given by the following, since a unit cost is incurred by each active connection and a penalty cost is incurred only when all circuits are busy. Then

$$z(x) = B(x)\lambda Q + \sum_{k=0}^{C} k P_k(x), \qquad \text{with} \qquad P_k(x) = \frac{\rho(x)^k / k!}{\sum_{j=0}^{C} \rho(x)^j / j!}, \; k = 0, \ldots, C.$$

$P_k(x)$ is the stationary probability of k sessions being in progress at a random time when the server is located at x. Now the probabilities $P_k(x)$ are all monotonically increasing in $\rho(x)$, or alternately $\bar{D}(x)$, therefore minimizing $\bar{D}(x)$ minimizes system cost.

7.3.2 Optimizing multiple FTP server locations and assignment

We briefly formulate and comment on the optimal location of *indistinguishable* multiple FTP servers on the network and simultaneous assignment of nodes to these servers. More precisely, in the simplest model, all requests from a node are served by its assigned server. If the server is busy, the request is assumed lost to the system, i.e. it does not spill to an available server. Server assignment is static and does not change over time.

Let N be the number of nodes in the network. Suppose $K < N$ servers are to be located on it,

since the case $K = N$ is trivial. Let $x_j \in N$ be the location of the jth server, $j = 1, \ldots, K$, and define

$$y_{ij} = \begin{cases} 1, & \text{if node } i \text{ assigned to server } j, \\ 0, & \text{otherwise.} \end{cases}$$

Define the vector notation $\mathbf{x} = (x_1, \ldots, x_K)^T$ and $\mathbf{Y} = [y_{ij}] \in \mathbb{R}^{N \times K}$. As before, let $D(x_j, i)$ be the end-to-end delay between the location of the jth server and node i. Now call $B_j(\mathbf{x}, \mathbf{Y})$ the stationary probability for an arriving call to be blocked at server j under the location and assignment policy $(\mathbf{x}, \mathbf{Y})$. Then we define the cost $z(\mathbf{x}, \mathbf{Y})$ of a policy as:

$$z(\mathbf{x}, \mathbf{Y}) = \sum_{j=1}^{K} \sum_{i=1}^{N} w_i y_{ij} \Big([1 - B_j(\mathbf{x}, \mathbf{Y})](\alpha \bar{F}_i + \beta \bar{D}(x_j, i)) + B_j(\mathbf{x}, \mathbf{Y}) Q \Big).$$

Our problem is now stated as follows, where $\mathbf{e}$ is a vector of 1's of appropriate dimension:

$$P_2 : \quad \min_{\mathbf{x}, \mathbf{Y}} \quad z(\mathbf{x}, \mathbf{Y}),$$
$$\text{s.t.} \quad \mathbf{Y}^T \mathbf{e} = \mathbf{e}.$$

We note comments on the formulation below, but do not pursue an in-depth investigation.

1. Note that in case of light traffic, i.e. $\lambda \to 0$, one expects $B_j(\mathbf{x}, \mathbf{Y})$ to go to 0 for all j and all location policies $(\mathbf{x}, \mathbf{Y})$ and the solution for P_2 must approach the deterministic K-median solution. Therefore P_2 must be at least as complex as the deterministic K-median problem [Hak64a], which is among the hardest NP-hard problems.
2. Similarly, in heavy traffic, if $B_j(\mathbf{x}, \mathbf{Y})$ approaches 1 for all locations, the problem reduces to an assignment problem.
3. In all other traffic ranges, the complexity of the problem suggests searching for heuristics or analyzing appropriately simplified instances to gain insight into the structure of the solutions. Such approaches have already been tried in stochastic facility location theory and might prove useful here (see, for instance [CCI72]).

7.4 Optimizing a Single WWW Server Location

In this section, we discuss the complexity of the optimal location problem for a single WWW server, but do not pursue in-depth analysis. The goal is to outline some promising formulations and interesting directions. RTTs are used a proxy measure for the file transfer times throughout this section.

7.4.1 Difficulties in optimizing web-server location

With the models for the network and the WWW server discussed in section 7.2.3, the optimal location problem for a WWW server results in an $M^X/G/k$ queuing model, with the service time distribution dependent on the location of the server.

Obtaining and evaluating performance measures for the $M^X/G/k$ model is complicated. For instance, even first-moment information for the time taken to serve a request is not available in closed form. To see this, consider the simpler case of Poisson arrivals and a general service time, the $M/G/k$ queue. This problem does not have a closed form solution.

Adding to the complication is the fact that we require the service times to be heavy tailed. For instance, we know that for the $M/G/1$ queue, the assumption of heavy tailed service time means the mean of the time spent in the system does not exist. This follows from the Pollaczek-Khintchine (P-K) formula for the mean system time which has the service time variance term in the numerator. Specifically, if S is the service time, and W the mean time spent in the system, then the P-K formula is (c.f. [GH85]): $W = E(S) + (\rho^2 + \lambda^2 \text{var}(S))/(2\lambda(1-\rho))$, where $\rho = \lambda E(S)$..

Other alternatives such as bounding the mean service time are also seldom available. Most bounds available from from literature (cf. Harris [GH85] on bounds for mean sojourn time for $GI/G/k$ queues, among others) are usually in terms of the first and second moments of the service time, rendering them inapplicable when the second moment of the service-time does not exist, such as for heavy-tailed distributions.

Finally, when one requires more than first-moment information, as in determining the probability of transfer time being below a threshold, things get even more complicated than before. This is expected to be the practical case since, because of the high variation in the service times, the mean transfer time may just not be useful.

In the absence of an exact closed form or a tractable algorithmic expression for the mean sojourn time in a general multi-server queue, several approximate analyses can be attempted. We consider an infinite capacity ($M^X/G/\infty$) approximation below.

7.4.2 The infinite-capacity approximation - $M^X/G/\infty$ system

In this section, we briefly comment on the formulations that might result when using an infinite capacity approximation for a WWW server. The plausibility of an infinite 'queue-servers' approximation can be argued from the usually large number of connections that can be simultaneously opened from a WWW server[7], but the bursty nature of the arrival process and the presence of heavy-tails (implying very large size file transfers) necessitates the need to test this assumption more carefully.

Call $G_{xi}(t)$ the distribution of a randomly requested file's transfer time from node x to i. We assume $G_{xi}(t)$ is easily available from the distribution of the RTTs between pairs of locations. Following are some comments on possible formulations.

1. If the objective is to maximize the probability of the file transfer time being less than some constant $\alpha \geq 0$, the problem reduces to a 1-median problem with the distance metric between pairs of locations $G_{xi}(\alpha)$. To see this, note that the CDF of the system-wide service time when the server is located at node x is $G_x(t) = \sum_{i=1}^{N} w_i G_{xi}(t)$. Therefore the optimal location $x^* = \arg\max \sum_{i=1}^{N} w_i G_{xi}(\alpha)$.

   One may be tempted to assign a penalty cost Q to the event that the service time of a request exceeds α. This again results in a 1-median problem if the cost is a weighted sum such as $G_x(\alpha) + Q(1 - G_x(\alpha)) = (1-Q)G_x(\alpha) + Q$. Clearly, whenever $Q \geq 1$, the optimal location is one that maximizes $G_x(\alpha)$, i.e. the 1-median location.

2. An capacitated formulation might be investigated if we are willing to make the following assumptions: (i) batch sizes are i.i.d independently of the location i, (ii) an arriving batch, if not fully accepted, is considered lost, with an associated penalty, (iii) batch sizes are finite random variables and (iv) the service time of all requests in a batch are exactly the same.

   Now supposing the capacity of the system is C and the largest possible batch size is K, one can view this system as one with multi-class arrivals. Here the arrival rate for class $k = 1 \ldots, K$, is λ_k. An arrival of class k requires k units of capacity to be admitted. When admitted, the service time of the request is distributed with the distribution function $G_x(t)$. This system

[7]The larger servers can have as many as 4000-5000 separate TCP connections open at the same time.

is now a stochastic knapsack and the steady-state distribution of the state of the knapsack is product-form and easily available [Ros95]. One may now be able to write appropriate formulations for an objective function and determine the optimal location. We do not pursue this further.

7.5 Proof of Theorem 1

We reproduce below, with notational changes, the proof of Theorem 1 as provided in Berman et al. [BCL+90]. Recall that $z(x) = [1 - B(x)]\bar{D}(x) + QB(x)$. We state the theorem again for convenience.

Theorem. The optimal location for P_1 coincides with the 1-median location when $Q \geq 0$.

Proof:

Since the 1-median location minimizes $\bar{D}(x)$, we need to show only that

$$\frac{dz(x)}{d\bar{D}(x)} \geq 0 \quad \text{for all} \quad x \in V.$$

We begin by noting the following relationships:

$$\begin{aligned}
&\frac{d\rho(x)}{d\bar{D}(x)} = \lambda\beta \quad \text{from} \quad \rho(x) = \lambda\alpha\bar{F} + \lambda\beta\bar{D}(x), \\
&\frac{dB(x)}{d\bar{D}(x)} = \frac{dB(x)}{d\rho(x)}\,\frac{d\rho(x)}{d\bar{D}(x)} = \lambda\beta\frac{dB(x)}{d\rho(x)}, \\
&\frac{dB(x)}{d\rho(x)} = \left[\sum_{k=0}^{C}\frac{\rho(x)^k}{k!}\right]^{-2}\frac{\rho(x)^{C-1}}{(C-1)!}\left[\sum_{k=0}^{C}\frac{\rho(x)^k(C-k)}{k!C}\right] \geq 0.
\end{aligned}$$

Then from $z(x) = [1 - B(x)]\bar{D}(x) + QB(x)$, we have:

$$\begin{aligned}
\frac{dz(x)}{d\bar{D}(x)} &= [1 - B(x)] + \lambda\beta[Q - \bar{D}(x)]\frac{dB(x)}{d\rho(x)} \\
&= [1 - B(x)] - \lambda\beta\bar{D}(x)\frac{dB(x)}{d\rho(x)} + \lambda\beta Q\frac{dB(x)}{d\rho(x)} \\
&\geq [1 - B(x)] - \lambda\beta\bar{D}(x)\frac{dB(x)}{d\rho(x)} \\
&\geq [1 - B(x)] - \rho(x)\frac{dB(x)}{d\rho(x)}.
\end{aligned}$$

The first inequality is true since $QdB(x)/d\rho(x) \geq 0$ and the second follows from $\rho(x) = \lambda\alpha\bar{F} + \lambda\beta\bar{D}(x) \geq \lambda\beta\bar{D}(x)$. Algebraic arguments are then used to show that $[1 - B(x)] - \rho(x)dB(x)/d\rho(x)$ is non-negative. This step is slightly more involved and a proof can be found in Chiu and Larson [CL85].

7.6 Summary

This chapter highlighted the role modeling and analysis can play in market segmentation activities related to Yield Management. Specifically, we considered segmenting the web-hosting services of a service provider through service-level guarantees. To accomplish this, however, one has to compute the guarantees with sufficient reliability based on network traffic patterns. This necessitates the

need for modeling and analysis. Our goal in this chapter was to formulate some interesting and promising models.

To do this, we proposed models for an FTP server, a WWW server and the network, and discussed in detail their underlying assumptions. We then used the framework of stochastic facility location theory to formulate location models for an FTP and a WWW server. We highlighted the nature of results obtainable from such models by describing the optimal location of a single FTP server. Finally, we commented on the multiple FTP servers location problem and discussed in detail the complexity of the single WWW server location problem.

7.6.1 Contributions

This chapter was intended to structure questions rather than attempt answers and that is its contribution. The idea of using service guarantees is not new but the modeling directions suggested here are fresh.

1. The use of stochastic facility location theory in this domain is new, to our knowledge, and might yield significant benefits by tapping into the large body of already available literature.

2. Contributions resulting from the formulated models are also expected. For instance, simple models for the server location problem on the Internet are not available (c.f. section 7.1.2). Further, the queuing-theoretic models that arise here reinforce interest in some of the emerging analytical directions, such as, for instance, in the analysis of multi-server queues with heavy-tailed service times (c.f. section 7.2.3).

3. Finally, even outside the context of Yield Management, several contributions are expected from a good model of server location/replication and caching. We refer the reader to section 7.1.1 for a relevance of this problem to the general Internet.

Chapter 8

Pricing: Quasi Real-time Pricing of Long-distance Service

This chapter demonstrates: (i) the role of pricing models in our proposed framework of section 2.2.4, and (ii) how these models might differ from traditional pricing models by focusing on factors relevant to revenue maximization, such as revenue cannibalization for instance. We do not attempt in-depth analyses but articulate a service, argue modeling assumptions, formulate a simple model to concretize our point-of-view and derive some simple insights from the model.

To make our case, we use a YM service which discounts network calls in *quasi real-time* to maximize revenue. It essentially consists of communicating to the subscribers, at fixed periods, the availability of discount rates for long-distance calls. The discounts offered and the destinations to which they apply are under the control of the network provider. Section 8.1 provides the details of the service. Variants of such services exist in different forms (cf. section 8.1.1) but are few and relatively unknown and vigorous debate usually surrounds their viability and market acceptability. In a speculative exercise, we hypothesize the most natural and consumer-acceptable (in our view) version of such a service to model its operation.

The chapter is organized as follows. Section 8.1 presents the service and a detailed discussion of practical issues, such as marketing and the expected benefit of discount pricing from an Economics viewpoint. It also explains our modeling approach and presents a brief literature review on pricing. Section 8.2 discusses and formulates a model for the discounting decision, to make the case for some interesting lines of analysis. A summary of the chapter and its contributions are presented in section 8.4.

8.1 The Service

Consider a service where at fixed intervals of time, say every 15-30 minutes, given some information about the state of the network[1] we wish to announce available discount rates for selected origin-destination pairs during a future interval, to maximize revenue from the network. For instance, suppose that the usual rate for calls from Chicago to Boston is 35 cents/min during peak periods and 10 cents/min at off-peak hours plus possibly a fixed charge per call. *In addition*, suppose these subscribers are made aware of %age discount *off this base tariff* from Chicago to Boston, possibly by means of postings on a web page, or email (similar to fare-watcher e-mails by airlines) or by means of a ticker tape on the phone-set, in the following format:

[1] Typically utilization, past demand history and the current prices.

Chicago to Boston: Regular - $0.35/min, for discount rate $.20/min during 3-4pm, dial 1-800-977-1212
*Chicago to Boston: Regular - $0.10/min, for discount rate $0.02cents/min during 12-2am, dial 091-472-098**

The network provider's decision problem is to manage these discount offerings to increase revenue from the network. Some comments on the features and intent of this service follow.

- One expects users to be receptive to such a service because of possible savings off their regular monthly bills. Subscribing to a discount notification service for a nominal charge - say $1 per month, simply gives a user the option to use the discount prices if desired, without a substantial increase in her monthly bill.
- The use of lengthy dialing codes for the discount service is an attempt at service differentiation[2] by adding a level of inconvenience. One can actually require more than one step where each step requires the dialing of a lengthy code.
- The target market here is the most price-sensitive portion of the subscriber-ship. The idea is similar to coupon offerings in retail sales. The most sensitive customers are likely to pay the most attention to these offerings and be willing to dial one or more extra codes to reach their destination.

The idea of these coupon-like discounts is based on the hope that benefit from customer reaction will be significant, based on experiences with other products. Studies on coupons show that about 20-30% of the market bothers to regularly clip, save and use coupons when they go shopping [PR89]. Statistical studies confirm their higher price-sensitivity. In the YM context, for example, airlines have known for a while that demand for excursion fares is about four times as price-elastic as first-class service[3].

However, one has to be careful of the possibility of cannibalizing revenue from otherwise higher paying demand. Airlines, for instance, attempt to minimize such cannibalization by segregating demand into leisure, business and frequent travelers, in addition to closely regulating discount fares. This problem is expected to be more acute for telephone networks as there is usually a fixed subscriber base that makes up the majority of usage on the network. More or less, demand has to come from a fraction of these subscribers, and one needs working mechanisms to ensure that their regular usage is not unduly displaced to discount periods. On the positive side, a side advantage of this service could be the prevention of loss of revenue, such as to dial-around calls and to customer switching during long-distance price-wars. These and other practical factors are discussed in section 8.1.1.

8.1.1 Practical issues

Several recent incidents confirm that demand for network services increases with a lowering in price. A case in point was when Sprint^{TM} Canada offered unlimited calling anywhere in Canada for $20 per month. This resulted in excessive network congestion forcing them to put a cap of 800 minutes on the plan [All98]. Another classic and well-publicized case was that of America-Online^{TM}. On December 1st 1996, America On-Line decided to offer unlimited connection for $19.95 a month. However flat rate Internet access in a country where local calls are often free proved to be a fatal attraction. Users logged on and stayed on, tying up the phone lines and drowning the service provider.

[2]People sometimes prefer not to use a lengthy code simply to avoid the inconvenience of dialing it. This idea was proposed by an actual telephone company in one of the author's intern-ship experiences.

[3]Estimates from a study by J. M. Cigliano, "Price and Income Elasticities for Airline Travel: The North American Market", *Business Economics*, **15**, September 1980:17-21. *Cited in [PR89].*

In spite of the above, it is not clear that there is a substantial market for real-time discounted long-distance calls. The chief reasons cited against the acceptability of such a scheme are consumer preferences for a flat-pricing structure and the relatively low revenue per-call, which means there may not be enough price-sensitivity to discounts. This is in contrast to airlines, for example, where seats cost hundreds of dollars each. It is also argued that with a fixed subscriber base, if customers were rational agents and were operating according to a fixed budget, they would always make sure they pay only as much as before when using discount rates. In other words, discounting can never pay for itself in the current environment[4].

The issue of sensitivity to discount rates for telephone calls can be countered: the first argument is the market use and acceptance of coupons for low-price items which are often less than a couple of dollars, such as soap, gelatin, cake-mixes etc. [PR89], a scheme very similar to our discount offerings. The second is a study on elasticity estimates for long-distance calls which reports price elasticity estimates of -0.54 for long-distance telephone calls for specific routes based on actual data [Dis96].

The points regarding consumer preference for a flat-pricing structure and the rational behavior of customers can also be countered somewhat. In the service we propose, the market pricing structure remains as it is, and the offered discounts are simply off the base tariff. One does not expect market acceptability problems with such a service. Regarding rational behavior of customers, we cite an interesting article from *Philadelphia Daily News*, 4 April, 1995, which reported the results of a study by the *Telecommunications Research & Action Center* citing that "most people could save 20% to 30% on long-distance charges by switching to a discount-rate plan, even without changing companies". Rational behavior is not always the case with large consumer bodies, and sometimes the presence of a large discounts is enough to encourage more spending than would otherwise occur.

Finally, to indicate the existence of a market for discount calls, we cite the emergence of the "dial-around" market for discount long-distance calls. The dial-around service consists of dialing a pre-assigned code before a long-distance call, which is then routed through an alternate network, without switching providers. The incentive for the customers is obviously savings off their phone bills. The most visible dial-around service is MCI^{TM}'s 10-10-321 number which had almost 50% of the dial-around market, estimated to be about $1.5 billion in 1997 and expected to reach $2 billion by the end of 1998[5]. This was a major concern for other companies like AT&T which saw its revenue affected significantly by such a service[6]. AT&T^{TM} finally launched its own dial-around company under the name "Luck Dog Phone Company" after many other marketing counter-attacks[7]. This clearly indicates not only that a market exists for discount calls but that it is significant. In such a world, discount offerings can not only serve as an instrument for increasing revenue from the network, but also to prevent the loss of revenue through competitor dial-around services, every dial-around call representing lost revenue.

However, even if a market exists and customers do react and use the offered discounts, the question still remains: does such discounting makes sense? We can reasonably expect from experience that discounts will result in immediate increase in call-minutes, but where would the demand come from? If the increase in volume of call-minutes is not sufficient for the decrease in revenue per call-minute because of discounting, one would lose net revenue. There is thus an essential trade-off between 'cannibalizing' revenue from higher paying calls and increased discount utilization since demand is not infinite. The model we formulate attempts to capture this trade-off using simple reasoning.

---

[4] Some of these points were raised by representatives of an actual phone company in one of the author's experiences as an intern.

[5] *New York Times* (national edition), 28 June, 1998.

[6] *Wall Street Journal*, 7 April, 1998.

[7] *Advertising Age*, vol. 69, no. 44, p. 17, 2 November, 1998.

8.1.2 An Economics perspective

Some economic reasoning might help illustrate the feasibility of the service and the cannibalization issue further. Say only 1% of the subscribers take up this offer. For a company such as AT&T^{TM}, with approximately 80 million customers[8], the figure is 800,000. That immediately gives $800,000 per month as the subscription fee for the service. In addition, if an increased revenue of $0.50 per month per subscriber is achieved on average, we have $400,000 million per month increase. These are annual increases of $14.4 million, with little added costs. Remember that we are assuming a market penetration of only 1%. Of course, if one loses more than $1 per person in call revenue, one achieves a net loss. For smaller companies, the effects may not be as dramatic, but are still significant. Sprint^{TM}, for instance, provides long-distance service to around 15 million homes, and the above calculations result in added revenue of $2.7 million annually. With these calculations, one needs around 0.6 million people to subscribe to such a service to get $1 million added revenue.

When is it reasonable to expect a revenue increase from such a service? Reasoning in an aggregate sense, consider a demand curve for total usage in call-minutes over a period of time from a set of subscribers to whom we wish to market such a service, as in figure 8-1. Price P is shown along the y axis and call minutes Q along the x axis. At current rate r, the total demanded call-minutes are Q_r and the total revenue rQ_r since there are no marginal costs. This is shown by area A in the figure. Now say we discount the rate to $d \leq r$. If d prevailed over the entire period, we would see usage Q_d, but since it prevails for controlled periods, we see usage $\bar{Q}$ instead, between Q_r and Q_d. Some fraction f of usage $\bar{Q}$ takes place at r and the rest at d. Now the total revenue is expressed as the sum of the areas B and C, equaling $[rf+d(1-f)]\bar{Q}$. It is clear from the figure that one gains revenue shown in area D while losing revenue shown by area F. This is the nature of revenue cannibalization. For instance, when the aggregate demand curve is convex, one expects such discounting to be more effective, since the gain in area D will typically be larger than if the function is concave.

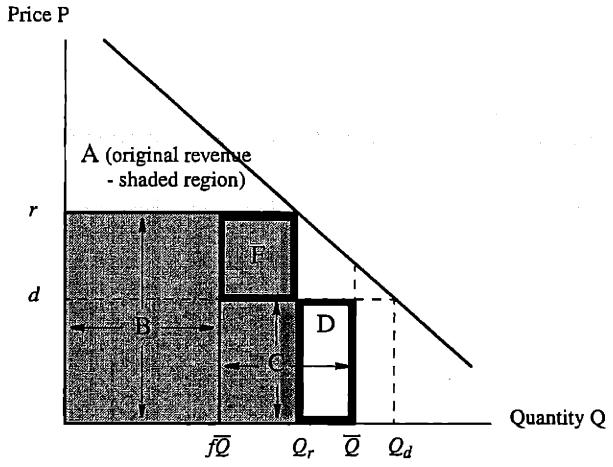


Figure 8-1: Demand curve showing revenue effects of discounting. A (the shaded area) was the original revenue before discounting, $B + C$ is the revenue after discounting, we've lost revenue F and gained D.

[8]Source: *Wall Street Journal* (3 Star, Eastern (Princeton, NJ) Edition), 7 April, 1998.

8.1.3 Modeling directions

The modeling directions we pursue are motivated by the debate of section 8.1.1 over the revenue generating potential of our service. The question of when one can gain revenue using real-time discounts motivates investigating conditions on demand necessary to achieve increased revenues. Questions of interest here, for instance, could be: under optimal decisions by the controller, what demand functions will achieve increased revenues? Do we achieve an increase in revenue because people call more during discount periods or because they call longer ? When demand is convex or concave ? More or less, these questions relate to qualitative behavior of demand. This is useful in the context of our work because we do not have real-data to assign cardinal properties to demand functions, which usually involves assigning absolute values to utility functions of consumers.

To illustrate the questions of interest here, consider the simplest case of a single consumer. At rate r per call-minute, say this person makes K_r calls on average, over some period of interest, to a certain location and each call is usually C_r minutes long, making her total average call-minutes $x_r = K_r C_r$ and her total bill rx_r. Now suppose one offers her a reduced rate d and she now calls K_d times and each call is C_d minutes long, on average. We expect that she calls at least as often with the reduced rate and at least as long on average, so $K_d \geq K_r$ and $C_d \geq C_r$. With $x_d = K_d C_d$, the total revenue is dx_d. Now clearly we make more money only when $dx_d \geq rx_r$ or $d/r \geq x_r/x_d$. This condition can be written in terms of the demand elasticity E_d for mean holding-times per call as $E_d \leq -1$ (elasticity is the preferred measure since it is ordinal rather than cardinal). For instance, for a linear demand curve, this indicates that we can only make money if the current rate is in the upper portion of the demand curve (see section 8.3 for a short reminder of demand elasticities).

This line of questioning is not unlike micro-economic theory, i.e. attempting to understand the structure of demand functions that result in increased revenue. Recall how in micro-economics, a usual first interest is in describing the qualitative behavior of the market. This leads to the investigation of more or less ordinal or qualitative properties of demand curves and utility functions. Usually Econometrics is used later to investigate their cardinal properties and validate the theory. Elasticity estimates, for instance, are usually obtained by surveys and interviews (see section 8.1.1 for an elasticity estimate of long-distance calls reported in literature). Similarly, we reason that once desirable qualitative behavior of demand has been described, marketing studies can be carried out to determine if consumer response is indeed expected to result in added revenue from such a service.

8.1.4 Literature review

Pricing literature in other industries is vast and is difficult to categorize in a short section. A recent paper of interest for telecommunications is by Paschalidis and Tsitsiklis [PT98] which studies optimal pricing of calls as a function of available capacity when the arrival rate is a function of the price and the arrival process is Poisson. It is shown that fixed price policies are not too far from optimal when the call-holding times are exponential and the system behaves as a loss system.

Other pricing research in telecommunications has focused more on usage-sensitive pricing of Internet services, rather than discounting or spot-pricing of long-distance calls. Examples are Kelly [KMT98] and Mackie-Mason [MMV95], for instance.

Related papers of interest, in the sense that the underlying models consider intensity control for optimal pricing, similar to [PT98] but not in a telecommunications setting, are Gallego et. al. [GVR94] where the pricing problem is studied in a perishable-inventory pricing setting – the traditional yield management problem. Gallego also provide a good review of pricing research,

We have so far not found anything that views demand in the way that we do and where the interest is on the trade-off between displacing revenue from higher-paying call-minutes to discount call-minutes.

8.2 Modeling Discount Offerings

In this section we first informally consider various factors relevant to determining the discounting decisions. We then develop a restricted model to make a case for lines of analysis different from traditional pricing models. Specifically, the model illustrates how call-holding times are much more important than call arrival rates, in the sense that one can actually lose more revenue by attempting to increase the call arrival rates if the holding times do not increase appreciably.

8.2.1 Discussion

Suppose that the general problem is to announce to each location at some interval, say every three hours, discounts to a set of destinations for each fifteen minute time-slot, for the next three hours. Consider the following factors that might influence such discounts.

1. *Utilization*: Physical factors are the most obvious. When utilization is expected to be high, discounts obviously do not make much sense, both because of the possibility of degrading the quality of service and the fear of cannibalizing revenue from already existing demand.

2. *The time-of-day effect on the 'take-up' rate*: The time of day is expected to influence the number of people who 'see' the advertised offerings. Since people are usually busier on weekdays than weekends and in the afternoons than in the evening, the take-up rate for an offering should depend on the time-of-day. This has some implications for the offering decision, as one might like to offer 'really deep' discounts only at times when relatively fewer and select people are at home, such as 'retirees' or 'students'.

3. *The effect of discount on immediate demand*: Possible effects to keep in mind here could include: (i) the increase in current utilization may not be sufficient to compensate for the loss of current revenue due to discounting and (ii) there may be a *future* loss of revenue because people who call, say once a week, might use the discount periods to satisfy their needs.

4. *Information about past discount responses*: Suppose one observes an unusually large number of discount calls from A to B over the past half a day or so. It is not too unreasonable to infer then that most of the 'discount demand' that could have taken place for that day has already done so, and further discounting will only lead to 'cannibalizing' revenue from other higher paying demand.

5. *Longer-term customer learning behavior*: Finally, one needs to consider how the presence of discounts will change the behavior of the consumers. For instance, too much predictability in discount offerings might prompt some users to never call unless there is a discount. Offerings should therefore be appropriately randomized both in time and in the amount of discount to resist the forming of regular expectations on part of the consumers. However, one stills needs some degree of regularity to maintain user perception that such a service is of some value to them.

8.2.2 A model

Since incorporating all of the above factors in a first model is difficult, we consider a very simplified case here. With minimal assumptions on the demand functions, we develop a plausible model for expressing the *additional revenue* in response to a discount. This allows us to understand conditions on demand that are necessary to achieve increased revenues from such discounting.

Consider the simplest case where the controller has only one decision: what discount rate to offer for the next Δt (say one hour) to a *single destination*. Assume no more discounts will follow in later periods and there are no other complicating factors. We focus on the effect of this discount on

the users and ignore the time-of-day effects (which can govern how many people actually 'see' the advertised discount offering). Let us also forget capacity constraints for the time being.

Suppose S subscribers have signed up for the service. This is the population of interest to us. Let the decision variable $d \in [0,1]$ be the fraction of the regular rate r offered to this destination, i.e. the discounted rate is $d.r$. Then for any offered discount d:

- Call the fraction of people who actually 'see' the discount q.
- Let the fraction of the ones who 'see' the discount and decide to use it be $p(d)$, which depends on the discount. We can reasonably expect that $p(d)$ should be monotonically non-increasing in d. $p(d)$ can also be viewed as the probability of a customer using the discount when all customer actions are viewed to be independent.
- Call $g(d) = 1/\mu(d)$ the mean holding time of calls made at discount d. The mean holding time of un-discounted calls is expressed as $g(1) = 1/\mu(1)$, and again we expect only that $g(d)$ should be monotone non-increasing in d, a reasonable assumption. All holding times are assumed independent random variables.

Then without any other assumptions on the functions $p(d)$ and $g(d)$ we can write the expected revenue from discount calls in period Δt as $R(d) = q\ S\ r\ d\ p(d)\ g(d)$. The quantity $q.p(d).S$ can be interpreted as Δt times the arrival rate of discount calls over Δt, allowing us to define $\lambda(d) = q.p(d).S/\Delta t$. All we have done so far is to write the expected revenue $R(d)$ from discount calls as $\lambda(d).g(d).d.r.\Delta t$, where the only assumptions on the arrival rate function $\lambda(d)$ and the mean holding time function $g(d)$ are that they are monotone non-increasing in d and further, that $\lambda(d) \leq S/\Delta t$ (an assumption that every subscriber makes at most one call).

Now the question is: where is this revenue $R(d)$ coming from? Simple reasoning can be used to apportion this revenue to one of several causes, again not assuming that one can determine exactly the cause of the call, but simply to derive an optimization problem. Each discount call that is actually made falls into one of the following categories:

- *Discretionary call*: This is a call that would not otherwise have occurred if there was no discount and does not affect future demand for this customer. Call the probability of a discount call being discretionary $f_1(d)$. Again, dependence on the discount seems reasonable, since people may not make discretionary calls if the discount is not enough. We assume only that $f_1(d)$ is monotone non-increasing in d. Any revenue from such a call is net positive revenue.
- *A "Dial-around" call*: This is a call that would otherwise have been made using a "dial-around" number. The customer, on seeing the discount, decides to use the offering instead of the dial-around. See section 8.1.1 for the importance of the dial-around market for discount calls. Call the probability of a call being of this type $f_2(d)$, which is monotone non-increasing in d. Again, any revenue from such a call is net positive revenue since we would have lost those call minutes to a dial-around provider.
- *A 'Need based' call*: This is a call the customer would have placed 'anyway', perhaps as part of her regular calling pattern, or for some other reason. The person saw the discount and decided to use it instead of paying the regular rate. Here, we distinguish two further cases:
  - With probability l, the call does not affect future calling pattern of this customer, as in she does not wait only for discounts, not calling at other times. In this case, loss or gain of revenue is limited to the current call.
  - With probability $1 - l$, It affects future calling patterns, in which case loss or gain of revenue is from this call and the future expected losses and gains from altered customer behavior. Here we need an approximation of the future costs in addition to the current revenue to decide whether we gain or lose over the long-term.

Since we are considering a discount only for one period and no more discounts, we can ignore the case of future costs due to changed customer behavior for now. Let $f(d) = f_1(d) + f_2(d)$, the probability of call being either "dial-around" or "discretionary". Also, when a call is "need-based", the "net additional revenue" $r(d)$ from such a call may be written as $q.p(d).S.[d.r.g(d) - r.g(1)] = R(d) - \lambda(d).g(1).r.\Delta t$. This leads to the following optimization problem for the controller:

$$\begin{aligned} \max \;\; & f(d)R(d) + \big[1 - f(d)\big]r(d) \;\; = \;\; R(d) - \big[1 - f(d)\big]\lambda(d)g(1)r\Delta t \\ \text{s.t.} \;\; & 0 \leq d \leq 1. \end{aligned}$$

With a few manipulations, we can write the objective function as $r.\Delta t.g(1)\lambda(d).[dg(d)/g(1) - 1 + f(d)]$. Since $r.\Delta t.g(1)$ is just a positive constant, we ignore it and end up re-writing the optimization problem as follows, with the objective function being a fixed fraction of the 'net' additional revenue due to discount d:

$$\begin{aligned} \max \;\; & \lambda(d) \left[d\frac{g(d)}{g(1)} - 1 + f(d) \right] \\ \text{s.t.} \;\; & 0 \leq d \leq 1. \end{aligned}$$

This formulation immediately leads to the following observations.

- The optimal value of the program above is positive only if there is a d in the $[0, 1]$ interval where $[dg(d)/g(1) - 1 + f(d)] > 0$, regardless of the arrival rate $\lambda(d)$, since the arrival rate is always non-negative. The interpretation of this is of some interest. It implies that arrival rate control is of no value unless the holding times increase proportionally to make up for the discounting in rate. This is contrary to most pricing models in literature where arrival rate control is used to increase revenue using spot-pricing. In fact, it is clear from our simple formulation that if one increases arrival rates while $[dg(d)/g(1) - 1 + f(d)] < 0$, one loses more in terms of net revenue due to additional 'cannibalization' of the regular call-minutes. For instance, we get no additional revenue at $d = 1$ (no discounting) since the objective function has value 0. Similarly, since $1 - f(0) \leq 0$, we always lose revenue at $d = 0$ since no money is being made – the rate of loss of revenue is then $\lambda(0)(1 - f(0))$.
- Consider the condition for getting positive revenue from a discount d, i.e. $[dg(d)/g(1) - 1 + f(d)] > 0$. We rewrite it in a few forms and note the interpretations:
  - Writing the condition as:
    $$d \geq \frac{\mu(d)[1 - f(d)]}{\mu(1)},$$
    we note the interpretation that one cannot discount more than the ratio of the mean holding times of regular calls vs. non-discretionary calls. This establishes a lower-bound on the discount, the ratio being a notion of additional demand 'volume' generated.
  - Writing it as
    $$E_d \leq \frac{1}{d}\left[\frac{f(d)}{1 - d} - 1\right],$$
    where E_d is the demand elasticity of the mean holding times $g(d)$, we have an upper-bound on the elasticity of the demand curve. Since $E_d \leq 0$ always, if $f(d)/(1 - d) > 1$, we always make a profit. The condition $f(d)/(1 - d) > 1$ expresses the fact that the proportion of discretionary calls is at least the same as the percentage discount offered, regardless of the ratio of the holding times.
  - Finally, questions that merit further attention relate to the behavior of the quantity $[dg(d)/g(1) - 1 + f(d)] > 0$ when the functions $g(d)$ and $f(d)$ are convex/concave etc. In this case, we do not pursue results but note that:

* When both the mean holding times $g(d)$ and the discretionary call probability $f(d)$ are linear differentiable (to exclude piecewise linear functions) on $0 \leq d \leq 1$, there always exists a discount d at which one can achieve positive revenue.
* When both $g(d)$ and $f(d)$ are exponential, in limited experiments, we could not find any value of $d \in [0, 1]$ for which the revenue came out positive.

8.2.3 Interesting research directions

The above section reveals the critical effect revenue cannibalization can have on the success of the proposed service. We make the case for further experimentation with the ideas presented here, for this issue has implications in other areas, such as airline revenue management, for example. If it is established that for some demand function, a benefit is achievable from such discount offerings, other directions such as the following can be explored.

- The impact of a discount on future revenues.
- Incorporation of capacity constraints.
- Extension to multiple destinations.
- Announcing discounts for multiple periods.

8.3 A Note on Demand Elasticities

Given a demand curve expressing the demanded quantity Q of a 'good' as a function of unit price P, the price elasticity of demand E_p is defined as:

$$E_p = \frac{\%\Delta Q}{\%\Delta P} = \frac{\Delta Q/Q}{\Delta P/P} = \frac{P}{Q}\frac{\Delta Q}{\Delta P} = \frac{\partial \ln Q}{\partial \ln P}$$

Elasticity is a preferred measure for dealing with demand curves because it indicates the percent change in quantity with a percent change in price, and is not the same as the slope of the curve. This is easiest to illustrate for a linear demand curve of the form $Q = a - bP$ as shown in figure 8-2, which has constant slope equaling $-b$, but it elasticity is not constant and in fact equals $-bP/Q$, a quantity that changes over the curve from $-\infty$ at $Q = 0$ to 0 at $Q = a$.

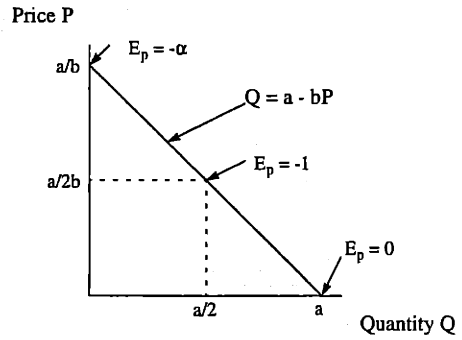


Figure 8-2: Elasticities of a linear demand curve.

8.4 Summary

This chapter demonstrates how pricing models fit in our modeling framework of section 2.2.4, using the context of a quasi real-time discounting service for long-distance calls. We articulated the idea for the service and argued its viability in some detail. We formulated a model to make the case for pursuing some non-traditional analytical directions when determining the discounting decisions – such as the possible cannibalization of existing demand and the recognition that not all demand behavior might achieve increase revenues.

8.4.1 Contributions

This chapter creates more questions than answers, but that is exactly its contribution. It makes the case for alternative lines of analysis for pricing research in revenue management, focusing on issues of practical interest. Several differences exist in the modeling approach of this chapter from existing pricing literature, both in telecommunications and otherwise. Existing research usually assumes known demand functions which express the arrival rate of calls as a function of price. The objective of interest is to determine optimal pricing policies given some assumptions about the arrival rate process (c.f. 8.1.4). By contrast, our approach is to first determine the factors that are relevant for revenue increase, whether arrival rates or holding times. This is because we do not view the demand for call minutes as as infinite, a reasonable situation considering the subscriber-ship structure of the current market.

Chapter 9

Summary, Contributions and Future work

Figure 9-1 is the final summary of thesis and a concise schematic of both the work and its context.

This thesis attempted to (i) show how the intuition and strategies of airlines Yield Management might be applied to telecom and (ii) highlight how Operations Research can help in the modeling and analysis of this new area. In spite of popular opinion that telecom shares the perishable inventory and negligible marginal cost characteristics of airlines, no one has yet coherently described how one might 'yield manage' telecom networks. A significant problem blocking this is the nature of telecom services, which are vastly different from airlines.

We proposed in chapter 1 that new services should be created to 'yield manage' telecom networks. These services must be designed explicitly to use only spare capacity and segment the market to generate extra revenue. We argued in detail the need for this approach and its relevance to the existing telecom industry, fully realizing that this will require an investment in network infrastructure for managing these new services. However, this is a familiar exercise for telecom operators and creation of new services is an active part of telecom operations.

The managing infrastructure for these new services will consist of software embedded in the network, with the purpose of minimizing their impact on existing traffic and maximizing revenue from available capacity. Much of the behavior of this software will be based on decision rules that incorporate network information to make intelligent choices. It is here that OR can play a strong role. Models that result in optimal decisions regarding pricing, capacity allocation, forecasting and so forth will be needed for each service's operation.

Recognizing the complicated models that could result from trying to incorporate all decisions in a single network-wide model for a service, we proposed using a framework to identify and decouple the decisions. This more or less parallels the airlines practice of breaking up the system-wide YM problem into chunks to make the problems tractable. A similar framework might help guide modeling choices for these new services.

To that end, we proposed borrowing the airlines modeling framework of *forecasting, over-booking, seat-inventory control, pricing* and *market segmentation* for telecom. Even though the framework classifies decision problems arising in airlines operations, we realized that the abstract versions of the problems are very relevant for telecom. We therefore translated the framework by describing how several decisions arising from the operations of each service can be identified and related to each framework category.

The majority of the thesis was the new services we proposed to illustrate our argument and the models we formulated to demonstrate our modeling framework. For each service idea, we attempted as much as possible to describe the likely markets and the most obvious manner in which the

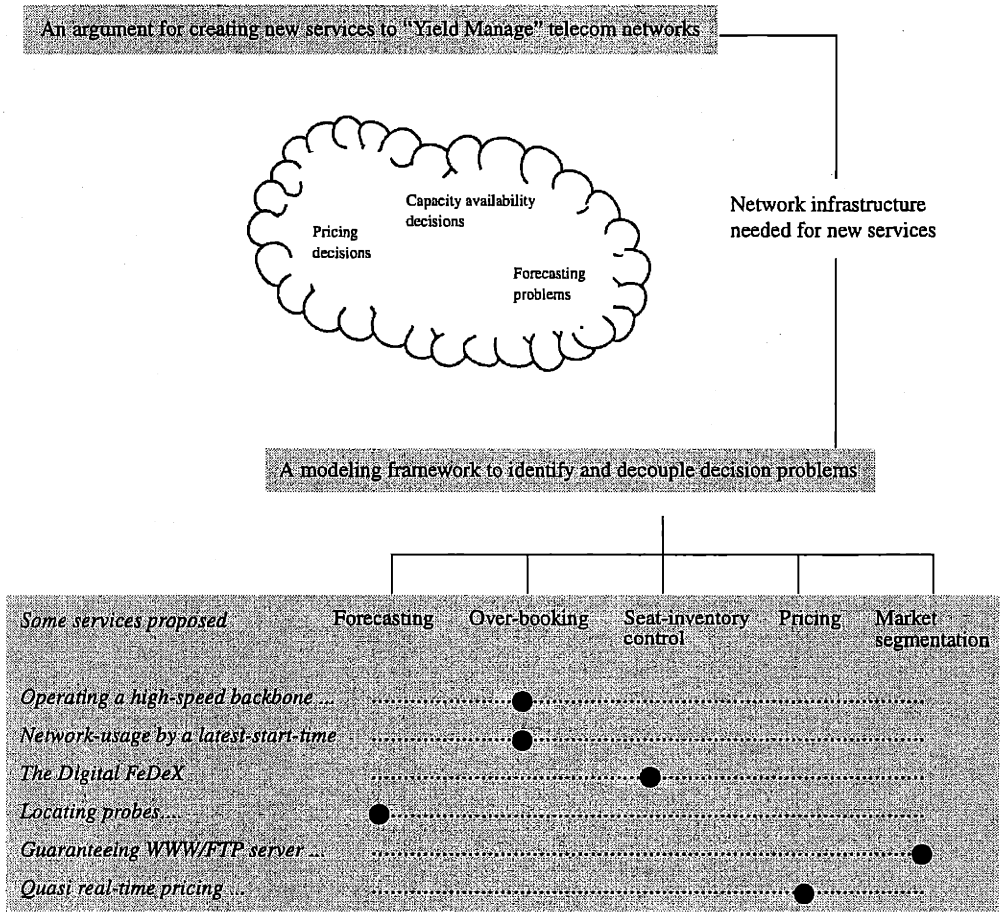


Figure 9-1: A summary of work in this thesis. The •'s are the models attempted.

offering might be practically managed. We then modeled one of the possible decisions identified by the framework. Most of our models were single-link and were meant to be illustrative of the analyses and results one might obtain from modeling.

It briefly deserves mention that significant effort has been expended in imagining the new services and in thinking of their practical operations. This looks obvious in hindsight, but actually takes a lot of ingenuity and understanding of both the technology and the markets, and research into the possibilities.

9.1 Contributions and Future Work

It does not make sense to list all possible contributions of this work here for two reasons. First, if it does indeed help motivate YM in telecom, there will be too many. If it does not, there will be few. Second, the breadth and texture of this work makes it more useful to list the contributions where they are most relevant. For these reasons, general contributions of this work to telecom YM and to OR are listed in sections 1 and 1.2. Contributions of the modeling work are mentioned the end of each modeling chapter. Here we briefly mention some last words on this issue.

It is, to our knowledge, the first work that outlines a coherent argument for making telecom YM practical. In fact, we are not aware of any other work of this scope related to revenue management

in telecom. In that sense, its value could also be in something as simple as a starting collection of ideas, exciting further work on the topic.

It is also useful to comment on what has been achieved in this thesis and more importantly, what remains undone. It is not our intention to claim that the services proposed and the models formulated here will be implemented exactly as prescribed. It should also be clear that the work here exemplifies only one aspect of the operational problems raised by offering YM services on a network. In that sense, our work is only illustrative of the services and models that could arise to support serious YM efforts in telecom, and does not address several major aspects such as network software design, marketing of the new services and the regulatory issues that will arise.

We see the future modeling work as refining and enriching the framework as well as validating the models. This will be tightly coupled to the marketing and infrastructure development, both of which will have a huge effect on the success of the ideas proposed here.

Bibliography

[All97] S. Allot. Network Management: A Core Skill for Future Telcos. *Telecommunications*, August 1997.

[All98] D. Allen. Special Canada Supplement: Interview with David Colville, vice chairman, CRTC. *Telecommunications*, December 1998.

[BAL94] M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand.* The MIT Press, Cambridge, Massachusetts, 1994.

[Bay96] K. Baynton. Managing the Growth of the Internet. *Telecommunications*, June 1996.

[BB58] M. J. Beckmann and F. Bobkowski. Airline Demand: An Analysis of Some Frequency Distributions. *Naval Research Logistics Quarterly*, **5**:43–51, 1958.

[BBL95] O. Berman, D. Bertsimas, and R. C. Larson. Locating Discretionary Facilities, II: Maximizing Market Size, Minimizing Inconvenience. *Operations Research*, Vol. 43, No. 4:pp. 623–632, July-August 1995.

[BC98] P. Barford and M. E. Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. In *Proceedings, 1998 ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, pages 151–?, Madison, Wisconsin, USA, June 1998.

[BCL+90] O. Berman, S. S. Chiu, R. C. Larson, R. Odoni, A, and R. Batta. Location of Mobile Units in a Stochastic Environment. In P. B. Mirchandani and R. L. Francis, editors, *Discrete Location Theory*, chapter **12**, pages 503–549. John Wiley & Sons, New York, 1990.

[Bel87] P. P. Belobaba. *Air Travel Demand and Airline Seat Inventory Management.* Ph.D. dissertation, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[Ber95] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. I.* Athena Scientific, Belmont, Massachusetts, 1995.

[BLF92] O. Berman, R. C. Larson, and N. Fouska. Optimal Location of Discretionary Service Facilities. *Transportation Science*, Vol. 26, No. 3:pp. 201–211, 1992.

[BM93] S. L. Brumelle and J. I. McGill. Airline Seat Allocation with Multiple Nested Fare Classes. *Operations Research*, **41**:127–137, 1993.

[Bor98] S. L. Borthik. Why We Can't Compare ISP Performance Yet. *Business Communications Review*, **28**(9):35–40, September 1998.

[Bot94] T. C. Botimer. *Airline Pricing and Fare Product Differentiation.* Ph.D. dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1994.

[BP73] A. V. Bhatia and S. C. Parekh. Optimal Allocation of Seats by Fare. Presentation to AGIFORS Reservations Study Group, Trans World Airlines, 1973.

[Bru97] M. R. Bruneau. You'll Never Guess Who Wants to be Your Phone Company. *Telecommunications*, October 1997.

[Bul97] J. Bulkeley. Doing Business on the Net. *Telecommunications*, January 1997.

[Cau99] B. Caulfield. Range of Customer Needs Reshapes a Young Industry: In growing market, hosting firms seek a way to stand out. *ISP World*, February 1, 1999. http://www.internetworld.com/print/1999/02/01/ ispworld/19990201-range.html.

[CB96] M. E. Crovella and A. Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. In *Proceedings, 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 160–169, Philadelphia, USA, May 1996.

[CC95] M. E. Crovella and R. L. Carter. Dynamic Server Selection in the Internet. In *Third IEEE Workshop on the Architecture and Implementation of High Performance Communications Systems.* (HPCS '95), 1995. Available from http://www.cs.bu.edu/faculty/crovella/papers.html.

[CC96] R. L. Carter and M. E. Crovella. Dynamic Server Selection using Bandwidth Probing in Wide-Area Networks. Technical Report BU-CS-96-007, Computer Science Department, Boston University, March 1996. Available from http://www.cs.bu.edu/faculty/crovella/papers.html.

[CCI72] G.M. Carter, J. M. Chaiken, and E. Ignall. Response Areas for Two Emergency Units. *Operations Research*, **20**:571–594, 1972.

[CL85] S. S. Chiu and R. C. Larson. Locating an n-Server Facility in a Stochastic Environment. *Computers and Operations Research*, Vol. 12:509–516, 1985.

[Cla97] D. D. Clark. An Internet Model for Cost Allocation and Pricing. In L. W. McKnight and J. P. Bailey, editors, *Internet Economics.* MIT Press, 1997.

[Cla99] K. C. Claffy. Internet Measurement and Data Analysis: Topology, Workload, Performance and Routing Statistics. In *NAE '99 workshop.* Cooperative Association for Internet Data Analysis (CAIDA), 1999. Available from http://www.caida.org/outreach/papers/Nae.

[CSEZ93] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang. Pricing in Computer Networks: Motivation, Formulation and Example. *IEEE/ACM Transactions on Networking*, **1**(6):614–627, December 1993.

[Dal76] D. J. Daley. Queueing Output Processes. *Advances in Applied Probability*, **8**(2):395–415, June 1976.

[Dis96] D. E. Dismukes. The Demand for Long Distance Telephone Communication: A Route-Specific Analysis of Short-Haul Service. *Studies in Economics and Finance*, Vol. 17(1), Fall 1996.

[Fal69] L. M. Falkson. Airline Overbooking: Some Comments. *Journal of Transport Economics and Policy*, **3**:352–354, 1969.

[For98] H. Ford. Building Intranets: The Business Case. *Telecommunications*, December 1998.

[Gal96] R. G. Gallager. *Discrete Stochastic Processes.* Kluwer Academic Publishers, Norwell, Massachusetts, 1996.

[GGLM82] F. Glover, R. Glover, J. Lorenzo, and C. McMillan. The Passenger Mix Problem in the Scheduled Airlines. *Interfaces*, **12**:73–79, 1982.

[GH85] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory.* John Wiley & Sons, New York, 2nd edition, 1985.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman and Company, San Francisco, 1979.

[GS95] J. D. Guyton and M. F. Schwartz. Locating Nearby Copies of Replicated Internet Servers. Technical Report CU-CS-762-95, Department of Computer Science, University of Colorado at Boulder, February 1995. Available via ftp from http://fermivista.math.jussieu.fr/ftp/ftp.cs.colorado.edu.html.

[GVR94] G. Gallego and G. J. Van Ryzin. Optimal Dynamic Pricing of Inventories with Stochastic Demand over Finite Horizons. *Management Science*, **40**(8):999–1020, August 1994.

[GVR97] G. Gallego and G. J. Van Ryzin. A Multi-Product Dynamic Pricing Problem and its Applications to Network Yield Management. *Operations Research*, **45**:24–41, 1997.

[Hak64a] S. L. Hakimi. Optimum Distribution of Switching Centers in a Communication Network and Some Related Graph Theoretic Problems. *Operations Research*, Vol. 13:462–475, 1964.

[Hak64b] S. L. Hakimi. Optimum Location of Switching Centers and Absolute Centers and Medians of a Graph. *Operations Research*, Vol. 12:450–459, 1964.

[Hal99] D. Hall. The Call of the Web. *Telecommunications*, January 1999.

[HM83] P. Harris and G. Marucci. A Short Term Forecasting Model. In *AGIFORS Symposium Proceedings*, **23**, Memphis, TN, 1983.

[Kau99] D. H. Kaufman. Delivering Quality of Service on the Internet. *Telecommunications*, February 1999.

[Kel79] F. P. Kelly. *Reversibility and Stochastic Networks.* John Wiley & Sons, New York, 1979.

[Kel91] F. P. Kelly. Loss Networks. *Annals of Applied Probability*, **1**:319–378, 1991.

[Kel96] F. P. Kelly. Notes on Effective Bandwidths. In F.P. Kelly, S. Zachary, and I.B. Ziedins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

[Kel97] F. P. Kelly. Charging Schemes for Multiservice Networks. In V. Ramaswami and P. E. Wirth, editors, *Teletraffic Contributions for the Information Age, ITC15*, pages 781–790. Elsevier, Amsterdam, 1997.

[Kle75] L. Kleinrock. *Queueing Systems, Vol I: Theory.* John Wiley & Sons, New York, 1975.

[KMT98] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability. *Journal of the Operational Research Society*, **49**:237–252, 1998.

[KVR98] I. Karaesman and G. J. Van Ryzin. Overbooking with Substitutable Inventory Classes. Working Paper, Graduate School of Business, Columbia University, New York, NY, 1998.

[Lee90] A. O. Lee. *Airline Reservations Forecasting: Probabilistic and Statistical Models of the Booking Process.* Ph.D. dissertation, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1990.

[L'H86] E. L'Heureux. A New Twist in Forecasting Short-Term Passenger Pickup. In *AGIFORS Symposium Proceedings,* **26**, Bowness-on-Windermere, England, 1986.

[Lit72] K. Littlewood. Forecasting and Control of Passenger Bookings. In *AGIFORS Symposium Proceedings,* **12**, Nathanya, Israel, 1972.

[LTWW94] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, Vol. 2(1):1–15, February 1994.

[May76] M. Mayer. Seat Allocation, or a Simple Model of Seat Allocation via Sophisticated Ones. In *AGIFORS Symposium Proceedings,* **16**, Key Biscayne, FL, 1976.

[MB97] L. W. McKnight and J. P. Bailey. *Internet Economics.* The MIT Press, Cambridge, Massachusetts, 1997.

[McG89] J. I. McGill. *Optimization and Estimation Problems in Airline Yield Management.* Ph.D. dissertation, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, BC, 1989.

[McG95] J. I. McGill. Censored Regression Analysis of Multiclass Demand Data Subject to Joint Capacity Constraints. *Annals of Operations Research*, **60**:209–240, 1995.

[MF90] P. B. Mirchandani and R. L. Francis. *Discrete Location Theory.* John Wiley & Sons, New York, 1990.

[MMV95] J. K. MacKie-Mason and H. R. Varian. Pricing the Internet. In B. Kahin and J. Keller, editors, *Public Access to the Internet*, pages 269–314. The MIT Press, Cambridge, MA, 1995.

[MVR99] J. I. McGill and G. J. Van Ryzin. Revenue Management: Research Overview and Prospects. *Transportation Science*, **33**(2):233–256, May 1999.

[MZH83] N. Megiddo, E. Zemel, and S. L. Hakimi. The Maximum Coverage Location Problem. *SIAM Journal of Algebraic and Discrete Methods*, Vol. 4, No. 2:pp. 253–261, June 1983.

[Nag79] K. V. Nagarajan. On an Auction Solution to the Problem of Airline Overbooking. *Transportation Research*, **13A**:111–114, 1979.

[Nol96] Micheal A. Noll. Sizing the Industry: the Numbers Tell the Story. *Telecommunications*, September 1996.

[NR78] S. Nozaki and S. Ross. Approximations in Finite-capacity Multi-server Queues with Poisson Arrivals. *Journal of Applied Probability*, No. 15:pp. 826–834, 1978.

[NW78] G. L. Nemhauser and L. A. Wolsey. Best Algorithms for Approximating the Maximum of a Sub-modular Set Function. *Mathematics of Operations Research*, Vol. 3:pp. 177–188, 1978.

[NW88] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization.* John Wiley and Sons, Inc., New York, 1988.

[NWF78] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An Analysis of Approximations for Maximizing Sub-modular Set Functions, I. *Mathematical Programming*, 14:pp. 265–294, 1978.

[Pax97] V. E. Paxson. *Measurements and Analysis of End-to-End Internet Dynamics.* PhD thesis, Computer Science Division, University of California, Berkeley, 1997.

[PF94] V. Paxson and S. Floyd. Wide-Area Traffic: The Failure of Poisson Modeling. In *Proceedings, 1994 SIGCOMM Conference*, pages 257–268, London, UK, August 31st - September 2nd 1994. ACM.

[PKC96] K. Park, T. G. Kim, and M. E. Crovella. On the Relationship Between File Sizes, Transport Protocols and Self-Similar Network Traffic. Technical Report BU-CS-96-016, Computer Science Department, Boston University, August 1996. Available from http://www.cs.bu.edu/faculty/crovella/papers.html.

[PR89] R. S. Pindyck and D. L. Rubinfeld. *Microeconomics.* Macmillan Publishing Company, New York, 1989. pp. 383-384.

[PT98] I. C. Paschalidis and J. N. Tsitsiklis. Congestion Dependent Pricing of Network Services. Technical report, Systems Group, Department of Manufacturing Engineering, Boston University, Boston, MA, October 1998. To appear in IEEE/ACM Transactions on Networking. Available via ftp from http://ionia.bu.edu/publications/main.html.

[Ric82] H. Richter. The Differential Revenue Method to Determine Optimal Seat Allotments by Fare Type. In *AGIFORS Symposium Proceedings,* **22**, Lagonissi, Greece, 1982.

[RMVe96] J. W. Roberts, U. Mocci, and J. Virtamo (editors). *Broadband Network Teletraffic: Performance Evaluation and Control of Broadband Multiservice Networks.* Springer-Verlag, Berlin Heidelberg, July 1996.

[Rob95] L. W. Robinson. Optimal and Approximate Control Policies for Airline Booking with Sequential Nonmonotonic Fare Classes. *Operations Research,* **43**:252–263, 1995.

[Ros95] K. W. Ross. *Multiservice Loss Models for Telecommunication Networks.* Springer-Verlag, London, 1995.

[Rot68] M. Rothstein. *Stochastic Models for Airline Booking Policies.* Ph.D. dissertation, Graduate School of Engineering and Science, New York University, New York, NY, 1968.

[Rot71] M. Rothstein. An Airline Overbooking Model. *Transportation Science,* **5**:180–192, 1971.

[Rot85] M. Rothstein. O.R. and the Airline Overbooking Problem. *Operations Research,* **33**:237–248, 1985.

[RS67] M. Rothstein and A. W. Stone. Passenger Booking Levels. In *AGIFORS Symposium Proceedings,* **7**, Noordwijk, The Netherlands, 1967.

[Sa87] J. Sa. Reservations Forecasting in Airline Yield Management. Master's thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[Sim68] J. Simon. An Almost Practical Solution to Airline Overbooking. *Journal of Transport Economics and Policy,* **2**:201–202, 1968.

[Sim72] J. Simon. Airline Overbooking, The State of the Art — A Reply. *Journal of Transport Economics and Policy,* **6**:254–256, 1972.

[SLD92] B. C. Smith, J. F. Leimkuhler, and R. M. Darrow. Yield management at American Airlines. *Interfaces,* **22**(1):8–31, January-February 1992.

[SP88] B. C. Smith and C. W. Penn. Analysis of Alternative Origin-Destination Control Strategies. In *AGIFORS Symposium Proceedings,* **28**, New Seabury, MA, 1988.

[Swa90] W. M. Swan. Revenue Management Forecasting Biases. Working Paper, Boeing Commercial Aircraft, Seattle, WA, 1990.

[Tay68] C. J. Taylor. The Determination of Passenger Booking Levels. In *AGIFORS Symposium Proceedings*, **2**, Fregene, Italy, 1968.

[TMW97] K. Thompson, G. J. Miller, and R. Wilder. Wide-Area Internet Traffic Patterns and Characteristics. *IEEE Network*, Vol. 11(6), November-December 1997. Extended version of the paper obtained from http://www.vbns.net/presentations/papers/index.html.

[Tra97] Lenore V. Tracey. 30 Years: A Brief History of the Communications Industry. *Telecommunications*, June 1997.

[Var95] H. Varian. Pricing Information Goods. Presented at Harvard Law School Research Libraries Symposium, May 2-3 1995.

[Vic72] W. Vickrey. Airline Overbooking: Some Further Solutions. *Journal of Transport Economics and Policy*, **6**(3):257–270, September 1972.

[VRM98] G. J. Van Ryzin and J. I. McGill. Revenue Management Without Forecasting or Optimization: An Adaptive Algorithm for Determining Seat Protection Levels. Working Paper, Graduate School of Business, Columbia University, New York, NY and Queen's University, Kingston, ON, 1998.

[WB92] L. R. Weatherford and S. E. Bodily. A Taxonomy and Research Overview of Perishable-Asset Revenue Management. *Operations Research*, **40**:831–844, 1992.

[Whi84a] W. Whitt. Approximations for Departure Processes and Queues in Series. *Naval Research Logistics Quarterly*, **31**(4):499–521, December 1984.

[Whi84b] W. Whitt. Departures from Queues with Many Busy Servers. *Mathematics of Operations Research*, **9**(4):534–44, November 1984.

[Whi88] W. Whitt. A Light-Traffic Approximation for Single-Class Departure Processes from Multi-Class Queues. *Management Science*, **34**(11):1333–46, November 1988.

[Wil88] E. L. Williamson. Comparison of Optimization Techniques for Origin-Destination Seat Inventory Control. Master's thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[Wil92] E. L. Williamson. *Airline Network Seat Inventory Control: Methodologies and Revenue Impacts.* Ph.D. thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1992.

[Wol86] R. D. Wollmer. A Hub-Spoke Seat Management Model. Unpublished company report, Douglas Aircraft Company, McDonnell Douglas Corporation, Long Beach, CA, 1986.

[Wol89] R. W. Wolff. *Stochastic Modeling and the Theory of Queues.* Prentice-Hall, Englewood Cliffs, NJ, 1989.

[Wol92] R. D. Wollmer. An Airline Seat Management Model for a Single Leg Route when Lower Fare Classes Book First. *Operations Research*, **40**:26–37, 1992.

[Won90] J. T. Wong. *Airline Network Seat Allocation.* Ph.D. thesis, Northwestern University, Evanston, IL, 1990.

[WP98] W. Willinger and V. Paxson. Where Mathematics Meets the Internet. *Notices of the AMS*, **45**(8):961–970 (see page 970), September 1998.

[WPS97] Q. Wang, J. M. Peha, and M. A. Sirbu. Optimal Pricing for Integrated Services Networks. In L. W. McKnight and J. P. Bailey, editors, *Internet Economics*, pages 353–378. MIT Press, Cambridge, MA, 1997.