

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

NOTE TO USERS

The original manuscript received by UMI contains indistinct and light print. All efforts were made to acquire the highest quality manuscript from the author or school. Microfilmed as received.

This reproduction is the best copy available

UMI

Carnegie Mellon University

**Economies of Scale in Information Dissemination
over the Internet**

**A dissertation submitted to the Graduate School
in partial fulfillment of the requirements for the degree of**

Doctor of Philosophy in Engineering and Public Policy

by

John Chung-I Chuang

**Pittsburgh, Pennsylvania
November 1998**

UMI Number: 9918561

**Copyright 1998 by
Chuang, John Chung-I**

All rights reserved.

**UMI Microform 9918561
Copyright 1999, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI

**300 North Zeeb Road
Ann Arbor, MI 48103**

© Copyright, 1998, John Chung-I Chuang. All rights reserved.

Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

THESIS


SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

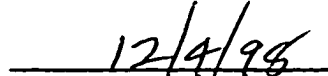
Economies of Scale in Information Dissemination over the Internet

by

John Chung-I Chuang

ACCEPTED BY THE DEPARTMENT OF ENGINEERING AND PUBLIC POLICY


Major Professor



Date


Department Head


Date

APPROVED BY THE COLLEGE COUNCIL


Dean


Date

Acknowledgments

I would like to thank my committee members, Professors Marvin Sirbu, Granger Morgan, Alex Hills and Hui Zhang, for their guidance and support. Specifically, I am most indebted to my advisor and mentor, Marvin, for his generosity and kindness. My sincere gratitude also goes to the many colleagues who have provided invaluable feedback on this work.

Financial support from the National Science Foundation (Grant IRI-9411299), the Council on Library and Information Resources A. R. Zipf Fellowship in Information Management, and the Department of Engineering and Public Policy at Carnegie Mellon University, is gratefully acknowledged.

To the many wonderful people of Pittsburgh that I have come to know, especially those of EPP and of the Sacred Heart community, I thank them for their friendship and their prayers.

I thank my family, especially my mom and dad, for their love and encouragement.

I thank my wife, Melissa, for her love and her smile.

I thank God. Amen.

Table of Contents

LIST OF FIGURES.....	IX
LIST OF TABLES	XI
ABSTRACT.....	XIII
1. INTRODUCTION.....	1
1.1 ECONOMICS OF INFORMATION AND INFORMATION DISSEMINATION	2
1.2 ECONOMIES OF SCALE: DIMENSIONS AND LEVELS	5
2. EOS IN OBJECTS - INFORMATION BUNDLING	11
2.1 BUNDLING AND UNBUNDLING OF INFORMATION GOODS	12
2.2 ECONOMICS OF BUNDLING	15
2.3 N-GOOD BUNDLING MODEL.....	17
<i>2.3.1 Modeling heterogeneity in consumer preferences.....</i>	<i>19</i>
2.3.1.1 Consumer choice in pure bundling.....	24
2.3.1.2 Consumer choice in pure unbundling.....	25
2.3.1.3 Consumer choice in mixed bundling.....	27
<i>2.3.2 Production costs and economies of scale.....</i>	<i>28</i>
2.4 ANALYSIS AND EMPIRICAL RESULTS	30
<i>2.4.1 Optimal pricing and revenue mix.....</i>	<i>33</i>
<i>2.4.2 Internet-based document delivery technology.....</i>	<i>35</i>
2.5 CONCLUSION	39
3. EOS IN RECEIVERS - MULTICAST COMMUNICATION.....	43
3.1 PRICING MULTICAST COMMUNICATION: A COST-BASED APPROACH.....	44
3.2 COST QUANTIFICATION.....	45
<i>3.2.1 Quantifying Multicast Tree Cost.....</i>	<i>47</i>
<i>3.2.2 Methodology.....</i>	<i>50</i>
<i>3.2.3 Results.....</i>	<i>53</i>
<i>3.2.4 Tree Saturation.....</i>	<i>55</i>
3.3 MULTICAST PRICING	57
<i>3.3.1 Membership Accounting.....</i>	<i>59</i>
<i>3.3.2 Other Issues.....</i>	<i>60</i>

3.4 DENSE VS. SPARSE MODE MULTICAST	61
3.5 CONCLUSION	67
4. EOS IN TIME - DISTRIBUTED NETWORK STORAGE.....	69
4.1 DISTRIBUTED NETWORK STORAGE INFRASTRUCTURE WITH QoS GUARANTEES.....	71
4.2 RELATED WORK.....	75
4.2.1 <i>Network Caching</i>	76
4.2.2 <i>Network Replication</i>	77
4.2.3 <i>Transmission-based QoS</i>	78
4.3 SERVICE SPECIFICATION.....	80
4.3.1 <i>Deterministic vs. Statistical Guarantees</i>	84
4.3.2 <i>Performance-Oriented vs. Placement-Oriented Services</i>	86
4.4 SERVICE PROVISION	87
4.4.1 <i>Resource Reservation Protocol</i>	87
4.4.2 <i>Resource Mapping</i>	88
4.4.3 <i>Admission Control</i>	90
4.4.4 <i>Service Provision Architecture</i>	92
4.5 REAL-TIME RESOURCE MANAGEMENT.....	96
4.5.1 <i>Local Storage Management</i>	97
4.5.2 <i>Traffic Policing</i>	98
4.5.3 <i>Hierarchical Resource Sharing</i>	98
4.5.4 <i>Global Storage Management</i>	101
4.6 ADDITIONAL MECHANISMS.....	101
4.6.1 <i>Resource Discovery</i>	102
4.6.2 <i>Accounting, Billing and Payment</i>	102
4.7 ECONOMICS	103
4.7.1 <i>QoS Pricing</i>	104
4.7.2 <i>Industrial Organization</i>	105
4.7.2.1 <i>Distributed Storage Economy</i>	105
4.7.2.2 <i>Spot Market, Futures Market and Supplemental Insurance</i>	105
4.7.2.3 <i>Vertical Integration and Component-Based Competition</i>	106
4.8 CONCLUSION	109
5. RESOURCE MAPPING FOR DISTRIBUTED NETWORK STORAGE SERVICES	111
5.1 MATHEMATICAL MODEL FOR RESOURCE MAPPING AND ADMISSION CONTROL.....	111
5.1.1 <i>Traffic Profile</i>	112

5.1.2 Performance Requirements.....	114
5.1.3 Resource Mapping.....	114
5.1.3.1 Service with Worst Case Delay Bound.....	115
5.1.3.2 Service with Average Delay Bound.....	116
5.1.3.3 Service with Average and Maximum Delay Bounds.....	118
5.1.3.4 Service with Stochastic Guarantees.....	118
5.1.4 Admission Control.....	119
5.2 RESOURCE MAPPING FOR ARPANET.....	119
5.2.1 Base Case: Uniform Demand Distribution, Unconstrained Replica Locations.....	121
5.2.2 Non-Uniform Spatial Demand Distribution.....	123
5.2.3 Partial Replication of Multi-Object Collection.....	124
5.3 NETWORK STORAGE CAPACITY PLANNING PROBLEM.....	127
5.3.1 Resource Mapping with Constrained Replication Server Sites.....	131
5.4 MAPPING INTO STORAGE AND TRANSMISSION RESOURCES.....	133
5.5 CONCLUSION	138
6. CONCLUSION.....	141
6.1 POLICY IMPLICATIONS AND LESSONS	141
6.2 CONTRIBUTIONS MADE IN THIS DISSERTATION	144
6.3 FUTURE WORK.....	146
APPENDIX 1. DERIVATION OF PRODUCER SURPLUS FOR ALTERNATIVE BUNDLING STRATEGIES.....	149
APPENDIX 2. SAMPLE LIST OF WEB-HOSTING SERVICE PROVIDERS.....	157
APPENDIX 3. SOURCE CODE FOR MULTICAST COST QUANTIFICATION.....	158
APPENDIX 4. TAXONOMY OF DATA DUPLICATION SCHEMES ACCORDING TO TRADITIONAL EX-POST VS. EX-ANTE DISTINCTION.....	164
APPENDIX 5. SOLUTION METHOD FOR RESOURCE MAPPING WITH PARTIAL REPLICATIONS OF MULTI-OBJECT COLLECTIONS	166
BIBLIOGRAPHY.....	169

List of Figures

Figure 1.1. Leveraging economies of scale along different dimensions (objects, receivers, time) at the information product level.....	7
Figure 1.2. Leveraging economies of scale along different dimensions (objects, receivers, time) at the bit transport level.....	8
Figure 2.1. Consumer choice regions for two-good bundling model. Alice and Bob will choose different product offerings under different bundling regimes.	16
Figure 2.2. Total outlay vs. number of articles consumed. An <i>ex post</i> two-part tariff (in bold) offers a predictable price cap on consumer expenditure.....	18
Figure 2.3. Article valuation by an individual reader indexed by $\{w_o, k\}$	21
Figure 2.4. This figure demonstrates the diversity of consumers that can be indexed by $\{w_o, k\}$	22
Figure 2.5. Consumer choice in pure bundling scenario.....	25
Figure 2.6. Optimal article consumption level in pure unbundling scenario.....	26
Figure 2.7. Consumer choice in mixed bundling scenario.....	28
Figure 2.8. Profit-maximizing bundling strategy: it is clear that mixed bundling is the dominant strategy across all marginal cost and economies of scale conditions.....	31
Figure 2.9. Optimal price ratio for mixed bundling strategy across various economies of scale and marginal cost conditions.....	34
Figure 2.10. Optimal revenue mix for mixed bundling strategy.....	35
Figure 2.11. Effect of transmission cost on journal subscription pricing.....	38
Figure 2.12. Effect of declining k_d (transmission cost) on economies of scale and revenue mix.....	39
Figure 3.1. Transmission vs. computing cost trends.....	46
Figure 3.2. Example network shows that degree of link savings achievable is strongly dependent on spatial distribution of receivers.	50
Figure 3.3. Quantifying economies of scale in multicast communication - a process overview.....	51
Figure 3.4. Normalized multicast tree length as a function of membership size - slope is constant (~ 0.8) across various network topological styles.....	54
Figure 3.5. Normalized multicast tree length as a function of membership size - slope is constant (~ 0.8) across various network sizes.	54
Figure 3.6. Normalized multicast tree length as a function of membership size - results confirmed with real networks.	55

Figure 3.7. An illustration of the “tree saturation” effect: it takes just ~500 randomly selected dial-in ports (or 0.5% of all ports) to subscribe to a multicast group before all 100 network nodes become part of the multicast tree. All subsequent subscribers can be served at no additional cost.....	57
Figure 3.8. Comparing alternatives for sending one data packet to receivers in the MBone network.	64
Figure 3.9. Comparing alternatives for sending a 5kbps data stream to receivers in the MBone network - there <i>appears</i> to be no difference between sparse and dense mode multicast.....	64
Figure 3.10. Comparing dense and sparse mode multicast for sending a 5kbps data stream to receivers in the MBone network - dense mode multicast clearly consumes more bandwidth when there are few receivers, but the two modes are comparable with subscription density as low as 4% (about 200 receivers).....	66
Figure 4.1. From performance requirements to performance realization: the process flow of establishing a network storage service with QoS guarantees. The components in bold are the key components of the infrastructure.	73
Figure 4.2. Service provision architectural alternatives.	94
Figure 4.3. Hierarchical resource sharing example.....	100
Figure 4.4. Example shows vertically integrated storage provider can internalize transmission cost savings not available to an independent storage provider.....	108
Figure 5.1. Network topology of early ARPANET.	120
Figure 5.2. Resource Mapping for ARPANET.	121
Figure 5.3. Resource mapping for non-uniform spatial distribution.....	123
Figure 5.4. Resource mapping for multi-object collections with non-uniform object distribution using full or partial replication.	125
Figure 5.5. Resource mapping for multi-object collections with non-uniform object distribution using partial replication.	126
Figure 5.6. Network capacity planning problem - different replication strategies may realize the lowest cost solution depending on the relative magnitudes of fixed (c_0) and variable (c_s) costs.....	129
Figure 5.7. Resource cost relative to storage-only solution (4 replicas) at different storage to transmission cost ratios.	136
Figure 5.8. Optimal mapping decision for storage and transmission resources (ARPANET with $\tau_{max} = 3$).....	137
Figure 5.9. Resource cost relative to storage-only solution (4 replicas).....	138

List of Tables

Table 1.1. The various dimensions and levels of EoS explored in this work.....	6
Table 2.1. Distribution of number of articles read in a journal.	23
Table 2.2. Consumer choice in mixed bundling scenario.....	27
Table 3.1. Networks used in this study.....	52
Table 3.2. Data and control/overhead for various options of sending data to multiple destinations.....	62
Table 4.1. Some examples of network storage services.....	81
Table 4.2. Entities involved in resource mapping and admission control functions in different service provision architectural alternatives.....	95
Table 5.1. ARPANET Statistics.....	120
Table 5.2. Resource Mapping for Service with Average Delay Bound.....	122
Table 5.3. Comparing mapping efficiencies for constrained versus unconstrained replication server sites.....	133

Abstract

This dissertation studies the different levels and dimensions along which economies of scale (EoS) savings may be realized when information is disseminated over the Internet. At the information product level, EoS savings may be realized along the object, consumer and temporal dimensions through strategies such as information bundling, site-licensing and subscriptions. At the information transport level, EoS savings may be realized along the same dimensions through just-in-time delivery, multicast, network caching and replication strategies. Each of these strategies is studied in this work.

Along the object dimension, a multi-product bundling model with multi-dimensional consumer taste characteristics is developed to study the optimal bundling and pricing strategy of information goods such as academic journals. Using empirical journal usage data and cost projections for information-delivery over the Internet, the model finds that metered usage (i.e., articles-on-demand) should account for a significant fraction of revenue when articles and subscriptions are optimally priced according to a mixed bundling strategy.

Along the receiver dimension, a communication cost model for multicast is developed. The model demonstrates that multicast group size can serve as an excellent proxy for multicast tree cost. Computer simulations show that, statistically, multicast tree length grows at the 0.8 power of the multicast group size until the point of tree saturation, beyond which additional receivers can be added to the group without further tree growth. In other words, the marginal cost of multicast declines according to an exponential decay function until it reaches zero at tree saturation. This result is validated

with both real and generated networks, and is robust across topological styles and network sizes. This suggests that a two-part tariff may be appropriate if providers choose to adopt a cost-based approach to multicast pricing.

Along the temporal dimension, economies of scale savings can be realized through network caching and replication. This work offers the vision of and motivation for a distributed network storage infrastructure with service guarantees. Caching and replication can be treated as different service classes within a unified Quality-of-Service (QoS) framework. Key components of the distributed network storage architecture include: service specification, resource reservation, resource mapping, admission control, real-time resource management and pricing. After establishing a research roadmap, this work focuses on the resource mapping problem and develops a formal mapping model, allowing services with different traffic profiles and performance specifications to be mapped into an optimal combination of storage and transmission resources. The model is also extended to tackle network storage capacity planning problems.

The work described in this dissertation promotes an understanding of how new network technologies have changed, and will continue to change, the economics of information dissemination. This understanding is essential to the design of engineering, economic and policy structures that will constitute the information infrastructure of the future.

1. Introduction

The Internet, designed as a packet-switched data network, excels in providing both point-to-point and point-to-multipoint communications. In addition to carrying email messages and phone conversations (both point-to-point applications), the Internet is also efficient in disseminating information to large numbers of geographically distributed recipients (a point-to-multipoint application). In this sense the Internet is radically different from the traditional public switched telephone network (PSTN) or the broadcast and cable television networks, all of which operate as either a point-to-point or a point-to-multipoint network, but not both.¹

This dissertation studies the various ways through which economies of scale savings may be realized when information is disseminated over the Internet. Economies of scale (EoS) are supply-side conditions where the average cost of a product declines as the quantity produced increases. The presence of EoS in information dissemination would result in a lower average cost when (i) multiple information objects are delivered at

¹ Multi-party teleconferencing is possible over the circuit-switched PSTN, though it is limited to a small number of participants and is extremely costly.

the same time (EoS in object dimension), (ii) an information object is delivered to multiple recipients at the same time (EoS in receiver dimension), or (iii) an information object can be reused or shared by one or more recipients at different times (EoS in temporal dimension).

Section 1.1 provides an overview of the economics of information and information dissemination, and the motivation for why we care about EoS in information dissemination. Section 1.2 identifies the different levels and dimensions along which EoS may be realized. This also serves as an outline for the rest of the dissertation.

1.1 Economics of Information and Information Dissemination

The fundamental difference between the economics of information and that of more traditional commodities is that the information economy faces high fixed costs and low marginal costs. Information production is often a labor-intensive process. In addition to input factors such as pencil and paper, it also requires much human toil and ingenuity. Yet once the information is produced, it can be duplicated and distributed with relative ease. Similarly, the construction of telecommunications and data networks requires huge up-front capital investments. But once the infrastructure is in place, the incremental cost of transmitting a data packet is very low.

Information in its pure form is non-exhaustible -- sharing a piece of information with others does not deprive the original owner of his/her ability to use the information.²

² Information does not quite qualify as a public good, however, since it is still excludable -- information usage can be limited to those who are willing to pay. Also, the value of some information may lie precisely in the fact that it is not known to others, but we do not worry about strategic value of information here.

However, information is often encapsulated within a container (e.g., paper, magnetic tape, compact disc) which is physically tangible and therefore exhaustible. When a person checks out this copy of the dissertation from the library, the next patron will have to wait for its return. When this dissertation becomes a runaway bestseller, printing presses and delivery trucks will have to be employed in its reproduction and distribution. In effect, the dissemination of information actually involves the physical duplication and distribution of the encapsulating containers.

With the digitization and networking of information, many herald the arrival of the era where the marginal cost of information dissemination is *virtually* zero. Information is no longer enclosed in physical containers, but is communicated over the information network as electronic signals. Semiconductor memory and network routers are substitutes for papyrus and delivery trucks; transmission links are substitutes for roadways. The Internet, being a packet-switched, store-and-forward network, makes multiple transient copies of each data packet as it makes its way from source to destination. The ease with which data may be duplicated and transported over the global Internet leads pundits to proclaim that “bandwidth is free” and “distance is dead” (Economist, 1995).

If information dissemination is truly costless, then there will be no need for bandwidth conservation, and any discussion on its economies of scale will be irrelevant. The cost of disseminating one gigabyte of information to a million recipients over the global network will be the same as the cost of sending a single byte of information to one recipient on the same local network -- zero!

In reality, bandwidth is not free. The marginal cost of information dissemination is only *virtually* zero, and its average cost is definitely not zero. First of all, the zero marginal cost assumption breaks down when the infrastructure is operating at or near its

capacity. Recall that optimal capacity planning dictates that network capacity to be set at less than or equal to peak usage, i.e., over-provisioning is a sub-optimal solution. Therefore, it should be expected that a well-engineered network is operating at or above capacity, at least during the peak usage periods. In the presence of an admission controller, transmission capacity can be treated as a scarce resource to be rationed. In this case, one can compute the shadow price of sending a data packet as the opportunity cost of not sending another data packet, and from that deduce the marginal cost. In the absence of an admission controller, as is the case of the current best-effort Internet, additional packets may continue to be injected into the saturated network, resulting in buffer overflows, dropped packets and performance degradation. These congestive effects similarly result in non-zero marginal costs. Finally, in the long run, persistent over-subscription of the network capacity that leads to unacceptable performance levels can only be corrected by the expansion of capacity, which requires additional fixed capital investment. This means that the average cost of information dissemination will never reach zero.

And distance is not dead either. Despite the prevalence of distance-insensitive pricing regimes, the day when distance becomes truly irrelevant on the Internet remains far in the future. Even though data packets are not charged according to the distance traveled, the delay and drop rate they experience is still strongly correlated with network distance. One of the major costs of network expansion today is the trenching cost necessary for the laying of the optical fiber channels, and it is strongly sensitive to distance. In fact, Internet service providers (ISPs) are dismissing the notion that long distance Internet connections will be available at the same price as local Internet connections (Sidgmore, 1998).

Therefore, even in the face of declining transmission costs, efficient link utilization and bandwidth conservation will remain important goals of network

engineering and design. Network architects and users alike will continue to take advantage of economies of scale in order to maximize efficiency and savings.

1.2 Economies of Scale: Dimensions and Levels

Network architects, in pursuit of efficient link utilization and bandwidth conservation, often seek out and take advantage of scale and scope economies in network design. Economies of scope savings are realized by supporting various data and application types over the same network. These are outside the scope of this work. Economies of scale savings, on the other hand, are realized by building a network that can be shared by many senders and receivers, exchanging multiple data objects at all times of the day.

This "sharing" of network resources accomplishes savings in two ways. First, statistical multiplexing reduces the stochastic variance of network traffic. This lowers the likelihood of network congestion in the short-run and the need to over-provision the network (to maintain a blocking rate) in the long-run. Second, artifacts specific to a particular network technology can often be leveraged to support shared use of the network. For example, the packet-switched nature of the Internet enables the use of multicast for efficient multipoint communication. Similarly, the ability to place storage elements within the network permits object reuse via network caching and replication. A key focus of this work is to identify and characterize some of these new opportunities that were not available in the paper-based environment.

It is possible to take advantage of economies of scale in information delivery from several dimensions, and at different levels. This dissertation explores the spatial and temporal dimensions of the scale economies at both the information product level and the bit-transport level.

Table 1.1. The various dimensions and levels of EoS explored in this work.

Dimension \ Level	Information Product	Bit Transport
Spatial - Object	mixed bundling	just-in-time delivery
Spatial - Consumer	site licensing	multicast
Temporal	subscriptions	caching & replication

Chapter 2 of the dissertation takes an information product level view. By constructing a multi-product bundling model with multi-dimensional consumer taste characteristics, it is shown that mixed bundling (i.e., selling goods in multiple package sizes) is the profit-maximizing strategy for an information producer such as a journal publisher. To engage in a mixed bundling strategy, the journal publisher would sell both the original journal bundle and the unbundled articles. This was not attractive in the paper-based journal industry because there were strong diseconomies of scale in unbundling the journals (with respect to both distribution and transaction costs). With the digitization and networking of information, however, it is now economically feasible and desirable to sell and deliver individual articles.

The dominance of the mixed bundling strategy can be extended in the consumer and temporal dimensions as well. In addition to selling information goods to individuals on a per-use basis, information producers can further expand their product line to include site licenses (multiple users at a site) and subscriptions (unlimited use during subscription period). In these cases, scale economy savings are realized principally through the aggregation of multiple transactions (across consumers and time) into a single transaction. The various economies of scale opportunities available to the information producer are represented in a three-dimensional object-receiver-time space in Figure 1.1.

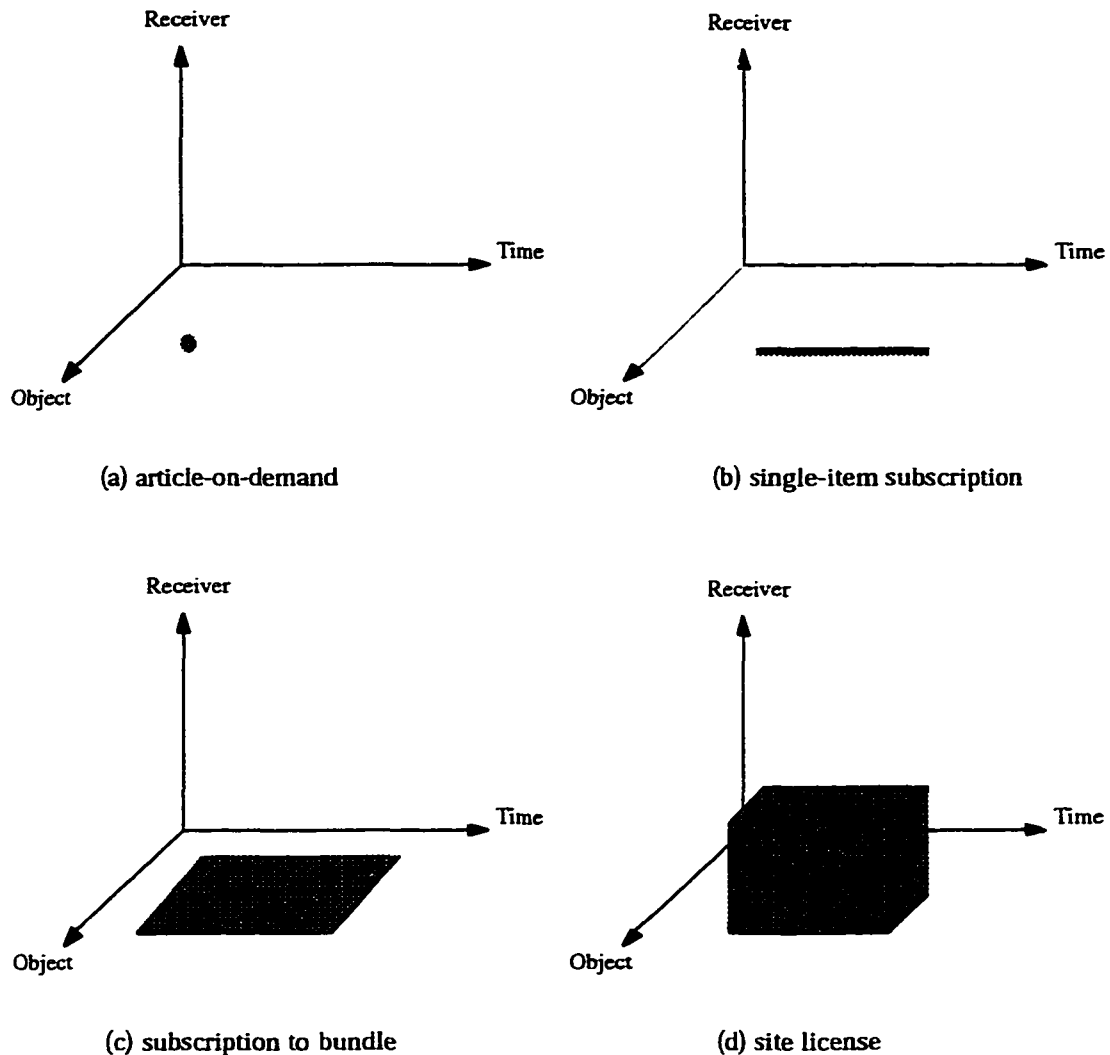


Figure 1.1. Leveraging economies of scale along different dimensions (objects, receivers, time) at the information product level.

Finally, Chapter 2 also investigates the economies of scale in the object-space dimension at the bit-transport level. In the traditional paper-based subscription model, all objects that constitute the subscription are bundled into a package and delivered to the subscriber, even though the subscriber might be interested in only a small subset of all the objects. The technologies of the Internet, however, allow us to shift from this “just-in-case” paradigm to a “just-in-time” paradigm. In this new arrangement, the subscribers are

entitled to unlimited access to all objects in the subscription, but will download only those objects that they are interested in. This is represented in Figure 1.2(a).

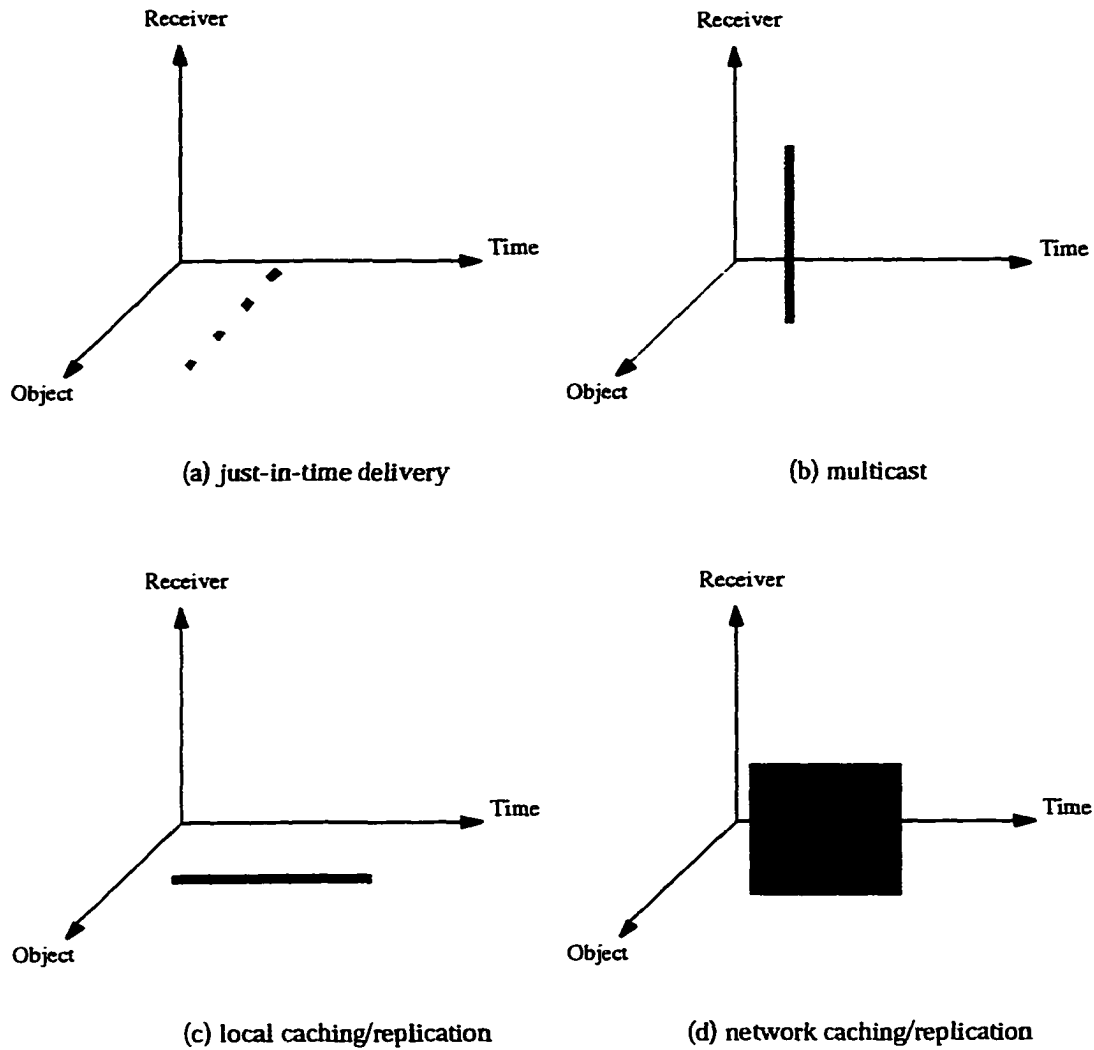


Figure 1.2. Leveraging economies of scale along different dimensions (objects, receivers, time) at the bit transport level.

Chapters 3 through 5 of the dissertation focus on the two remaining elements of the matrix in Table 1.1. Chapter 3 is concerned with multicast, the technology that allows efficient dissemination of data to multiple receivers at distributed locations in the network. This is graphically represented as Figure 1.2(b). Specifically, a cost-based approach to pricing this multicast service is proposed. The work shows that a tariff based on multicast group size is most closely aligned to the actual cost structure, and will result in minimal price distortion. This is an improvement over the current flat-rate scheme because it does not discriminate against those applications with few receivers, such as videoconferencing.

Chapters 4 and 5 explore the technology and economics of building a distributed network storage infrastructure for the Internet. Caching and replication of data achieves scale economies in the temporal dimension by allowing data reuse at a subsequent time (Figures 1.2(c) and (d)). For example, network caching takes advantage of the fact that popular objects are accessed more often and by more consumers than less popular ones. By optimally placing copies of these popular objects at distributed locations in the network, demand for these objects may be satisfied by a nearby copy rather than the original, resulting in bandwidth savings. Chapter 4 presents the architectural framework for a distributed network storage infrastructure, while Chapter 5 focuses on the formulation of the resource mapping problem, a key technical component of the infrastructure.

Chapter 6 discusses some of the policy implications and lessons that can be drawn from this research. A summary of contributions and future work concludes the dissertation.

2. EoS in Objects - Information Bundling

The digitization and networking of information goods necessitate a rethinking of their production and distribution economics. An N -good bundling model with multi-dimensional consumer preferences is developed to study the key factors that determine the optimal bundling strategy. Using analytical and empirical methods, mixed bundling is established as the dominant (i.e., profit maximizing) strategy. Pure unbundling is also shown to outperform pure bundling, even in the presence of some degree of economies of scale, if consumers positively value only a subset of the bundle components, which is the predominant case in the academic journal context. These results provide strong incentives for academic journal publishers to engage in mixed bundling, i.e., offer both individual articles and journal subscriptions, when selling and delivering over the Internet.³

³ An earlier version of this chapter was presented as (Chuang and Sirbu, 1997).

2.1 Bundling and unbundling of information goods

Academic journals have traditionally been sold in the form of subscriptions. Individual articles are bundled into journal issues; issues are bundled into subscriptions. This aggregation approach has worked well in the paper-based environment, because there exist strong economies of scale in the production, distribution and transaction of journals.

Yet, the demand for scholarly information is diverse, unique, and sometimes whimsical. Scholars are often willing to expend a great deal of effort to secure a copy of a specific article unavailable from their personal subscription staple. With the proliferation of journal titles, it is impossible for every scholar to subscribe to all journals relevant to his/her work. Libraries, through their institutional subscriptions to the journals, serve to satisfy the scholars' demand for individual articles. Ordover and Willig (1978) treat journals as "sometimes shared goods" in the study of their optimal provision. Under the fair-use provision⁴ of the Copyright Act, scholars are permitted to reproduce single copies of individual articles from the library subscription copy for non-commercial purposes. There are frequent occasions, however, when the scholar's information needs go beyond the scope of the library's journal collection. In such circumstances the library is permitted to duplicate and share articles with other member libraries of an inter-library loan (ILL) consortium, as long as such "borrowing" does not lead to copying "in such aggregate quantities as to substitute for a subscription".⁵ Empirical studies have found that libraries are incurring costs of up to \$20 per ILL item borrowed or loaned (King and Griffiths, 1995). This suggests that a potential market does exist for unbundled articles at both the individual and institutional levels.

⁴ 17 U.S.C. § 107 (1988 & Supp. V 1993).

⁵ 17 U.S.C. § 108(g)(2) (1988). Specifically, the CONTU Guidelines (reprinted in 1976 U.S.C.C.A.N. 5810, 5813-14) set forth a copying limit of five copies per year of articles from the most recent five years of any journal article.

The publishers, unable to directly appropriate charges for these forms of shared use, recompense for their loss of potential revenue by charging libraries an institutional subscription rate higher than that for individuals. This form of indirect appropriation constitutes price discrimination of the third degree.⁶ While the legality of such practices is seldom questioned⁷, effective third degree price discrimination requires clear demarcation of market segments and minimal leakage across the segments. Both the segmentation of market and the preclusion of effective resale channels are fairly easy to enforce in the academic journal market, since institutions cannot easily disguise themselves as individual subscribers. Along with the apparent inelasticity of demand exhibited by the subscribing institutions, these have been blamed for the escalation of journal prices in recent years.⁸

With the global expansion and rapid commercialization of the Internet, the economics of journal publishing is quickly changing. Many publishers are experimenting with various forms of on-line access to their journals. It is now technically feasible for the publisher to electronically deliver, and charge for, individual journal articles requested by a scholar sitting at his/her desktop. The establishment of a ubiquitous electronic payment infrastructure, and the deployment of micropayment services in particular, will dramatically lower the cost of purchasing digital information goods over the Internet.

⁶ See Liebowitz (1985) and Besen and Kirby (1989) for detailed treatment of journal photocopying and indirect appropriability; Joyce and Merz (1985) study the extent of price discrimination by journals across various academic disciplines.

⁷ Dyl (1983) muses upon the applicability and antitrust implications of the Robinson-Patman Act to price discrimination by academic journals.

⁸ Interested readers can consult Lewis (1989), Byrd (1990), Metz and Gherman (1991), Spigai (1991) and Stoller, Christopherson and Miranda (1996) for works on the economics of scholarly publishing and serials pricing from the library and information sciences communities' perspective.

From the scholars' perspective, this form of access is instantaneous, on-demand, and avoids the costs associated with traditional library access, such as travel to the library, physical duplication of the article, and congestion due to shared use of journals.

Given the market demand for unbundled journal articles, it is somewhat puzzling to see several recent and independent works calling for the bundling of information goods (Bakos and Brynjolfsson, 1997, Fishburn, Odlyzko and Siders, 1997, Varian, 1995). We believe that the confusion and apparent contradiction is the result of an incomplete understanding of the economics of bundling. By identifying the different flavors of bundling and quantifying their relative performance under different supply and demand conditions, this work seeks to demonstrate that not all forms of bundling are appropriate for information goods.

Specifically, by developing an N -good bundling model with multi-dimensional consumer preferences, this work is able to establish that mixed bundling is the dominant strategy, outperforming pure bundling and pure unbundling in maximizing producer surplus. This implies that profit-maximizing publishers should expand their product-line and sell individual articles in addition to journal subscriptions. By extension, a publisher with multiple journal titles should also offer site-licenses that are effectively 'super-bundles' in addition to single-title subscriptions and individual articles.

Section 2.2 provides a short survey of the product bundling literature. The N -good bundling model is developed in Section 2.3, first the demand side in Section 2.3.1, followed by the supply side in Section 2.3.2. In Section 2.4, the model is applied to the academic journal industry for empirical results and analysis. Specifically, we look at how technology trends in distribution and transaction may change the supply side of the model but not change the fundamental result. We conclude with Section 2.5.

2.2 Economics of bundling

A multi-product monopolist may choose to bundle its goods for a variety of reasons. On the supply side, commodity bundling can result in cost savings due to the presence of economies of scale. On the demand side, bundling can be used as an effective tool for extracting consumer surplus. Both factors must be taken into account in the design of optimal bundle prices. Additionally, producers in imperfectly competitive markets may choose to bundle their products for strategic reasons. However, bundling for strategic leverage has no direct implications on pricing design and is outside the scope of this work.⁹

Burnstein (1960) and Stigler (1963) are generally credited with the first references to the bundling phenomenon in the economics literature. Adams and Yellen (1976) operationalize the model for a bundle consisting of two goods, and identify three modes of bundling strategies, namely pure bundling, mixed bundling, and component selling (or pure unbundling). In pure bundling, consumers are restricted to purchasing either the entire bundle or nothing at all. In pure unbundling, no bundle is offered but consumers can put together their own bundle by buying both the component goods. Finally, a monopolist who chooses to engage in mixed bundling will allow the consumers to purchase the bundle or either one of the individual components. Consumers who choose to purchase the bundle will usually pay less than they otherwise would if they had purchased both component goods separately.

Figure 2.1 illustrates the consumer choice regions under each of the three bundling strategies. The axes in each plot represent the consumers' valuation for each of the two component goods, G_1 and G_2 . An individual consumer who is willing to pay w_1 for G_1

⁹ Carbajo, de Meza and Seidmann (1990) and Whinston (1990) provide further treatment of this topic.

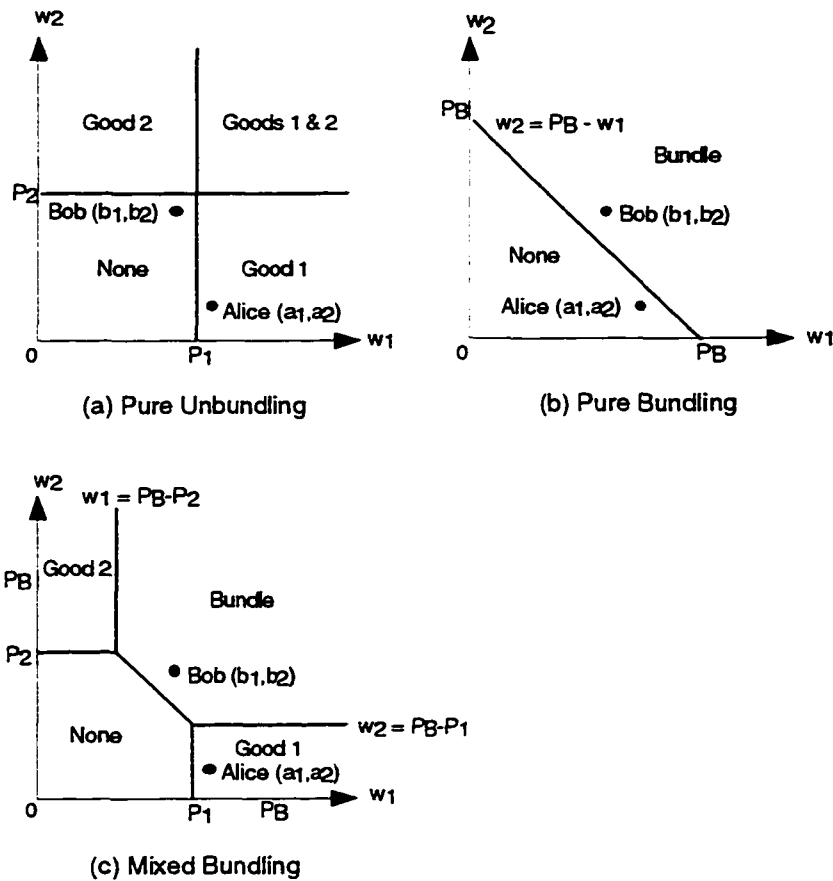


Figure 2.1. Consumer choice regions for two-good bundling model. Alice and Bob will choose different product offerings under different bundling regimes.

and w_2 for G_2 can thus be represented as a point (w_1, w_2) in this consumer space. Depending on the type of bundling strategy employed by the producer, the consumer will make the appropriate purchasing decision based upon his/her position in this two-dimensional $\{W_1, W_2\}$ space. For example, consumer Alice at (a_1, a_2) will purchase only G_1 in Figure 2.1(a) because her willingness-to-pay (WTP) for G_1 , a_1 , is greater than its offer price P_1 , but her WTP for G_2 , a_2 , is less than the offer price P_2 . Bob, on the other hand, purchases nothing under this pure unbundling scenario since both b_1 and b_2 are less than P_1 and P_2 respectively. Interestingly, the situation almost reverses itself if the producer switches to pure bundling instead, as in Figure 2.1(b). Alice rationally chooses to purchase nothing since her aggregate WTP $(a_1 + a_2)$ is less than the price of the bundle, P_B . Bob now purchases the bundle since the sum of b_1 and b_2 is greater than P_B . Using

similar logic, Alice consumes G_1 and Bob consumes the bundle in the mixed bundling case, as illustrated in Figure 2.1(c). This simple, yet powerful illustration shows that the choice of the optimal bundling strategy and the selection of the optimal prices are strongly dependent on the distribution of the consumer population in this $\{W_1, W_2\}$ space.

Schmalensee (1982) and McAfee, McMillan and Whinston (1989) build upon the Adams/Yellen framework, with careful treatment of the consumers' correlation of value between the two components. Among other results, they show that both pure bundling and pure unbundling are boundary cases of mixed bundling and are weakly dominated by the latter strategy in general. Chae (1992) applies the commodity bundling model to information goods in his study of the subscription TV market. He concludes that the bundling of CATV channels is practiced not to extract consumer surplus, but simply because there are economies of scope in the distribution technology.

2.3 N-good bundling model

All of the above-mentioned works are limited to bundles consisting of only two components. A typical academic journal, on the other hand, has between 80 to 100 articles per subscription period. An appropriate N -good bundling model is needed for this context. Unfortunately, a complete N -good model with 2^N bundle combinations and N -dimensional consumer preferences quickly becomes computationally unwieldy as N gets large. Hanson and Martin (1990), by formulating the model as a mixed integer linear programming problem, manage to attack a bundle pricing problem with $N=21$.¹⁰

¹⁰ Armstrong (1997) shows that an approximate solution for the optimal tariff problem is a cost-based two-part tariff, i.e., a fixed up-front membership fee plus a per-article charge set equal to the marginal cost. However, this approximation reasonably converges only for N in the 'several thousands' range, and the

Recognizing the need to balance profit-maximization and consumer rejection of a complex pricing schedule, we opt for a simpler model where no sub-bundles are available. The consumer either purchases the journal subscription for a price P_J or individual articles at a price of P_A apiece. This simplifies the model from that of setting 2^N optimal prices to setting only two prices, P_A and P_J . This is reminiscent of setting a menu of optional two-part tariffs in the nonlinear pricing literature (Willig, 1978 and Wilson, 1993).¹¹ Low-demand readers purchase articles individually, while high-demand readers pay the flat fee P_J and enjoy unlimited access to all articles (Figure 2.2).

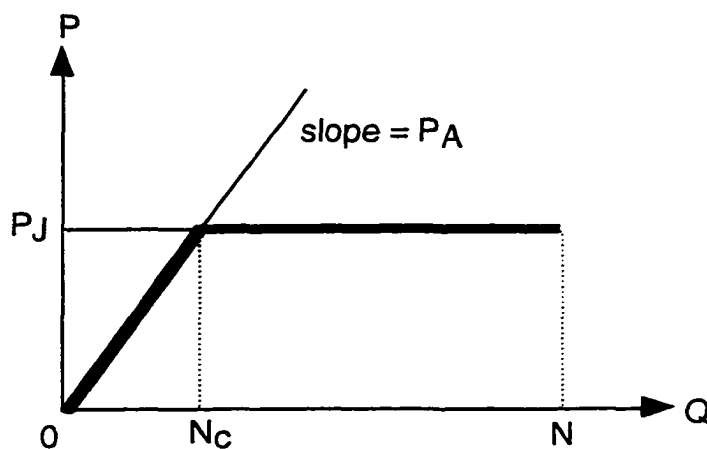


Figure 2.2. Total outlay vs. number of articles consumed. An *ex post* two-part tariff (in bold) offers a predictable price cap on consumer expenditure.

absence of a price cap may make it unacceptable to consumers who are used to the traditional subscription model.

¹¹ Technically, an $(n+1)$ -part tariff can be made to be Pareto-superior to an n -part tariff. Indeed, Laffont, Maskin and Rochet (1985) derived the optimal nonlinear tariff for consumers with two dimensional characteristics, which has a gradually declining marginal price schedule. Using the same argument in the bundling context, any mixed bundling strategy with more than two prices (up to 2^n) will necessarily perform better than a mixed bundling strategy with two prices, and thus pure bundling and pure unbundling strategies as well. Again, the extent to which a publisher chooses to offer multiple prices is clearly dependent upon its multi-variate optimization capabilities, and more importantly, consumer acceptance/rejection of a complex pricing structure.

Optional two-part tariffs can be either *ex ante* or *ex post* in nature (Mitchell and Vogelsang, 1991, page 95). In an *ex ante* arrangement, readers elect to join either the subscriber group or the “article-on-demand” group prior to consumption. Knowing one’s expected consumption behavior is critical in making the “right” decision. An “article-on-demand” reader who expects to read only a few articles but ends up reading more than N_c ($= P_j/P_A$) articles would have to pay more than if he/she had become a subscriber in the first place. Many consumers (especially those with fixed budgeting and fund allocation considerations) are reluctant to sign up for these pay-per-use arrangements precisely because of this uncertainty factor. An *ex post* approach eliminates this problem by allowing the consumer to choose the cheaper of the two pricing schemes at the end of the billing period, thereby placing a predictable upper bound on the final bill. However, the need for a final settlement incurs an administrative and metering overhead over true pay-per-use models.

2.3.1 Modeling heterogeneity in consumer preferences

The N -good bundling model departs from the traditional nonlinear pricing model in that consumers are not choosing to purchase n units of non-distinguishable articles, as if purchasing x kilowatt-hours of electricity or y minutes of cellular-phone air-time. Instead, each of the N articles is unique and distinct from one another. Consumers may value one article dramatically differently from the next. Unfortunately, a complete description of consumer heterogeneity using an N -dimensional vector $\{w_1, w_2, \dots, w_N\}$ again leads to intractability. We seek a concise way to capture the essence of consumer’s willingness-to-pay across the different articles.

Zahray and Sirbu (1990) attempt to capture the heterogeneity in consumer preferences for academic journals, albeit in one variable, the reservation price for the

journal. A similar approach is taken by Bakos and Brynjolfsson (1997), where consumers are characterized by a single type variable w , and consumer valuations of goods are i.i.d. (independent and identically distributed). By employing a single variable, both models can only capture consumer valuations for the bundle in its aggregate. This is adequate in the pure bundling context, where journals are sold only in the form of subscriptions. In the mixed bundling context, however, it is important to account for the correlation of values across the components as well.

Consider, for example, a publisher selling a two-article journal in a market with only two consumers, our friends Alice and Bob. Alice is willing to pay \$10 for the first article and \$0 for the second, while Bob values the articles at \$7 and \$5 respectively. A publisher engaging exclusively in pure bundling (i.e. subscription only) is only interested in the aggregate willingness-to-pay of the two consumers. He/she will price the subscription at \$10 for a total revenue of \$20. A mixed bundling strategist, on the other hand, will desire additional information on the correlation of values for the component articles. In this example, the publisher will price individual articles at \$10 and raise the subscription price to \$12, thereby realizing a revenue of \$22 and completely extracting the consumer surplus in the process. In effect, the publisher has managed to separate the market into two – the segment with high correlation of value across articles (Bob) is sold the subscription; the segment with low correlation (Alice) is offered individual articles.

The present work employs two variables, w_0 and k , to describe the N -dimensional consumer preference. We allow each journal reader to rank the N articles in the journal in decreasing order of preference, such that his/her favorite article is ranked first, the least favorite is ranked last, and weak monotonicity is observed. The reader may place zero value on any number of the N articles. By assuming a linear demand function for all positive-valued articles, we can plot an individual reader's valuation of all the articles in the journal in Figure 2.3. Each of the articles are positionally ranked between 0 and N

along the horizontal axis. The individual's most highly valued article has $n = 0$, and so the y-intercept, w_0 , represents the WTP for his/her most favored article. The valuation for the subsequently ranked articles is assumed to fall off at a constant rate until it reaches zero at $n = k \cdot N$. No articles have negative value with the assumption of free disposal -- readers are free to discard unwanted articles at zero cost. The variable k dictates the slope of the demand curve, and it also indicates the fraction of articles in the journal that has non-zero value to the individual. For example, a reader with $k=0.01$ is willing to pay a non-zero amount for only one article in a journal with a hundred articles, while another reader who positively values half of the articles in the journal will have a k of 0.5. If an individual's k is greater than unity, that means he/she places positive value on all N articles in the journal and the demand curve will never cross the horizontal axis. Figure 2.4 shows a diverse range of consumer preferences that can be described using this two-dimensional $\{w_0, k\}$ index.

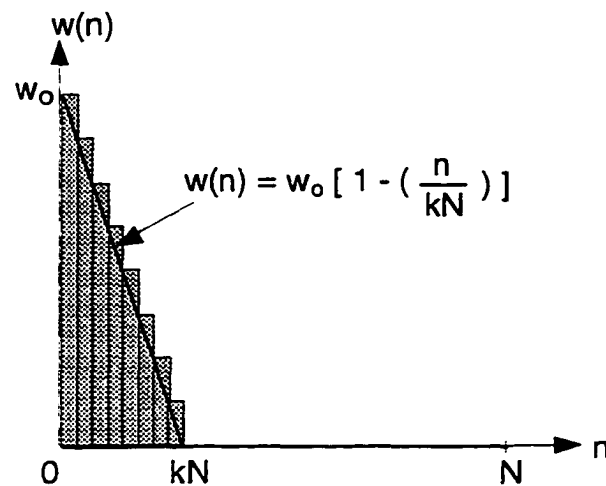


Figure 2.3. Article valuation by an individual reader indexed by $\{w_0, k\}$.

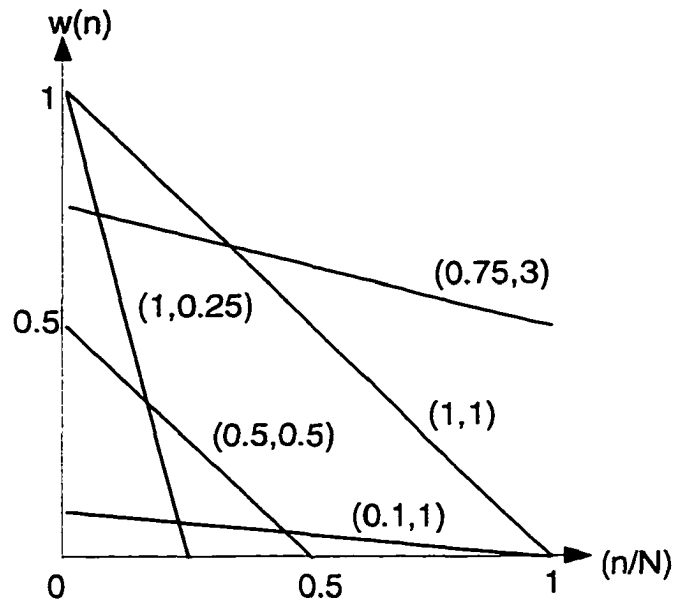


Figure 2.4. This figure demonstrates the diversity of consumers that can be indexed by $\{w, k\}$.

Empirical studies performed by King and Griffiths (1995) indicate that the correlation of article valuations is not very high for academic journals (Table 2.1). Out of the 80 to 100 articles (per subscription period) in an average journal, over 40% of readers surveyed read no more than five articles. Only 0.9% of readers read more than 50, or about half of all articles in the subscription period. This suggests that a majority of readers have small k 's, and very few readers have their k -value close to or exceeding unity. This result is incorporated into our analysis below as a fitted probability distribution for k , $f_k(k)$.

Table 2.1. Distribution of number of articles read in a journal.

Number of Articles Read in a Journal	Proportion of Readers (%)	Cumulative Proportion of Readers (%)
1 to 5	43.60	43.60
6 to 10	34.40	78.00
11 to 15	8.21	86.21
16 to 20	5.50	91.71
21 to 25	3.37	95.08
26 to 30	1.97	97.05
31 to 40	1.23	98.28
41 to 50	0.82	99.10
more than 50	0.90	100.00

Formally, an individual's valuation for the n -th article can be expressed as:

$$w(n) = \max \left\{ 0, w_0 \cdot \left[1 - \frac{1}{k} \left(\frac{n}{N} \right) \right] \right\}, \quad 0 \leq n \leq N-1, \quad (2.1)$$

with $w_0 \geq 0$ and $k \geq 1/N$. Using this formulation, we can proceed to determine the individual's reservation price of the journal, his/her consumption decision in face of the prices P_A and P_J , and the optimal number of articles consumed in each of the three bundling scenarios.

2.3.1.1 Consumer choice in pure bundling

In pure bundling, potential readers can only choose to subscribe to the journal or buy nothing at all. Purchasing individual articles is not an available option. Therefore, an individual's decision is based solely on the price of the subscription, P_J , and his/her reservation price of the journal bundle in the aggregate. This reservation price, W_J , is simply the summation¹² (or integration if we approximate n as a continuous variable) of his/her reservation prices for all the individual articles:

$$W_J = \int_0^N w(n) \cdot dn \quad (2.2)$$

The net benefit U_J derived from subscribing is the difference of the reservation price W_J and the actual subscription price P_J :

$$U_J = W_J - P_J. \quad (2.3)$$

A potential reader will only choose to subscribe if the subscription results in a positive net benefit $U_J > 0$. The $U_J = 0$ curve, plotted in $\{w_0, k\}$ space in Figure 2.5, separates the readership population into two regions. Those that fall in the region R_J will choose to subscribe, while those in region R_0 will opt out. Please refer to Appendix 1 for derivation of this and subsequent results.

¹² We assume here and in subsequent sections that there are no economies of scope in demand, i.e., the marginal benefits of individual articles are additive but not superadditive.

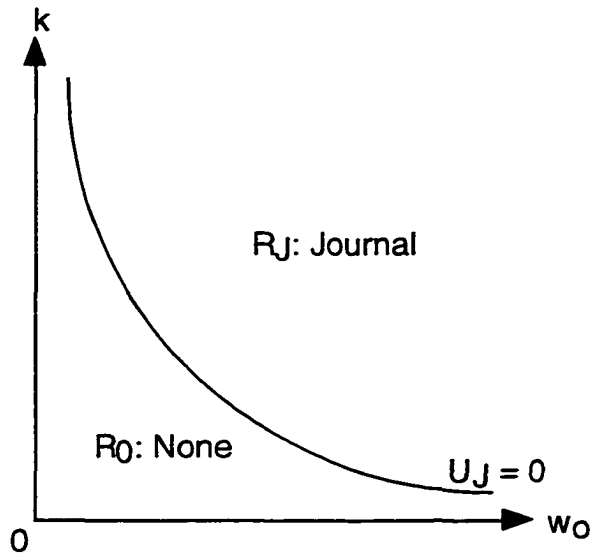


Figure 2.5. Consumer choice in pure bundling scenario.

2.3.1.2 Consumer choice in pure unbundling

In the pure unbundling scenario, all articles are available individually at a unit price of P_A . Consumers are free to purchase as many or as few articles as they desire, up to and including all N articles in the journal. A rational-choice utility-maximizing consumer will consume only those articles with $w(n) \geq P_A$, realizing a net benefit of $w(n) - P_A$ for each of those articles. The marginal article consumed by the consumer, n^* , has a benefit $w(n^*) = P_A$. Therefore, for $w_0 \geq P_A$, the optimal number of articles read by an individual indexed by $\{w_0, k\}$ can be expressed as

$$n^* = \min \left\{ N, \frac{k \cdot N \cdot (w_0 - P_A)}{w_0} \right\}, \quad (2.4)$$

with the maximum capped at N , the total number of articles available in the journal. On the other hand, for an individual with $w_0 < P_A$, even the most favored article is deemed unworthy of the price tag P_A . In this case, n^* would be equal to zero and no articles will

be purchased. Figure 2.6 presents the optimal article consumption level in $\{w_0, k\}$ space. In addition to the optimal consumption level, the net benefit derived from consuming n^* articles, U_A , can also be expressed as

$$U_A = W_A - n^* \cdot P_A, \quad (2.5)$$

where the gross benefit, W_A , is itself a function of n^* :

$$W_A = \int_0^{n^*} w(n) \cdot dn. \quad (2.6)$$

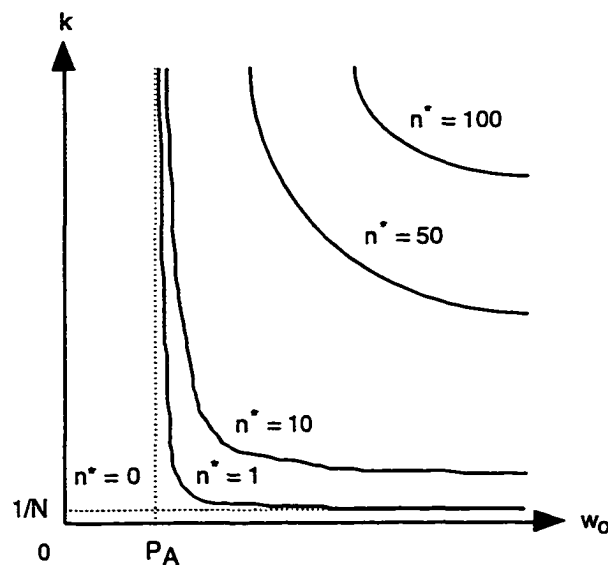


Figure 2.6. Optimal article consumption level in pure unbundling scenario.

2.3.1.3 Consumer choice in mixed bundling

In mixed bundling, consumers seek to maximize their utility by choosing one of three options: subscribe to journal, purchase individual articles, or neither. Depending on each individual's U_J and U_A measures, as defined above, he/she can fall into one of five regions in Table 2.2. This is illustrated in the consumer choice diagram, Figure 2.7. For example, individuals who value their most favored article at less than the article price (i.e., $w_0 < P_A$) have a negative U_A and will not purchase any articles in unbundled form. If their valuation of all the articles in the aggregate is less than the subscription price P_J , they will not subscribe to the journal either. These individuals fall in the R_0 region. On the other hand, if their aggregate valuation is greater than P_J , they will fall in the R_{J1} region and will choose to subscribe to the journal. Individuals with high w_0 and low k tend to value only a few articles highly, and will be best off purchasing individual articles. These consumers are found in region R_{A1} . Finally, consumers in R_{A2} and R_{J2} receive positive benefits from either journal subscription or article purchase, and make their respective purchasing decisions based on the relative magnitudes of their U_J and U_A .

Table 2.2. Consumer choice in mixed bundling scenario.

Region	U_J	U_A	$U_J > U_A ?$	Purchase
R_0	< 0	< 0	--	Nothing
R_{A1}	< 0	> 0	No	Article(s)
R_{J1}	> 0	< 0	Yes	Journal
R_{A2}	> 0	> 0	No	Article(s)
R_{J2}	> 0	> 0	Yes	Journal

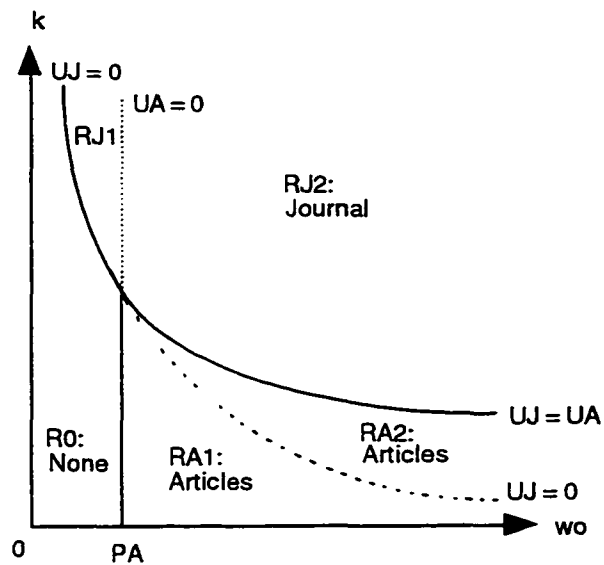


Figure 2.7. Consumer choice in mixed bundling scenario.

2.3.2 Production costs and economies of scale

Thus far we have focused on the demand side of the problem. We now turn to the supply side, specifically to the underlying technology and production functions of academic journals. As previously noted, the information industry in general and the journal publishing industry in particular are characterized with high fixed costs (FC) and low marginal costs (MC). A producer will only stay in the market if gross margin (gross revenue minus variable cost) is enough to cover fixed cost. As long as the total revenue is greater than total cost, the optimal pricing decision is then independent of FC . (Alternatively, we can think of the fixed cost as either zero or sunk). This assumption allows the treatment of FC as an exogenous variable in the present model.

We incorporate the presence (or absence) of economies of scale (EoS) in the production function by establishing the following relationship between the marginal costs MC_J and MC_A :

$$MC_J = N^\gamma \cdot MC_A. \quad (2.7)$$

N is the number of articles in the journal and γ is the economies of scale index. When $\gamma < 1$, economies of scale are present and a subscription bundle of N articles is cheaper to produce and sell than N individual articles. Therefore the publisher can realize cost savings via bundling. When $\gamma = 1$, there are no economies of scale in journal production or distribution. No cost savings can be realized by bundling. Finally, if there are diseconomies of scale in the production function, it can be described with $\gamma > 1$. Prior work in bundling almost invariably assumes no cost savings from bundling, i.e., $\gamma = 1$. Chae's assumption of extreme economies of scope in the CATV delivery technology translates to a special case of $\gamma = 0$. By treating the extent of economies of scale as an endogenous variable, this model allows a parametric analysis of its influence on the producer's optimal bundling strategy.

Based upon the distribution of consumers in the $\{w_o, k\}$ space and the underlying cost structure of journal production, the publisher proceeds to optimize P_A and P_J to maximize gross margin Π :

$$\Pi(P_J, P_A) = \iint_{R_J} [P_J - MC_J] f(w_o, k) \cdot dw_o \cdot dk + \iint_{R_A} n^* [P_A - MC_A] f(w_o, k) \cdot dw_o \cdot dk, \quad (2.8)$$

where the term $f(w_o, k)$ is the joint probability density function (p.d.f.) of the readership described in $\{w_o, k\}$ space. It is worthwhile to note that, in the case where the optimal

strategy turns out to be pure bundling (pure unbundling), the second (first) integral component will be zero.

2.4 Analysis and empirical results

The N -good bundling model is used to quantify how the choice of the optimal bundling strategy and optimal pricing are affected by MC and γ on the supply side, and $f(w_0, k)$ on the demand side. Recalling that w_0 is an individual's valuation of his/her most favored article in the journal, and k is the fraction of articles in the journal that have non-zero value to the individual, we assume independent distributions for w_0 and k . We normalize w_0 to be uniformly distributed between 0 and 1. Using the King/Griffiths data in Table 2.1, k is fitted to an exponential distribution with $\lambda = 13.8758$ or $\mu = 1/\lambda = 0.072$ ($R^2 = 0.97117$). This means that the average reader reads only 7.2% of all articles in a typical journal. Figure 2.8 shows, for a journal with $N = 100$ articles, the producer surplus (as measured by gross margin) attainable via each bundling alternative as a function of MC and γ . The marginal cost of a single article, MC , is restricted to be no greater than the highest individual valuation, $\max[w_0]$ (which we normalize to unity without loss of generality). There would be no market participation if it were more costly to produce an article than anyone is willing to pay. Given our interest in scenarios where MC is small but non-zero, these and subsequent figures are plotted on semi-log scale.¹³

¹³ In the degenerate case where marginal cost is zero, the value of γ (economies of scale factor) becomes irrelevant, and we would naturally expect mixed bundling, pure bundling and pure unbundling to perform equally well in terms of maximizing producer surplus.

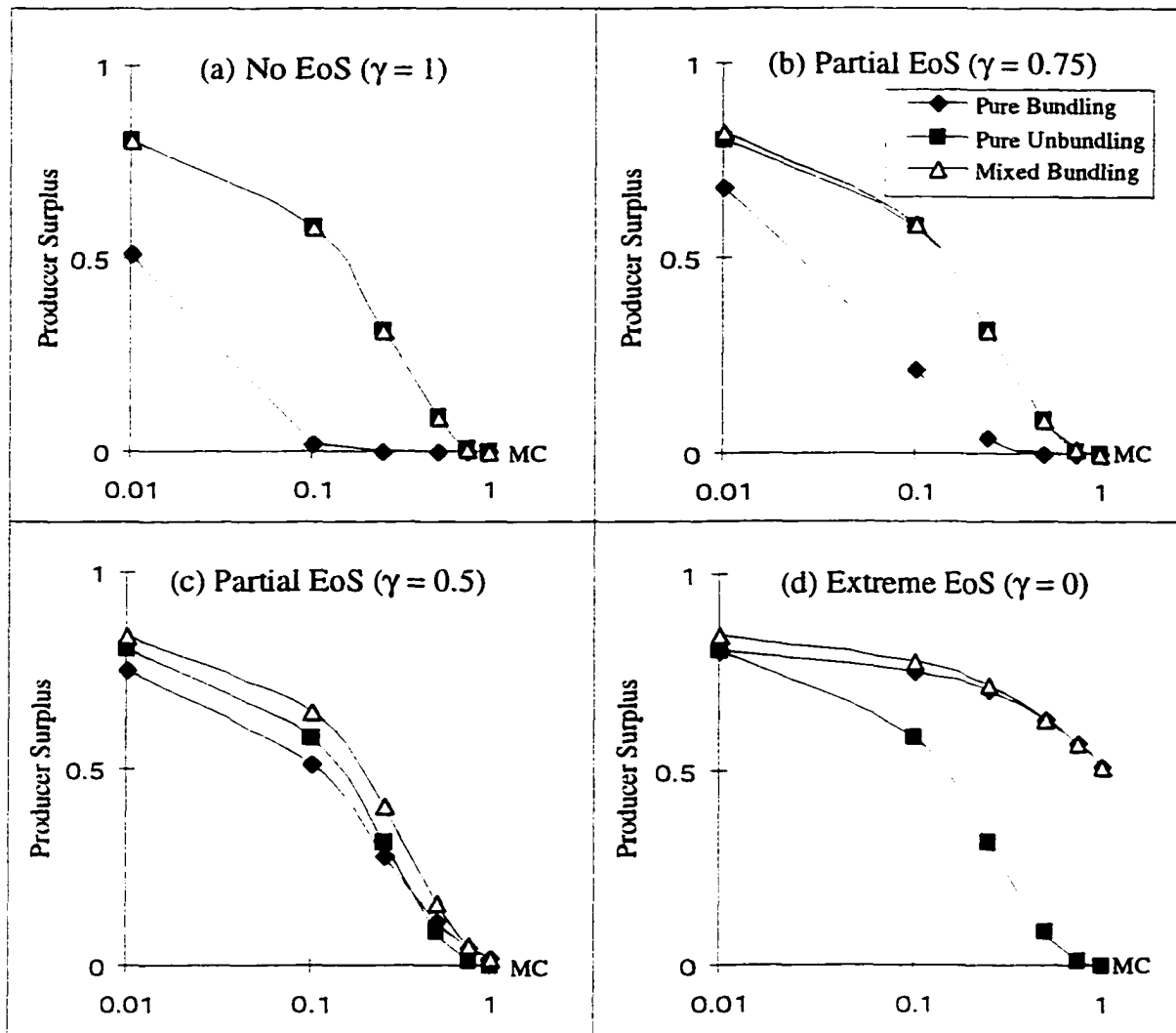


Figure 2.8. Profit-maximizing bundling strategy: it is clear that mixed bundling is the dominant strategy across all marginal cost and economies of scale conditions.

In Figure 2.8(a), there are no economies of scale and the EoS factor $\gamma = 1$. The marginal cost of the journal is N times that of a single article, and so no cost saving is realizable from bundling. The pure bundling strategy is clearly dominated by the other two strategies. The mixed bundling and the pure unbundling alternatives are essentially identical to each other. This suggests that even if the publisher opts for a mixed bundling strategy, virtually the entire revenue will come from article sales. As the production function begins to exhibit some economies of scale, cost-related bundling incentives begin to appear. Yet, in Figure 2.8(b), where $\gamma = 0.75$, the situation remains unchanged. Mixed

bundling continues to be the optimal strategy, with pure unbundling slightly inferior at low MC levels. As γ continues to fall in the face of stronger economies of scale, mixed bundling becomes the strictly dominant strategy. In Figure 2.8(c), where $\gamma = 0.5$, pure bundling and pure unbundling trade dominance depending on the magnitude of MC , but both are dominated by mixed bundling. Finally, in the case of extreme economies of scale, where $\gamma = 0$ in Figure 2.8(d), it costs as much (or as little) to produce and sell an entire journal as it does a single article. Mixed bundling strictly outperforms pure unbundling at all MC levels, while pure bundling approaches mixed bundling in capturing producer surplus as MC approaches $\max[w_0]$ or unity. In this case most of the publisher's revenue will be derived from journal subscriptions.

The first observation is that mixed bundling is superior to pure bundling and pure unbundling across all values of MC and γ . This extends earlier results for two-good models to the present N -good model. This result makes intuitive sense since both pure bundling and pure unbundling are boundary cases of mixed bundling, and therefore can do no better than the mixed bundling strategy. The price discrimination mechanism is at work here, as the mixed bundling strategy creates an incentive-compatible condition, inducing the high and low-demanders to reveal their preferences by self-selecting into the appropriate consumption groups.

Additionally, we observe that pure bundling does not necessarily dominate over pure unbundling in the N -good scenario. Specifically, the model identifies plausible conditions under which unbundling is actually superior to bundling (in pure forms). When marginal cost is non-zero, pure bundling is undesirable not only in the absence of economies of scale ($\gamma = 1$), but also if the degree of EoS is too weak (as illustrated by $\gamma = 0.75$) for the cost-saving bundling incentive to become a dominating factor. Even in the presence of strong economies of scale ($\gamma = 0.5, 0$), the relative merits of pure bundling and

unbundling are still dependent on the magnitude of the marginal cost relative to consumer valuations of the articles. Inefficiency in resource allocation (and loss of surplus) will result if individuals are forced to purchase the bundle and consume some articles which they value below marginal cost. Adams and Yellen label this condition where consumption occurs at sub- MC levels as a violation of the 'Exclusion' assumption. This is of real concern to journal publishers since the distribution of k (as fitted to empirical data from King and Griffiths) is such that most readers actually place zero value on most of the articles in an average journal that they read. Except for the case of $MC = 0$, or the case of $\gamma = 0$, where the marginal costs for all but the first article are effectively zero, exclusion is always violated for those readers with $k < 1$. In our numerical analysis, where k is exponentially distributed with a mean $\mu = 0.072$, the probability of $k \geq 1$, i.e., a reader having positive valuations for all articles in the journal, is on the order of 10^{-6} , or one out of one million readers. (To place this number in context, Science, one of the most widely read academic journals, has a circulation of 165,000; IEEE Spectrum and American Economic Review, two mainstream periodicals in the electrical engineering and economics disciplines, have circulation of 30,000 and 27,000 respectively.¹⁴) Therefore, the choice of optimal bundling strategy lies in the balance between cost-savings from bundling and loss of surplus due to exclusion violation. The proposition by Adams and Yellen (p. 488) that pure unbundling "is a more desirable strategy the greater the cost of violating Exclusion" holds true here.

2.4.1 Optimal pricing and revenue mix

The mixed bundling publisher is interested in the optimal pricing of its articles and subscriptions. Figures 2.9 and 2.10 show the optimal pricing ratio (P_J/P_A) and the

¹⁴ Circulation data from Ulrich's International Periodicals Directory, 34 ed. R.R. Bowker Publishing, 1996.

corresponding revenue mix for various marginal cost and EoS conditions, respectively. While the semi-log scales preclude the plotting of data at $MC = 0$, we note that when marginal cost is zero, the subscription (to a bundle of 100 articles) should be priced at approximately ten times that of an individual article, and this optimal pricing ratio would result in a revenue stream that is well balanced between the sale of articles (56%) and subscriptions (44%). When the marginal cost is non-negligible, however, the optimal ratio becomes sensitive to the economies of scale condition. If there are extreme economies of scale ($\gamma = 0$), the cost-saving incentive induces the publisher to rely more heavily on the sale of bundled subscriptions as MC increases. With strong economies of scale ($\gamma = 0.5$), the optimal pricing ratio stays constant but the revenue mix shifts decisively towards subscription sales with increasing cost. On the other hand, when the economies of scale are absent or weak ($\gamma = 1, 0.75$), the publisher is best served by increasing the price ratio, thereby realizing most or all of its revenue through individual article sales.

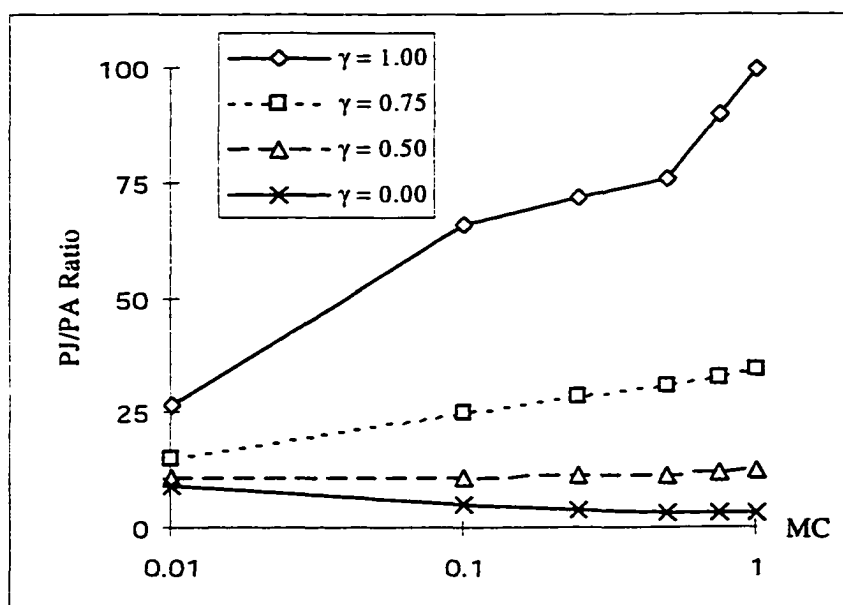


Figure 2.9. Optimal price ratio for mixed bundling strategy across various economies of scale and marginal cost conditions.

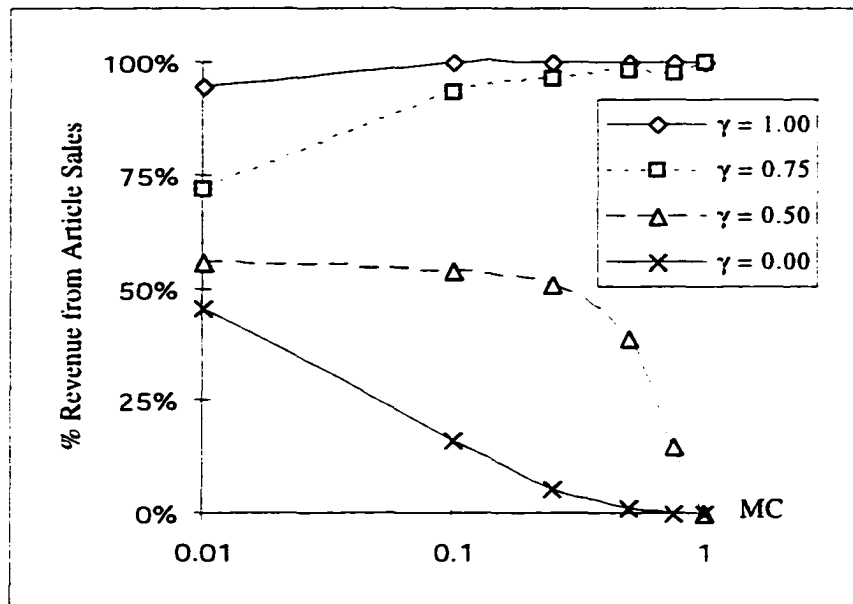


Figure 2.10. Optimal revenue mix for mixed bundling strategy.

2.4.2 Internet-based document delivery technology

It is possible to characterize the extent to which economies of scale are present in the current set of network-based document delivery technologies. Specifically, we ask what is a reasonable value of γ , and how might it change with technology? We identify two major components to the marginal cost of delivering a journal or an article. These are the cost to transmit raw data bits and transaction costs. Production and data storage are fixed costs to the publisher and should be excluded from consideration in this context.

We consider the scenario where the publisher outsources both data transmission and fee-collection functions to specialized services. Web hosting services are offered by a multitude of Internet presence providers. Entire digitized archives of journal articles can be hosted on a web server and made accessible for downloading by scholars. Several

micropayment systems are also available to facilitate electronic payment for articles or other information goods sold via the Internet.¹⁵

The marginal costs MC_J and MC_A are characterized using three cost coefficients:

$$\begin{cases} MC_J = \kappa_f + \kappa_v \cdot P_J + \mu_s \cdot N \cdot \kappa_d \\ MC_A = \kappa_f + \kappa_v \cdot P_A + \kappa_d \end{cases} \quad (2.9)$$

where κ_f , κ_v , κ_d are cost coefficients and μ_s is the expected fraction of articles downloaded by a subscriber. We discuss each variable in turn. Transactional costs are modeled after the two-part fee structure of credit-card transactions. κ_f is a fixed fee levied for each transaction, while κ_v is the variable component charged in proportion to the value of the transaction (P_J and P_A respectively).¹⁶ This implies, significantly, that the marginal costs are no longer constants as we have assumed thus far, but have become functions of P_J and P_A , respectively.

The variable κ_d is the cost of transmitting or downloading one journal article. Web hosting services currently charge between \$0.05 and \$0.50 per MB (megabyte) of data accessed by a client from the server.¹⁷ A journal page, scanned at 600 dpi and compressed in Group 4 Fax/TIFF format, takes up about 100kB (kilobytes). Assuming a typical journal article has ten pages, downloading a journal article requires the

¹⁵ See MacKie-Mason and White (1996) and Sirbu (1997) for surveys of digital payment mechanisms.

¹⁶ A typical credit card operation has κ_f and κ_v set at \$0.30 and 1.66% and is not suited for small value transactions because of this high κ_f . NetBill (<http://www.netbill.com>), an experimental electronic micropayment system developed at Carnegie Mellon University, has $\kappa_f = \$0.02$ and $\kappa_v = 5\%$, enabling it to support transactions down to 5-10 cents. This latter set of cost coefficients is used for this analysis. See Sirbu and Tygar (1995) for a description of the NetBill electronic micropayment system.

¹⁷ Price schedules for incremental data downloads obtained from a website survey of web hosting service providers, January 1997. See Appendix 2.

transmission of 1MB of data. This translates to a κ_d of between \$0.05 and \$0.50. With continued improvements in data transmission and compression technologies, it is reasonable to expect further declines in κ_d .

Most providers sell downloads at a fixed cost per bit, so the publisher enjoys no economies of scale in data transmission per se. However, selling a journal subscription on-line does not necessarily require the transmission of all N articles to the subscriber. The subscribers are free to download all N articles, but most will choose to download only a fraction of all articles. This “just-in-time” (as opposed to “just-in-case”) delivery paradigm results in an expected transmission cost of $\mu_s \cdot N \cdot \kappa_d$ instead of $N \cdot \kappa_d$ for each journal subscription. We can quantify μ_s as the conditional expectation of the fraction of articles read by the subscribing sub-population (the region R_j in Figure 2.7),

$$\mu_s = \frac{\iint_{R_j} k \cdot f(w_o, k) \cdot dw_o \cdot dk}{\iint_{R_j} f(w_o, k) \cdot dw_o \cdot dk}. \quad (2.10)$$

We have shown that the area of integration R_j is a function of the prices set by the publisher. Therefore μ_s is dependent on the prices as well. Substituting equations (2.9) and (2.10) into equation (2.8) with the appropriate values for the κ coefficients and re-optimizing, we can gain insight into how μ_s and γ are affected by a decline in transmission cost κ_d , which in turn determine the optimal pricing and revenue mix decisions. Figure 2.11 shows that the optimal subscription price P_j (right hand axis) varies significantly in the current range of κ_d . The expected fraction of articles read from a subscription copy μ_s (left hand axis) follows a similar trend, which is not surprising given its dependency on P_j . The higher the price of a subscription, the more articles one will have to read in order to justify becoming a subscriber. It is interesting to note that, even when transmission costs

become negligible ($\kappa_d = 0$), μ_s is still significantly greater than μ of 0.072 for the overall journal readership.

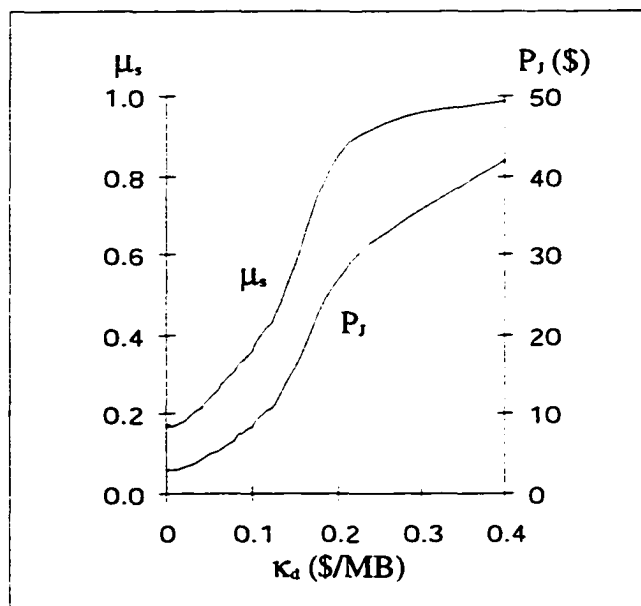


Figure 2.11. Effect of transmission cost on journal subscription pricing.

Figure 2.12 further illustrates how the economies of scale (γ) and optimal revenue mix are likely to be impacted by a declining κ_d . For κ_d greater than \$0.20/MB, there are essentially no economies of scale and most of the revenue is derived from article sales. At $\kappa_d = \$0.05/\text{MB}$, the current low-end estimate, γ falls to 0.6 and we begin to see a well balanced revenue mix between article and journal sales. But even when $\kappa_d = 0$, we see that γ will not fall below 0.3, and 30% of the revenue is still derived from selling individual articles. Under no circumstance should we expect the entire publishing revenue to come from subscription sales alone.

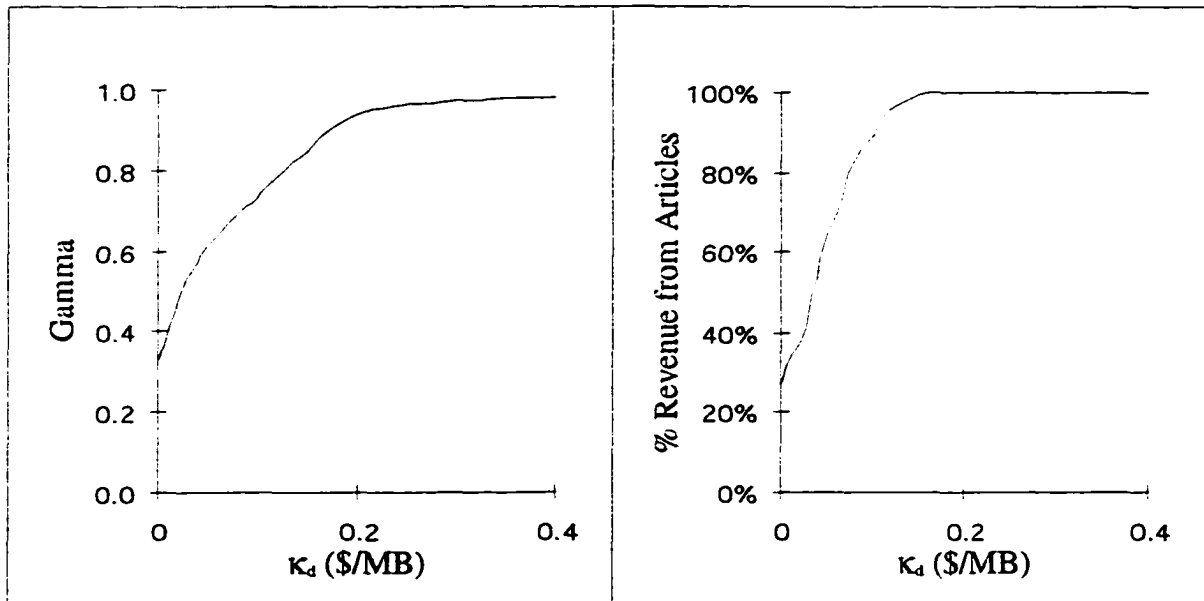


Figure 2.12. Effect of declining κ_d (transmission cost) on economies of scale and revenue mix.

While we have held κ_r and κ_v constant in our analysis, it is reasonable to expect a decline in these coefficients as well. The Millicent protocol, for example, proposes a light-weight micropayment mechanism with cryptographic operations that cost one-tenth to one-hundredth of a cent (Manasse, 1995). Yet one should not expect κ_r and κ_v to fall at a similar rate as κ_d . This is because transaction costs are not solely dictated by progress in hardware technology or the state of the art in cryptography. Other sources of payment system costs such as customer service, fraud protection, chargebacks and back-office accounting may decline only slowly over time, if at all.

2.5 Conclusion

Several recent independent works suggest that bundling is desirable for information goods (Bakos and Brynjolfsson, 1997, Fishburn, Odlyzko and Siders, 1997, Varian, 1995). The current work demonstrates, however, that a different conclusion may

be drawn when the important distinction between mixed and pure bundling is made. While mixed bundling is always the dominant strategy, our results also show that pure bundling may, under certain conditions, be inferior to pure unbundling. We therefore caution against any wholesale adoption of pure bundling without a thorough analysis of the supply and demand of the information product in question. Specifically, for information goods that presently exist in bundled form (e.g., academic journals), unbundling (i.e., switching from pure bundling to mixed bundling) can actually increase producer surplus. This result suggests that an academic journal publisher should expand its on-line product offering to include unbundled articles in addition to traditional subscriptions. By offering a menu of choice that includes both the original bundle and the components, the publisher can extract consumer surplus more completely via consumer self-selection. By extension, the publisher can do even better by simultaneously bundling and unbundling the journal, adding “super-bundles” of multiple journal subscriptions or site-licenses to the product mix. Mackie-Mason and Riveros (1997) offer another bundle option in addition to unbundled articles and the traditional subscription, namely the generalized subscription. In this arrangement, the user purchases unlimited access to N units of articles, and is free to select any N articles from the entire archive of M articles (with $M \gg N$).

Our model assumes that a journal is made up of N individual articles. In reality there are other separable components to a journal subscription, such as the table of content, indices, abstracts and other announcements. Readers can assign different valuations for each of these components just as they do for the individual articles. Therefore these components can be candidates for unbundling as well. RevealAlert, a recent product offered by CARL, delivers via email the tables of contents of up to fifty user-selected journal titles.

A casual survey will reveal that all the major players in the academic journal publishing industry are actively pursuing the possibility of network access to their journal products. Many have made impressive strides in a very short period of time. Some publishers provide on-line access to article abstracts, tables of content and indices to their journal titles; others offer fully searchable text, complete with images and mark-up tags, of the journal articles. Most publishers have installed (or plan to install) some form of access control and billing mechanism so that charges can be appropriated for the usage of these materials. However, lessons learnt from various research/demonstration projects indicate that significant economic, behavioral and institutional barriers need to be crossed before on-demand network delivery of academic journals can become ubiquitous.¹⁸ Intelligent pricing designs must take into consideration the information needs and usage behavior patterns of the journal reading population, as well as the economies-of-scale characteristics of the underlying technologies.

¹⁸ Okerson and O'Donnell (1995) edit an interesting forum discussion, which took place entirely on the Internet, on the future of scholarly journals; drawing experience from various electronic journal endeavors such as Psycology, Chicago Journal of Theoretical Computer Science , and the electronic pre-print archive for high-energy physics at Los Alamos National Labs.

3. EoS in Receivers - Multicast Communication

The Internet was designed as, and remains primarily, a hop-by-hop packet-forwarding network. Network nodes perform nothing more than forward packets towards their destinations. Other functionalities are pushed out to the end hosts, keeping the core of the network as simple as possible. This "end-to-end" philosophy (Saltzer, Reed and Clark, 1984) has proved to be a major contributor to the robustness and scalability of the Internet.

Now, as the Internet evolves into a medium for information dissemination, there is increasing need for network support for efficient one-to-many communication. Multicast, for example, is enabled by intelligent, on-demand packet duplication at the network nodes. Similarly, the addition of network storage elements allows data caching and replication to be performed in the network itself.

3.1 Pricing Multicast Communication: A Cost-based Approach

Multicast has been a proposed IETF standard for over a decade (Deering, 1986), and the experimental MBone network has been operational since 1992 (Casner and Deering, 1992, Eriksson, 1994). There is strong industry push to deploy IP multicast in the Internet proper, yet the single biggest economic concern remains: how should multicast be priced (Shenker et al., 1996)?¹⁹

It is important to recognize at the outset that multicast as a network service will be used by many different applications. These could range from multimedia teleconferencing and distributed interactive simulation to software distribution, webcasting and other "push" applications. These applications have very different bandwidth/latency requirements and scaling characteristics. They compete for network resources not just against one another, but against unicast traffic as well. Therefore, any resource allocation scheme will have to be non-discriminatory between applications and traffic types (unicast vs. multicast).

This paper advocates a cost-based approach to multicast pricing. When prices are set to reflect actual network resource consumption, they minimize market distortion and result in efficient and equitable resource allocation. Additionally, this paper calls for pricing multicast relative to the corresponding unicast service. If unicast is subject to a flat-rate pricing scheme, multicast should also be subject to a flat-rate pricing scheme; if unicast traffic becomes subject to a usage-based pricing regime, then multicast should be priced according to usage as well. As long as multicast is priced relative to unicast, all the results in this work are valid under either pricing regime. More importantly, economic theory reminds us that prices serve as market signals to the users, providing feedback

¹⁹ An earlier version of this chapter was presented as (Chuang and Sirbu, 1998).

regarding their usage of network resources. Given a tariff structure where multicast and unicast services are priced consistently with each other, the end user will correctly choose multicast over unicast when it is indeed the cheaper (and more efficient) alternative.

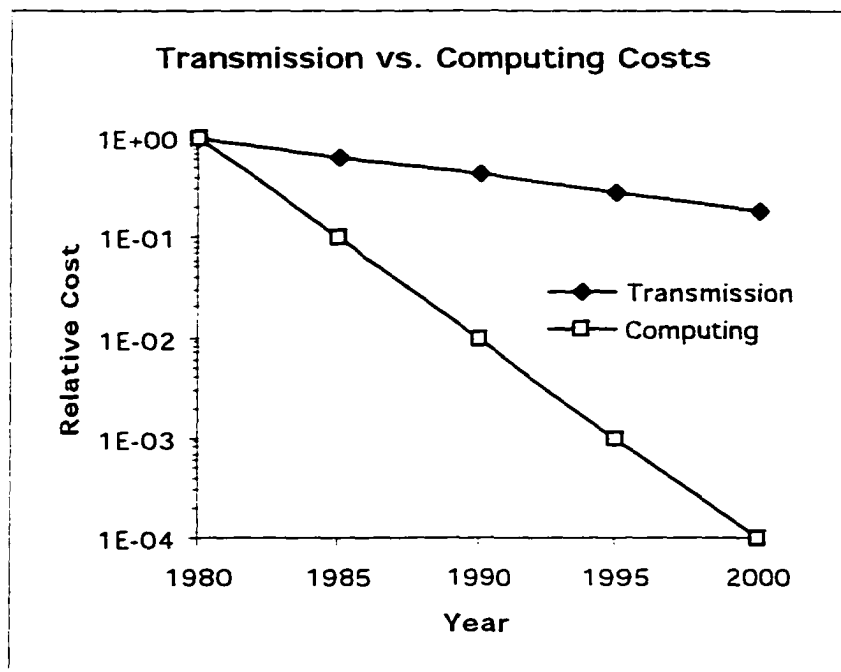
The structure of this chapter is as follows. We begin in Section 3.2 with the quantification of multicast link usage. This allows us to capture the economies of scale realizable by multicast. The cost structure thus established is then applied to the pricing design in Section 3.3. Finally, Section 3.4 looks at how dense and sparse mode multicast should be priced to reflect the difference in bandwidth usage between the two modes.

3.2 Cost Quantification

Multicast achieves bandwidth savings over unicast by duplicating packets (to multiple destinations) only when routing paths diverge. By avoiding the transmission of duplicate packets over any link, significant economies of scale over unicast can be realized.

This work focuses on the quantification of bandwidth usage, i.e., link cost, as opposed to node costs, such as routing table memory, CPU usage, etc. After all, multicast can be thought of as the result of an engineering-economic optimization, where significant bandwidth savings is realized at the expense of control and processing overhead at the routing nodes. This tradeoff is justified because the cost of transmission has, historically, declined at a slower rate than the cost of processing/memory (Figure 3.1). It is debatable whether this cost gap between transmission and processing will persist, given the breakthroughs in optical amplification and wave-division multiplex (WDM) technologies, and the diminishing returns from further transistor size shrinkage. Node costs may also become significant for multicast for another reason: multicast employs logical addresses in a flat addressing space, and hence CIDR-style route

aggregation (Rekhter and Li, 1993) is not possible. Address depletion will not be the direct limit to scalability, especially if the Internet moves to IPv6 (Deering and Hinden, 1995). Instead, routing table entries will become the scarce resource since every single multicast group will require its own separate entry. For source-based multicast trees, this cost will have to be multiplied M times for each multicast group with M active senders. At some point in the future it may become necessary to institute some market-driven or administrative mechanism for multicast address allocation (Estrin et al., 1997, Rekhter and Li, 1996, Rekhter, Resnick and Bellovin, 1996).



(Source: Spragins, Hammond and Pawlikowski, 1991, page 25.)

Figure 3.1. Transmission vs. computing cost trends.

There are several studies that compare the performance and resource costs of various multicast routing protocols (Billhartz et al., 1997, Doar and Leslie, 1993, Salama, Reeves and Viniotis, 1997, Wei and Estrin, 1994, Wei and Estrin, 1995). In addition to link usage (as measured by tree costs), the different protocols are also evaluated in terms of delay and traffic concentration metrics. All of these studies, however, only compare

multicast protocols against one another, rather than against a unicast baseline. This precludes any direct computation of how much bandwidth savings is realizable if one switches from unicast to multicast.

A recent measurement study on Mbone traffic provides some empirical data on the nature and characteristics of multicast traffic (Almeroth and Ammar, 1997). The study finds that “while there is a direct relationship between the number of unicast packet-hops and the number of receivers, the number of multicast packet-hops remains nearly flat. Even when the number of group members increases, the number of packet hops increases only slightly”. This result is limited, however, by the narrow range of membership size (50-200 receivers out of ~5000 Mbone nodes, or 1-4% subscription rate) sampled in the study. As this paper shall demonstrate, multicast cost does indeed rise with membership size, albeit at a slower rate than unicast.

3.2.1 Quantifying Multicast Tree Cost

A network provider offering multicast service would be interested in quantifying the link usage of a multicast delivery tree. Specifically, for a multicast group of membership size N , we can express the (normalized) multicast tree cost as:

$$L_m/L_u = N^k \quad (3.1)$$

where

- L_m : total length of multicast distribution tree;
- L_u : average length of unicast routing path;
- N : multicast group size;
- k : economies of scale (EoS) factor, ranging between 0 and 1.

The total length of a multicast tree, L_m , is simply the summation of edge costs of all links that make up the tree. These edge costs may have weight metrics that are hop-based and/or distance-based. In this study we choose the hop-based approach, i.e., setting the cost of all edges at unity. This is consistent with general Internet routing today, where hop-count is the widely-used metric for route cost calculations.²⁰

Without any loss of generality, we normalize L_m by L_u . This means that the normalized multicast tree cost, L_m/L_u , is a dimensionless parameter. L_u is the expected path length between any two nodes in the network. Equivalently it is the average distance a unicast packet will have to travel from the source to the destination in this network. L_u is clearly network-specific -- it is influenced by topological factors such as the number of nodes and links in the network, average node degree, network diameter, etc. Its value, however, should be relatively static and well understood by the network provider.

N is the number of receivers in the multicast group. It is important to realize that we are referring to the network routing nodes rather than the individual hosts in this context. There may be one or more hosts, and therefore one or more potential multicast group members, attached to each leaf router. However, for a variety of reasons²¹, the leaf

²⁰ As pointed out by Pejhan, Schwartz and Anastassiou (1996), results based on hop-based metrics are generalizable to both source-based shortest-path trees and minimal spanning trees. Our results will not be significantly different even if we adopt a hop-distance hybrid metric (using a rule of thumb by Mahdavi (1997) that 100 kilometers of link distance have equivalent cost to one hop), as the majority of links are short-haul links.

²¹ According to version 2 of the IGMP protocol (Fenner, 1997), "multicast group membership means the presence of at least one member of a multicast group on a given attached network, not a list of all of the members". When a host wishes to join a group, it should transmit a 'REPORT' message (and up to two additional 'REPORT' messages for redundancy) in case it is the first member of that group on the network. However, a host is not required to send a 'LEAVE' message when it leaves the group. Furthermore, to avoid report implosion, multiple responses to periodic 'General Query' messages are suppressed.

router may not have an accurate count of the total number of hosts belonging to a group. Indeed, such an accurate count is not required. The leaf router will join (or remain on) the multicast tree as long as one or more local hosts is in the multicast group. It does not know (or care) who or how many hosts are in the group. The number of routing nodes that have subscribed hosts -- rather than the actual number of subscribed hosts -- is the more meaningful definition of multicast group size, because it is the former which determines resource consumption in the provider's network. This definition of membership size has some interesting ramifications, as we shall see in Section 3.2.4.

We use the factor k to capture the extent of economies of scale realizable via multicast. In addition to quantifying this EoS factor, it is also important to study and characterize its dependence on (or independence of) different network variables, such as network size, topology, membership size, distribution.

It is trivial to come up with extreme spatial distributions of receivers that will result in scenarios of $k = 0$, $k = 1$, or anything in between. Consider the simple network in Figure 3.2, where a sender is sending data to two separate multicast addresses. For the first multicast group, the receivers are downstream from the source router via different links, and so no link savings are realizable ($k \approx 1$). On the other hand, the receivers in the second multicast group lie in a common distribution path, and significant link savings are realizable ($k \approx 0$). For generality, this study assumes that receivers are randomly distributed throughout the network.

Therefore, the router cannot infer, from the accounting of 'REPORT' and 'LEAVE' messages, the local multicast group size.

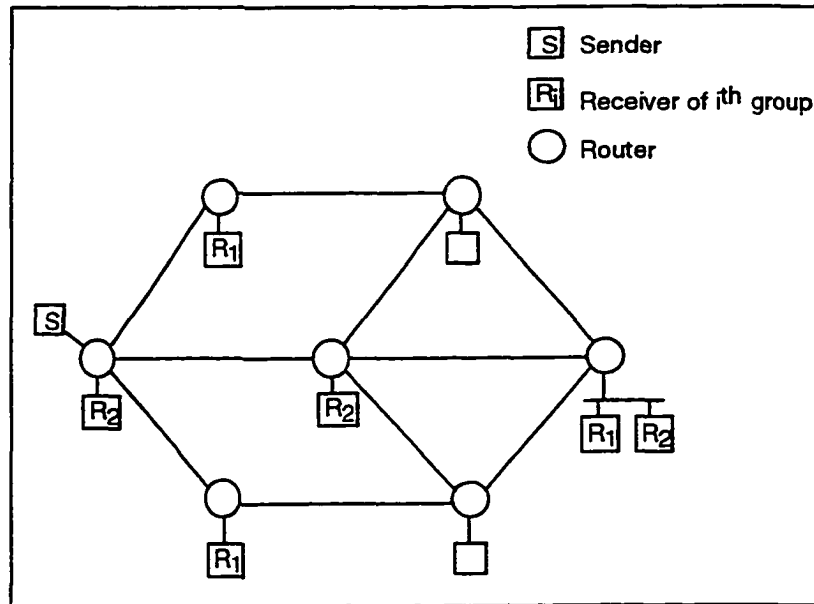


Figure 3.2. Example network shows that degree of link savings achievable is strongly dependent on spatial distribution of receivers.

3.2.2 Methodology

Figure 3.3 provides a pictorial overview of the methodology used in this study. Shortest-path multicast trees are constructed, using Dijkstra's algorithm (Dijkstra, 1959), over a variety of networks and receiver sets. This allows us to quantify the cost of multicast trees, validate the relationship of Equation (3.1), and estimate the EoS factor k .

To determine if the size and topological style of the networks affect multicast tree costs, we employ real and generated networks that are representative of inter-domain routing topologies of the Internet. These networks consist of routing nodes and interconnecting links. Real network topologies are gathered from the MBone²² and the early ARPANET. Network generation tools (GT-ITM (Zegura, Calvert and

²² MBone network topology from 7/30/1996; available from <http://www.nlanr.net/Caidants/Mrwatch.data.tar.gz>

Bhattacharjee, 1996, Calvert, Doar and Zegura, 1997) and tiers (Doar, 1996)) are utilized to produce realistic networks of different topological styles, as illustrated in Table 3.1.

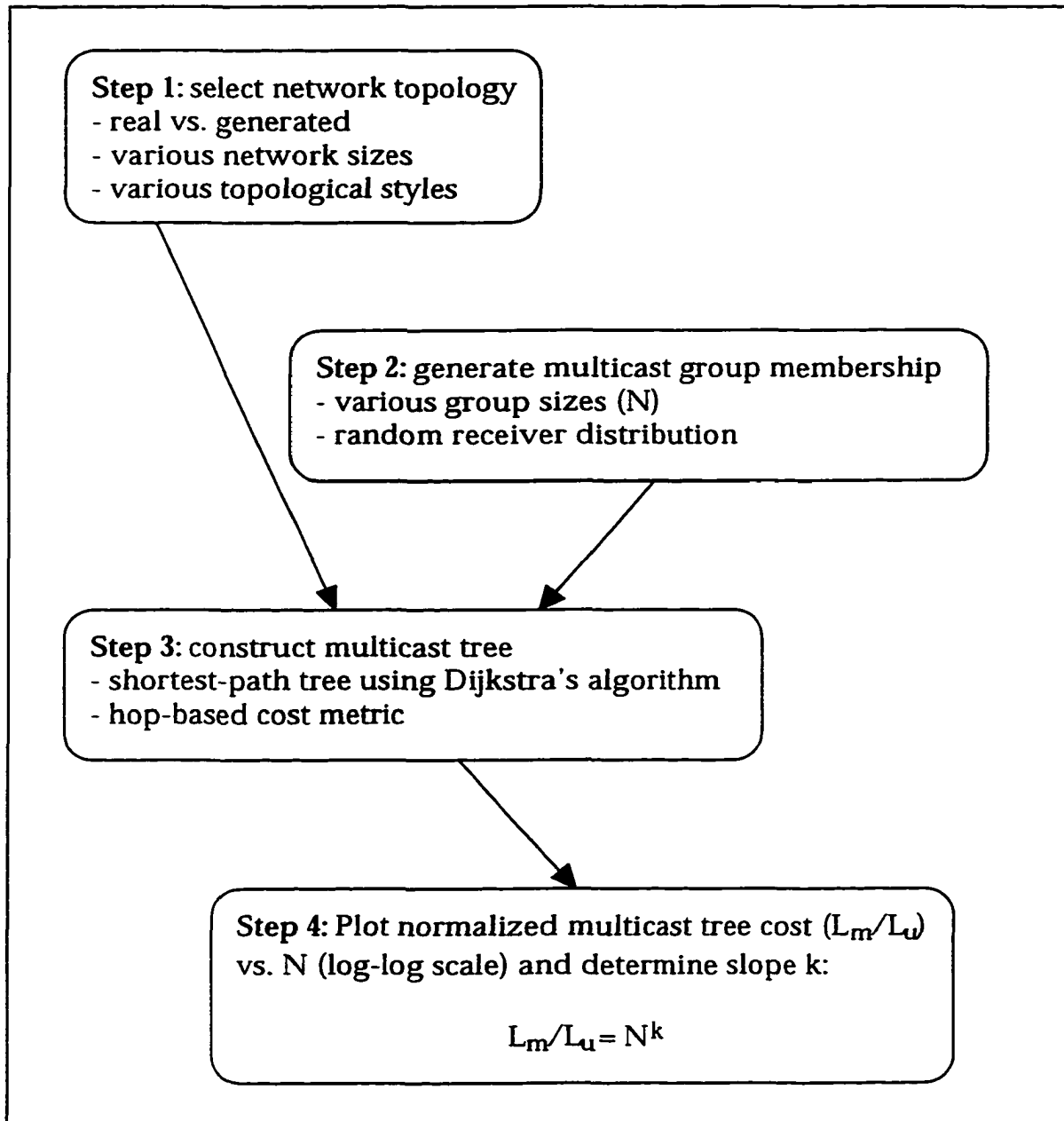


Figure 3.3. Quantifying economies of scale in multicast communication - a process overview.

Through user-specified parameters, we can control the style and size of the networks generated by the network generation tools. Specifically, by controlling parameters on edge probabilities, we are able to generate networks with average node degrees consistently in the range of 3 to 4, which is typical of present-day networks. (For a network with N nodes and M links, its average node degree is $2*M/N$.) Please refer to (Doar, 1996, Zegura, Calvert and Bhattacharjee, 1996, Calvert, Doar and Zegura, 1997) for more details on the use of these tools.

Ten different topologies are created for each of the five generated network styles; ten sets of receiver distributions are generated for each group membership size. For the arpa and mbone networks, where single real topologies are available, a hundred sets of receiver distributions are generated. Therefore, all data points in the following plots have sample size of 100. Table 3.1 lists the topologies used in this study.

Table 3.1. Networks used in this study.

Name	Type	Source/ Tool used	Topological Style	# of Nodes	# of Links	Avg. Node Degree
arpa	real	ARPANET	-	47	68	2.89
mbone	real	MBone	-	5019	9310	3.71
r100	generated	GT-ITM	random	100	169.4	3.39
ts100	generated	GT-ITM	transit-stub	100	181.1	3.62
ts1000	generated	GT-ITM	transit-stub	1000	1819.0	3.64
ti1000	generated	tiers	hierarchical	1000	1681.5	3.36
ti5000	generated	tiers	hierarchical	5000	8837.0	3.35

3.2.3 Results

The results of our analysis confirm that the cost of multicast trees can indeed be approximated by Equation (3.1), and that the economies of scale (EoS) factor k falls within a narrow range for reasonable network conditions. This implies that the L_m/L_u ratio is an exponential function of the number of receivers in the multicast group, N . Figure 3.4 shows that this exponential relationship applies for all three topological styles (random, transit-stub, hierarchical) of generated networks, and the value of k lies in the 0.8 range (standard deviations are shown with error bars).

Figure 3.5 shows the same L_m/L_u ratio for two networks, one with 1,000 nodes and the other with 5,000 nodes. From this plot it is apparent that the slope k (again ~ 0.8) is independent of the total number of network nodes. For example, a 500-member multicast group in a 1,000-node network (50% subscription rate) will realize a similar EoS factor as a 500-member group in a 5,000-node network (10% subscription rate). This important result shows that it is the *absolute number* of nodes that are receivers in a group, not the percentage of nodes that are receivers, that should be used as an indicator for multicast tree cost.

Figure 3.6 confirms that the relationship holds for real network topologies of vastly different sizes, namely the early ARPANET and the Mbone topology of 1996. The EoS factor k is again closely bounded in the 0.8 range.

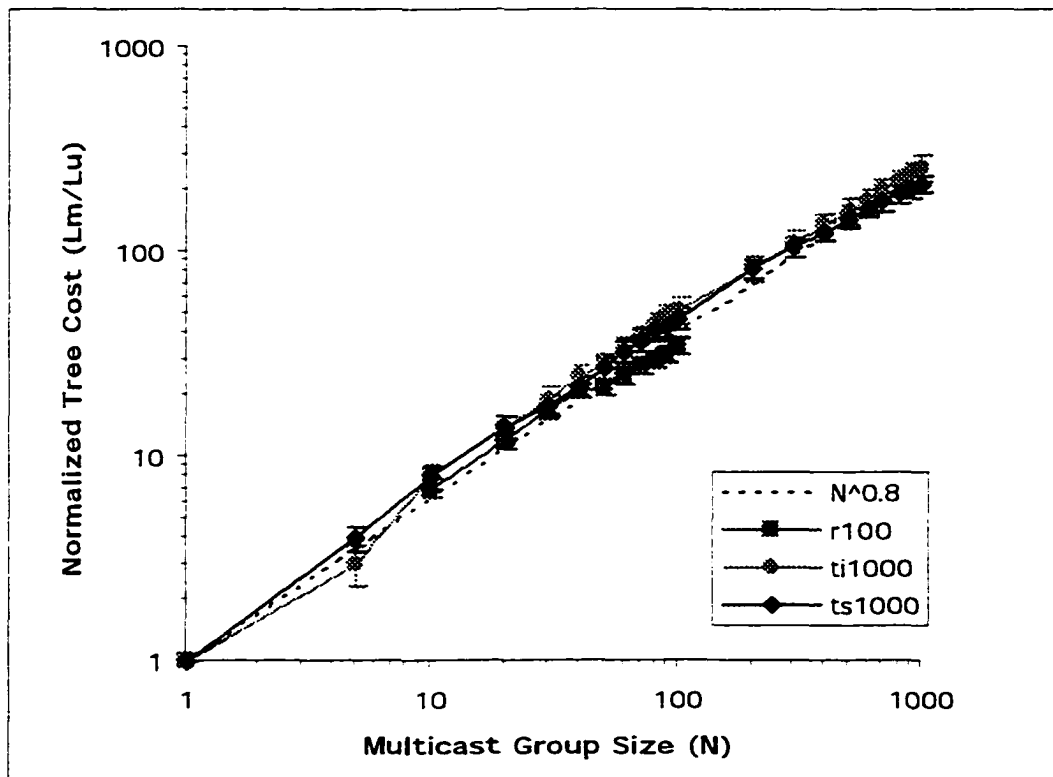


Figure 3.4. Normalized multicast tree length as a function of membership size - slope is constant (-0.8) across various network topological styles.

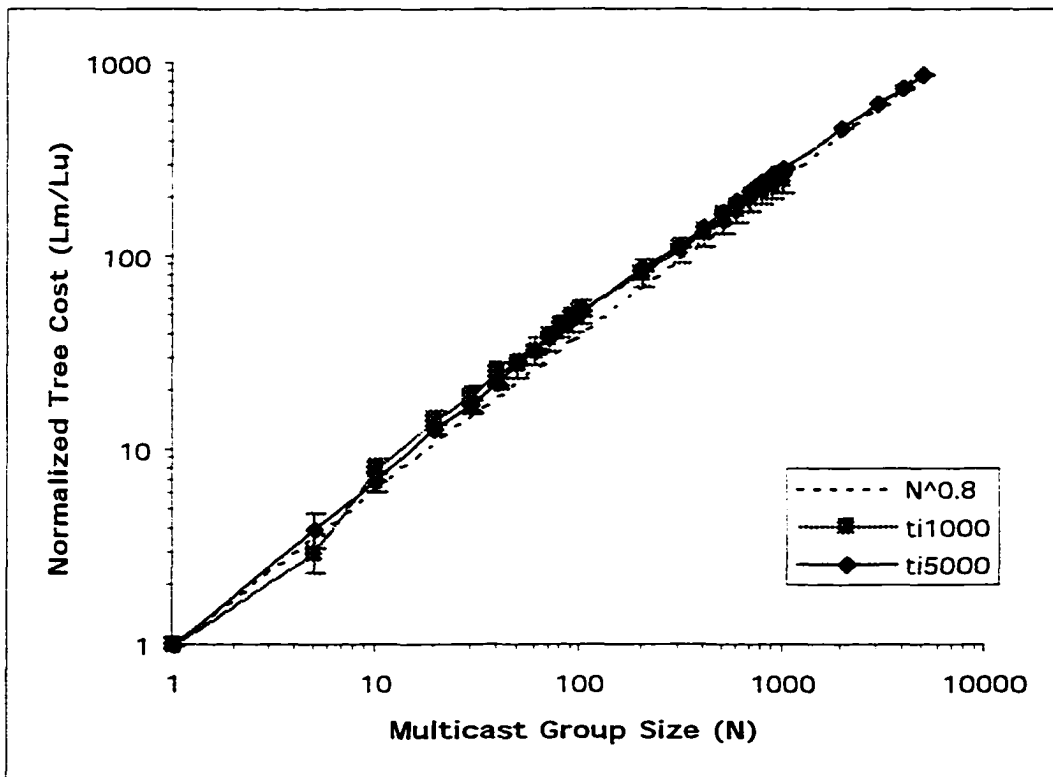


Figure 3.5. Normalized multicast tree length as a function of membership size - slope is constant (-0.8) across various network sizes.

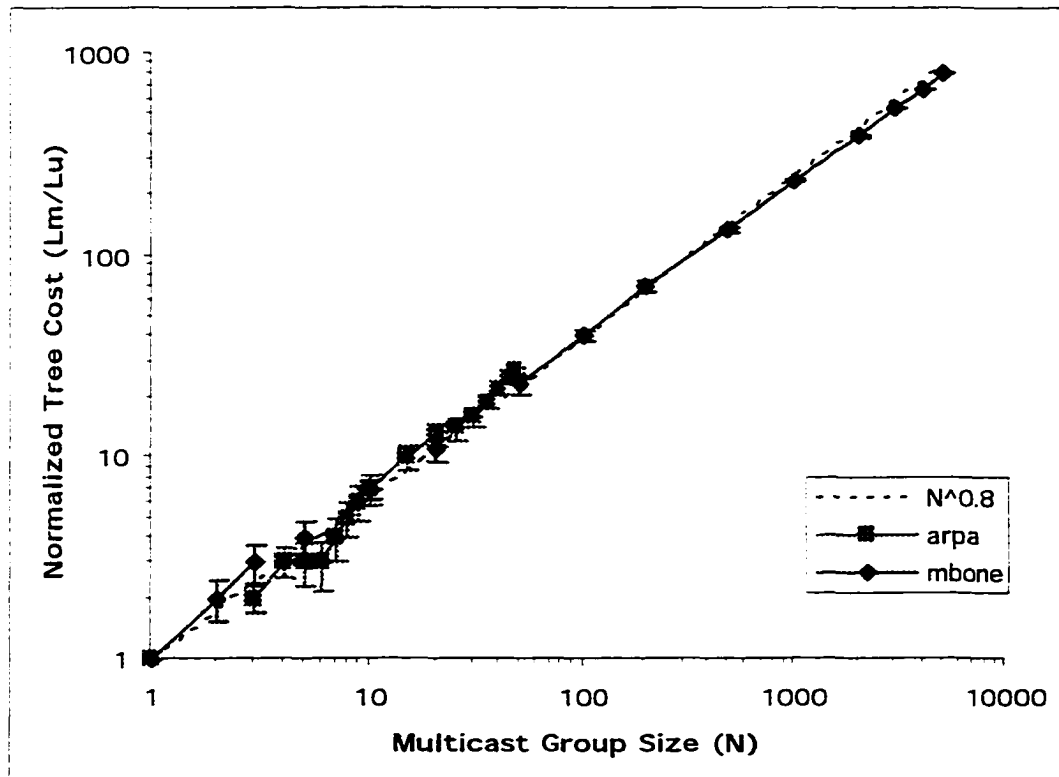


Figure 3.6. Normalized multicast tree length as a function of membership size - results confirmed with real networks.

3.2.4 Tree Saturation

As we have indicated in Section 3.2.1, multiple hosts on a same subnet attached to a leaf router may all be part of a multicast group, but they only count as one receiver from the router's point of view. From a cost-accounting perspective, this result is actually desirable, since the incremental cost of serving additional receivers on a shared broadcast capable subnet is zero. Even where the subnet is non-broadcast, as with ISP POPs, the subnet costs are typically covered by direct subscriber network access charges. However, the presence of multiple hosts per leaf router also leads to the tree saturation effect, which manifests itself in topologies with large local fanouts.

Tree saturation is best illustrated by the example of a realistic national ISP, which has 1,000 dial-in ports at each of its 100 points-of-presence (POPs). This means that the ISP can have up to 100,000 individual hosts connected to the network at any given time. Probabilistically, it takes just ~500 randomly distributed hosts (0.5% of total host population) to join a multicast group before all the POPs have at least one group member.²³ At this point, the multicast delivery tree is “fully grown” or “saturated”, and additional group members can be served at essentially zero incremental cost. Figure 3.7 gives an illustration of this tree saturation effect. Note that the x-axis is now the number of subscribing hosts, rather than the number of POPs with subscribers.

²³ The expected number of subscribing hosts (sampling with replacement, for simplicity) needed to place all M points-of-presence on the tree is:

$$E[N] = M \sum_{k=1}^M \frac{1}{k}.$$

For $M=100$, $E[N] = 519$. For the actual case where we have sampling without replacement, the expected number would be even lower.

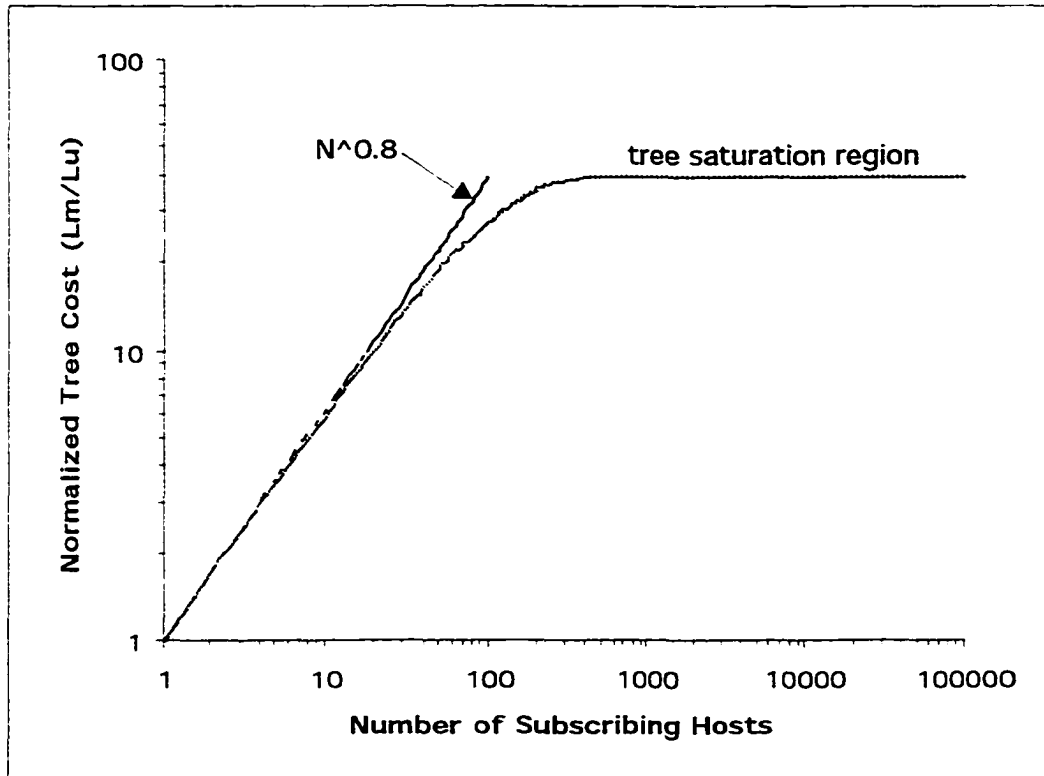


Figure 3.7. An illustration of the “tree saturation” effect: it takes just ~500 randomly selected dial-in ports (or 0.5% of all ports) to subscribe to a multicast group before all 100 network nodes become part of the multicast tree. All subsequent subscribers can be served at no additional cost.

3.3 Multicast Pricing

Since multicast tree cost can be accurately predicted from its membership size, we can directly apply the cost expression of Equation (3.1) into a simple price relationship:

$$P_m/P_u = N^{k'} \quad (3.2)$$

where P_m : price of multicast stream to N nodes, relative to
 P_u : price of unicast stream to a single receiver;
 k' : network-specific EoS factor (empirically derived).

This relationship holds regardless of whether unicast traffic is subject to per-packet (usage-sensitive) or per-month (flat-rate) pricing. Clearly, this gives us a very strong motivation to price multicast as a function of N , the multicast group size.

The price relationship of Equation (3.2) is applicable even if we are operating in the “tree saturation” regime. In this case, N should simply be set to N_{TOT} , the total number of nodes in the network. For example, an ISP with 100 POPs and $k' = 0.8$ would set the price ceiling of its multicast service to be $100^{0.8}$ or 40 times that of its unicast service.

It is important to realize that this pricing approach is different from a flat-rate pricing approach.²⁴ In the latter case, all multicast streams are priced at a flat rate, even if there is only a small number of receivers, and the tree is far from reaching saturation. Of course, a flat-rate pricing approach avoids the accounting overhead associated with traffic metering. However, this pricing scheme would favor applications with large numbers of receivers, at the expense of other applications with fewer receivers. Consequently, applications with fewer than P_m/P_u receivers (e.g., teleconferencing between several parties) will not opt for multicast even though it is more bandwidth efficient.

This proposed pricing scheme can be characterized as a two-part pricing scheme. Multicast traffic is charged according to N (raised to the 0.8 power) until N reaches N_{TOT} , at which point the price ceiling takes effect. This two-tiered approach would ensure that the multicast service is made available to all traffic types in a non-discriminatory manner.

²⁴ UUNET, the pioneer in IP multicast deployment, currently adopts a flat-rate pricing approach with a P_m/P_u ratio of ~400 (UUNET Technologies, 1997). But its president is also predicting an Internet-wide switch to usage-sensitive pricing in the near future.

3.3.1 Membership Accounting

One practical question remains: how can N be determined for multicast traffic, and at what accounting cost? We first recognize that the receiver-initiated nature of IP multicast precludes centralized knowledge of membership size. Examination of multicast packets is fruitless because the destination address in the packet header is a logical one, revealing nothing about the number and locations of the receivers. Secondly, multicast group membership can change in real time. Receivers may join or leave the group at any point in time, and the multicast tree will be dynamically grafted or pruned accordingly. Any snapshot at the beginning or end of a multicast session will not necessarily yield an accurate picture of the group membership.

As pointed out by Shenker et al. (1996), multicast pricing is an inherently non-local problem. Therefore membership accounting has to be achieved via distributed metering. One approach might be to count the number of multicast routers that are part of the multicast tree. In the case of reservation-based traffic, membership accounting might be achieved by the monitoring of QoS reservation signaling (e.g., RSVP). Network measurement and accounting software are commercially available²⁵ for installation as edge-metering devices to capture the necessary information for accounting and billing purposes.

It is worth reiterating that the membership size thus captured will only indicate the number of network nodes with subscribing hosts, N_R , not the total number of subscribing hosts, N_H . From the network's point of view, it is concerned with the bandwidth usage of the multicast tree, whose cost is dependent on N_R . Therefore, it is economically efficient for the network provider to price its multicast service as a function

²⁵ One example is Cisco's NetFlow software (Cisco, 1997), which can be configured to support various charging schemes such as QoS-based charging and distance-sensitive charging.

of N_R . The end user, on the other hand, has to take N_H into account when comparing the multicast and unicast alternatives. In the case of unicast, the sender has to transmit a duplicate copy of data to each of the N_H destination hosts.²⁶ For multicast, the membership size will be N_R , with $N_R < N_H$, since some of the hosts may be attached to common routing nodes. The sender would choose multicast over unicast as long as P_m , which is equal to $(N_R^{0.8}) * P_u$, is less than $N_H * P_u$.

3.3.2 Other Issues

This chapter does not address the issue of cost allocation and settlement (Herzog, Shenker and Estrin, 1995), except by noting that receiver-initiation does not necessarily imply that the charges have to be split among the receivers. There are many instances in telephony, for example, where the payment party is different from the initiating party. Assigning all multicast charges to the sender would result in a simpler billing system because (i) it is consistent with the unicast paradigm, i.e. meter and charge at the network entry point, and (ii) it avoids the ambiguities involved in equitably splitting the charges among multiple receivers. Out-of-band settlements are always available if needed.

This chapter addresses multicast pricing at the network layer (layer 3). When we look at reliable multicast at the transport layer (layer 4), we recognize that multicast retransmission traffic patterns may not be as predictable as unicast. There are various multicast transport protocols being proposed and developed (Obraczka, 1998). Depending on the number of receivers and the protocol used, the number of retransmitted packets may be comparable, if not more than the number of original data packets, even at

²⁶ We do not consider the case where proxy caches are installed at the edges of the network, in which case subsequent requests from the same subnet may be satisfied from the local copy. This would mean that the sender would only need to transmit N_R copies even in unicast mode.

a low packet-loss rate. Since reliable multicast is still in the development and definition phase, it is important to ensure that multicast pricing schemes at the network layer properly influence the choice of reliable multicast protocol at the transport layer.

3.4 Dense vs. Sparse Mode Multicast

Many different flavors and generations of multicast routing protocols have been proposed (Waitzman, Partridge and Deering, 1988, Ballardie, Francis and Crowcroft, 1993, Moy, 1994, Deering et al., 1996, Thaler, Estrin and Meyer, 1997), of which several have been implemented on the MBone. They can be classified into either dense mode (DVMRP, PIM-DM) or sparse mode (CBT, PIM-SM). In addition to bandwidth usage, there are many other dimensions to the tradeoff between DM and SM multicast, and these have been extensively studied elsewhere (Billhartz et al., 1997, Doar and Leslie, 1993, Salama, Reeves and Viniotis, 1997, Wei and Estrin, 1994, Wei and Estrin, 1995). As a general rule of thumb, DM multicast is perceived to be appropriate for mass-dissemination applications such as webcasting, whereas SM multicast is more suited for teleconferencing and other applications with just a few receivers. The question is: should dense and sparse mode multicast be priced differently?

Dense and sparse mode protocols differ primarily in their tree-construction techniques. Dense mode protocols take a flood-and-prune approach, where data packets are periodically flooded to the entire network, and branches are pruned where there are no downstream receivers. Sparse mode protocols, on the other hand, grow the distribution tree on a branch-by-branch basis as new nodes join the multicast group. Dense mode protocols work well when most nodes in the network are receivers, but are extremely bandwidth inefficient when the group members are few and sparsely located throughout the network.

We can incorporate the control overhead into the cost model we have developed, thus allowing a comparison of the total bandwidth usage between SM and DM multicast (and unicast and broadcast as well). Table 3.2 shows the link cost (measured in packet-hops per second) for transmitting data to a set of N receivers at a data rate of α packets/second. L_m and L_u are the multicast tree lengths and unicast path lengths as before:

Table 3.2. Data and control/overhead for various options of sending data to multiple destinations.

Type	Data	Control Overhead
unicast	$\alpha * N * L_u$	-
multicast (sparse mode)	$\alpha * L_m$	L_m / τ_{sm}
multicast (dense mode)	$\alpha * L_m$	$2(L_m' - L_m) / \tau_{dm}$
broadcast	$\alpha * L_m$	$\alpha * (L_m' - L_m)$

For unicast, no control overhead is necessary to coordinate the multiple receivers; the sender simply transmits one packet to each of the N receivers, and each receiver is on average L_u hops away from the sender. For sparse mode multicast, a tree of length L_m is constructed for packet delivery. The maintenance of this tree requires a periodic transmission of control messages. Once every τ_{sm} seconds, receivers have to reannounce their intention to remain on the tree by sending out a refresh message. Otherwise the link from which incoming packets are received will be pruned from the multicast tree. Regardless of how many receivers are downstream of a link, only one refresh message is required for each of the L_m links per refresh period. Dense mode multicast, on the other hand, takes a flood-and-prune approach. Periodically (once every τ_{dm} seconds) all multicast forwarding states time out and the data packet is broadcast to all nodes in the

network. Then those nodes who have no downstream receivers will send a 'prune' message to remove itself from the tree. If L_m' is the length of the broadcast tree, then there will be $(L_m' - L_m)$ links on which an unwanted data packet will trigger the transmission of a 'prune' message in the reverse direction. Finally, for broadcast communication, each of the L_m' links will carry a copy of the data packet. However, for $(L_m' - L_m)$ of these links, the data packet will be discarded, and hence these are classified as overhead in Table 3.2.

The impact and significance of the control overhead is dependent on the data rate α and the timeout periods. For the current multicast protocols, both τ_{sm} and τ_{dm} are on the order of minutes. Figure 3.8 shows the total link cost (data and control) needed to transmit a single data packet to N receivers in the MBone network using the various alternatives. As expected, the link cost for unicast is linearly proportional to the number of receivers, and the link cost for broadcast is constant. We also observe a crossover from sparse to dense mode multicast as the number of receivers increases. However, we make the interesting observation that dense mode multicast is never the least-cost option in this scenario except when all nodes are receivers. In fact, for the transmission of a single data packet, broadcast is the preferred approach when more than 40% of nodes are receivers. This suggests the cost of control messages may be prohibitively high for very low data-rate applications.

Figure 3.9 shows the same cost comparison when we move to a data rate of 5kbps (which is a conservative lower bound for most file transfer and multimedia applications). Unicast and broadcast are both unattractive except at the boundaries. Dense and sparse mode multicast *appear* to do equally well at all subscription density levels.

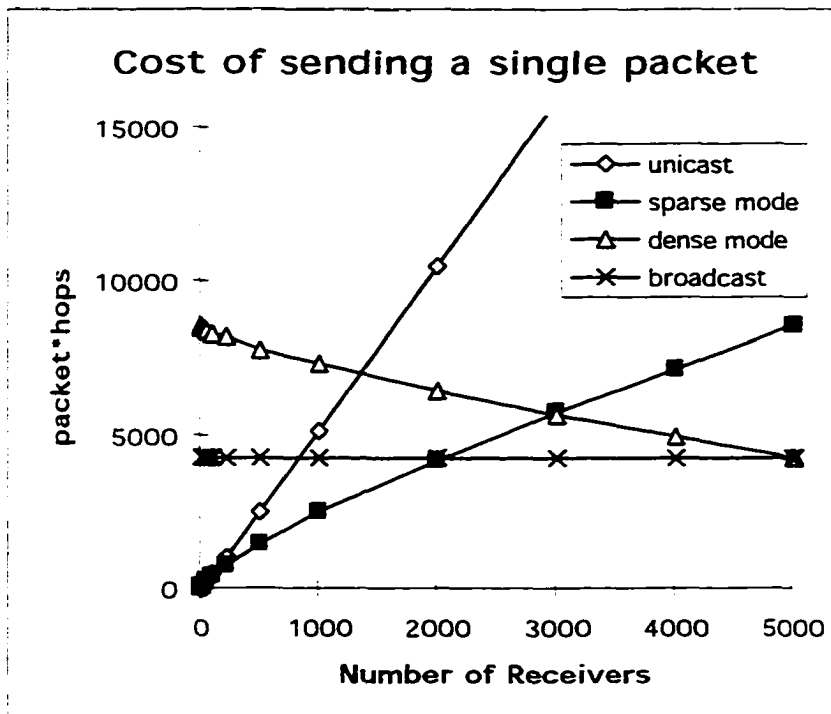


Figure 3.8. Comparing alternatives for sending one data packet to receivers in the MBone network.

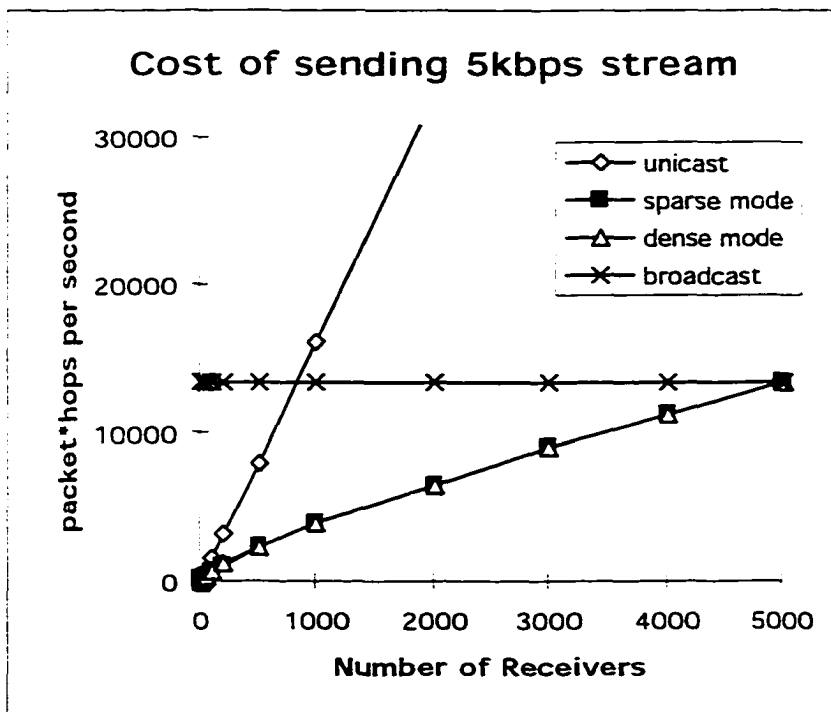


Figure 3.9. Comparing alternatives for sending a 5kbps data stream to receivers in the MBone network - there appears to be no difference between sparse and dense mode multicast.

Upon normalization to unicast cost, and replotting in log-log scale, Figure 3.10 reveals that there is a significant overhead associated with DM multicast at low subscription density levels. In fact, DM multicast is worse than unicast if there are less than ten receivers in the group (or 0.2% subscription density). On the other hand, when all network nodes are receivers (an unambiguously “dense” situation), DM does not perform significantly better than SM.²⁷ We observe that the SM cost curve maintains a slope of ~ 0.8 . This echoes our earlier results, and corroborates previous findings that overhead traffic amounts to no more than 1% of total traffic for SM multicast (Billhartz et al., 1997).

These results confirm that teleconferencing type applications can only be efficiently supported by sparse mode multicast. Dense mode multicast, on the other hand, will likely serve the webcasting market, where the groups are typically larger and less dynamic. Therefore, dense and sparse mode multicast services should be priced such that users select the appropriate mode based on their expectation of the group membership size.

One possible pricing approach is to offer DM multicast at a flat-rate while pricing SM multicast according to membership size. This way, applications with large numbers of receivers would opt for DM and its flat charge, while those with few receivers would choose the membership-sensitive tariff of SM. This approach might be justified by the fact that tree saturation is more likely to occur for webcasting scenarios. Furthermore, if we expect metering costs to be proportional to the number of receivers, then it may make economic sense to meter SM but not DM group memberships.

²⁷ The precise crossover point between SM and DM is highly variable from one topology to the next, but the two curves always approach the $k=0.8$ slope asymptotically. This suggests that there can be no meaningful formula or numerical expression for the “sparseness” or “denseness” of a multicast group.

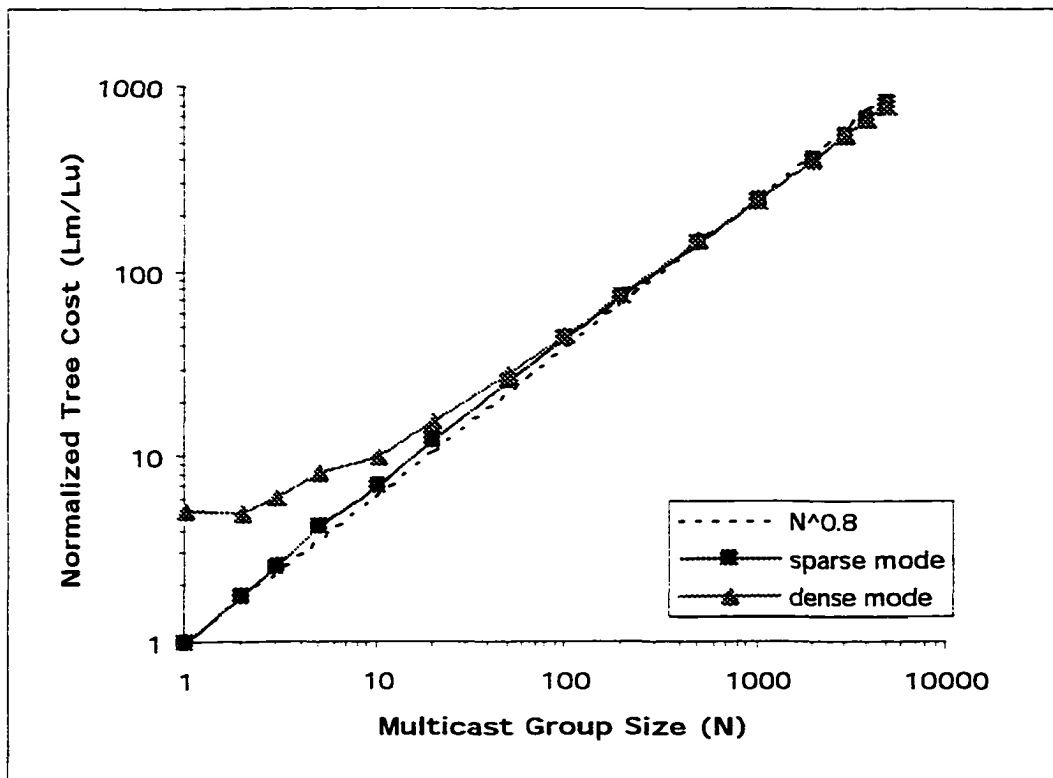


Figure 3.10. Comparing dense and sparse mode multicast for sending a 5kbps data stream to receivers in the Mbone network - dense mode multicast clearly consumes more bandwidth when there are few receivers, but the two modes are comparable with subscription density as low as 4% (about 200 receivers).

On the other hand, it is also entirely possible that SM multicast will become the general-purpose multicast vehicle, displacing DM multicast altogether. As illustrated by Figure 3.10, DM has little if any competitive advantage over SM multicast on a strictly link-usage basis. If this scenario occurs, SM multicast should be priced according to a two-tier approach as described in Section 3.3. This is the only way to ensure that multicast is available to both teleconferencing and webcasting application-types in a non-discriminatory fashion.

3.5 Conclusion

Through the quantification of multicast link usage, this work has demonstrated that the cost of a multicast tree varies at the 0.8 power of the multicast group size. This result is validated with both real and generated networks, and is robust across topological styles and network sizes. Practically, this means that the cost of a multicast tree can be accurately predicted given its membership size.

If a network provider takes a cost-based approach to multicast pricing, as advocated by this work, the above result provides a strong motivation to price multicast according to group size. Recognizing the effect of tree-saturation, a price ceiling should be incorporated into the price schedule, with the ceiling set precisely at the tree saturation level. This two-part tariff structure is superior to either a purely membership-based or a flat-rate pricing scheme, since it reflects the actual link usage at all group membership levels. Undesired subsidies between mass-dissemination applications (e.g., webcasting) and those with few receivers (e.g., teleconferencing) are eliminated, allowing the multicast service to be available to all applications in a non-discriminatory manner.

Explicit accounting of the control overhead allows a comparison of dense and sparse mode multicast within our cost framework. We find that sparse mode multicast maintains the exponential relationship between group size and cost, while dense mode multicast is inefficient at extremely low membership levels. This suggests that, in the event when both multicast modes co-exist to serve different markets, dense mode multicast is a good candidate for flat-rate pricing and the mass-dissemination market, while sparse mode multicast is a good candidate for pricing based on membership size and the teleconferencing market.

4. EoS in Time - Distributed Network Storage

Whereas multicast communication achieves economies of scale in delivering data to multiple receivers, network storage services such as caching and replication realize economies of scale in the temporal dimension. By storing copies of data objects in distributed locations throughout the network, accesses to data can be satisfied by nearby copies, saving the need to go all the way back to the original source. This results in four significant benefits:

- reduced access latency
- reduced bandwidth consumption
- server load balancing
- improved data availability/redundancy

The traditional distinction between caching and replication is that of ex-post versus ex-ante data duplication. An initial data request is needed to trigger the caching of the data object, and subsequent requests for the same object are served from the cached

copy until it is purged from the cache. A replicated copy of an object, on the other hand, is made in anticipation of its use at some future time. This anticipatory execution of replication can be based on a highly selective and speculative prefetching algorithm, or a complete duplication of the entire object-space (e.g., a mirror site).²⁸

This work suggests that new and important insights can be gained by looking at caching and replication from a Quality-of-Service (QoS) perspective. The QoS concept is not new; it comes from the transmission domain of data networking. What is new is the treatment of caching and replication as different QoS services within a unified network storage framework.

When applied to data transmission, QoS introduces the distinction between guaranteed service and best effort service. Best effort service is the default service used by most applications, and it does not offer any guarantees regarding packet delivery. Guaranteed service, on the other hand, offers performance guarantees based on latency, jitter and/or loss rates.

When the same QoS concept is applied to network storage, we recognize that caching can be considered to be a best-effort service, in contrast to replication which is a guaranteed service. Network caches perform best-effort service by storing a local copy of each object it sees (except those explicitly tagged uncacheable). Since caches have finite storage capacity, they have to evict old objects to make room for new ones. The fact that objects may be purged at any time means caches cannot provide any guarantees of data persistence. The possibility of a cache miss introduces uncertainty in the access latency of an object, similar to the introduction of jitter in data transmission.

²⁸ See Appendix 4 for a brief discussion and taxonomy of data duplication schemes.

Replication, on the other hand, represents a service commitment to keep a persistent copy of the object. There can be no misses at the replica (transmission analogy: no packet drops) in this guaranteed service. In order to provide guarantees, replication requires some form of resource reservation. Today, replication setups are mostly ad hoc and require manual intervention for want of a standardized reservation protocol. It is therefore extremely expensive, if not impossible, for these static replicas to respond to changing traffic patterns and network conditions.

The need for service guarantees in data transmission is driven by real-time network applications that cannot tolerate variations in packet delay. We believe that the demand for network storage services with guarantees will similarly come from applications that cannot tolerate the performance variations inherent in all caching schemes. These applications may be mission-critical, have stringent performance and/or availability requirements, or place high value in consistent data access latency, thereby requiring data objects to be kept in persistent storage even if they do not exhibit reference locality, or are rarely accessed at all. No amount of intelligent or adaptive caching, or overprovisioning (short of infinite cache size) can address the needs of these applications.

4.1 Distributed Network Storage Infrastructure with QoS Guarantees

The work in this chapter calls for the building of a distributed network storage infrastructure with QoS guarantees. This infrastructure will support, in one integrated framework, network storage services ranging from best-effort caching to replication with performance guarantees. Content owners can, through the use of standardized protocols, reserve network storage resources to satisfy their application-specific performance requirements. They can specify either the number and/or placement of the replicas, or higher-level performance goals based on access latency, bandwidth usage or data

availability. The network storage provider will optimally allocate storage resources to meet the service commitments, using leftover capacity for best-effort caching. Content consumers retrieve the nearest copy of the data object, be it from a replica, cache, or the original source, in a completely transparent manner.

The network storage resource thus reserved will be available for housing objects ranging from web pages, audio and video files, to databases, applets and executables. These storage nodes can also support scripts and processes that generate dynamic objects and maintain logs of access statistics.

Furthermore, this distributed network storage infrastructure can be integrated with the existing transmission-based QoS framework so that applications can select the optimal combination of storage and transmission resources to satisfy their performance requirements. While the focus of this chapter is on services based on network storage resources, we have explicitly adopted and adapted design philosophies and terminology from the transmission domain so as to facilitate a seamless integration of the two infrastructures in the future.

Figure 4.1 illustrates the process of turning performance requirements into performance realization via intelligent allocation of network storage resources. Based on its application-specific performance requirements and *a-priori* information about the probable pattern of information access by consumers (across objects, space, and time), the publisher uses some standardized semantics to express its formal QoS requirements. These requirements are then conveyed to the network storage service provider using some well established resource reservation protocol.

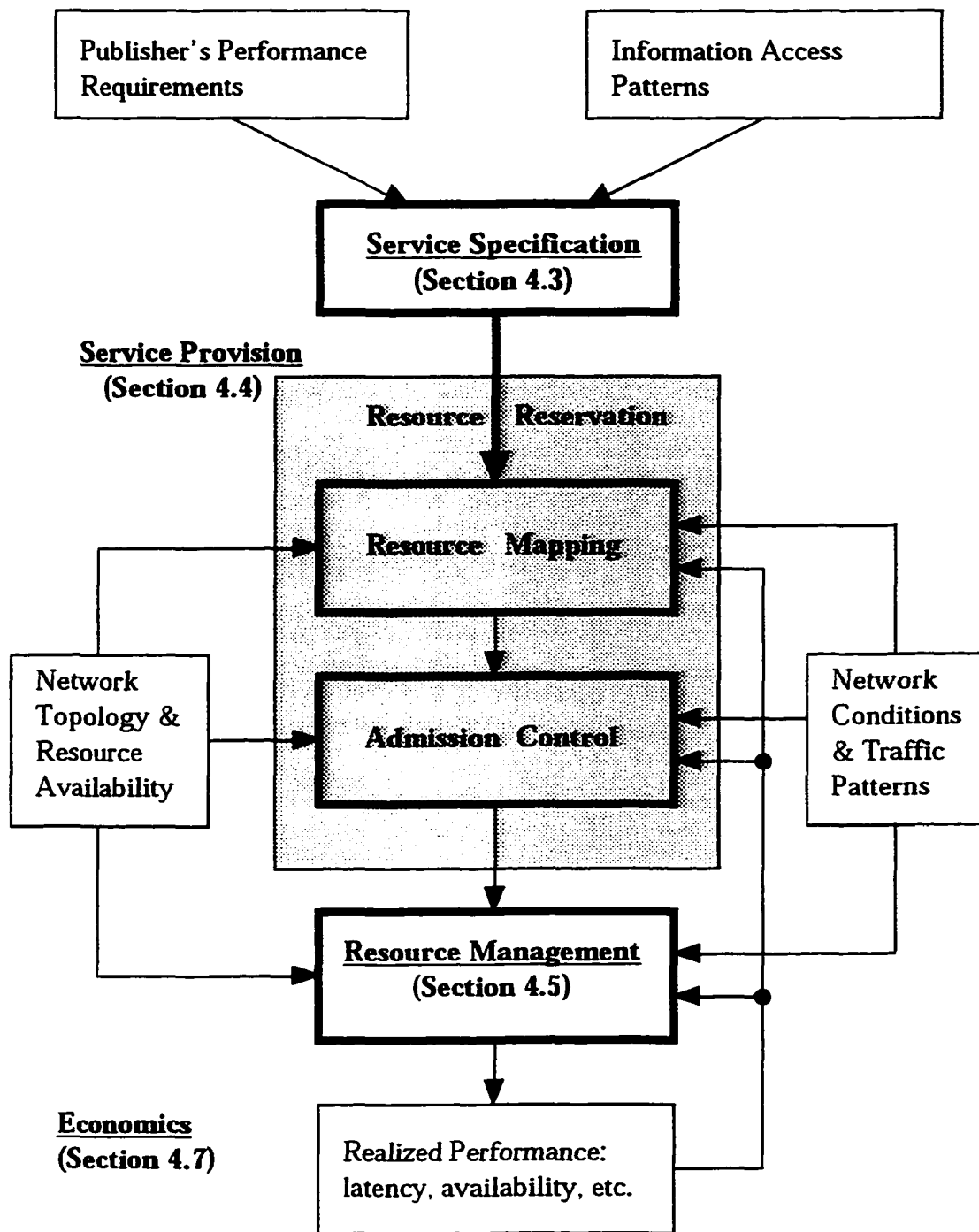


Figure 4.1. From performance requirements to performance realization: the process flow of establishing a network storage service with QoS guarantees. The components in bold are the key components of the infrastructure.

The provider, using information available to it regarding network topology, storage resources, and other current and projected resource demands, maps the QoS requirements into an optimal set of specific resource requirements for both distributed storage and transmission capacity to meet the QoS requirements at minimum resource cost.

Having calculated this optimal resource mapping, the service provider attempts to reserve the specific link and storage resources as determined by the mapping. Depending upon the extent of previous resource reservations, individual transmission and storage facilities may admit or deny the reservation request. If the requests are denied, then an alternative resource mapping must be computed, and the process repeated.

Individual storage nodes execute real-time resource management policies (e.g., local cache replacement, replica update policies, etc.) to maximize local resource utilization while meeting all service commitments.

The result is a performance realization, measured using the same metrics as those used to express the QoS requirements in the first place.

Finally, like all other multi-class service infrastructures, appropriate incentive mechanisms must be put in place, so that prices can act as market signals to moderate the use of the network storage resources.

Given this framework, we can identify the key components of the distributed network storage infrastructure:

- service specification (Section 4.3)
- service provision (Section 4.4)
 - resource reservation (Section 4.4.1)

- resource mapping (Section 4.4.2)
- admission control (Section 4.4.3)
- real-time resource management (Section 4.5)
- other mechanisms (Section 4.6)
- economics (Section 4.7)

Each of the components and their associated research problems will be described in the sections indicated above, following a review of relevant literature in Section 4.2. In addition, the resource mapping problem is studied in greater detail in Chapter 5. This work is intended to serve as a QoS framework upon which community discussion of this distributed network storage infrastructure can proceed.

4.2 Related Work

This work benefits from the cross-fertilization of two fields, namely (i) network caching and replication, and (ii) transmission-based QoS. While active research continues apace in both fields, this is the first proposal that introduces the notion of QoS-based services to the network storage domain. Caching alone is a “best effort” service that cannot provide service guarantees to a publisher.

The ideas of caching, replication, and memory hierarchy in general are well established in computer hardware, operating systems, distributed databases and distributed file systems design. The growth of Internet traffic (FTP and HTTP in particular) spurred the expansion of the memory hierarchy into the network itself. Caching and replication started out at the edge of the network. Caching proxies (Luotinen and Altis, 1994, Abrams et al., 1995) are installed at campus gateways and at the ISP’s

metropolitan points-of-presence (POPs); mirror sites are installed, with manual intervention, as replicated servers of popular FTP/HTTP sites.

4.2.1 Network Caching

Before long, caching progressed into the wide-area network (WAN) itself. Motivated by strong references of locality observed in wide-area data access patterns (Danzig, Hall and Schwartz, 1993, Almeida et al., 1996, Huberman et al. 1998), network caches are organized into hierarchies (Chankhunthod et al., 1995). Despite a dynamic hierarchical caching proposal (Blaze and Alonso, 1992), the dominant network-caching infrastructure is still a manually configured hierarchy of object caches. The static hierarchy is limited to no more than three levels for latency considerations. Some proposals call for object sharing among neighboring caches via inter-cache communication protocols (Malpani, Lorch and Berger, 1995, Wessels and Claffy 1997, Wessels and Claffy 1998, Fan et al., 1998); others call for network caches that can handle dynamic objects (Iyengar and Challenger 1997, Cao, Zhang and Beach, 1998). There is currently a flurry of adaptive, self-organizing caching proposals that promises intelligence, scalability and adaptability for network caching (Bhattacharjee, Calvert and Zegura, 1997, Heddaya, Mirdad and Yates, 1997, Wang and Crowcroft, 1997, Zhang, Floyd and Jacobson, 1997, Michel et al., 1998). Under the rubric of "active networking", DARPA has recently funded a project looking at the adaptation of protocols developed for cache management in network attached storage devices to the larger problem of unreliable WANs (Nagle et al., 1998).

There has been a proliferation of novel replacement policies for network caches. These policies are all variants of the Least Recently Used (LRU) or Least Frequently Used (LFU) policies. But in addition to object popularity, these policies also incorporate

object size, distance, latency and/or cost factors into the decision-making process (Cao and Irani, 1997). On the other hand, there are applications whose objects are managed independent of how frequently or recently they were accessed. For example, objects that have stringent performance requirements or are mission-critical need to be kept in persistent storage even if they do not exhibit reference locality, or are rarely accessed. In these instances, the object owners would seek to secure network storage resources with QoS guarantees not available with network caching.

4.2.2 Network Replication

There have been many different proposals for network replication, though the only ubiquitous scheme is also one of the earliest: NNTP (Kantor and Lapsley, 1986). Network News Transfer Protocol involves massive replication of news-articles to NNTP servers (on the order of 10,000's) throughout the network. However, NNTP only offers weak consistency, providing no guarantees regarding on-time replication of articles. Subsequent work based on this massive replication concept uses either multicast (Lidl, Osborne and Malcolm, 1994) or the hierarchical organization of servers (Danzig, Delucia and Obraczka, 1994, Obraczka, 1994). The Internet2 initiative is proposing a distributed storage infrastructure which allows massive replication to a system of replicated servers, though the content publisher would have no control over the placement of the objects nor receive any performance guarantees (Beck and Moore, 1998). Wolfson, Jajodia and Huang (1997) examine adaptive replication algorithms for selecting web or database replica sites, but does not consider the problem of meeting specific QoS criteria.

Other proposals for network replication tend to focus on the ex-ante vs. ex-post distinction of data duplication (Baentsch et al., 1997). Therefore, they should be more accurately characterized as proactive or push caching (Gwertzman and Seltzer 1995),

selective pre-fetching (Kroeger, Long and Mogul, 1997, Wang and Crowcroft, 1996), or demand-driven replication (Bestavros, 1995). The prior research most closely aligned with our proposal is that of Bestavros and Cunha (1996) and Bestavros (1997). However, this work focuses on relatively stable data, and does not examine algorithms for optimal placement, alternatives along the spectrum from guaranteed service (mirrors) to best effort (caches), nor tradeoffs between reserving storage and reserving transmission capacity. Markatos and Chronaki (1998) argue for a hierarchical combination of selective replication (pre-fetching) and caching.

Today, mirroring and contracting to web-hosting services remain the only viable replication options that provide some form of persistence guarantee to a publisher. Both involve high degrees of customization and human intervention, and are therefore limited to static, long-term arrangements,²⁹ involving entire sites as opposed to individual data objects. Responding to changing traffic patterns and network conditions is extremely costly, if not impossible, in these cases.³⁰

4.2.3 Transmission-based QoS

The need for network support for multiple service levels has been long recognized (Clark and Tennenhouse, 1990, Ferrari and Verma, 1990, Ferrari, 1992). Real-time network applications require some form of performance guarantees that are not available from a single-class best-effort infrastructure. Therefore, the concept of QoS was introduced at the IETF and ATM Forum organizations and became embodied in standards

²⁹ Official sites for the Olympic Games, World Cup are notable exceptions.

³⁰ There is a recent proposal to bring differential service to web servers and content-hosting servers (Almeida et al., 1998). This scheme calls for the preferential scheduling and processing of requests, but does not offer any guarantees with regard to object persistence. In this it is similar to notions of differential service for network transmission.

such as the *intserv* framework (Braden, Clark and Shenker, 1994) and the Traffic Management Specification (ATM Forum, 1996). These standards specify the different service classes and the service guarantees available to network applications. The realization of these schemes requires advances in traffic specification (Ohnishi, Okada and Noguchi, 1988, Ferrari and Verma, 1990), resource reservation (Zhang et al., 1993), resource mapping (Hui, 1988, Guérin, Ahmadi and Naghshineh, 1991, Kelly, 1991) and admission control (Hyman, Lazar and Pacifici, 1993, Jamin et al., 1997), scheduling algorithms³¹ (Demers, Keshav and Shenker, 1990, Ferrari and Verma, 1990, Parekh, 1992, Floyd and Jacobson, 1995) and queue management (Floyd and Jacobson, 1993), along with various other control and management mechanisms such as traffic policing. Pricing design for multi-service networks has also witnessed a flurry of research activity (Cocchi et al., 1991, Low and Varaiya, 1993, Honig and Steiglitz, 1995, Sairamesh, Ferguson and Yemini, 1995, Shenker 1995, Clark 1997, Gupta, Stahl and Whinston, 1997, Songhurst and Kelly, 1997, Wang, Peha and Sirbu, 1997, de Veciana and Baldick, 1998). While the transmission QoS literature provides a useful starting point for identifying the mechanisms needed for a distributed network storage infrastructure, we believe that there are fundamental differences between the two infrastructures that require more than simple adaptation of designs and architectures.³²

Having identified the relevant literature in network caching, replication and transmission-based QoS, we are now ready to describe the key components of the distributed network storage infrastructure.

³¹ Zhang (1995) provides a comprehensive survey of packet scheduling disciplines.

³² For example, transmission buffers generally follow a FIFO discipline (or some variant of FIFO), but network storage is usually random access. Therefore, the nature and cost of congestion is not the same. Also, we expect the nature and degree of traffic burstiness to be different between data transmission and network storage demand.

4.3 Service Specification

The first step towards creating a useful distributed network storage infrastructure is to identify service classes that may be of value to applications. In the previous sections we have simply identified caching and replication as the two basic service classes. In reality, different applications have diverse needs and performance goals and will therefore demand different flavors of network storage services. A service specification standard or API (application programming interface) will allow the content owners and the network storage providers to communicate, using unambiguous metrics, the requirements and expectations of a service commitment.

There are two chief elements to a service specification: *traffic profile* and *performance requirements*. In data transmission, the traffic profile of the source is usually expressed as some combination of peak and average rates, maximum burst length, token bucket filter rate, etc. (Ohnishi, Okada and Noguchi, 1988, Ferrari and Verma, 1990). Performance requirements, on the other hand, are usually specified in delay bounds, acceptable loss rates, etc. When a service contract is established, the network is responsible for meeting the performance requirements, so long as the source transmits data within the prescribed traffic profile.

The specification of a network storage service also consists of a traffic profile and performance requirements. The traffic profile declares the amount of storage capacity to be reserved, the time and duration of the reservation, and the distribution of data accesses, if known. The performance requirements can be expressed along one or more of the following (sometimes overlapping) dimensions:

- data access latency (mini-sum, mini-max)
- data access jitter

- acceptable miss rate (including 0%)
- data availability/redundancy
- coverage area
- bandwidth savings
- cost

The distributed network storage infrastructure has to be able to accommodate new service classes and new performance metrics as the market demands them. We provide some example services here for illustrative purposes (Table 4.1).

Table 4.1. Some examples of network storage services.³³

Service	Description (traffic profile, performance requirements)
#1 Deterministic	1GB storage capacity for 1 hour, 100ms maximum latency
#2 Deterministic	1GB storage capacity for 1 hour, 50ms maximum latency
#3 Average	1GB storage capacity for 1 hour, 50ms average latency
#4 Combination	1GB storage capacity for 1 hour, 50ms average latency, 100ms worst case latency
#5 Stochastic	1GB storage capacity for 1 hour, $\text{Probability}[\text{latency} > 100\text{ms}] \leq \epsilon$
#6 Geographic	1GB storage capacity for 1 hour, 100ms latency bound for all receivers in specific domain or region, or to specific set of receivers
#7 Budget-constrained	1GB storage capacity for 1 hour, minimizing worst-case latency, subject to budget constraint of no more than K replicas
#8 Placement-oriented	1GB storage capacity for 1 hour, at N specific nodes
#9 Advance reservation	1GB storage capacity from 2330hr, December 31 1999 to 0029hr, January 1 2000, 100ms latency bound

³³ While these example services have performance requirements in terms of worst-case data access latency, similar services may also be specified with average latency, jitter, or other performance metrics.

For the first six services, the metric used for data access latency is milliseconds (ms). Alternatively, latency might be measured in terms of network hops. For example, instead of requiring a 100ms latency bound, service #1 might specify that all data accesses have to be served by a replica no more than four hops away. Unless the radius of the network is less than or equal to four hops (or 100ms), this service will require the reservation of one gigabyte (1 GB) of storage capacity each at multiple nodes. In fact, it is up to the service provider to decide which set of nodes need to be hosting replicas in order to achieve the latency bound.

Service #2 is identical to service #1 except for a more stringent latency requirement. Given a general network topology, we would expect that the number of replicas needed for this service to be more than twice that of service #1. This suggests that the cost of service is a non-linear function with respect to latency.

Service #3 specifies an average latency bound rather than a worst-case latency bound, and therefore the distribution of demand is important. For example, if the majority of requests originate from a small cluster of nodes, the service may be satisfied by having a single replica close to these nodes. Service #4 simply combines the performance requirements of services #1 and #3.

Service #5 provides a statistical guarantee that no more than a fraction ϵ of all data accesses will miss the latency bound. A secondary performance guarantee on miss latency (e.g., 100% of data accesses within 500ms) may or may not be provided. This class of services allows more efficient utilization of the storage resources via statistical multiplexing, and should be priced more cheaply than service #1.

Some data objects may be of limited geographic scope, and service #6 allows this information to be applied in determining the placement of the replicas.

Service #7 gives the service provider the responsibility to optimize latency performance given a budget constraint. In place of minimizing worst-case latency, services in this class may also optimize for other criteria such as average latency, jitter, minimum-cut, etc.

In some cases, the content owner may decide the exact locations at which to replicate the objects. Service #8, for example, would allow the content owner to place one replica in each of the major continents, thereby avoiding the need for any transoceanic transmission.

Finally, service #9 is identical to service #1 in all aspects except for the start and end times of the service. By adding a *start-time* field to the *duration* field in traffic profile, the content owner can make reservations for storage capacity in advance, rather than wait until the need for storage becomes imminent. In order to support this service, the infrastructure will have to keep track of resource availability into the future, so that resource mapping and admission control may be performed correctly. In return, this service allows forward planning by both the service provider and the service consumer. The tradeoff between the ability to plan into the future and the cost of maintaining scheduling information will determine the optimal planning horizon for this service offering. Two very recent proposals of advance reservation mechanisms for transmission-based services (Berson, Lindell and Braden, 1998, Schelén and Pink, 1998) reaffirm the importance of this class of services.

These services are just a small sample of the many possible services that may be offered over the distributed network storage infrastructure. Various other performance goals may be substituted for worst-case latency in many of these examples. Clearly, the more types of services to support, the richer the specification semantics need to be. The

challenge, as always, is in achieving the right balance between simplicity and flexibility. While these example services offer a glimpse into the many dimensions along which services may be classified, we choose to highlight two particular dimensions in the following two sub-sections.

4.3.1 Deterministic vs. Statistical Guarantees

Firstly, services can be differentiated by the "firmness" of their guarantees. The QoS work in the data transmission arena provides ample illustrations (Ferrari, 1990, Clark, Shenker and Zhang, 1992). The IETF (Internet Engineering Task Force), for example, has specified three classes of services as part of their integrated-services framework: guaranteed service (GS), controlled load service (CLS) and best effort service (BES) (Braden, Clark and Shenker, 1994, Shenker and Wroclawski, 1997, Shenker, Partridge and Guérin, 1997, Wroclawski, 1997). Similarly, the ATM Forum (1996) has specified four classes: constant bit rate (CBR), variable bit rate (VBR), available bit rate (ABR) and unspecified bit rate (UBR). These service classes can be characterized as providing one of the following performance guarantees: deterministic, statistical, or no guarantee. Services with deterministic guarantees, such as GS and CBR, provide lossless packet transmission with a worst-case latency bound. Those with statistical guarantees permit a small fraction of packets to arrive outside the latency bound, which may be acceptable to some adaptive applications, in exchange for a greater degree of statistical multiplexing, higher resource utilization, and therefore lower cost. Finally, best effort services (e.g., BES and UBR) offer no guarantees whatsoever.

Applying this to our example services, we see that services #1, #2, #6 and #9 provide deterministic guarantees on access latency. All data accesses are guaranteed to experience no more than the stipulated 50ms or 100ms delay. Services #3 and #5, on the

other hand, offer statistical guarantees. Service #3 makes latency guarantees only for data accesses in the aggregate, but not for individual data accesses. For service #5, up to ϵ of data accesses may fall outside the latency bound without violation of the commitment. Service #4 offers a combination of deterministic and statistical guarantees. Finally, the best effort service of network caching corresponds to the base case of offering no guarantees.

It is important to recognize that services #3-5 do not necessarily represent the full range of services with statistical guarantees. The exact specification of statistical guarantee services may be dependent on the stochastic nature or the source of burstiness of the traffic load in question.

There are two sources of burstiness in terms of demand for network storage capacity. First, it is conceivable that some content owners may experience fluctuations in the size of their corpus. News publishers, for example, may have a relatively stable corpus size for ordinary news days but an explosion of additional news articles on days with extraordinary world events or stockmarket activity. These publishers may wish to characterize their traffic load with average and peak capacity numbers. Second, data access patterns may be bursty with respect to the objects requested, the geographic locations of the consumers, etc. These patterns may or may not be amenable to characterization using some demand distribution function (across objects, space and time).

To the extent that these stochastic behaviors or burstiness can be accurately characterized and made available to the network, appropriate statistical multiplexing techniques can be applied to improve storage utilization. On the other hand, for those applications with no burstiness in storage demand, they cannot hope to realize any statistical multiplexing gains, and are better off with deterministic-guarantee services.

Finally, in addition to these guaranteed services, there is also effort at the IETF to introduce differential or differentiated service to the Internet (Diffserv Working Group, 1998). This service provides no performance guarantees, but offers some notion of a premium service where packets are given preferential treatment over best effort packets. If we wish to apply this differential service concept to network storage, then some form of cache replacement policy that takes priority into account will be required. Alternatively, “premium” data objects may be tagged and initiated with a negative number in its *age* field when it is first cached.

4.3.2 Performance-Oriented vs. Placement-Oriented Services

Storage services can also be classified as either performance-oriented or placement-oriented. Performance-oriented services offer high-level performance guarantees such as latency bounds, but hide the exact number and placement of replicas from the service requester. Content owners who do not wish to concern themselves with network topologies, but care only about overall performance, would subscribe to this category of services. Example services #1-5 and #9 fall into this category.

On the other hand, some content owners may wish to exercise complete control over the number and placement of the replicas. They are willing to bear the cost of learning about the topology of storage nodes in the network. These content owners would request placement-oriented services that are similar to example service #8.

Services #6 and #7 allow the requester to impose some geographic and/or budget constraints, but still leave the final replica placement decision to the service provider. Therefore these services should be considered as performance-oriented.

This performance versus placement distinction has important architectural and economic ramifications, as we shall see in Section 4.4 and throughout the rest of this chapter. Ultimately, the distributed network storage infrastructure has to be flexible enough to support a wide range of services, including those yet to be specified.

4.4 Service Provision

Having identified some possible network storage service classes, we turn to the mechanisms for providing these services. As in transmission-based QoS provision, there are three main components of network storage service provision: *resource reservation*, *resource mapping* and *admission control* (Aurrecochea, Campbell and Hauw, 1998).

4.4.1 Resource Reservation Protocol

A resource reservation protocol allows the service requester and the service provider to communicate and negotiate the reservation of transmission and storage resources according to the service specifications. The protocol must be able to support the various types of services to be offered, including both performance-oriented and placement-oriented services. It would also be desirable for the protocol to include provisions for returning to the requester delivery logs and other indications that service level agreements are being met.

The resource reservation protocol for network transmission services, RSVP (Zhang et al., 1993), serves as a useful starting point for discussion. One possibility might be to extend the current RSVP protocol so that it can support reservation requests for storage resources as well as transmission resources. However, we foresee some difficulties with this approach. First of all, the concepts of the routing path and end-to-

end reservation do not apply to storage. Secondly, in the case of replication, the “receivers” or the content consumers may not be known at reservation time.³⁴ Whereas both sender and receiver(s) are involved in transmission-based resource reservation, only the content owner is involved in the storage-based case. This goes against the fundamental design philosophy of receiver-initiation in RSVP. The specification of the resource reservation protocol is outside the scope of this work, and should be postponed until the overall network storage service provisioning architecture has been defined.

4.4.2 Resource Mapping

Resource mapping is the translation of high-level service specifications into low-level resource requirements. To be able to make optimal resource allocation decisions, the resource mapping entity has to be constantly updated with the status and availability of a heterogeneous set of resources at a global level. It may need to maintain a knowledge-database with information such as network topology, storage capacity, link capacity, link delay, network condition, and predictions of future traffic patterns (possibly based on measurements of current traffic patterns).

For a storage-based QoS infrastructure, the resource mapper will map QoS requirements into storage resources only. It does so by assuming that only best effort transmission service is available, and this service is characterized by some delay distribution on each link. On the other hand, for a unified transmission-storage QoS infrastructure, the resource mapper may map QoS requirements into a combination of

³⁴ Conversely, the installation and use of local caches by the end user (or organization) may be considered a form of receiver-based storage resource reservation, but it is usually performed without explicit involvement of the content owners (senders). Finally, network caching may be performed by the network provider in complete transparency to both senders and receivers, and without the need for resource reservation.

storage and transmission resources. These transmission resources may range from dedicated transmission capacity (e.g., leased lines), QoS services based on intserv, diffserv, to IP “overnet” services that provide single-hop connectivity between specified end points.³⁵

For placement-oriented services, resource mapping is trivial since the exact storage nodes involved are explicitly specified. In fact, we can say that “resource mapping” has already been performed by the service requester prior to resource reservation.

For storage services with deterministic guarantees, resource mapping has to be performed based upon the peak or worst-case resource requirements. The demand distribution of data accesses is irrelevant; the resource mapper simply identifies the set of network nodes at which storage capacity needs to be reserved in order to meet latency and/or other performance requirements for any object requested by any consumer.

For storage services with statistical guarantees, the resource mapper can take into consideration the probability distribution of data accesses when determining the optimal set of network nodes. To the extent that demand for network storage can be characterized as Markovian, it may be possible to apply the effective bandwidth or equivalent capacity concepts from the data transmission domain (Kelly, 1991, Guérin, Ahmadi and Naghshineh, 1991). However, in Chapter 5, we shall show that the problem can be characterized and solved as a weighted k -center problem, as in the location theory literature (Labbé, Peeters and Thisse, 1995).

³⁵ Digital Island, an Internet Service Provider, offers single hop connectivity between major network access points throughout the world by selective provisioning of network capacity. This service is used by online publishers, for example, to achieve performance targets for their information dissemination applications (Rendleman, 1997).

The output of a resource mapping process, i.e., the amount of resources needed to satisfy a particular service, is strongly dependent on the probability distribution of data accesses. Consider two data collections, each with 1,000 objects, each of size 1MB, resulting in a total corpus size of 1GB each. Suppose individual objects in collection #1 are equally likely to be accessed, while those in collection #2 exhibit different degrees of popularity in accordance to Zipf's Law (Zipf, 1949), such that 10% of individual objects account for 90% of all data accesses. The content owner requests, for each of the two collections, a service for 1GB capacity, and a 100ms delay bound to be met by at least 90% of all data accesses. If the resource mapping entity is furnished with the above demand distributions, it will compute the equivalent storage capacity to be 0.9GB and 0.1GB for the two collections respectively.

4.4.3 Admission Control

Because network transmission and storage capacities are finite, not all service requests can be accepted without adversely degrading the performance of the network. Therefore, admission control is needed to reject those requests whose service contracts could not be fulfilled by the resources available at the time.

Admission control occurs in two stages. First, individual resource nodes (network switches or storage nodes) make local decisions as to whether a service request can be accommodated given the current availability of local resources. If all local decisions are positive, then a global check on aggregate requirements (e.g., aggregate delay bound) is performed (if necessary) before the final accept/reject decision is made.

In the case of transmission, admission control occurs along the routing path between sender and receiver (or receivers in the case of multicast). Switching nodes make

local conditional acceptances and forward the request downstream, or send a reject message back to the sender. If a conditional acceptance is made, the switching node is obliged to set aside the requested capacity until the aggregate admission control decision is made, at which point the capacity is either fully committed or returned to the available pool. Therefore, the local admission control decisions have to occur sequentially on a hop-by-hop basis, and are finally followed by the aggregate decision.

In the case of storage, there is no notion of a path within a service request, and so all of the local admission control decisions can occur independently and in parallel. Furthermore, there is no need for an aggregate admission control decision, since there are no end-to-end requirements to be met. Therefore, all that is needed is a central entity to transmit admission control queries to and collect responses from the storage nodes. This role may be played by the resource mapper, or in the case of placement-oriented services, by the service requester itself.

There is clearly a tightly-coupled relationship between admission control and resource mapping. Therefore, it is important to recognize and leverage the possible synergy that may exist between the two entities. When resource utilization level is high, and the likelihood of a service request being rejected by the individual resource nodes is high, the resource mapping and admission control process may be iterated several times before a success is finally encountered. In this situation, it may be appropriate for the resource mapping and admission control functions to switch to a “greedy” algorithm or a quorum-based algorithm.

Both approaches reduce the number of possible iterations by sending admission control queries to more than enough nodes at the first attempt. In the “greedy” algorithm, the resource mapper will provide multiple sets of nodes that can satisfy a particular service request. The sets may or may not have common elements. The admission

controller will send queries to the union of the sets, and declares the request admitted as soon as it receives positive responses from all the nodes of any given set. In the quorum-based algorithm, the resource mapper will provide a set of candidate nodes to which queries are sent. The service request will be declared admitted as soon as a quorum number of nodes returns a positive response.

4.4.4 Service Provision Architecture

Consider a population of service requesters, demanding network resources from a set of storage nodes, operated by one or more service providers. In addition to these three entities, there may also be resource brokers who play the role of intermediary or reseller in the system. The service provision architecture defines the logical organization and placement of the resource mapping and admission control functions among these various entities, and the communication (of resource reservation messages) between them.

One important requirement for this service provision architecture is the support of both performance and placement-oriented services. As we have already seen, performance-oriented and placement-oriented services are very different in nature. The architecture that will not only accommodate, but efficiently handle, both classes of services must exhibit characteristics of openness and flexibility.

For placement-oriented services, resource mapping is trivial since the exact storage nodes involved are explicitly specified. In fact, we can say that "resource mapping" has already been performed by the service requester prior to resource reservation. If all services were of the placement-oriented flavor, a fully distributed architecture would be most appropriate.

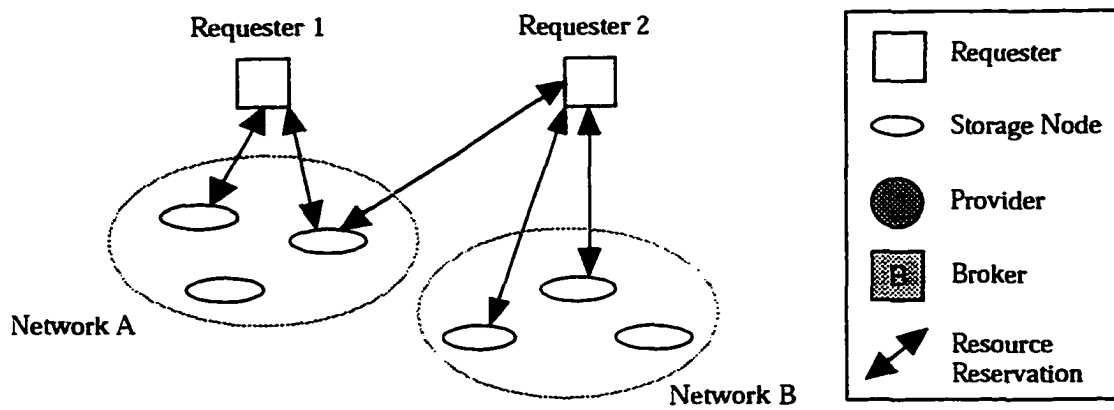
In this fully distributed architecture (Figure 4.2(a)), a service requester makes resource reservation directly with the individual storage nodes. The distributed storage nodes make independent admission control decisions based upon local resource availability. If the local admission control decisions in the aggregate satisfy the service specification, then the service can be established. In this arrangement, local storage nodes have full autonomy over storage allocation, and the service provider does not need to be involved in the process of service provision.³⁶ Unfortunately, a fully distributed architecture would have difficulty supporting performance-oriented services.

In order to support performance-oriented services, some entity in the network must take on the responsibility of resource mapping. In fact, if all services are of the performance-oriented flavor, a fully centralized architecture might be appropriate.

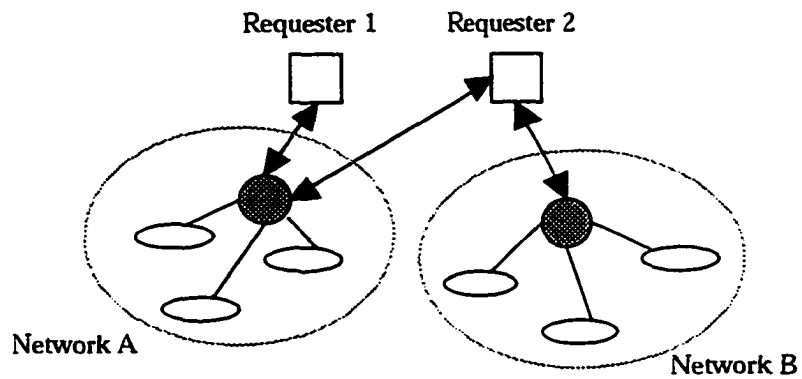
In a fully centralized architecture (Figure 4.2(b)), the service requester would send the reservation request message to the centralized entity (possibly the network storage provider itself). This entity performs resource mapping and queries the storage nodes for resource availability. If the storage resources are placed under centralized control, then resource mapping and admission control can be performed as a single integrated function. Key advantages to this approach include: cost savings through maintaining only a single database of resource availability, and the possibility of scheduling optimizations and statistical multiplexing across multiple services. However, requiring all reservation requests to go through a centralized entity creates two problems: (i) inefficiency for placement-oriented services, and (ii) performance bottleneck and non-scalability.

We propose an open distributed architecture with brokers (Figure 4.2(c)) as the appropriate model for supporting both placement-oriented and performance-oriented

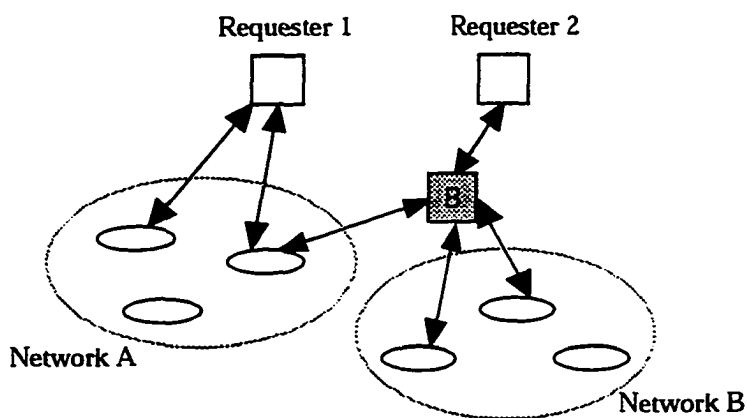
³⁶ The service provider may still be involved in other global functions such as billing, etc.



(a) fully distributed architecture



(b) fully centralized architecture



(c) distributed architecture with brokers

Figure 4.2. Service provision architectural alternatives.

services. In this model, storage nodes continue to offer placement-oriented services, while performance-oriented services are offered as value-added services through the brokers.

Table 4.2. Entities involved in resource mapping and admission control functions in different service provision architectural alternatives.

Architecture Service Class	Function	Entities involved			
		Requester	Broker	Storage Node(s)	Provider
<u>1. Fully Distributed</u>	Resource Mapping Admission Control	x		x	
<u>2. Fully Centralized</u>	Resource Mapping Admission Control [†]				x x
<u>3. Distributed with Brokers</u>					
(a) Placement-oriented (wholesale to broker)	Resource Mapping Admission Control		x	x	
(b) Performance-oriented (retail by broker)	Resource Mapping Admission Control		x x		
(c) Placement-oriented (retail by storage nodes)	Resource Mapping Admission Control	x		x	

The brokers first purchase capacity from the storage nodes, on a wholesale basis, as if they are requesting placement-oriented services for themselves (Table 4.2, class 3(a)). They then resell the capacity on a retail basis to requesters as performance-oriented services (Table 4.2, class 3(b)). In this case, the broker performs resource mapping and admission control (based on its own stockpile of storage resources), just like the provider performs resource mapping and admission control in the fully centralized

[†] In this arrangement, the storage nodes are placed under centralized control, and so the resource mapping and admission control functions may be integrated. Alternatively, if the storage nodes retain their allocative autonomy, then the provider will have to query them for admission control decisions.

architecture. Finally, for those content owners who are interested in placement-oriented services, they will contact the storage nodes directly (Table 4.2, class 3(c)), just as they did in the fully distributed architecture.

In supporting brokerage and resale of network storage resources, the open distributed architecture encourages competition in resource mapping. Indeed, the storage providers can choose to become brokers themselves in this architecture. However, independent brokers can secure resources from multiple storage providers, thus offering services with greater coverage area, higher redundancy, etc. The brokers are in effect adding value by maintaining global state, performing resource mapping, aggregating admission control, and optimizing resource usage, thereby turning placement-oriented services into performance-oriented services.

4.5 Real-Time Resource Management

After the establishment of network storage services, the service provider has to perform real-time resource management in order to meet and enforce all service commitments.

In network transmission, resource management crudely means deciding which packets to transmit next (scheduling management) and which packets to drop (buffer management). The simplest queue discipline is FIFO (first-in first-out), which results in best effort transmission. To accomplish QoS guarantees, a combination of packet scheduling such as fair-weighted queuing (Keshav, 1991) and traffic shaping at the edge of the network (e.g., token bucket with leaky bucket rate control) is necessary (Parekh and Gallager, 1994). This chapter will not deal with resource management in the data transmission context.

In network storage, resource management means deciding which data objects to keep in memory, which objects to purge. The most common replacement policy is LRU (least recently used) and it results in the implementation of best effort caching. To support QoS in network storage, we need to support the coexistence of data objects from both best-effort caching and guaranteed-service replication. Replicated objects have to be kept in memory for the entire duration of their service contract, while cached objects are aged and purged according to some object replacement policy. In addition to the variety of network cache replacement heuristics being proposed (Lorenzetti, Rizzo and Vicisano, 1996, Williams et al., 1996, Cao and Irani, 1997), cache replacement strategies can also include directives from the publisher (HTTP 1.1's *no-cache* pragma), and ad hoc rules for identifying dynamic pages (Inktomi, 1998). The techniques for marking and keeping replicated objects in memory might be adapted from virtual memory management (e.g., page locking) or distributed file system design (e.g., hoarding) (Kistler and Satyanarayanan, 1992). Finally, cache consistency mechanisms and replication update policies have to be put in place, and techniques for accomplishing these are readily available from distributed databases and file systems design.

4.5.1 Local Storage Management

There are several important research questions that have to be addressed with regards to local storage management. First, is there an optimal mix between replicated and cached objects in a network storage node? If so, what is the optimal mix? Alternatively, should a minimum fraction of storage be dedicated to caching? Intuitively, it makes sense not to commit all resources to replication, even though replication is expected to generate higher revenue than caching. A healthy supply of caching capacity will better deal with the burstiness in traffic and minimize the likelihood of thrashing.

4.5.2 Traffic Policing

Another local storage management issue is traffic policing. What happens when the content owner sends content in excess of the reserved amount? The storage manager exercises jurisdiction over this “non-conformant” traffic, and decides whether these objects should be discarded immediately, put into cache space (if available), or replace some existing objects in replication memory. Alternatively, the content owner may be sending an updated version of an object, in which case the stale object has to be identified and replaced.

The concept of *committed information rate* (CIR) from frame relay may be applied here. In data transmission, performance guarantees are provided for traffic transmitted at up to the committed information rate, while traffic in excess of the CIR are delivered as best-effort traffic. This guarantees each sender a minimum share of a link resource, while allowing them to send additional traffic when other senders are idle. An analogous concept of a *committed storage rate* (CSR) may be developed, such that a publisher is guaranteed a minimum fraction of a multi-publisher storage facility, and can store additional objects if free space is available. An alternate service might guarantee a minimum object lifetime before cache swap out. The feasibility of these alternatives will have to be verified through modeling and simulation using cache trace data.

4.5.3 Hierarchical Resource Sharing

Hierarchical resource sharing or dynamic storage allocation also finds its analogy in link-sharing in the network transmission context (Floyd and Jacobson, 1995, Bennet and Zhang, 1997). A content owner may have different classes of objects in its corpus, and

wishes to assign different QoS levels for the different classes. The owner can make separate storage reservations, each with different performance requirements, for the different object classes. Alternatively, it can make a single storage reservation that allows real-time control over the allocation of reserved storage resources to different classes of data objects.

Consider the example of a popular news web-site (Figure 4.3). The size of the entire corpus is 2.5GB, and the publisher classifies the objects into one of three groups. The first group comprises of objects deemed critical by the publisher, such as the homepage and its navigational bars, the headline news articles, and the advertising banners. While its current size is 250MB, the publisher expects the size to fluctuate, but not to exceed 500MB. The bulk of the news content (2GB) makes up the second group. Finally, 250MB of corporate information (e.g., press releases, job openings, mugshots of CEO and VP's) constitute the third group.

The publisher reserves 1GB of storage capacity and specifies the proportion to which storage will be allocated among the three groups. The publisher wants 100% of the group 1 objects to be in memory, even if the size of the group grows to 500MB. Therefore, group 1 is allotted 500MB or 50% of the storage quota. Groups 2 and 3 are then assigned 48% and 2% of the quota respectively.

Since there are currently only 250MB of group 1 objects, all of these objects are guaranteed to be in memory. The extra 250MB of group 1's quota will be proportionately shared (at a ratio of 24:1) between groups 2 and 3. Therefore, group 2 gets $480 + 240 = 720\text{MB}$ of storage and group 3 gets $20 + 10 = 30\text{MB}$ of storage. Should additional objects be added to group 1, storage capacity will be reclaimed from groups 2 and 3. This ensures that group 1 objects are always in memory, up to 500MB. Without this resource sharing scheme, the publisher would have to reserve and dedicate 500MB of

storage capacity to group 1 objects, even when there are less than 500MB of objects most of the time.

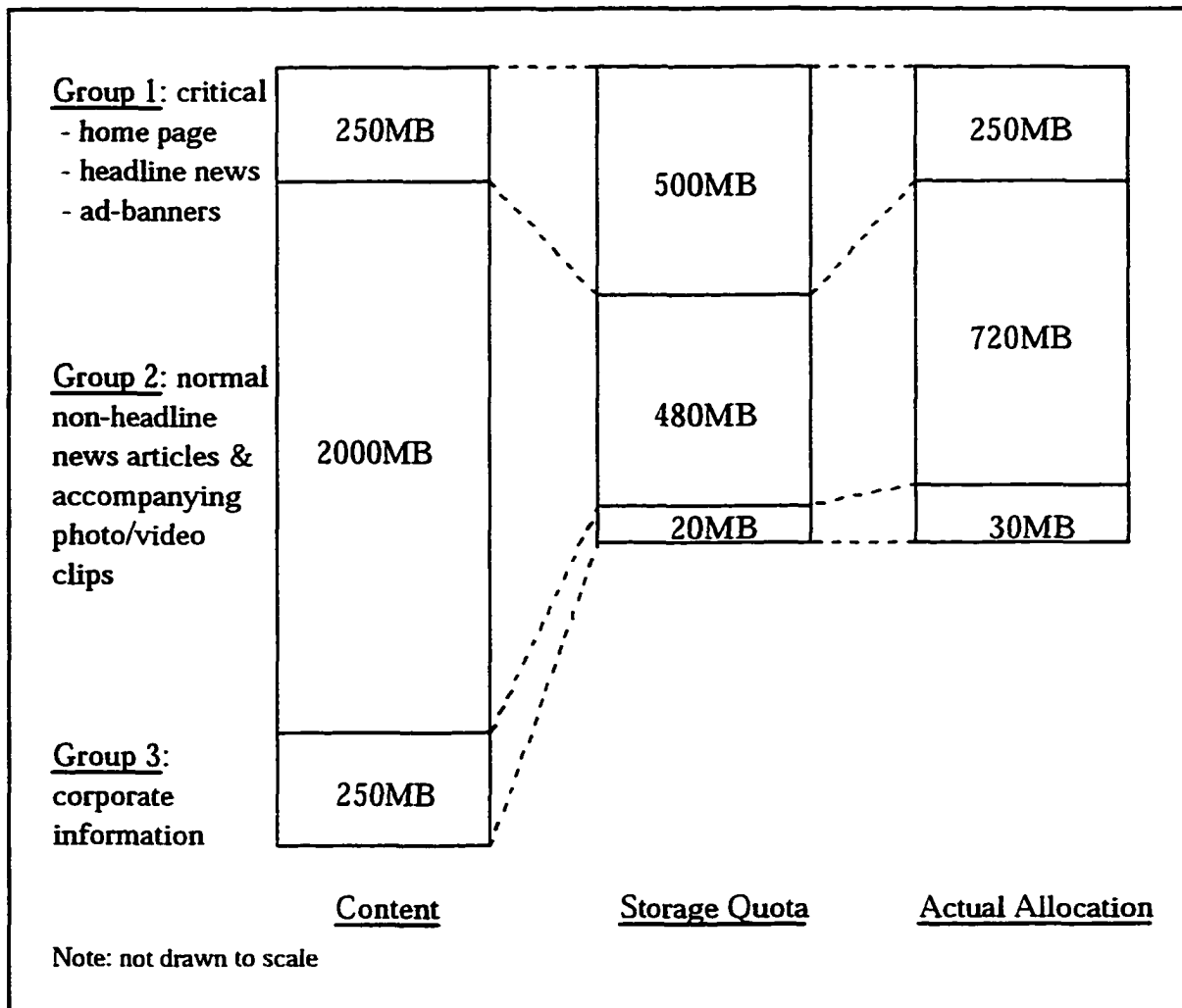


Figure 4.3. Hierarchical resource sharing example.

Using this resource sharing scheme, the publisher can also control the degree of statistical multiplexing to take advantage of reference localities in data access patterns. In the same example, the publisher is able to achieve 100% coverage of group 1 objects (no statistical multiplexing), 36% coverage for group 2 objects, and 12% coverage for group 3 objects. The publisher can increase or decrease the storage quota for groups 2 and 3 to control the respective hit rates.

From this example, it is clear that hierarchical resource sharing is attractive because it gracefully absorbs the "burstiness" in object-class-sizes and facilitates user-controlled statistical multiplexing.

4.5.4 Global Storage Management

While the previous subsections deal with management issues local to the storage nodes, there are also global storage management issues that require study. In the normal operation of the distributed network storage infrastructure, there may be situations that require movement of data objects between storage nodes even after resource mapping and reservation. For example, changes in network status (e.g., network congestion, down nodes or links) may necessitate the movement of objects to maintain the existing service commitments. Alternatively, there may arise opportunities (e.g., termination of existing commitments, addition of new capacity) where data movement can lead to improved resource utility or load balancing. The scheduling of data migration, replication and de-replication constitutes the scope of global storage management (Schill, 1992).

4.6 Additional Mechanisms

There are two additional mechanisms necessary for the distributed network storage infrastructure: (i) resource discovery and (ii) accounting, billing and payment. These mechanisms are not specific to the present vision, and have been the subjects of substantial research efforts elsewhere. Therefore this section will provide brief descriptions and pointers to the appropriate works.

4.6.1 Resource Discovery

A resource discovery mechanism allows a client to locate a data object or to identify the "best" copy among multiple instances of the data object. The selection of the best copy might be based on distance, latency, cost, and/or other criteria, and it may change over time as network conditions change or data objects migrate. The parallel problem (to resource discovery) is location transparency, where clients are automatically pointed to the best available copy without the client knowing or worrying about the actual location of the copy.

There are two components to resource discovery: naming and name resolution. The Internet community has been working on the Uniform Resource Name (URN) framework (Sollins and Masinter, 1994), which facilitates the assignment of a globally unique, persistent identifier to a resource independent of its location. A URN resolver (Sollins, 1998) in turn translates the URN into a uniform resource locator (URL) (Berners-Lee, Masinter and McCahill, 1994) with location information. The location of the nearest object or server is a much researched problem (Guyton and Schwartz, 1995, van Steen, Hauck and Tanenbaum, 1996, Plaxton, Rajaraman and Richa, 1997, Amir, Peterson and Shaw, 1998, van Steen et al., 1998). DNS mapping (Braun and Claffy, 1994, Daniel and Mealling, 1997) and anycasting (Partridge, Mendez and Milliken, 1993) are a few of the proposed mechanisms for tackling this problem.

4.6.2 Accounting, Billing and Payment

The migration to usage-based pricing and the introduction of QoS to the Internet both require the deployment of accounting, billing and payment mechanisms. Efforts are underway at the IETF to define the appropriate accounting architecture (Hirsh, Mills and

Ruth, 1991, Ruth, 1997). Additional work would be needed to identify the storage-specific metrics and include them in the unified Internet accounting framework. Billing systems have been developed and deployed in experiments to study end-user behavior in face of usage-sensitive pricing (Edell, McKeown and Varaiya, 1995, Varaiya, Edell and Chand, 1998). Sirbu (1997) provides a survey of various payment protocols developed to support electronic transfer of funds over the Internet.

From the storage providers' perspective, accounting is necessary for two purposes, namely billing and policing. From the service requesters' perspective, accounting serves two purposes as well, namely performance assurance and hit statistics reporting. Publishers have been having difficulty obtaining accurate hit statistics from web caches, and some have resorted to cache-busting practices in order to keep track of accesses to their content. The availability of hit statistics is therefore critical to gaining the support of the publishers.

4.7 Economics

The distributed network storage infrastructure represents a completely new economy with its unique set of cost structure, market agents, industrial organization and economic rules. Therefore its architects and designers have to be cognizant of the economic implications of different technical design choices, and consistently select the alternatives that promote competition, efficiency and equitability. This section will identify QoS pricing as the key economic mechanism, and discuss issues and implications relating to the industrial organization of this new infrastructure.

4.7.1 QoS Pricing

A priority pricing scheme is essential for all multi-class network services. If different service classes were priced identically, all users would choose the highest-grade service to maximize their utility. Then the situation reverts back to that of a single-class service. Price differentiation between the service classes provides incentive for users to declare their performance requirements. High-demand users would pay a premium for better service quality, while adaptive users are rewarded with having to pay less to use a lower-quality service.

The fundamental concepts of QoS pricing are similar between transmission and storage. In practice, pricing for storage-based services should be more straightforward, since resource usage is localized (at the storage nodes) and therefore easier to quantify.

If one were to take a cost-based approach, network storage services should be priced according to amount of storage capacity reserved. The shadow price is the opportunity cost of reserving a unit of memory space. Under this arrangement, there should be no charge for best-effort caching since no storage capacity is reserved beforehand. This cost-based pricing approach is socially optimal because there is no price distortion.

In reality, storage service providers may choose a demand-based pricing approach for profit maximization. They recognize that users place value in the persistence guarantee, freshness guarantee, access statistics reporting, etc., of a guaranteed-service. To the extent the providers can estimate the demand-curves for the different service classes, they can extract the surplus from the users.

4.7.2 Industrial Organization

4.7.2.1 Distributed Storage Economy

There have been several proposals for organizing distributed storage as a market-based economy (Drexler and Miller, 1988, Ferguson, Nikolaou and Yemini, 1993, Stonebraker et al., 1994, Narayan, Losleben and Cheong, 1995). A property-rental analogy can be used to describe these proposals, with the storage space as rental property, service providers as landlords, consumers as tenants, and service contracts as leases. Using this analogy, we can describe a service contract as a lease agreement between landlord and tenant for the occupation of a rental property for the duration of the lease. Best-effort caching, on the other hand, requires no leases. But the landlord can evict a tenant at any time when a higher-valued tenant is found. The main advantage of this market-based organization of network storage is that resources are allocated efficiently without centralized control. Furthermore, this construct assists the storage providers in framing and answering questions such as "what is the optimal lease duration?" or "how far ahead should I accept reservations?" or "when should storage capacity be added or removed from the market?"

4.7.2.2 Spot Market, Futures Market and Supplemental Insurance

In section 4.3 the notion of services for future storage capacity was introduced. This suggests that a futures market might be established for network storage.

As compared to a spot market, a priority-service market structure which supports forward contracts can be operated at lower cost if the resources in question are perishable, transaction costs are significant, and customers' valuations are stable over time (Chao and Wilson, 1987). Furthermore, this market structure reveals consumer willingness-to-pay

under different service reliability conditions, and this information can be used for capacity planning purposes.

Finally, in the presence of consumer risk-aversion, supplemental insurance provisions may be included in the service contracts. This will enable efficient risk-sharing between the service consumers and providers.

4.7.2.3 Vertical Integration and Component-Based Competition

Network storage is a parallel infrastructure to network transmission. For some applications, the desired performance level may be achieved by reserving a combination of transmission and storage resources. The need to architect the distributed network storage infrastructure such that it can be integrated with the transmission-based QoS infrastructure has been emphasized throughout this chapter. Indeed, it is not unreasonable to expect that economies of scope savings may be realized by the simultaneous provision of network transmission and storage services. So, should network transmission providers (e.g., the ISPs) be allowed to own and operate network storage services?

The answer depends on the assessment of whether network transmission is a competitive or monopolistic market. If network transmission is offered on a competitive basis, then ISPs should be free to compete against other entities in the network storage service arena. On the other hand, if ISPs are assessed to be monopolists in network transmission, then they should not be allowed to make use of their monopolist position to gain unfair advantage in the storage domain.

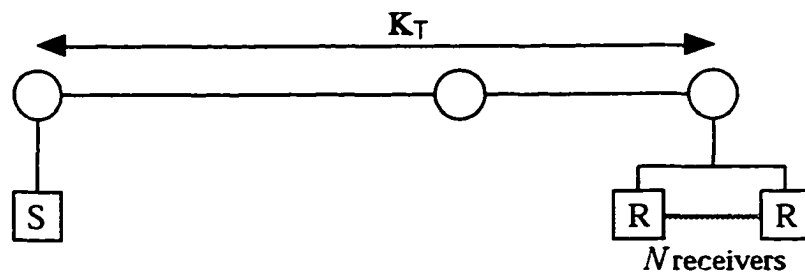
While economists have assumed a monopolistic market structure to facilitate their economic analysis (e.g., Mackie-Mason, Shenker and Varian, 1996), conflicting evidence

of competition and monopoly abound. On the one hand, there are numerous ISPs of varying sizes providing access and transport services at competitive prices. The presence of new entrants such as Qwest and Level 3 suggests that the market does not have high barriers to entry. On the other hand, consolidation in this industry is widely anticipated, and the strategic posturing by the major ISPs with regards to peering agreements seem to point to a future with only a very small number of ISPs able to offer global connectivity.

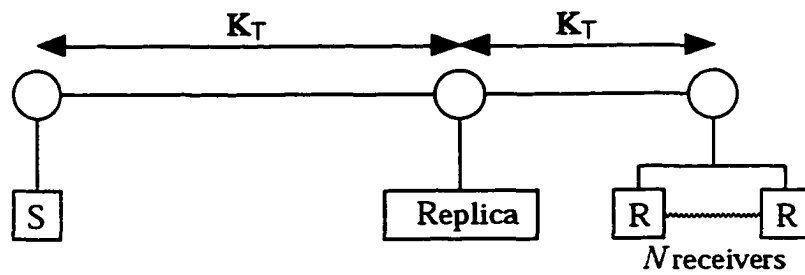
It is also instructive to study the tariff structures of network transmission. Today, network transmission is predominantly priced at a flat rate, with some providers moving to a usage-based pricing scheme. Yet, when people talk about usage-sensitive pricing, they are referring to the number of packets injected into the network, but not how near or far the packets have to travel. Under a distance insensitive pricing regime, no strong business case can be made for an independent third party replication service.

Consider the example in Figure 4.3, where a content owner wishes to deliver a data packet to N receivers. Under scenario (a), the content owner simply transmits N copies of the packet to the receivers, incurring a transmission charge of $N * K_T$ (where K_T is the cost of transmitting one packet, regardless of distance traveled). In scenario (b), the content owner arranges for a copy to be placed at the independent replica server, bringing the object closer to the receivers. But the total transmission charge goes up to $(N+1) * K_T$, not to mention the additional storage costs incurred. In scenario (c), the storage and transmission services are vertically integrated, and the true cost of transmission is internalized. It can be computed as $(1-\alpha) * K_T + N * \alpha * K_T$. As α approaches zero, the transmission cost approaches K_T . To the extent transmission providers can control the pricing of transmission, they can create economies of scope advantage not available to independent storage providers.

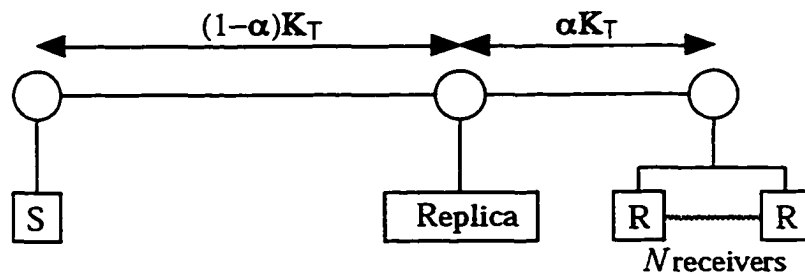
To the extent that transmission is a competitive market, vertical integration of the transmission and storage markets does not present any anti-trust concerns. Component-based competition allows firms to compete, on equal footing, in either one or both of the markets. However, in the absence of competition in the transmission market, regulators will have to weigh the merits of vertical integration (i.e., economies of scope savings) against its costs (i.e., potential for anti-competitive behavior).



(a) no replication



(b) replication at independent 3rd party server



(c) replication at vertically integrated server

Figure 4.4. Example shows vertically integrated storage provider can internalize transmission cost savings not available to an independent storage provider.

4.8 Conclusion

This chapter motivates a distributed network storage infrastructure with quality-of-service guarantees, and describes its technical and economic mechanisms. When fully realized, this infrastructure will support, in one integrated framework, network storage services ranging from best-effort caching to replication with performance guarantees. Content owners can, through the use of standardized protocols, reserve network storage resources to satisfy their application-specific performance requirements. They can specify either the number and/or placement of the replicas, or higher-level performance goals based on access latency, bandwidth usage or data availability. The network storage provider will optimally allocate storage resources to meet the service commitments, using leftover capacity for best-effort caching. Content consumers retrieve the nearest copy of the data object, be it from a replica, cache, or the original source, in a completely transparent manner.

This chapter establishes a QoS framework upon which community discussion on this vision can proceed. It also identifies key research areas and problems that need to be tackled, including those in service specification, resource mapping, admission control, resource reservation, storage management, location transparency, accounting, pricing and industrial organization.

5. Resource Mapping For Distributed Network Storage Services

Resource mapping is one of the fundamental components of a distributed network storage infrastructure with Quality-of-Service guarantees. It translates high-level service specifications (which includes the traffic profile and performance requirements) into low-level resource requirements, such as storage and transmission capacities. This chapter provides a formal model of resource mapping upon which the process can be applied and analyzed.

5.1 Mathematical Model for Resource Mapping and Admission Control

Consider a network $G(V,E)$ with vertices V and edges E . Each of the vertices is a demand point, i.e., it has one or more content consumers that issue requests for data objects. Let $S \subseteq V$ be the set of supply points, i.e., network nodes where replication

servers are installed, with storage capacity available for replication and placement of data objects. The number, location and capacity of these replication servers are determined by a network capacity planning process, and are therefore treated as exogenous parameters to the resource mapping problem. While the fixed cost associated with setting up these servers is assumed sunk for the purpose of resource mapping, it will be properly accounted for when we tackle the capacity planning problem in Section 5.3. Presently, we shall assume that each storage node $s_j \in S$ has a variable storage cost of $c_S(j)$ per unit of storage per unit time.

Given the lengths of the links, it is possible to compute the shortest distance between a demand point $v_i \in V$ and a supply point $s_j \in S$ as $d(i,j)$. This distance may be measured in terms of hop count or other distance metric. Alternatively, $d(i,j)$ can represent the network delay between nodes v_i and s_j . If we consider the effects of network congestion, i.e., link delay may vary according to changing traffic load conditions, then $d(i,j)$ becomes a random variable. In this case an expected value of network delay may become appropriate. We assume that the choice of replica sites does not affect the aggregate traffic flow pattern, and therefore has no impact on the link delays, i.e., network storage providers are “delay-takers” rather than “delay-makers”.

5.1.1 Traffic Profile

Consider a storage service request for a collection of objects Q , starting at time T_s , for a duration of T_d . Each object $q_k \in Q$ is of size $b(k)$ octets, so the total size of the corpus is

$$B_{\text{corpus}} = \sum_{q \in Q} b(k). \quad (5.1)$$

We define $g(i,k)$ as the conditional probability that object q_k is requested by some content consumer at vertex v_i given that there is an object request. The marginal probability distribution functions (p.d.f.'s) across data objects in the collection and across network nodes are:

$$g_q(k) = \sum_{v_i \in V} g(i,k) \quad (5.2)$$

and

$$g_v(i) = \sum_{q \in Q} g(i,k) \quad (5.3)$$

respectively. The joint and marginal p.d.f.'s are such that

$$\sum_{v_i \in V} \sum_{q \in Q} g(i,k) = \sum_{q \in Q} g_q(k) = \sum_{v_i \in V} g_v(i) = 1. \quad (5.4)$$

If no *a-priori* information is available for the demand distribution, then it should be assumed that

$$g(i,k) = \frac{1}{|V| \cdot |Q|} \forall i,k. \quad (5.5)$$

Let λ be the total number of requests for objects in collection Q in the network G per unit time. Then the expected number of requests for object q_k at node v_i within a specific time interval T_d is equal to the product of $g(i,k)$, λ and T_d . Without loss of generality we can normalize T_d to one.

5.1.2 Performance Requirements

In section 4.2, we note that storage service requests can specify performance requirements based on delay/distance, availability or other performance measures. From an end-to-end perspective, network delay and server processing delay are two delay components experienced by the user. Network delay is that experienced by data packets as they are delivered from storage server to end client. It is dependent on the network distance between source and destination, the transmission capacity of the links and routers, and the traffic load. Server processing delay, on the other hand, is dependent on the processing capacity, queuing discipline, and the arrival rate of data requests at the storage servers. In this model, we focus on network distance, and ignore effects of heterogeneous transmission capacity, server processing capacity, and changing traffic conditions.

Different resource mapping functions are required for services with different delay requirements. We will describe the mapping problem for each of the following four classes of services:

- maximum (worst case) delay bound: $D_{\max} \leq \tau_{\max}$
- average delay bound: $D_{\text{avg}} \leq \tau_{\text{avg}}$
- average and worst case delay bounds: $D_{\text{avg}} \leq \tau_{\text{avg}}$ and $D_{\max} \leq \tau_{\max}$
- stochastic guarantee: $\text{Probability}[d > \tau_{\text{threshold}}] \leq \epsilon$

5.1.3 Resource Mapping

The resource mapper needs to determine the minimal set of storage servers $X_h \subseteq S$ needed to satisfy the traffic profile and performance requirement of the request.

5.1.3.1 Service with Worst Case Delay Bound

For this service, the set of storage servers X_h must be chosen such that the delay bound for data access is met for requests from any demand point $v_i \in V$ for any object $q_k \in Q$. This implies that all storage nodes in X_h must maintain a full replication of the collection Q . Therefore, the amount of storage capacity to be reserved at each node $x \in X_h$ is $B(x) = B_{corpus}$.

If $d(i, x)$ is the shortest-path distance between vertex v_i and storage server x , $x \in X_h$, then the distance from v_i to the nearest storage server is

$$d(i, X_h) = \min_{x \in X_h} d(i, x). \quad (5.6)$$

It follows that the worst-case distance between any demand point in the network and its closest server is

$$D_{\max}(X_h) = \max_{i \in V} d(i, X_h). \quad (5.7)$$

Then the resource mapping problem can be expressed as the inverse of the k -center problem (Hakimi, 1964, Hakimi, 1965):

$$\min \sum_{x \in X_h} B(x) \cdot cs(x) \quad (5.8)$$

subject to

$$D_{\max}(X_h) \leq \tau_{\max}; \quad (5.8a)$$

$$X_h \subseteq S. \quad (5.8b)$$

If all storage nodes have identical variable costs c_s , then the problem can be simplified to the minimization of the required storage capacity:

$$RSC = \sum_{x \in X_h} B(x). \quad (5.9)$$

But since $B(x)$ is equal to B_{corpus} for $x \in X_h$, we can further simplify the problem to minimizing the number of replicas needed:

$$h_\tau = \min\{h: X_h \subseteq S; D_{\max}(X_h) \leq \tau_{\max}; h \geq 0 \text{ and integer}\}. \quad (5.10)$$

Kariv and Hakimi (1979) showed that the k -center problem is *NP*-hard, even for simple networks. Algorithms to tackle the problem are presented in (Halpern and Maimon, 1982, Labbé, Peeters and Thisse, 1995).

5.1.3.2 Service with Average Delay Bound

Consider a service which specifies that the average delay of data accesses not exceed τ_{avg} . In this case, based on the demand distribution $g(i,k)$, the resource mapper has to determine the optimal set of storage nodes X_h , and the optimal subset of objects Q_x to be replicated at each node $x \in X_h$, such that the delay requirement is satisfied.

Let $X_k \subseteq X_h$ be the set of storage servers that will keep a copy of object q_k . We can specify the distance from node v_i to the nearest storage server $x \in X_k$ as $d(i, X_k)$. Then the average delay can be computed as:

$$D_{\text{avg}}(X_h) = \sum_{v \in V} \sum_{q \in Q} g(i, k) \cdot d(i, X_k). \quad (5.11)$$

At each storage server $x \in X_h$, it will have a subset Q_x of the collection Q . The amount of storage capacity required at node x can be computed as

$$B(x) = \sum_{q \in Q_x} b(k). \quad (5.12)$$

The mapping problem can be expressed as a variant to the inverse k -median problem:

$$\min \sum_{x \in X_h} B(x) \cdot cs(x) \quad (5.13)$$

subject to

$$D_{\text{avg}}(X_h) \leq \tau_{\text{avg}}; \quad (5.13a)$$

$$X_h \subseteq S. \quad (5.13b)$$

Again, if all storage nodes have identical variable costs c_s , then the problem can be simplified to the minimization of the required storage capacity as stated in equation (5.9).

Intuitively, we expect that the replicas are placed closer to the nodes with the largest numbers of data requests. Furthermore, we may also expect to find that the most popular objects in the collection are most widely replicated. However, these intuitions are not always true. If the demand distribution $g(i, k)$ is not independent in i and k , and there is high correlation between popularity and nodes, then the demand for a highly popular object may originate from a limited geographic area. In this case, a few local

copies (with sufficient server processing capacity) may suffice in servicing most of the data requests.

For those applications where partial replication of the collection is not possible, we need to impose the additional constraint: $X_k = X_h$ for all k , or equivalently, $Q_x = Q$ for all $x \in X_h$. This implies that there will be equal number of copies of each of the individual objects, regardless of their difference in access frequency.

5.1.3.3 Service with Average and Maximum Delay Bounds

A service may specify that the average delay of data accesses not exceed τ_{avg} , and further stipulate a maximum delay bound of τ_{max} . The mapping problem can be stated as

$$\min \sum_{x \in X_k} B(x) \cdot cs(x) \quad (5.14)$$

subject to

$$D_{\text{avg}}(X_h) \leq \tau_{\text{avg}} ; \quad (5.14a)$$

$$D_{\text{max}}(X_h) \leq \tau_{\text{max}} ; \quad (5.14b)$$

$$X_h \subseteq S. \quad (5.14c)$$

5.1.3.4 Service with Stochastic Guarantees

Finally, the mapping problem for a service with stochastic guarantees on delay bounds may be expressed as

$$\min \sum_{x \in X_h} B(x) \cdot cs(x) \quad (5.15)$$

subject to

$$\sum_{v_i \in V} \sum_{q_k \in Q} g(i, k) \Big|_{d(i, X, t) > \tau_{\text{threshold}}} \leq \epsilon ; \quad (5.15a)$$

$$X_h \subseteq S. \quad (5.15b)$$

5.1.4 Admission Control

Each storage node $s_j \in S$ has total storage capacity $TSC(j, t)$ and committed storage capacity $B_0(j, t)$ at time t . For each $x \in X_h$, a local admission control decision is made to accept storage request if, for $T_s \leq t \leq (T_s + T_d)$,

$$B(x) + B_0(x, t) \leq TSC(x, t). \quad (5.16)$$

In the event that one or more of the storage nodes in X_h return a rejection, the resource mapping process may be repeated. These nodes, however, will have to be excluded from the candidate pool for future iterations of the mapping process for the same request.

5.2 Resource Mapping for ARPANET

As an illustrative exercise, we will consider the resource mapping problem for the early ARPANET, a network derived from the real world (Table 5.1 and Figure 5.1).

Given the modest size of the network, we can apply the simple enumeration (exhaustive search) technique to the mapping problem. This allows us to explore the various different facets and dimensions of resource mapping.

Table 5.1. ARPANET Statistics.

Network	ARPANET
Number of nodes	47
Number of links	68
Average node degree	2.89
Network diameter (hops)	9

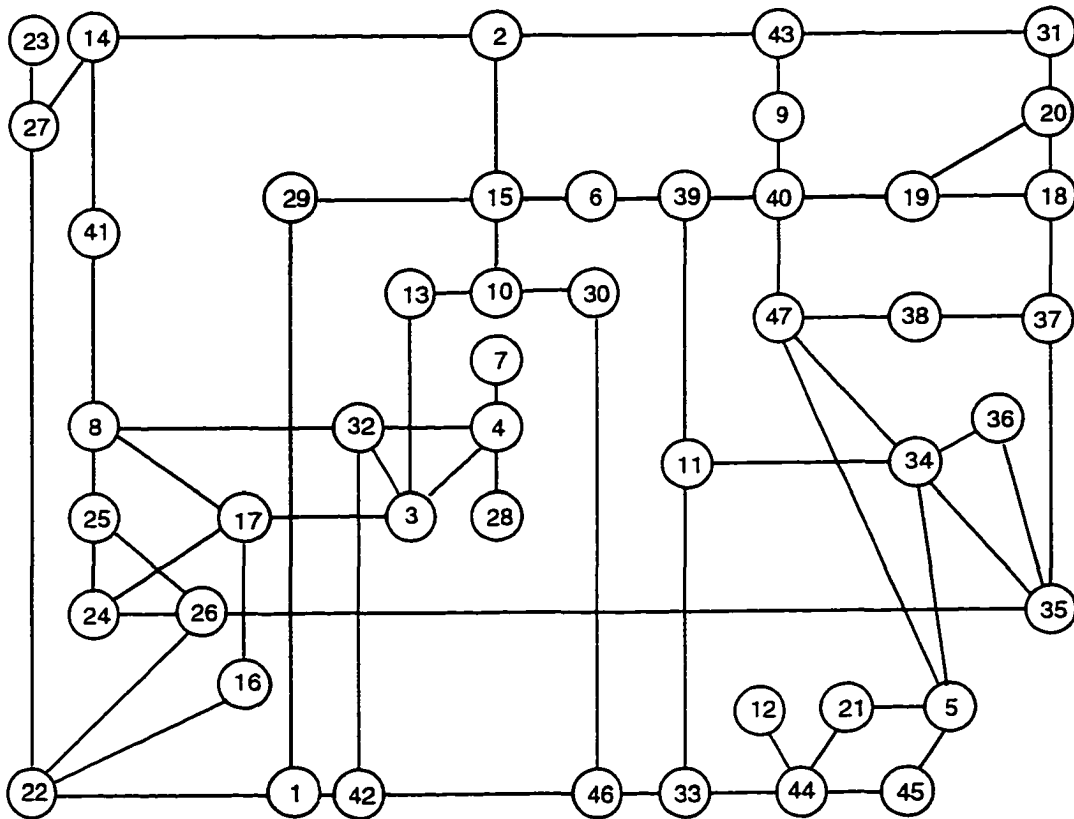


Figure 5.1. Network topology of early ARPANET.

5.2.1 Base Case: Uniform Demand Distribution, Unconstrained Replica Locations

The simplest case is to consider the resource mapping of a collection Q with uniform demand distribution across space and objects, i.e., $g(i,k)$ is described by equation (5.5). Additionally, there is no constraint on the geographic placement of the replicas, i.e., $S = V$ and replicas may be placed at any node v_i in the network. Finally, all nodes are assumed to have identical per-unit storage costs. The objective is to solve the cost-minimization problems as stated in (5.8) and (5.13) for mapping services with maximum and average delay bounds, respectively.

Figure 5.2 shows the results for both the maximum and average delay bound problems. The discontinuities are a direct result of the fact that only integer numbers of replicas are possible. As expected, a service with a smaller delay bound will require a

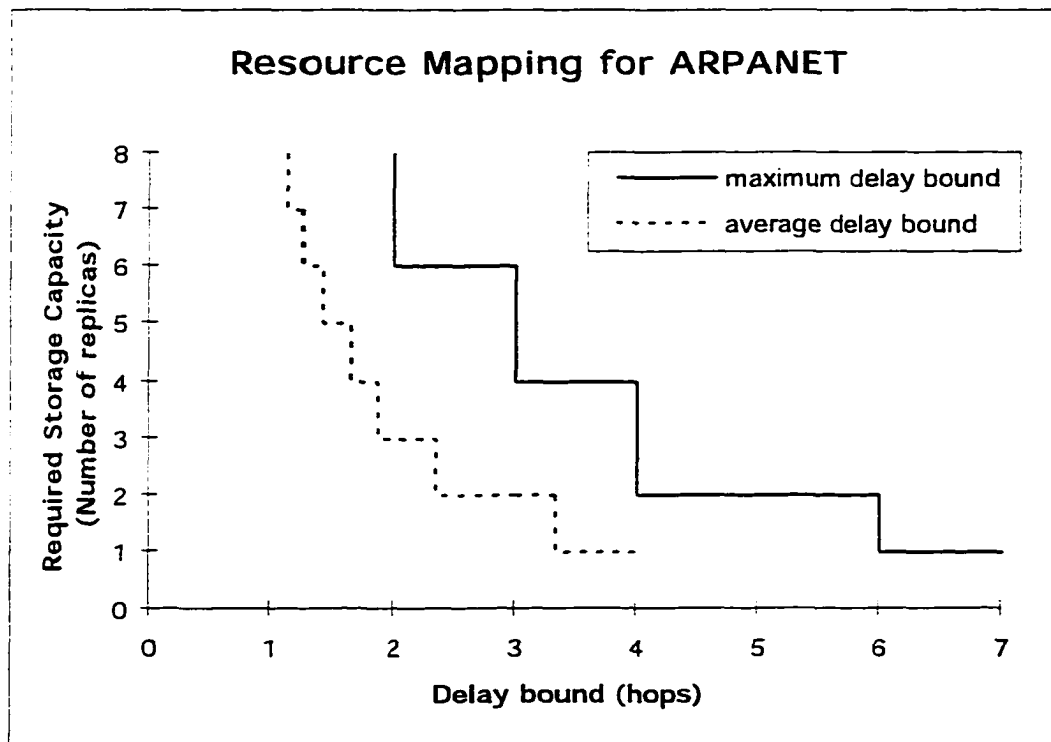


Figure 5.2. Resource Mapping for ARPANET.

larger number of replicas. For example, a service with $\tau_{max} = 4$ hops requires two replicas, whereas a service with $\tau_{max} = 2$ hops will require six replicas. From the plot we also observe that given the number of replicas, the achieved average delay bound is always lower than the achieved maximum delay bound.

Table 5.2 shows additional results for the average delay bound mapping problem. Specifically, it shows the delay reductions (in hops) achieved by each additional replica, and the optimal replica locations. It is important to point out that moving from h replicas to $h+1$ replicas does not involve the mere addition of a new replica site. Instead, the optimal locations of the $h+1$ replicas can be completely different from those of the h replicas. For example, the replica at node 26 is not retained, but replaced by those at nodes 8 and 47, when we move from an $h=1$ to an $h=2$ solution. On the other hand, we see that only 15 out of the 47 nodes in the network are potential replica sites for solutions with up to seven replicas.

Table 5.2. Resource Mapping for Service with Average Delay Bound.

# of replicas (h)	Avg. Delay (D_{avg})	Delay Reduction (ΔD_{avg})	Replica Locations (X_h)
1	3.32		26
2	2.34	0.98	08,47
3	1.87	0.47	02,03,34
4	1.64	0.23	02,03,33,35
5	1.40	0.23	15,26,32,40,44
6	1.23	0.17	15,19,22,32,24,44
7	1.13	0.11	04,08,15,19,22,34,44

5.2.2 Non-Uniform Spatial Demand Distribution

Any *a-priori* knowledge of the spatial distribution of object accesses may be leveraged to improve the utilization of storage resources for services with average delay bounds. An example of a non-uniform spatial distribution may be:

$$g_t(i) = \begin{cases} 10 \cdot C_i & i = 1, \dots, 6 \\ C_i & i = 7, \dots, 47 \end{cases} \quad (5.17)$$

where C_i is a constant such that condition (5.4) is satisfied. This means that nodes 1 through 6 each experience ten times more requests than nodes 7 through 47. Figure 5.3 shows the improvements in average delay for this distribution over a uniform spatial distribution. Specifically, a service with $\tau_{avg} = 1.5$ hops will only require three replicas, rather than five replicas in the uniform spatial distribution case.

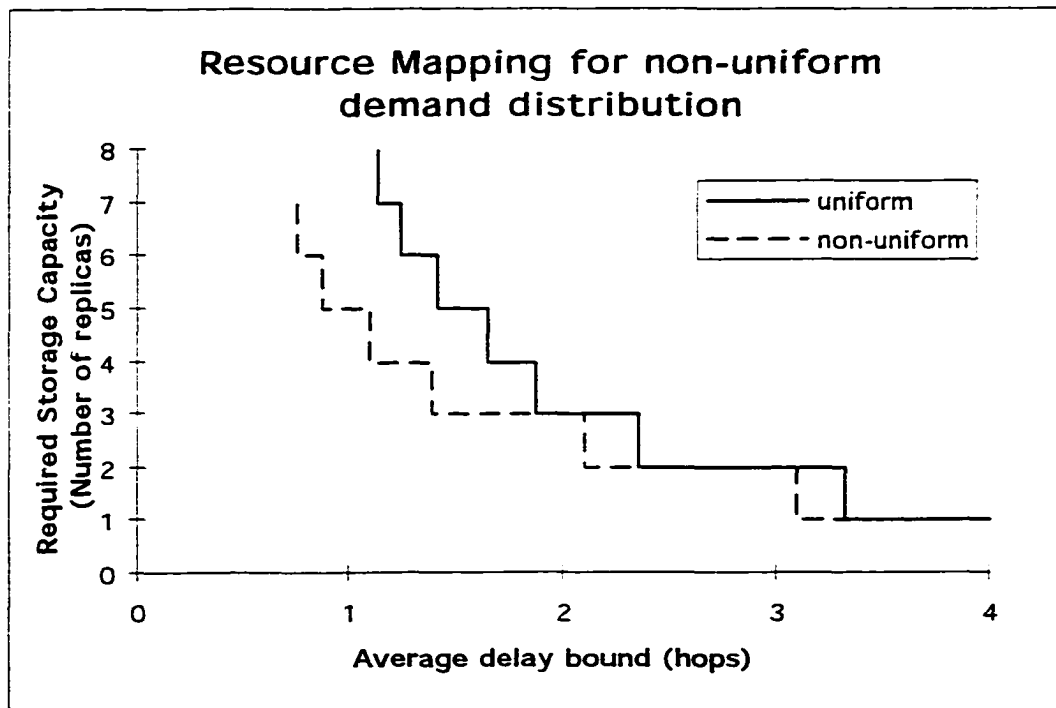


Figure 5.3. Resource mapping for non-uniform spatial distribution.

5.2.3 Partial Replication of Multi-Object Collection

A multi-object collection may have a non-uniform demand distribution $g_q(k)$ across its individual objects, i.e., some objects in the collection are more frequently accessed than others. If the collection owner can convey this distribution function to the resource mapper, the latter can leverage this information to improve the resource utilization for services with average delay bounds. Specifically, by allowing partial replication of the collection, the resource mapper can independently determine the optimal number of copies of each individual object. In particular, if $g(i,k)$ is independent across i and k , then there will be more copies of the more frequently accessed objects.

Consider a multi-object collection Q , where all objects $q_k \in Q$ are of identical size.³⁷ The collection has a uniform spatial demand distribution, i.e.,

$$g_v(i) = \frac{1}{\|V\|} \forall v_i \in V. \quad (5.18)$$

However, the collection has a non-uniform demand distribution across its objects. Specifically, the object access pattern obeys Zipf's distribution (Zipf, 1949):

$$g_q(k) = \frac{C_2}{k} \quad (5.19)$$

where C_2 is a constant such that condition (5.4) is satisfied. Figure 5.4 shows, for a multi-object collection, the efficiency gains of a mapping solution based on partial

³⁷ The mapping problem becomes more complex if the objects are of different sizes, but only slight modifications are required of the solution method presented in Appendix 5 to take object sizes into account. It is interesting to note that the mapping algorithm will favor smaller objects over larger ones unless the latencies are measured on a per-byte basis (rather than a per-object basis).

replication over one based on full replication. The full replication solution is constrained in two ways: (i) the addition of each new full replica necessarily means the addition of $\|Q\|$ object-copies, (ii) there has to be equal number of copies of each object.

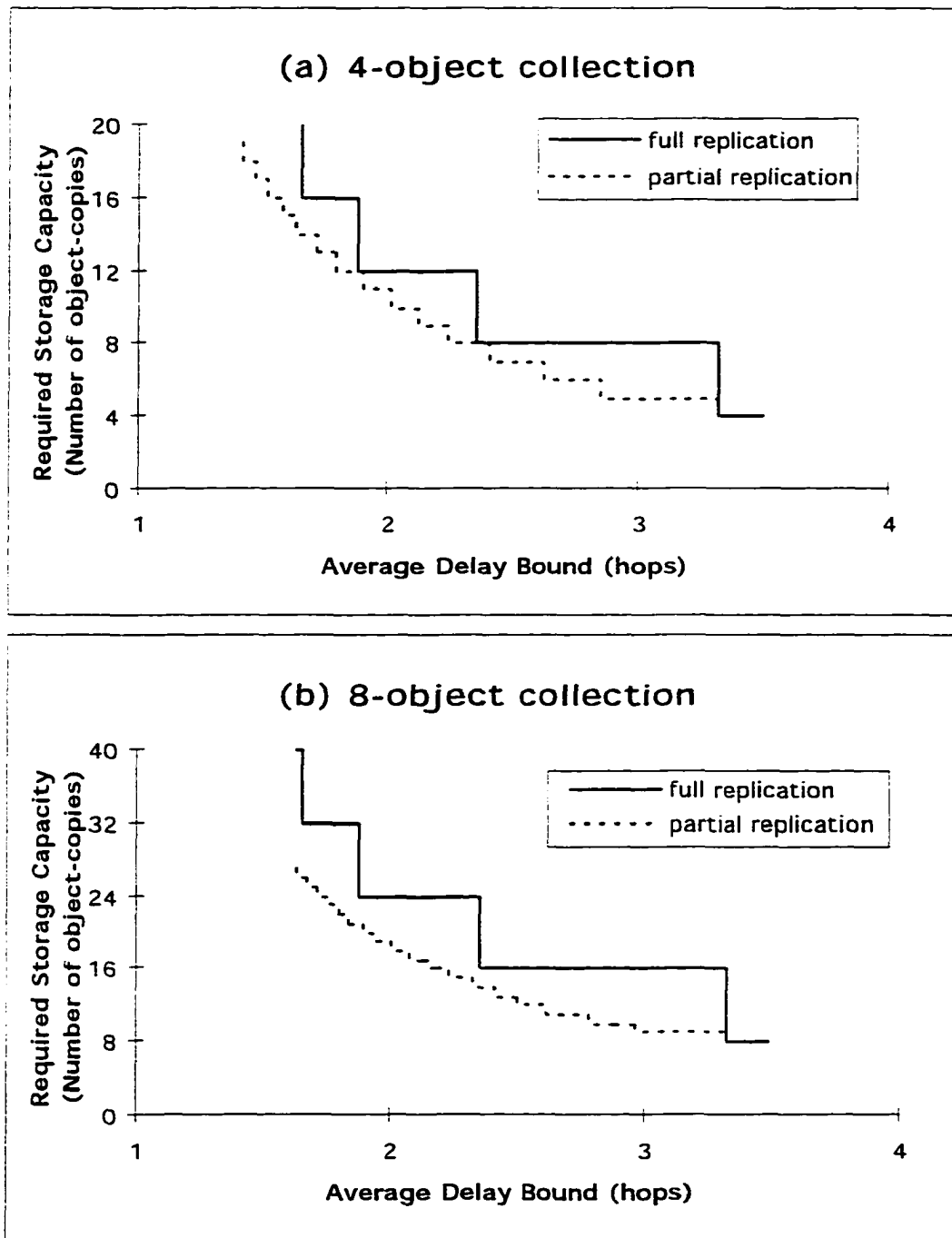


Figure 5.4. Resource mapping for multi-object collections with non-uniform object distribution using full or partial replication.

Let us consider a service for a four-object collection with τ_{avg} of 2.30 hops. The mapping solution based on full replication will require three full replicas (of four objects each, resulting in a total of twelve object copies) at nodes 2, 3, and 34 (see Table 5.2). On the other hand, the mapping solution based on partial replication will only require eight object copies (see Appendix 5 for solution method and full results). The eight object copies include: three copies of q_1 , at nodes 2, 3, and 34; two copies each of q_2 and q_3 , at nodes 8 and 47; and one copy of q_4 at node 26. For purposes of admission control, this translates into $B(2) = B(3) = B(34) = b(q_1)$, and $B(8) = B(47) = b(q_2) + b(q_3)$, $B(26) = b(q_4)$, and zero otherwise.

Figure 5.5 shows, for different sizes of collection Q , the resource mapping solution based on partial replication.

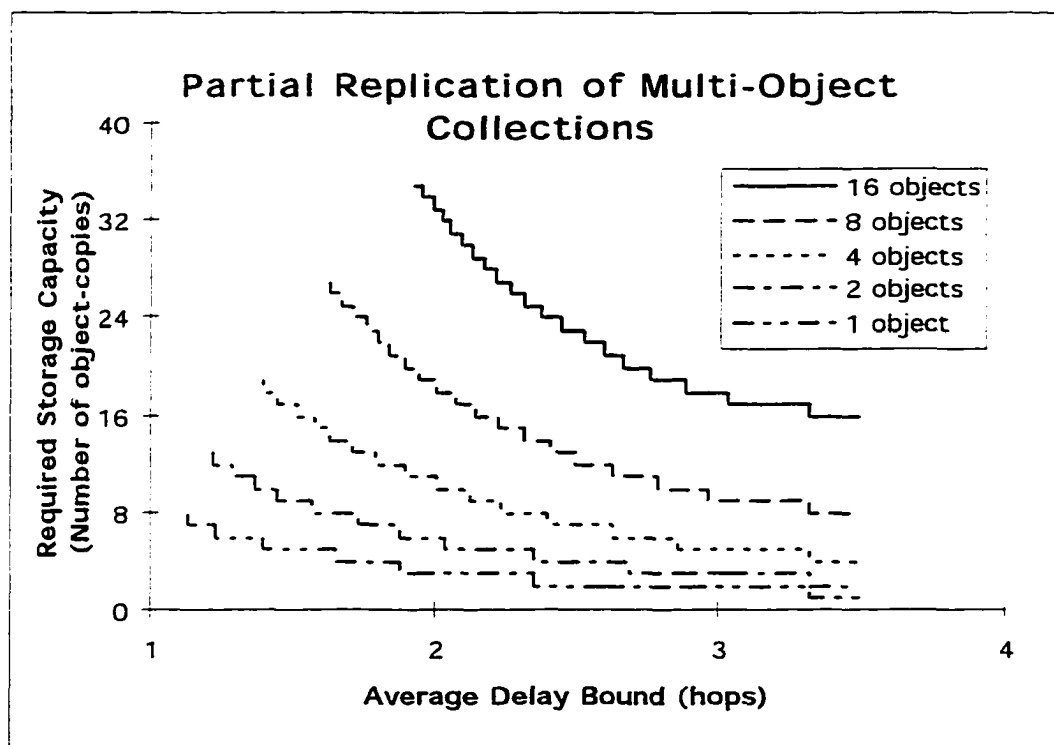


Figure 5.5. Resource mapping for multi-object collections with non-uniform object distribution using partial replication.

5.3 Network Storage Capacity Planning Problem

So far we have concentrated on the resource mapping problem (as opposed to the long-term capacity planning problem), treating the fixed cost of setting up a replication server as sunk, and focusing on the marginal cost of storage. This implies that there are no economies of scale in network storage costs, and the cost of reserving 1GB of storage at each of ten nodes is equal to the cost of reserving 10GB of storage at a single node.

In the long run, however, the network storage service provider does have to take the fixed cost into consideration. If this cost were negligible, the network storage service provider may choose to install replication servers at every network node, i.e., $S = V$, and the resource mapper will have the freedom to place replicas at any location within the network. On the other hand, if the fixed cost is substantial, the network storage service provider may want to install replication servers at only a subset of all nodes in the network, i.e., $S \subset V$.

The long term network storage capacity planning problem is really an extension to the short term resource mapping problem. Whereas we were interested in determining the optimal number and placement of the object replicas, i.e., h and X_h , in the resource mapping problem, we are now concerned with the optimal number, placement and capacity of the replication servers, i.e., $\|S\|$, S and $TSC(s_j) \forall s_j \in S$. Furthermore, the cost function to be minimized now includes a fixed cost c_0 that is incurred each time a new replication server is installed. For a network storage provider who wishes to achieve an average delay bound τ_{avg} for its aggregate traffic $Q_{aggregate}$, the network storage capacity planning problem can be stated as:

$$\min \sum_{s_j \in S} c_0(s_j) + TSC(s_j) \cdot c(s_j) \quad (5.20)$$

subject to

$$D_{\text{avg}}(\mathcal{S}) \leq \tau_{\text{avg}} ; \quad (5.20a)$$

$$\mathcal{S} \subseteq V. \quad (5.20b)$$

As a starting point, it appears reasonable for the provider to determine the optimal replication server set based upon the demand patterns aggregated across all collections in its target market. If the number of collections is large enough, and the traffic patterns across different collections are not strongly correlated, then the rate of change of the aggregate $g(i,k)$ should not necessitate frequent and rapid changes in \mathcal{S} .

We propose three heuristic solutions to the capacity planning problem, namely full replication, greedy partial replication and conservative partial replication. In the full replication strategy, the network storage provider installs $\|\mathcal{S}\|$ replication servers, and each server has the same amount of storage capacity, i.e., $TSC(s_j) = \|Q_{\text{aggregate}}\|$ for all s_j . Both of the partial replication strategies allow different total storage capacities to be installed at different servers, and $TSC(s_j) \leq \|Q_{\text{aggregate}}\|$. In the greedy strategy, the goal is to minimize the global storage capacity (summation of $TSC(s_j)$ across all replication servers), regardless of the number of replication servers needed. In the conservative strategy, the primary goal is to minimize the number of replication servers, and the secondary goal is to minimize global storage capacity. Note that the different heuristics may yield different solutions of $\|\mathcal{S}\|$ and global storage capacity for the same delay target.

These heuristics are evaluated using an ARPANET example. Assume that the aggregate $g(i,k)$ has a uniform spatial distribution across nodes v_i and a Zipfian distribution across objects q_k (i.e., similar to that in Section 5.2.3). Figure 5.6 shows, for different fixed costs c_0 (relative to $c_{\mathcal{S}}$), the total storage cost incurred by the different strategies.

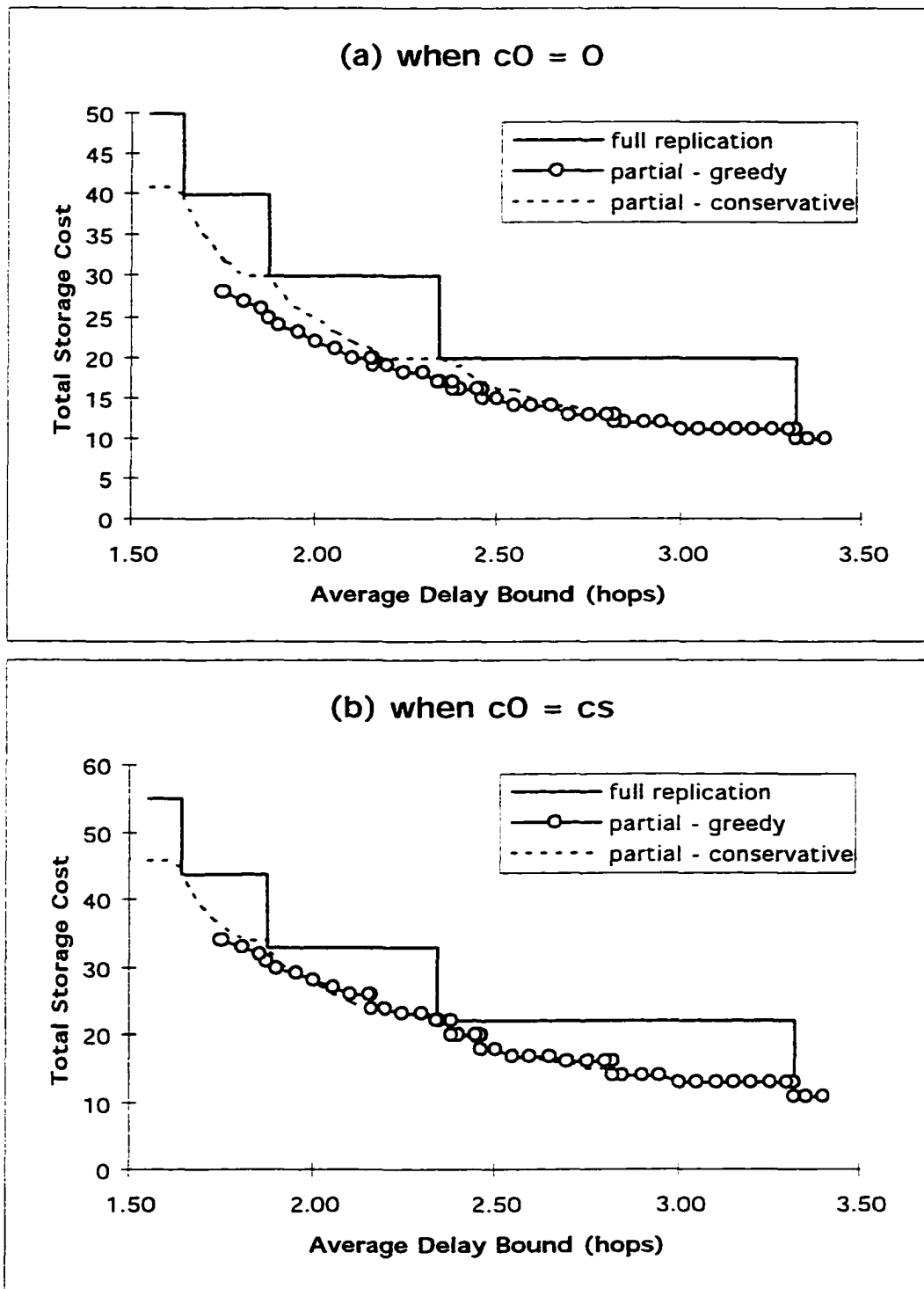


Figure 5.6. Network capacity planning problem - different replication strategies may realize the lowest cost solution depending on the relative magnitudes of fixed (c_0) and variable (c_s) costs.

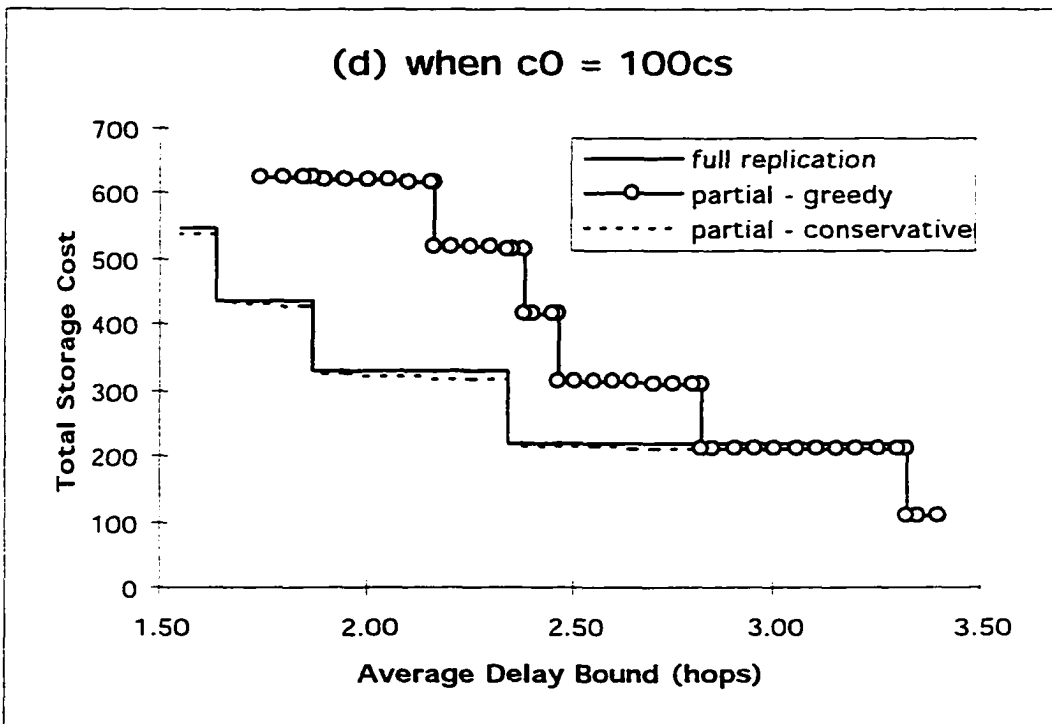
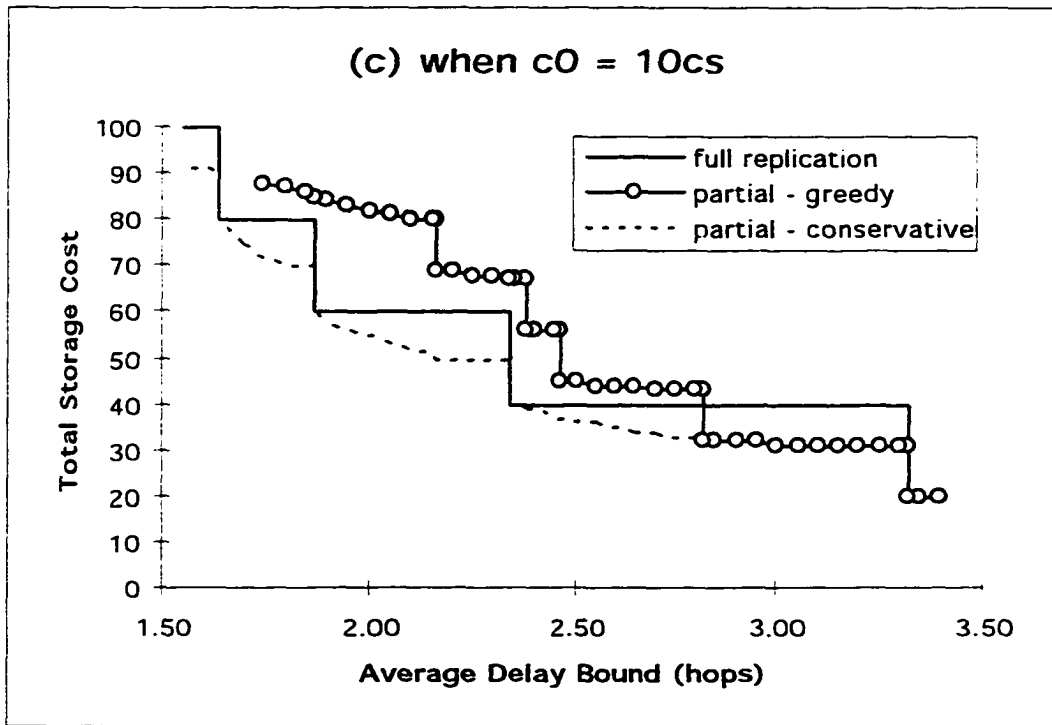


Figure 5.6. Network capacity planning problem - continued.

When the fixed cost of installing a new replication server is zero, the greedy partial replication strategy will realize the lowest cost solution among the three approaches (Figure 5.6(a)). In fact this solution is also the optimal solution to the planning problem when $c_0 = 0$. The similarity between Figure 5.6(a) and Figure 5.4 reminds us that this solution is basically that of the short-term resource mapping problem, where fixed cost was assumed sunk.

When the fixed cost of installing a replication server is non-zero, the conservative strategy becomes a more attractive heuristic, since it attempts to minimize the number of servers before minimizing global storage capacity. This strategy will yield a solution that is reasonably close to, if not equal to the optimal solution, especially in high fixed cost conditions. Finally, we note that at very high fixed costs, the variable cost of storage becomes negligible, and the full replication strategy will perform almost as well as the conservative partial replication strategy (Figure 5.6(d)).

5.3.1 Resource Mapping with Constrained Replication Server Sites

In the previous section we stated the network capacity planning problem and proposed solution strategies based upon the demand distribution of the traffic in its aggregate. However, each individual collection may have its collection-specific $g(i,k)$ similar to or different from the aggregate $g(i,k)$. How will the choice of the constrained replication server set S affect the mapping efficiency of specific collections?

Consider the scenario where $c_0/c_S = 100$, and the network storage provider chooses to maintain five replication servers to achieve an aggregate average delay bound of 1.50 hops (Figure 5.6(d)). The results from Table 5.2 (reproduced as the first three columns of

Table 5.3(a)) indicate the optimal locations of the servers are at nodes 15, 26, 32, 40 and 44.

With this new constraint of $S = \{15,26,32,40,44\}$, it is possible to perform resource mapping for services with spatial demand distributions similar to or different from the aggregate distribution. The right three columns of Table 5.3(a) summarize the mapping result for a collection Q_1 that has a uniform spatial demand distribution as described by (5.18), i.e., the collection-specific distribution is identical to the aggregate distribution. We observe that the $h=1$ and $h=5$ solutions are identical to the unconstrained scenario, and therefore incur no penalty in realized average delay. On the other hand, mapping solutions with two to four replicas will incur a delay penalty of between two and five percent. It is important to note, further, that a small delay may translate into a significant storage penalty. For example, a service with $\tau_{avg} = 1.90$ hops will require three replicas ($X_h = \{2,3,34\}$) in the unconstrained case, four replicas ($X_h = \{15,26,32,40\}$) in the constrained case, representing a 33% increase in storage capacity requirement.

Now consider a second collection Q_2 whose spatial demand distribution is not uniform, but as described by (5.17). The collection-specific demand distribution is now different from the aggregate, and this results in significant mapping inefficiencies as shown in Table 5.3(b). For example, an $h=5$ solution with a constrained S will incur a 57% delay penalty over the unconstrained case. Specifically, a service with $\tau_{avg} = 1.50$ hops will require three replicas ($X_h = \{3,5,15\}$) in the unconstrained case, five replicas ($X_h = \{15,26,32,40,44\}$) in the constrained case. This translates into a 67% increase in storage capacity requirement.

Table 5.3. Comparing mapping efficiencies for constrained versus unconstrained replication server sites.

Number of replicas	Unconstrained: $S = V$		Constrained: $S = \{15,26,32,40,44\}$		
	Replica Locations (X_h)	Average Delay	Replica Locations (X_h)	Average Delay	% penalty in delay
1	26	3.32	26	3.32	0%
2	08,47	2.34	32,40	2.38	2%
3	02,03,34	1.87	26,32,40	1.96	5%
4	02,03,33,35	1.64	15,26,32,40	1.70	4%
5	15,26,32,40,44	1.40	15,26,32,40,44	1.40	0%

(a) collection Q_1 has uniform spatial demand distribution as described by (5.18)

Number of replicas	Unconstrained: $S = V$		Constrained: $S = \{15,26,32,40,44\}$		
	Replica Locations (X_h)	Average Delay	Replica Locations (X_h)	Average Delay	% penalty in delay
1	15	3.09	15	3.09	0%
2	03,40	2.09	32,40	2.09	0%
3	03,05,15	1.38	15,32,40	1.68	22%
4	01,02,03,05	1.09	15,26,32,40	1.51	38%
5	01,02,03,05,06	0.87	15,26,32,40,44	1.37	57%

(b) collection Q_2 has non-uniform spatial demand distribution as described by (5.17)

The above example demonstrates that a constrained set of replication servers may result in significant loss of mapping efficiency, even for a collection whose spatial demand distribution is identical to that used for determining the optimal set in the first place. Service providers need to take this fact into account when performing capacity planning. For the content providers, this result should also serve as a reminder that co-location of multiple collections with very different spatial demand distributions may result in a solution that is far from optimal.

5.4 Mapping into Storage and Transmission Resources

It is possible to perform resource mapping into both storage and transmission resources, though it is a considerably more complex problem. Assume that some form of

transmission-based QoS service (e.g., intserv or diffserv) is available in the network, such that the delay between vertex v_i and storage server s_j can be reduced from $d(i,j)$ to $d_R(i,j)$ at an additional cost of $c_T(i,j)$ per octet transmitted. Then the reduced delay between vertex v_i and its nearest storage server is

$$d_R(i, X_h) = \min_{x \in X_h} d_R(i, x) \quad (5.21)$$

at a cost of $c_T(i, X_h)$.

Consider a service mapping problem with worst-case delay guarantees. For a given X_h , we can determine the set $V_L \subseteq V$ such that

$$\max_{i \in V_L} d_R(i, X_h) \leq \tau_{\max} \quad (5.22a)$$

and

$$\max_{i \in V \setminus V_L} d(i, X_h) \leq \tau_{\max}. \quad (5.22b)$$

Our objective is then to minimize total cost³⁸:

$$\min_{\substack{X_h \subseteq S \\ V_L \subseteq V}} \left[\int_{t=T_s}^{T_s+T_d} \left\{ \sum_{x \in X_h} B(x) \cdot c_d(x) + \sum_{i \in V_L} \sum_{k \in Q} \lambda \cdot g(i, k) \cdot b(k) \cdot c_T(i, X_k) \right\} \cdot dt \right]. \quad (5.23)$$

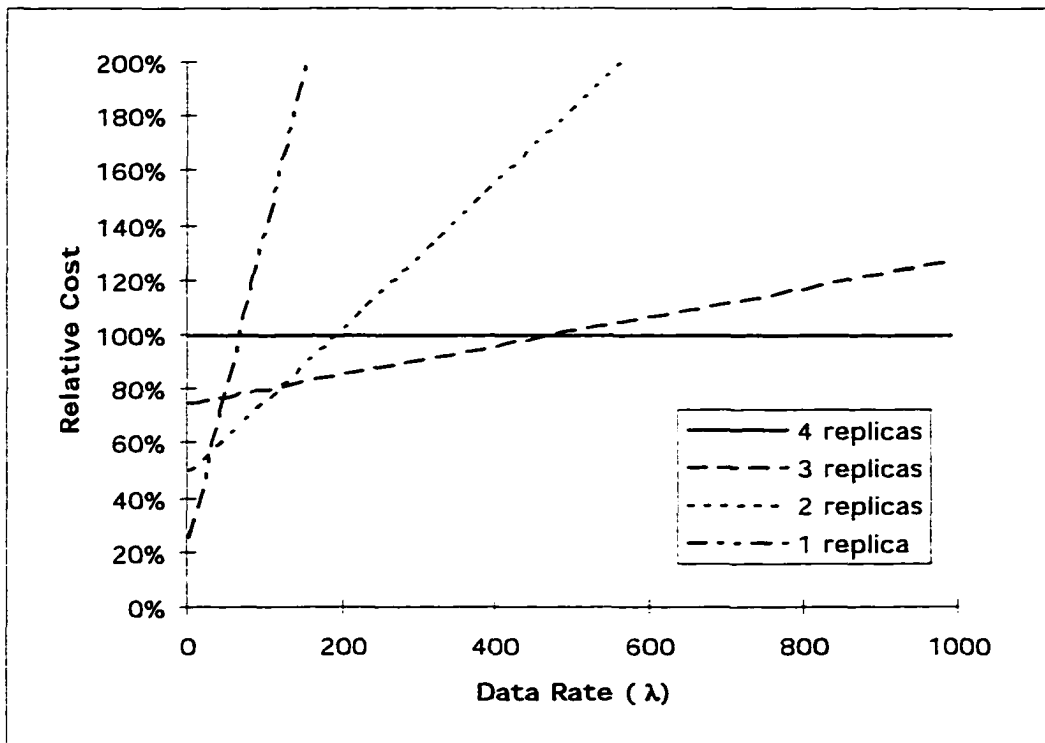
³⁸ We do not include the transmission cost for initial population and subsequent updates of the replicas here, though it may become significant if the data write-to-read ratio is high.

From equation (5.23) it is clear that there are two cost components associated with storage and transmission respectively. Storage cost is calculated as before: the amount of storage capacity at each node $x \in X_h$ multiplied by the per-unit storage cost $c_s(x)$. Additional transmission cost is incurred for each node $v_i \in V_L$. The cost is calculated by multiplying the amount of requested data by that node within the time period $[T_s, T_s + T_d]$ and the per octet incremental transmission cost c_T .

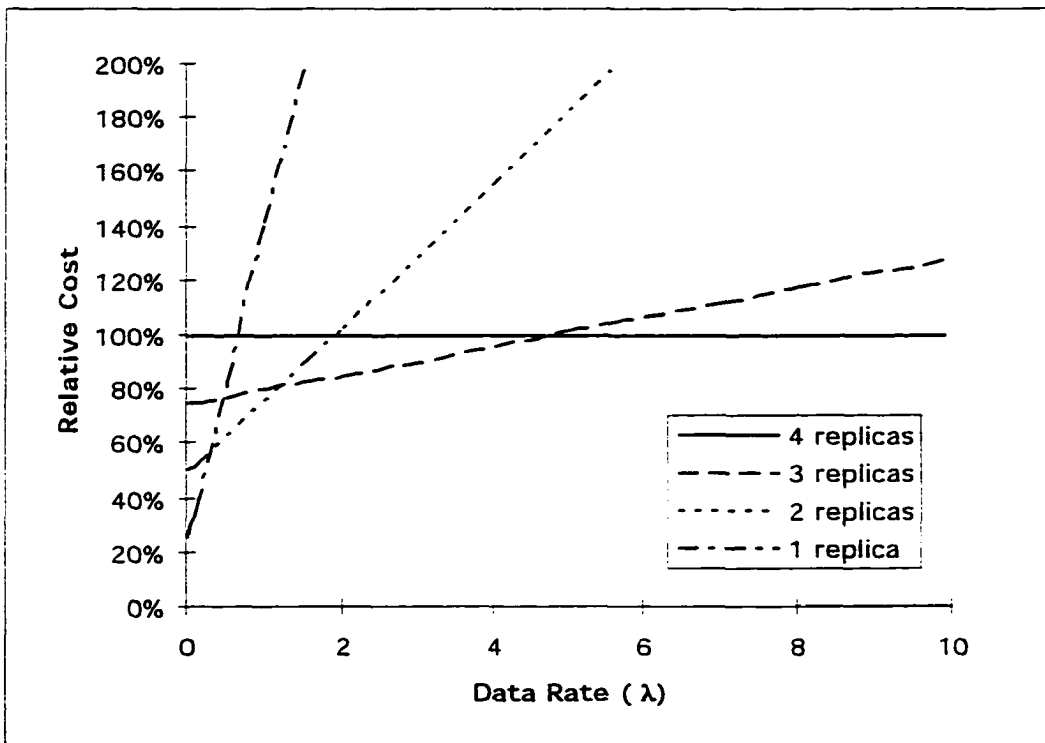
Consider a resource mapping problem for the ARPANET with $\tau_{max} = 3$ hops. Since we are concerned with worst case delay, we cannot take advantage of any non-uniformity in demand distribution. From Figure 5.2 we see that this service may be mapped into a solution with four replicas.

Alternatively, this service may be satisfied with a combination of storage and transmission resources, such that those data requests that originate from further than three hops away from the closest replica are serviced with additional transmission resources, allowing them to experience service quality comparable to those requests that originate from within three hops of the closest replica.

Figure 5.7 shows, for two different storage to transmission cost ratios, the cost of various transmission-plus-storage alternatives relative to the storage-only (four replicas) solution. In both cases, we see that each of the four solutions is the optimal solution for a given range of data access rate λ . The storage-only solution is optimal for the frequently accessed objects, since all nodes can be served from less than three hops away and no additional transmission cost is incurred. However, as the frequency of access declines, it becomes more economical to have fewer replicas and pay transmission charges for each data access that originates from more than three hops away. Substantial savings (as much as 70% over storage-only solution in this example) may be realized. It is worthwhile to point out, however, that accurate information on data access rate is crucial when choosing



(a) $c_s/c_T = 10$



(b) $c_s/c_T = 0.1$

Figure 5.7. Resource cost relative to storage-only solution (4 replicas) at different storage to transmission cost ratios.

a transmission-plus-storage solution. A higher than expected data access rate may quickly turn the optimized solution into a highly sub-optimal one.

Figure 5.8 shows the decision diagram for the ARPANET resource mapping problem with $\tau_{max} = 3$ hops, across data rate λ and storage to transmission cost ratio c_s/c_T . Consistent with our expectations, a higher data rate leads to an optimal solution with more replicas, while a higher storage to transmission cost ratio leads to an optimal solution with fewer replicas.

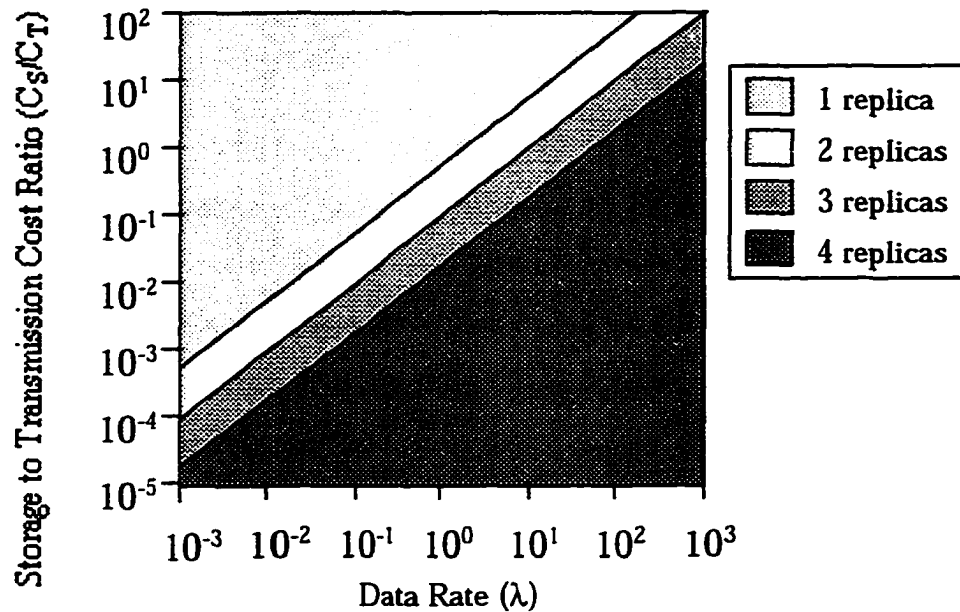


Figure 5.8. Optimal mapping decision for storage and transmission resources (ARPANET with $\tau_{max} = 3$).

Finally, Figure 5.9 is a three-dimensional representation of the cost of the various solutions relative to the storage-only solution. It is once again evident that significant cost savings may be achieved if the cost factors and the data rate is well understood and made known to the resource mapping entity.

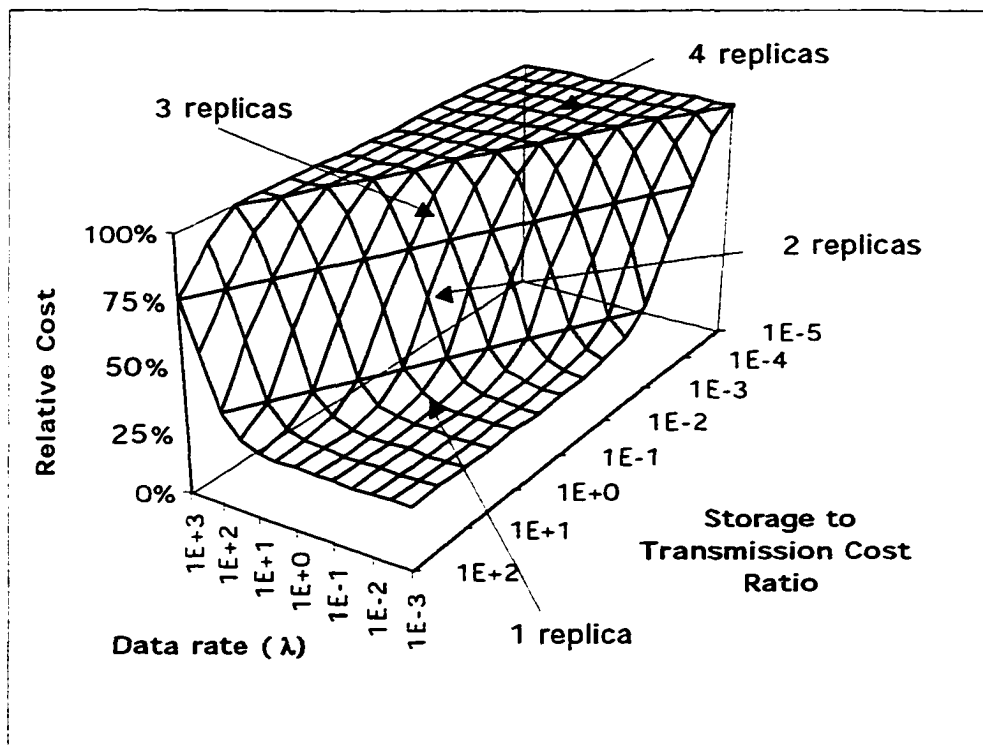


Figure 5.9. Resource cost relative to storage-only solution (4 replicas).

5.5 Conclusion

A formal model of resource mapping for network storage services is developed in this chapter, allowing the exploration of various flavors of mapping problems that may be encountered in a fully realized distributed network storage infrastructure. This model can be used for mapping services with different traffic profiles and performance specifications. The model is also extended to solve network capacity planning problems. Through the application of the model to the ARPANET network, it is demonstrated that (i) knowledge of non-uniformity in data access patterns may be exploited to achieve more efficient usage of network resources, and (ii) the number and placement of replication servers in the network can significantly affect mapping efficiency for individual collections.

The basic model performs mapping into storage resources only, assuming the presence of an underlying best-effort transmission service. The model is extended to handle mapping into an optimal combination of storage and transmission resources, where the transmission resource provides explicit performance improvements over best-effort service. These transmission resources may be physical transmission capacity (e.g., dedicated leased lines) or may be QoS services based on intserv, diffserv, or IP "overnet" services such as those offered by Digital Island. Depending on the relative costs of storage versus transmission, and the data access rates of the collection, a transmission-plus-storage solution may result in significant cost savings over a storage-only solution.

6. Conclusion

This dissertation offers a systematic characterization of the many facets of scale economies in network dissemination of information. It promotes an understanding of how new network technologies have changed, and will continue to change, the economics of information dissemination. This understanding is essential to the design of engineering, economic and policy structures that will constitute the information infrastructure of the future.

This chapter will discuss some of the policy lessons that can be drawn from this work. A summary of the contributions made in the dissertation, and the identification of future research directions, will conclude the chapter.

6.1 Policy Implications and Lessons

It has been emphasized throughout this dissertation that economic mechanisms are at least as important as technical mechanisms in achieving efficient utilization of network resources. Specifically, prices serve as market signals to the users, providing feedback

regarding their resource consumption patterns. Prices that reflect the actual costs of information dissemination will encourage optimal consumption behavior and maximize efficiency in resource allocation. On the other hand, if a significant lag develops between technology and economics, the price structure is unable to reflect the actual cost structure. Then, misaligned prices will lead to market distortions, hidden subsidies, inefficiencies and welfare losses.

The information networking industry is one characterized by rapid changes in technology. If technology is a moving target, then what hope is there for the economics to stay in step with the technology, or for the price to stay consistent with the cost? Should economists and policymakers be put in a position of constantly trying to catch up with the technologists? Thankfully, the answer is no. In a competitive environment, market pressures will force producers to price at cost. In this sense, competition ensures that the industry is both efficient and self-regulating. Therefore, from the perspective of a policymaker, ensuring the competitiveness of the information infrastructure is of the highest priority.

The distributed network storage infrastructure described in Chapter 4 offers an excellent example of the desirability of competition. In Section 4.7, it is suggested that economies of scope savings may be realized if network storage services were offered in conjunction with network transmission services. This implies that a single integrated ISP would be able to offer both storage and transmission based services more efficiently than two separate and independent entities, each offering storage-based and transmission-based services respectively. On the other hand, if one of the two market segments (e.g., the transmission segment) turns out to be monopolistic or oligopolistic in nature, then vertical integration by an ISP would open the door to anti-competitive behavior in the other, competitive market segment (e.g., the storage segment). For example, the integrated ISP may use the monopoly rents extracted from its transmission business to subsidize its

competitive storage business, and price its storage services at below cost. It does so with the knowledge that it is foregoing profit maximization in the short run (Fry et al., 1995, Hausman and Tardiff, 1995, Williamson, 1981). However, this predatory pricing strategy serves to drive out competition, increase market share, and ultimately achieve lock-in for the integrated ISP. By raising the barriers of entry and re-entry to potential competitors in the storage segment of the market, vertical integration may actually retard technological innovation in the network storage domain. An innovator who comes up with a revolutionary technology in network storage, but has no expertise in network transmission, would have tremendous difficulty entering the integrated market.

This painful and uncertain choice between economies of scope savings and component-based competition can be avoided if both market segments are competitive. A competitive market structure will offer a level playing field to transmission-only providers, storage-only providers, and integrated providers alike. Integrated providers are free to exploit efficiency gains from bundling; others are free to specialize in one component and compete strictly in that market segment.

Indeed, this is reminiscent of the spirit and rationale behind mixed bundling, which we discussed in Chapter 2. The dominance of mixed bundling may be extended to an integrated ISP, such that it is in its best interest to offer transmission and storage services both individually and in a bundle. Consumers (including resellers) are free to mix and match different service components from different providers to build their own service, or purchase the integrated service as a whole.

While this dissertation has been singing the praises of cost-based pricing, it must be pointed out that cost-based pricing should not be carried to the extreme. At some point of granularity, the cost of usage-metering and accounting will begin to outweigh the efficiency gains from fine-grained pricing. For example, first-class postage in the U.S. is

distance insensitive, because the cost of tracking and verification far outweighs the benefits of distance-sensitive pricing. Similarly, Internet traffic is not priced according to distance today, even though cost is clearly correlated with distance. Section 4.7 identifies a consequence of distance insensitive pricing -- that it may become an impediment to component-based competition for network storage service provision. This consideration, along with future changes in metering and accounting technologies, may tilt the cost-benefit calculus towards some form of distance sensitive pricing with varying degrees of granularity.

6.2 Contributions Made in this Dissertation

This dissertation identifies different levels and dimensions along which economies of scale conditions may exist for network-based information dissemination technologies and applications. It then proceeds to study these conditions along the object dimension (Chapter 2), receiver dimension (Chapter 3) and temporal dimension (Chapters 4 and 5).

Along the object dimension, a multi-product bundling model with multi-dimensional consumer taste characteristics is developed to study the optimal bundling and pricing strategy of information goods such as academic journals. Using empirical journal usage data and cost projections for information-delivery over the Internet, the model finds that metered usage (i.e., articles-on-demand) should account for a significant fraction of revenue when articles and subscriptions are optimally priced according to a mixed bundling strategy.

Along the receiver dimension, a communication cost model for multicast is developed. The model demonstrates that multicast group size can serve as an excellent proxy for multicast tree cost. Computer simulations show that, statistically, multicast tree length grows at the 0.8 power of the multicast group size until the point of tree

saturation, beyond which additional receivers can be added to the group without further tree growth. In other words, the marginal cost of multicast declines according to an exponential decay function until it reaches zero at tree saturation. This result is validated with both real and generated networks, and is robust across topological styles and network sizes. This suggests that a two-part tariff may be appropriate if providers choose to adopt a cost-based approach to multicast pricing.

Along the temporal dimension, economies of scale savings can be realized through network caching and replication. Chapter 4 offers the vision of and motivation for a distributed network storage infrastructure with service guarantees. Caching and replication can be treated as different service classes within a unified QoS framework. A research agenda is proposed, outlining key research areas and problems, including those in service specification, resource mapping, admission control, resource reservation, real-time storage management, location transparency, accounting, pricing and industrial organization.

A formal model of resource mapping for network storage services is developed in Chapter 5. This model can be used for mapping services with different traffic profiles and performance specifications. The model is also extended to solve network storage capacity planning problems. Through the application of the model to the ARPANET network, it is demonstrated that (i) knowledge of non-uniformity in data access patterns may be exploited to achieve more efficient usage of network resources, (ii) the number and placement of replication servers in the network can significantly affect mapping efficiency for individual collections. Finally, the model is extended to handle the mapping of services into an optimal combination of storage and transmission resources.

6.3 Future Work

The bundling model in Chapter 2 assumes that a journal is made up of N individual articles. This assumption may be relaxed to include other separable components to a journal subscription, such as the table of content, indices, abstracts and other announcements. Readers can assign different valuations for each of these components just as they do for the individual articles. Therefore these components can be candidates for unbundling as well. The optimal pricing strategy of these components may be of interest to information producers. Also, the bundling model can be extended to maximize total surplus (the sum of consumer and producer surpluses) rather than producer surplus only.

In Chapter 3, the power relationship between multicast group size and multicast tree length is established by means of computer simulation, where Dijkstra's algorithm is used to construct multicast trees over real and generated network topologies. It may be worthwhile to pursue a theoretical basis for this result, possibly based on graph theory and/or combinatorial analysis.

Chapter 4 provides a research roadmap for the distributed network storage infrastructure, and outlines the key mechanisms necessary for the vision to become a reality. These mechanisms include: service specification, resource mapping, admission control, resource reservation protocols, real-time storage management, pricing, etc. While each of these mechanisms has to be developed for the infrastructure, there is significant amount of knowledge from the fields of transmission-based QoS, distributed file systems design, distributed database design, etc., that can be leveraged in this effort.

The formal mathematical model constructed in Chapter 5 is a first step towards resource mapping for the distributed network storage infrastructure. The mapping problems are identified to be variants of the k -center and k -median problems, which have

been established by operations researchers to be NP -hard. Therefore, additional work is needed to develop algorithms (approximations and heuristics) that are efficient in and appropriate for the distributed network storage environment.

The field of information dissemination and data networking continues its rapid pace of change and transformation. New technologies and business models may unveil new levels and dimensions along which economies of scale opportunities could emerge. For example, the development of an ubiquitous electronic marketplace may result in scale economies in the seller dimension. The aggregation of multiple sellers in a common market exchange will reduce both transactional and search costs, which may or may not lead to a reduction in price dispersion (Economist, 1998). Economies of scale conditions in the demand side, i.e., network externalities, may also become important in the information dissemination context. Given the deluge of information sources, consumers may find greater value in those sources with larger subscription bases. Finally, it may be fruitful to examine scale economies in other data networking applications that are not specific to information dissemination, e.g., real-time interactive applications.

Appendix 1. Derivation of Producer Surplus for Alternative Bundling Strategies

A profit-maximizing journal publisher will seek to optimize the prices P_J and P_A by maximizing the objective function Π , restated here from equation (2.8):

$$\Pi(P_J, P_A) = \iint_{R_J} [P_J - MC_J] f(w_o, k) \cdot dw_o \cdot dk + \iint_{R_A} n^* [P_A - MC_A] f(w_o, k) \cdot dw_o \cdot dk \quad (\text{A1.1})$$

To derive the gross margin or producer surplus attainable from each of the three alternative bundling strategies, we need to identify the regions R_J and/or R_A in each scenario. This allow the limits of integration for the definite integrals to be quantitatively specified.

Additionally, the p.d.f. of the journal reading population in $\{w_o, k\}$ space has to be specified. The assumption of independence between random variables w_o and k , and the choice of the probability distributions gives $f(w_o, k) = f_{w_o}(w_o) \cdot f_k(k)$, where

$$f_{w_o}(w_o) = \begin{cases} 1 & 0 \leq w_o \leq 1; \\ 0 & \text{elsewhere;} \end{cases} \quad (\text{A1.2})$$

$$f_k(k) = \lambda \cdot e^{-\lambda \left(k - \frac{1}{N}\right)} \quad k \geq \frac{1}{N}. \quad (\text{A1.3})$$

From equation (2.4), n^* , the optimal number of articles consumed, is

$$n^* = \min \left\{ N, \frac{k \cdot N \cdot (w_o - P_A)}{w_o} \right\}. \quad (\text{A1.4})$$

Finally, from equation (2.7), we have the marginal cost per journal (MC_j) expressed in terms of the marginal cost per article (MC): $MC_j = N \cdot MC$. With these substitutions, producer surplus under each bundling alternative (Π_{PB} , Π_{PU} and Π_{MB}) can be expressed as functions of P_j , P_A , MC , γ and the model parameters N and λ .

A1.1. Pure Bundling

The limits of integration for the pure bundling strategy are based on the boundaries of the region R_j , as defined by the $U_j = 0$ curve in Figure 2.5. Solving for $U_j = 0$ requires the quantification of W_j . Integrating $w(n)$ over all articles ($0 \leq n \leq N-1$) results in

$$W_j = \int_0^N w(n) \cdot dn + \Delta_c \quad (A1.5)$$

or

$$W_j = \begin{cases} \frac{kNw_o}{2} + \Delta_c & \text{if } k \leq 1, \\ \frac{(2k-1)Nw_o}{2k} + \Delta_c & \text{if } k > 1, \end{cases} \quad (A1.6)$$

The compensating term Δ_c is the sum of all triangular areas not integrated under the demand curve $w(n)$ in Figure 2.3. There are kN (or N if $k > 1$) triangles, each with an area of $w_o/2kN$. Therefore Δ_c is independent of N (and k if $k \leq 1$):

$$\Delta_c = \begin{cases} \left(\frac{w_o}{2}\right) & \text{if } k \leq 1, \\ \left(\frac{1}{k}\right)\left(\frac{w_o}{2}\right) & \text{if } k > 1. \end{cases} \quad (\text{A1.7})$$

Substituting these results into $U_j = W_j - P_j = 0$ yields

$$w_o = \begin{cases} w_{o1} = \min\left[1, \frac{2P_j}{Nk+1}\right] & k \leq 1; \\ w_{o2} = \frac{2kP_j}{2kN - N + 1} & k > 1. \end{cases} \quad (\text{A1.8})$$

Using w_{o1} and w_{o2} as the limits of integration for equation (A1.1), we can express producer surplus under the pure bundling scenario as

$$\Pi_{PB} = \int_{k=\frac{1}{N}}^1 \int_{w_o=w_{o1}}^1 [P_j - N^T \cdot MC] f(w_o, k) \cdot dw_o \cdot dk + \int_{k=1}^{\infty} \int_{w_o=w_{o2}}^1 [P_j - N^T \cdot MC] f(w_o, k) \cdot dw_o \cdot dk$$

or

$$\Pi_{PB} = [P_j - N^T \cdot MC] \cdot \left\{ \int_{k=\frac{1}{N}}^1 \int_{w_o=w_{o1}}^1 \lambda e^{-\lambda \left(k - \frac{1}{N}\right)} \cdot dw_o \cdot dk + \int_{k=1}^{\infty} \int_{w_o=w_{o2}}^1 \lambda e^{-\lambda \left(k - \frac{1}{N}\right)} \cdot dw_o \cdot dk \right\}. \quad (\text{A1.9})$$

Note that the absence of a R_A region in pure bundling means that the second term of equation (A1.1) can be dropped. Differentiating Π_{PB} with respect to P_j and setting it to zero, the optimal bundle subscription price P_j and the corresponding Π_{PB} can be computed numerically.

A1.2. Pure Unbundling

Under pure unbundling, there is no region R_J since no subscription bundles are sold. Instead, the region of integration is R_A , the area to the right of the line $w_o = P_A$:

$$\Pi_{PU} = \int_{k=\frac{1}{N}}^1 \int_{w_o=P_A}^1 [P_A - MC] \cdot n^* \cdot f(w_o, k) \cdot dw_o \cdot dk. \quad (A1.10)$$

However, $n^*(w_o, k)$ has a discontinuity at $n^* = N$. This mandates the integration to be carried out in two parts. We can locate the region where $n^* = N$, which is northeast of the $n^* = 100$ curve in Figure 2.6, by solving $w(N-1) = P_A$. This yields $k = w_o/(w_o - P_A)$. Therefore, producer surplus under pure unbundling can be expressed as:

$$\Pi_{PU} = [P_A - MC] \left\{ \int_{w_o=P_A}^1 \int_{k=\frac{1}{N}}^{\frac{w_o - P_A}{w_o}} \left(\frac{kN(w_o - P_A)}{w_o} \right) \lambda e^{-\lambda \left(k - \frac{1}{N} \right)} dw_o dk + \int_{w_o=P_A}^1 \int_{k=\frac{w_o}{w_o - P_A}}^1 N \lambda e^{-\lambda \left(k - \frac{1}{N} \right)} dw_o dk \right\}. \quad (A1.11)$$

Again, Π_{PU} can be differentiated with respect to P_A and set to zero, and the optimal article price P_A and the maximum Π_{PU} can be solved numerically.

The actual utility gained from the purchase of n^* articles, W_A , can be determined by summing (or integrating) all the $w(n)$'s for $0 \leq n \leq n^*$.

$$W_A = \int_0^{n^*} w(n) \cdot dn + \Delta_c^*; \quad (A1.12)$$

or

$$W_A = (kN)(w_o - P_A) - \left(\frac{kN}{2w_o}\right)(w_o - P_A)^2 + \Delta_c^* \quad (\text{A1.13})$$

In this case, Δ_c^* is not the entire area Δ_c , only a fraction proportional to n^* .

$$\Delta_c = \begin{cases} \left(\frac{n^*}{N}\right)\left(\frac{w_o}{2}\right) & \text{if } k \leq 1, \\ \left(\frac{n^*}{N}\right)\left(\frac{1}{k}\right)\left(\frac{w_o}{2}\right) & \text{if } k > 1, \end{cases}$$

which can be reduced to:

$$\Delta_c = \min[1, k] \cdot \left(\frac{w_o - P_A}{2}\right) \quad (\text{A1.14})$$

From here, the net benefit derived from purchasing n^* individual articles under pure unbundling can be calculated as $U_A = W_A - n^* \cdot P_A$, or

$$U_A = \left(kN + \frac{\min[1, k]}{2}\right)(w_o - P_A) - \left(\frac{kN}{2w_o}\right)(w_o - P_A)^2 - \frac{kN(w_o - P_A)P_A}{w_o} \quad (\text{A1.15})$$

A1.3. Mixed Bundling

The consumer choice regions under mixed bundling may take on one of two slightly different shapes and boundaries depending on the P_J/P_A ratio. Figure A1.1 illustrates the two alternate scenarios. In each case, we need to solve $\{U_J = 0, U_A = 0, U_J = U_A\}$ to establish the boundaries between the different regions. The solution to $U_J = 0$ is

the same as that in the pure bundling scenario. The solution to $U_A = 0$ is simply $w_o = P_A$. Solving for $U_J = U_A$ yields (for $k \leq 1$ and $w_o \geq P_A$):

$$\bar{k} = \frac{w_o(2P_J - w_o)}{N(P_A^2 - 2w_oP_A) + w_o(w_o - P_A)}, \quad (\text{A1.16})$$

or equivalently,

$$\bar{w}_o = \frac{(kP_A + 2kNP_A - 2P_J) + \sqrt{(4 - 4k)kNP_A^2 + (kP_A + 2kNP_A - 2P_J)^2}}{2(k - 1)}. \quad (\text{A1.17})$$

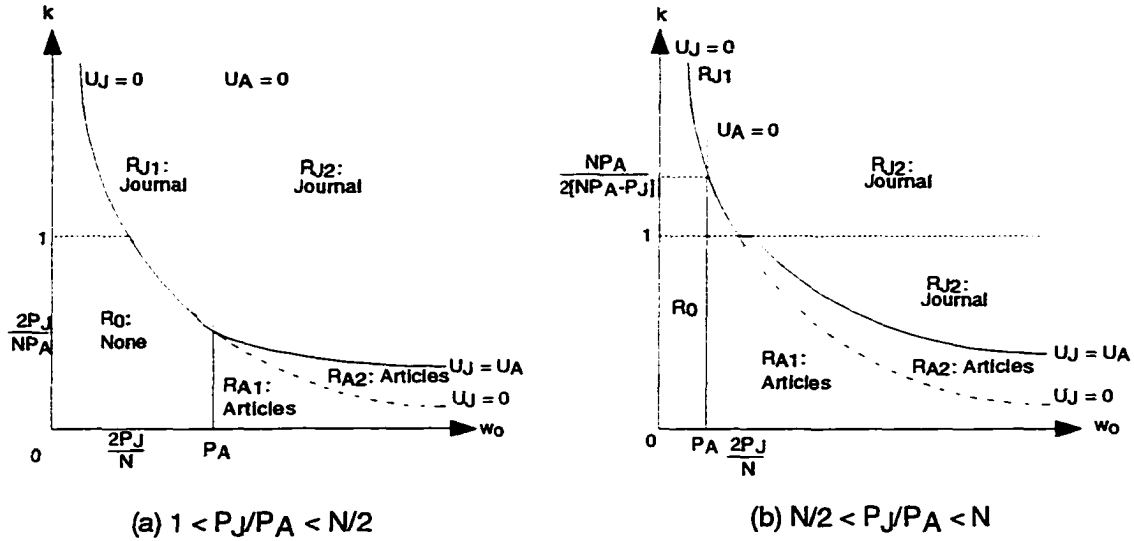


Figure A1.1. Consumer choice in mixed bundling scenario

For $P_J/P_A \leq N/2$, as illustrated by Figure A1.1 (a), we can express producer surplus as

$\Pi_{MB(S)}$:

$$\Pi_{MB(S)} = \Pi_{MB(S),A} + \Pi_{MB(S),J} \quad (\text{A1.18})$$

where

$$\Pi_{MB(S),A} = [P_A - MC] \cdot \left\{ \int_{w_0=P_A}^1 \int_{k=\frac{1}{N}}^{\bar{k}} n^* \cdot f(w_0, k) \cdot dw_0 \cdot dk \right\} \quad (A1.19)$$

and

$$\Pi_{MB(S),J} = [P_J - N^* MC] \left\{ \int_{w_0=P_A}^1 \int_{k=\bar{k}}^1 f(w_0, k) dw_0 dk + \int_{k=\frac{2P_J}{NP_A}}^1 \int_{w_0=w_0^1}^{P_A} f(w_0, k) dw_0 dk + \int_{k=1}^{\bar{k}} \int_{w_0=w_0^2}^{P_A} f(w_0, k) dw_0 dk \right\} \quad (A1.20)$$

For $N/2 \leq P_J/P_A \leq N$, we have, instead, $\Pi_{MB(B)}$:

$$\Pi_{MB(B)} = \Pi_{MB(B),A} + \Pi_{MB(B),J} \quad (A1.21)$$

where

$$\Pi_{MB(B),A} = [P_A - MC] \cdot \left\{ \int_{w_0=P_A}^{w_{k=1}^1} \int_{k=\frac{1}{N}}^1 n^* f(w_0, k) dw_0 dk + \int_{w_0=w_{k=1}^1}^1 \int_{k=\frac{1}{N}}^{\bar{k}} n^* f(w_0, k) dw_0 dk + \int_{w_0=P_A}^{w_0^2} \int_{k=1}^{\frac{NP_A}{2[NP_A-P]}} n^* f(w_0, k) dw_0 dk \right\} \quad (A1.22)$$

and

$$\Pi_{MB(B),J} = [P_J - N^* MC] \left\{ \int_{w_0=w_0^1}^1 \int_{k=\bar{k}}^1 f(w_0, k) dw_0 dk + \int_{k=1}^{\bar{k}} \int_{w_0=w_0^2}^1 f(w_0, k) dw_0 dk \right\} \quad (A1.23)$$

Since we do not know, *a priori*, the ratio P_j/P_A , we need to compute both $\Pi_{MB(S)}$ and $\Pi_{MB(B)}$, obtain the two sets of optimal prices, and by inspection of the ratios determine which set of the results is valid.

Appendix 2. Sample List of Web-Hosting Service Providers

<u>Service Provider</u>	<u>URL</u>	<u>\$/MB download</u>
AT&T Easy World Wide Web	http://www.att.com/	\$0.50
Cowboy.Net	http://cowboy.net/commercial_prices.html	\$0.05
Citizens Internet Service	http://www.swva.net/citizen/services/webprice.html	\$1.00
DC-AdNet	http://www.dc-adnet.com/prices.htm	\$1.00
Internet Industries Web Hosting	http://www.industries.net/webhosting.html	\$0.05
Internet Video Services' netvideo	http://www.netvideo.com/netvideo/price.html	\$0.02-\$0.08
Multiboard Communications	http://www.multiboard.com/services.html	\$0.07-\$0.10
PreciseNet Web Site Hosting	http://www.precisenet.com/host.htm	\$0.20
Pro-NetMedia Creations, Inc.	http://www.pcinc.com/pricing.htm	\$0.25
Serview Premium Webhosting	http://serview.com/pricing.html	\$0.10
Sustance	http://www.he.net/~sustance/prices.html	\$0.039-\$0.10

Compiled: January 1997

Appendix 3. Source Code for Multicast Cost Quantification

```
/*-----*
 *      Multicast Cost Quantification (filename: spt.c)      *
 *                                                         *
 *      John Chung-I Chuang                                *
 *      written 10/2/97                                    *
 *                                                         *
 *      Approach: construct source-based shortest path trees (in both *
 *      unicast and multicast modes) using adjacency list structure *
 *      and priority-first search algorithm (Sedgewick, 199x). *
 *                                                         *
 *      usage: spt <#receivers> <grp_list> <net_list> *
 *-----*/

#include <stdio.h>

/*-----*
 *      adjacency list structure (Sedgewick, p 421)      *
 *-----*/

#define maxV 40000

struct node {int v; int w; struct node *next; };
struct node *t, *z;
struct node *adj[maxV];

int j, x, y, ec, a, V, E;

FILE *ifp1;

void adjlist()
{
    fscanf(ifp1, "%d %d\n", &V, &E);
    z = (struct node *) malloc(sizeof *z);
    z->next = z;
    for (j = 1; j <= V; j++) adj[j] = z;
    for (j = 1; j <= E; j++)
    {
        fscanf(ifp1, "%d %d %d %d\n", &x, &y, &ec, &a);
        /* Set ec = 1 if computing hop-based shortest path tree */
        /* ec = 1; */
        t = (struct node *) malloc(sizeof *t);
        t->v = x; t->w = ec; t->next = adj[y]; adj[y] = t;
        t = (struct node *) malloc(sizeof *t);
        t->v = y; t->w = ec; t->next = adj[x]; adj[x] = t;
    }
}

void printadjlist()
```

```

{
    printf("Printing adjlist:\n");
    for (j = 1; j <= V; j++)
        for (t = adj[j]; t != z ; t = t->next)
            printf("%d %d %d\n", j, t->v, t->w);
}

/*-----*
 *       Reading in source and receiver membership list       *
 *-----*/

struct memb {int v; struct memb *next; };
struct memb *t1, *t2, *z2, *ontree;

int numr, des;
int src = 0;

FILE *ifp2;

void memblist()
{
    z2 = (struct memb *) malloc(sizeof *z2);
    z2->v = 0;
    z2->next = z2;
    ontree = z2;

    for (j = 1; j <= numr; j++)
    {
        fscanf(ifp2, "%d\n", &des);
        t2 = (struct memb *) malloc(sizeof *t2);
        t2->v = des; t2->next = ontree;
        ontree = t2;
    }
}

void printmemblist()
{
    printf("Number of receivers = %d\n", numr);
    printf("The receivers are \n");
    for (t1 = ontree; t1 != z2 ; t1 = t1->next)
        printf("%d\n", t1->v);
}

/*-----*
 *       priority-queue utilities       *
 *-----*/

#define maxQ 40000

static struct pqueue {int ver; int pri; };
static struct pqueue q[maxQ+1];
static int head,tail;

void pqinitialize()

```

```

{
    head = 0;
    tail = 0;
}

int pqempty()
{
    return head == tail;
}

void pqinsert(int v, int p)
{
    q[tail].ver = v;
    q[tail++].pri = p;
}

int pqremove()
{
    int j, hi, temp;
    hi = head;
    for (j = head+1; j < tail; j++)
        if (q[j].pri < q[hi].pri) hi = j;
    temp = q[hi].ver;
    q[hi].ver = q[--tail].ver;
    q[hi].pri = q[tail].pri;
    q[tail].ver = 0;
    q[tail].pri = 0;
    return temp;
}

int pqupdate(int v, int p)
{
    int j;
    int done = 0;
    int changed = 0;
    for (j = head; j <= tail && !done; j++)
    {
        if ((q[j].ver == v) && (p < q[j].pri))
        {
            q[j].pri = p;
            done = 1;
            changed = 1;
        }
        else if (q[j].ver == v) done = 1;
    }
    if (!done)
    {
        pqinsert(v,p);
        changed = 1;
    }
    return changed;
}

void pqprint()

```

```

{
    printf("pqprint\n");
    for (j = head; j <= tail; j++)
        printf("%d %d\n", q[j].ver, q[j].pri);
}

/*-----*
 *      priority-first search algorithm (Sedgewick, p 455)      *
 *-----*/

#define unseen 100000

/* use the following priority for finding shortest path tree */
#define priority (val[k]+(t->w))

/* use the following priority for finding minimum spanning tree */
/* #define priority (t->w) */

int val[maxV], dad[maxV];
int id = 0;

void visit(int k)
{
    if (pqupdate(k, unseen) != 0) dad[k] = 0;
    while (!pqempty())
    {
        id++; k = pqremove(); val[k] = -val[k];
        if (val[k] == unseen) val[k] = 0;
        for (t = adj[k]; t != z; t = t->next)
            if (val[t->v] < 0)
                if (pqupdate(t->v, priority))
                {
                    val[t->v] = -(priority);
                    dad[t->v] = k;
                }
    }
}

void constructfulltree()
{
    int k;
    pqinitialize();
    for (k = 1; k <= V; k++) val[k] = -unseen;
    for (k = 1; k <= V; k++)
        if (val[k] == -unseen) visit(k);
}

/*-----*/

int dist(int a, int b)
{
    int cc;
    if (a == b || b == 0) return 0;

```

```

        else
        {
            for (t = adj[a]; t != z ; t = t->next)
                if (t->v == b) cc = t->w;
            return cc;
        }
    }

void printfulltree()
{
    int mm;
    int fsptlen = 0;
    int funilen = 0;

    for (mm = 1; mm <= V; mm++)
    {
        funilen = funilen + val[mm];
        fsptlen = fsptlen + dist(mm,dad[mm]);
    }

    printf("full unicast tree length = %d\n", funilen);
    printf("full SPT length = %d\n", fsptlen);
    printf("ratio = %d%%\n", 100*fsptlen/funilen);
}

/*-----*
 *      finding tree lengths for mcast group      *
 *-----*/

struct memb *t3, *t4, *t5, *t6;

int ingroup(int k)
{
    int match = 0;
    for (t3 = ontree; t3 != z2 && !match; t3 = t3->next)
        if (k == t3->v) match = 1;
    return match;
}

void add2group(int k)
{
    t4 = (struct memb *) malloc(sizeof *t4);
    t4->v = k; t4->next = ontree; ontree = t4;
}

void constructgrouptree()
{
    int k,now,new;
    float f0, f1, f2;
    int gsptlen = 0;
    int gunilen = 0;
    int count = numr;
    int nextcount = 0;

```

```

int complete = 0;

for (k = 1; k <= V; k++)
    if (ingroup(k)) gunilen = gunilen + val[k];

while (!complete)
{
    complete = 1; nextcount = 0;
    for (t5 = ontree; t5 != z2 && count != 0; t5 = t5->next)
    {
        now = t5->v;
        new = dad[now];
        gsptlen = gsptlen + dist(now,new);
        count--;
        if (new != src && !ingroup(new))
            { add2group(new); nextcount++; complete = 0; }
    }
    count = nextcount;
}

f0 = numr * gsptlen / gunilen;
f1 = 100 * numr / V;
f2 = 100 * gsptlen / gunilen;

printf("%6d %6.2f %6.2f %6d %6d %6.2f\n",
        numr, f0, f1, gsptlen, gunilen, f2);
}

/*-----*
 *      main program                               *
 *-----*/

main(int argc, char ** argv)
{

    ifp1 = fopen(argv[3], "r");
    ifp2 = fopen(argv[2], "r");
    numr = atoi(argv[1]);

    adjlist();
    fclose(ifp1);

    memblist();
    fclose(ifp2);

    constructfulltree();

    constructgroupstree();

}

/*-----*/

```

Appendix 4. Taxonomy of Data Duplication Schemes According to Traditional Ex-Post vs. Ex-Ante Distinction

Table A4.1 presents a taxonomy of data duplication schemes according to the traditional ex-post vs. ex-ante distinction:

Table A4.1. Taxonomy of data duplication schemes.³⁹

copy location	ex-post (caching)	ex-ante (selective prefetch) <-----> (full replication)
client	browser cache (C)	pull: prefetch adjacent objects (C) push: webcast-enabled client (S)
organization	proxy cache (O)	webcast proxy server (S) netnews/NNTP (O)
network/provider	network cache (N) reverse proxy cache (S)	web hosting (N+S) distributed network storage service (N+S)
source	--	mirror site (S)

Caching and replication can be performed at any location between the source and destination. Web browsers maintain a local cache on the client's host; organizations run proxy caches at the network edges (typically at their firewalls); experimental network caches are placed at key nodes in the network. For the case of replication, we observe increasing selectivity as we move from the data source to the client. This is consistent with the basic principles and motivations of computer memory hierarchy design.

³⁹ The letter in parenthesis (C: client, O: organization, N: network or S: source) indicates the entity responsible for managing the copies.

The letter in parenthesis (C: client, O: organization, N: network or S: source) indicates the entity responsible for managing the copies. In the case of caching, the managing entity is apparently dictated by the location of the caches. In the case of replication, however, the source appears to take on a greater role in replication management at all locations. Network news is by no means an exception: news messages are composed by authors throughout the network and so each of the NNTP news servers at the network edges are both sender and receiver at the same time.

In the case where the network takes on the management of caches and replicas, we begin to see the prototype of a distributed network storage service. If the use of network caches and replications become prevalent, we may expect the network provider to begin engaged in the provision of both transmission-based and storage-based services. Then the network infrastructure will consist of not just the links and switches, but storage elements as well.

Appendix 5. Solution Method for Resource Mapping with Partial Replications of Multi-Object Collections

In Section 5.3.3, we considered the resource mapping problem where it is possible to create partial replicas of collections with multiple objects. Specifically, we assumed that all objects in the collection are of equal size, and per-unit storage cost is identical across all network nodes. This allows us to develop a solution method for the mapping problem stated in (5.11), which is presented in this appendix.

There are four inputs to this solution method, namely the delay target τ_{avg} , the demand distribution $g_q(k)$ across individual objects in the collection, the initial delay $D_{avg}(h=1)$ which is the realized average delay when there is a single, optimally placed full replica of the collection, and the delay reduction $\Delta D_{avg}(h, h+1)$ when we move from a solution with h full replicas to one with $(h+1)$ full replicas. The first input is service-specific, the second input is collection specific, while the third and fourth inputs are specific to the network $G(V, E)$ and the spatial demand distribution $g_v(i)$.

From the demand distribution $g_q(k)$ it is possible to determine the probability of object access: $P(q_k) \forall q_k \in Q$. Then for all objects q_k with $P(q_k) > 0$, set $h(q_k) = 1$, i.e., we need at least one copy of object q_k that has non-zero access probability.

From this point on, copies of individual objects are added, incrementally one object-copy at a time, until the average delay D_{avg} is less than or equal to the delay target τ_{avg} . The object chosen at each round is the one that will result in the largest reduction in delay, $\Delta D_p(q_k)$, where

$$\Delta D_p(q_k) = P(q_k) * \Delta D_{avg}(h(q_k), h(q_k)+1). \quad (A5.1)$$

This methodology is summarized in Table A5.1.

Table A5.1 Methodology for solution method

Step 1: For given network $G(V,E)$ and $g_v(i)$:

- a) set $D_{avg} = D_{avg}(h=1)$
- b) find $\Delta D_{avg}(h, h+1)$

Step 2: For given collection Q and demand distribution $g_q(k)$:

- a) find $P(q_k) \forall q_k \in Q$
- b) for all objects q_k with $P(q_k) > 0$, set $h(q_k) = 1$

Step 3: Add copy of object q_k with maximum $\Delta D_p(q_k)$, where

$$\Delta D_p(q_k) = P(q_k) * \Delta D_{avg}(h(q_k), h(q_k)+1)$$

Step 4: Increment $h(q_k)$

Step 5: $D_{avg} = D_{avg} - \Delta D_p(q_k)$

Step 6: If $(D_{avg} > \tau_{avg})$ go to step 4, else done.

Applying this solution method to the numerical example in Section 5.3.3, we have a service specification of $\tau_{avg} = 2.30$ hops for a four-object collection whose access pattern is characterized by Zipf's distribution. The resource mapper computes the delay reduction table (Table A5.2), whose matrix elements are the product of $P(q_k)$ and $\Delta D_{avg}(h, h+1)$. This table is used for the actual determination of the number of copies needed for each object in the collection. Starting with one copy of each of the four objects (second row of Table A5.3), the realized average delay $D_{avg}(h=1)$ is 3.32 hops. The first object-copy to be added is that of object q_1 , since it has the largest delay reduction of

0.4704 hops (third row of Table A5.3). This brings the current delay D_{avg} down to 2.85 hops. Next, a copy of object q_2 is added, with a delay reduction of 0.2352 hops. The third copy to be added is again of object q_1 , and finally with the addition of the fourth object copy (of object q_3), the realized delay D_{avg} is less than the delay target of 2.30 hops. The total number of object-copies required to achieve the average delay bound is therefore eight. These copies are placed according to Table 5.2, resulting in three copies of q_1 , at nodes 2, 3, and 34; two copies each of q_2 and q_3 , at nodes 8 and 47; and a single copy of q_4 at node 26.

Table A5.2 Delay Reduction Table Computed by Resource Mapper

h	$\Delta D_{avg}(h, h+1)$	$\Delta D_p(q_k) = P(q_k) * \Delta D_{avg}$			
		Object 1 $P(q_1) = 0.48$	Object 2 $P(q_2) = 0.24$	Object 3 $P(q_3) = 0.16$	Object 4 $P(q_4) = 0.12$
1	0.98	0.4704	0.2352	0.1568	0.1176
2	0.47	0.2256	0.1128	0.0752	0.0564
3	0.23	0.1104	0.0552	0.0368	0.0276
4	0.23	0.1104	0.0552	0.0368	0.0276
5	0.17	0.0816	0.0408	0.0272	0.0204
6	0.11	0.0528	0.0264	0.0176	0.0132

Table A5.3 Stages of Resource Mapping Process

Action	Number of object-copies					delay reduction	current delay (D_{avg})
	q_1	q_2	q_3	q_4	total		
Begin	1	1	1	1	4		3.32
Add object 1	2	1	1	1	5	0.4704	2.85
Add object 2	2	2	1	1	6	0.2352	2.61
Add object 1	3	2	1	1	7	0.2256	2.39
Add object 3	3	2	2	1	8	0.1568	2.23

Bibliography

Abrams, M., C.R. Standridge, G. Abdulla, S. Williams, and E.A. Fox. "Caching proxies: limitations and potentials." 4th International World Wide Web Conference, Boston MA, December 1995.

Adams, W.J., and J.L. Yellen. "Commodity bundling and the burden of monopoly." *Quarterly Journal of Economics* 90 (1976): 475-498.

Almeida, A., A. Bestavros, M. Crovella, and A. de Oliveira. "Characterizing reference locality in the WWW." IEEE Conference on Parallel and Distributed Information Systems, Miami Beach FL, December 1996.

Almeida, J., M. Daby, A. Manikutty, and P. Cao. "Providing differentiated levels of service in web content hosting." Sigmetrics Workshop on Internet Server Performance, 1998.

Almeroth, K., and Ammar. M.H. "Multicast group behavior in the Internet's multicast backbone (MBone)." *IEEE Communications Magazine*, June 1997.

Amir, Y., A. Peterson, and D. Shaw. "Seamlessly selecting the best copy from Internet-wide replicated web servers." Mimeograph, 1998.

Armstrong, M. "Price discrimination by a many-product firm." Mimeograph, 1997.

ATM Forum. "Traffic management specification version 4.0." ATM Forum Technical Committee, 1996.

Aurrecochea, C., A. Campbell, and L. Hauw. "A survey of QoS architectures." *Multimedia Systems Journal*, Special Issue on QoS Architecture, May 1998.

Baentsch, M., L. Baum, G. Molter, S. Rothkugel, and P. Sturm. "Enhancing the web's infrastructure from caching to replication." *IEEE Internet Computing* 1, no. 2 (1997): 18-27.

Baentsch, M. et al. "Quantifying the overall impact of caching and replication in the web." University of Kaiserslautern, 1997.

Bakos, Y., and E. Brynjolfsson. "Bundling information goods: pricing, profits and efficiency." Conference on Internet Publishing and Beyond: Economics of Digital

Information and Intellectual Property, Cambridge MA, January 23-25 1997.

Bakos, Y., and E. Brynjolfsson. "Aggregation and disaggregation of information goods: implications for bundling, site licensing and micropayment systems." In *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*, ed. D. Hurley, B. Kahin and H. Varian, MIT Press, 1998, in print.

Ballardie, T., P. Francis, and J. Crowcroft. "Core based trees (CBT) an architecture for scalable multicast routing." ACM SIGCOMM, 1993.

Beck, M., and T. Moore. "The Internet2 distributed storage infrastructure project: an architecture for Internet content channels." Third International WWW Caching Workshop, Manchester England, June 1998.

Bennett, J., and H. Zhang. "Hierarchical packet fair queueing algorithms." *IEEE/ACM Transactions on Networking* 5, no. 5 (1997): 675-689.

Berners-Lee, T., L. Masinter, and M. McCahill. "Uniform resource locators (URL)." RFC 1738, 1994.

Berson, S., R. Lindell, and R. Braden. "An architecture for advance reservations in the Internet." Mimeograph, 1998.

Besen, S.M., and S.N. Kirby. "Private copying, appropriability, and optimal copying royalties." *Journal of Law and Economics* 32, no. 2 part 1 (1989): 255-280.

Bestavros, A. "Demand-based document dissemination to reduce traffic and balance load in distributed information systems." IEEE Symposium on Parallel and Distributed Processing, San Antonio TX, October 1995.

Bestavros, A., and C. Cunha. "Server-initiated document dissemination for the WWW." *IEEE Data Engineering Bulletin*, September 1996, 3-11.

Bestavros, A. "WWW traffic reduction and load balancing through server-based caching." *IEEE Concurrency*, Special Issue on Parallel and Distributed Technology, Jan-Mar 1997, 56-67.

Bhattacharjee, S., K.L. Calvert, and E. Zegura. "Self-organizing wide-area network caches." Georgia Institute of Technology, GIT-CC-97/31, 1997.

Billhartz, T., J. B. Cain, E. Farrey-Goudreau, D. Fieg, and S. Batsell. "Performance and resource cost comparisons for the CBT and PIM multicast routing protocols." *IEEE Journal on Selected Areas in Communications* 15, no. 3 (1997).

Blaze, M.A., and R. Alonso. "Dynamic hierarchical caching in large-scale distributed file systems." *12th International Conference on Distributed Computing Systems*, Yokohama Japan, 1992.

Braden, R., D. Clark, and S. Shenker. "Integrated services in the Internet architecture: an overview." RFC 1633, 1994.

Braun, H.-W., and K. Claffy. "Web traffic characterization: an assessment of the impact of caching documents from the NCSA's web server." *Second International World Wide Web (WWW) Conference*, Chicago IL, October 15-17 1994.

Burnstein, M.L. "The economics of tie-in sales." *Review of Economics and Statistics* 42 (1960): 68-73.

Byrd, G.D. "An economic "commons" tragedy for research libraries: scholarly journal publishing and pricing trends." *College & Research Libraries* 51 (1990): 184-195.

Calvert, K., M. Doar, and E. W. Zegura. "Modeling Internet topology." *IEEE Communications Magazine*, June 1997.

Cao, P., and S. Irani. "Cost-aware WWW proxy caching algorithms." *USENIX Symposium on Internet Technologies and Systems*, 1997.

Cao, P., J. Zhang, and K. Beach. "Active cache: caching dynamic contents on the web." *Middleware '98*, 1998.

Carbajo, J., D. De Meza, and D.J. Seidmann. "A strategic motivation for commodity bundling." *Journal of Industrial Economics* 38 (1990): 283-298.

Casner, S., and S. Deering. "First IETF internet audiocast." *ACM Computer Communication Review*, July 1992, 92-97.

Chae, S. "Bundling subscription TV channels: a case of natural bundling." *International Journal of Industrial Organization* 10 (1992): 213-230.

Chankhunthod, A., P.B. Danzig, C. Neerdaels, M.F. Schwartz, and K. Worrell. "A hierarchical internet object cache." *University of Southern California*, 95-611, 1995.

Chao, H.P., and R. Wilson. "Priority services: pricing, investment, and market organization." *American Economic Review* 77, no. 5 (1987): 899-916.

Chuang, J.C.-I., and M.A. Sirbu. "Optimal bundling strategy of digital information goods: network delivery of articles and subscriptions." Conference on Internet Publishing and Beyond: Economics of Digital Information and Intellectual Property, Cambridge MA, January 23-25 1997. In *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*, ed. D. Hurley, B. Kahin and H. Varian, MIT Press, in print.

Chuang, J.C.-I., and M.A. Sirbu. "Pricing multicast communication: a cost-based approach." Internet Society INET'98, Geneva Switzerland, July 1998.

Cisco. "NetFlow." 1997.

Clark, D.D., and D.L. Tennenhouse. "Architectural considerations for a new generation of protocols." ACM SIGCOMM, 1990.

Clark, D., S. Shenker, and L. Zhang. "Supporting real-time applications in an integrated services packet network: architecture and mechanism." ACM SIGCOMM, 1992.

Clark, D. "Adding service discrimination to the Internet." *Telecommunications Policy* 20, no. 3 (1996): 169-181.

Clark, D. "Internet cost allocation and pricing." In *Internet Economics*, ed. L. McKnight and J. Bailey, 215-252: MIT Press, 1997.

Cocchi, R., D. Estrin, S. Shenker, and L. Zhang. "A study of priority pricing in multiple service class networks." ACM SIGCOMM, 1991.

Cocchi, R., S. Shenker, D. Estrin, and L. Zhang. "Pricing in computer networks: motivation, formulation and example." *IEEE/ACM Transactions on Networking* 1, no. 6 (1993): 614-627.

Daniel, R., and M. Mealling. "Resolution of uniform resource identifiers using the domain name system." Work in progress, 1997.

Danzig, P.B., R.S. Hall, and M.F. Schwartz. "A case for caching file objects inside internetworks." ACM SIGCOMM, 1993.

Danzig, P.B., D. Delucia, and K. Obraczka. "Massively replicating services in wide-area internetworks." University of Southern California, 1994.

de Veciana, G., and R. Baldick. "Resource allocation in multi-service networks via pricing: statistical multiplexing." *Computer Networks and ISDN Systems* 30 (1998): 951-962.

Deering, S. "Host extensions for IP multicasting." RFC 1112, 1989.

Deering, S., and R. Hinden. "Internet protocol, version 6 (IPv6) specification." RFC 1883, 1995.

Deering, S., D. Estrin, D. Farinacci, V. Jacobson, C.-G. Liu, and L. Wei. "The PIM architecture for wide-area multicast routing." *IEEE/ACM Transactions on Networking* 4, no. 2 (1996): 153-162.

Demers, A., S. Keshav, and S. Shenker. "Analysis and simulation of a fair queueing algorithm." *Journal of Internetworking: Research and Experience*, October 1990, 3-26.

Diffserv. "An architecture for differentiated services." IETF Diffserv Working Group, Work in progress, 1998.

Dijkstra, E.N. "A note on two problems in connection with graphs." *Numerical Math* 1 (1959): 269-271.

Doar, M., and I. Leslie. "How bad is naive multicast routing?" IEEE INFOCOM, San Francisco CA, 1993.

Doar, M. "A Better Model for Generating Test Networks." Globecom, 1996.

Drexler, K.E., and M.S. Miller. "Incentive engineering for computational resource management." In *The Ecology of Computation*, ed. B.A. Huberman: Elsevier Science Publishers B.V. (North-Holland), 1988.

Dyl, E.A. "A note on price discrimination by academic journals." *Library Quarterly* 53, no. 2 (1983): 161-168.

Economist. "The death of distance." *The Economist*, September 30 1995, 64-.

Economist. "Is this the end of sticky prices?" *The Economist*, May 16th 1998, 86.

Edell, R.J., N. McKeown, and P.P. Varaiya. "Billing users and pricing for TCP." *IEEE Journal on Selected Areas in Communications* 13, no. 7 (1995): 1162-1175.

Eriksson, H. "MBone: the multicast backbone." *Communications of the ACM* 37, no. 8 (1994): 54-60.

Estrin, D., and L. Zhang. "Design considerations for usage accounting and feedback in internetworks." *Computer Communications Review* 20, no. 5 (1990): 56-66.

Estrin, D., M. Handley, S. Kumar, and D. Thaler. "The multicast address set claim (MASC) protocol." Work in progress, 1997.

Fan, L., P. Cao, J. Almeida, and A.Z. Broder. "Summary cache: a scalable wide-area web cache sharing protocol." ACM SIGCOMM, 1998.

Fenner, W. "Internet group management protocol (IGMP), version 2." RFC 2236, 1997.

Ferguson, D., C. Nikolaou, and Y. Yemini. "An economy for managing replicated data in autonomous decentralized systems." International Symposium on Autonomous and Decentralized Systems, 1993.

Ferrari, D., and D.C. Verma. "A scheme for real-time channel establishment in wide-area networks." *IEEE Journal on Selected Areas in Communications* 8, no. 3 (1990): 368-379.

Ferrari, D. "Client requirements for real-time communication services." *IEEE Communications Magazine* 28, no. 11 (1990): 65-72.

Fishburn, P.C., A.M. Odlyzko, and R.C. Siders. "Fixed fee versus unit pricing for information goods: competition, equilibria, and price wars." *First Monday* 2, no. 7 (1997). In *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property*, ed. D. Hurley, B. Kahin and H. Varian, MIT Press, in print.

Floyd, S., and V. Jacobson. "Random early detection gateways for congestion avoidance." *IEEE/ACM Transaction on Networking* 1, no. 4 (1993): 397-413.

Floyd, S., and V. Jacobson. "Link-sharing and resource management models for packet networks." *IEEE/ACM Transactions on Network* 3, no. 4 (1995).

Fry, C.L., J.M. Griffin, D.R. House, and T.R. Saving. "The economics of structural separation from the perspective of economic efficiency." Final Report prepared for US WEST Communications by RRC, Inc., April 4 1995.

Guérin, R., H. Ahmadi, and M. Naghshineh. "Equivalent capacity and its application to bandwidth allocation in high-speed networks." *IEEE Journal on Selected Areas in Communications* 9, no. 7 (1991): 968-981.

Guérin, R., and V. Peris. "Quality-of-service in packet networks: basic mechanisms and directions." IBM, RC 21089, 1998.

Gupta, A., D.O. Stahl, and A.B. Whinston. "Priority pricing of integrated services networks." In *Internet Economics*, ed. L. McKnight and J. Bailey, 323-352: MIT Press, 1997.

Guyton, J., and M. Schwartz. "Locating nearby copies of replicated Internet servers." ACM SIGCOMM, 1995.

Gwertzman, J., and M. Seltzer. "The case for geographical push-caching." 5th Annual Workshop on Hot Operating Systems, May 1995.

Gwertzman, J., and M. Seltzer. "Autonomous Replication Across Wide-Area Internetworks." *SOSP* (1995): 234-.

Hakimi, S.L. "Optimum locations of switching centers and the absolute centers and medians of a graph." *Operations Research* 12 (1964): 450-459.

Hakimi, S.L. "Optimum distribution of switching centers in a communication network and some related graph theoretic problems." *Operations Research* 13 (1965): 462-475.

Halpern, J., and O. Maimon. "Algorithms for the m-center problems: a survey." *European Journal of Operation Research* 10 (1982): 90-99.

Hanson, W., and R.K. Martin. "Optimal bundle pricing." *Management Science* 36, no. 2 (1990): 155-174.

Hausman, J.A., and T.J. Tardiff. "Costs and benefits of vertical integration of basic and enhanced telecommunications services." Mimeograph, March 29 1995.

Heddaya, A., S. Mirdad, and D. Yates. "Diffusion-based caching along routing paths." NLANL Web Caching Workshop, Boulder CO, June 1997.

Herzog, S., S. Shenker, and D. Estrin. "Sharing the "cost" of multicast trees: an axiomatic analysis." ACM SIGCOMM, 1995.

Hirsh, D., C. Mills, and G. Ruth. "Internet accounting: background." RFC 1272, 1991.

Honig, M., and K. Steiglitz. "Usage-based pricing and quality of service in data networks." IEEE INFOCOM, 1995.

Huberman, B.A., P.L.T. Pirolli, J.E. Pitkow, and R.M. Lukose. "Strong regularities in world wide web surfing." *Science*, April 3 1998.

Hui, J.Y. "Resource allocation for broadband networks." *IEEE Journal on Selected Areas in Communications* 6, no. 9 (1988): 1598-1608.

Hyman, J.M., A.A. Lazar, and G. Pacifici. "A separation principle between scheduling and admission control for broadband switching." *IEEE Journal on Selected Areas in Communications* 11, no. 4 (1993): 605-616.

INI. "Development plan for an electronic library system - Final Report." Information Networking Institute, Carnegie Mellon University, 1990-1, 1990.

Inktomi. "Traffic server's compatibility with advertising and dynamic content." 1998.

Inktomi. "Reverse proxy caching with traffic server: the benefits of caching to web hosting providers." 1998.

Iyengar, A., and J. Challenger. "Improving Web server performance by caching dynamic data." USENIX Symposium on Internet Technologies and Systems, 1997.

Jamin, S., P. Danzig, S. Shenker, and L. Zhang. "A measurement-based admission control algorithm for integrated service packet networks." *IEEE/ACM Transactions on Networking* 5, no. 1 (1997): 56-70.

Joyce, P., and T.E. Merz. "Price discrimination in academic journals." *Library Quarterly* 55, no. 3 (1985): 273-283.

Kantor, B., and P. Lapsley. "Network News Transfer Protocol." RFC 977, 1986.

Kariv, O., and S.L. Hakimi. "An algorithmic approach to network location problems, I: the p-centers." *SIAM Journal of Applied Mathematics* 37 (1979): 513-538.

Kelly, F. "Effective bandwidth at multi-class queues." *Queueing Systems* 9 (1991): 5-16.

Kelly, F.P. "Charging and accounting for bursty connections." In *Internet Economics*, ed. L. McKnight and J. Bailey, 253-278: MIT Press, 1997.

Keshav, S. "On the efficient implementation of fair queueing." *Internetworking: Research and Experiences* 2 (1991): 157-173.

King, D.W., D.D. McDonald, and N.K. Roderer. *Scientific journals in the United States: their production, use, and economics*. Stroudsburg PA: Hutchinson Ross Publishing, 1981.

King, D.W., and J.M. Griffiths. "Economic issues concerning electronic publishing and distribution of scholarly articles." *Library Trends* 43, no. 4 (1995): 713-740.

Kistler, J.J., and M. Satyanarayanan. "Disconnected operation in the Coda file system."

ACM Transactions on Computer Systems 10, no. 1 (1992): 3-25.

Kroeger, T.M., D.D.E. Long, and J.C. Mogul. "Exploring the bounds of Web latency reduction from caching and prefetching." *USENIX Symposium on Internet Technologies and Systems*, 1997.

Labbé, M., D. Peeters, and J.-F. Thisse. "Location on networks." In *Network Routing*, ed. M.O. Ball et al., 8: Elsevier Science B.V., 1995.

Laffont, J.-J., E. Maskin, and J.-C. Rochet. "Optimal nonlinear pricing with two-dimensional characteristics." In *Information, Incentives and Economic Mechanisms*, ed. Groves, Radner, and Reiter, 1985.

Lewis, D.W. "Economics of the scholarly journal." *College & Research Libraries* 50 (1989): 674-688.

Lidl, K., J. Osborne, and J. Malcolm. "Drinking from the firehose: multicast USENET news." *USENIX 1994 Winter Conference*, 1994.

Liebowitz, S.J. "Copying and indirect appropriability: photocopying of journals." *Journal of Political Economy* 93, no. 5 (1985): 945-957.

Lorenzetti, P., L. Rizzo, and L. Vicisano. "Replacement policies for a proxy cache." *Universita di Pisa*, 1996.

Low, S.H., and P.P. Varaiya. "A new approach to service provisioning in ATM networks." *IEEE/ACM Transactions on Networking* 1, no. 5 (1993): 547-553.

Luotenen, A., and K. Altis. "World-wide web proxies." *1st International Conference on the WWW*, May 1994.

Mackie-Mason, J.K., and H. Varian. "Pricing congestible network resources." *IEEE Journal on Selected Areas in Communications* 13, no. 7 (1995): 1141-1149.

Mackie-Mason, J.K., and K. White. "Evaluating and selecting digital payment mechanisms." *Telecommunications Policy Research Conference*, 1996.

Mackie-Mason, J.F., S. Shenker, and H. Varian. "Network architecture and content provision: an economic analysis." *Mimeograph*, 1996.

Mackie-Mason, J.K., and J. Riveros. "Economics and electronic access to scholarly information." *Conference on Internet Publishing and Beyond: Economics of Digital Information and Intellectual Property*, Cambridge MA, January 23-25 1997. In *Internet*

Publishing and Beyond: The Economics of Digital Information and Intellectual Property, ed. D. Hurley, B. Kahin and H. Varian, MIT Press, in print.

Mahdavi, J. "Personal communication." , November 1997.

Malpani, R., J. Lorch, and D. Berger. "Making world wide web caching servers cooperate." Fourth International World Wide Web Conference, Boston MA, December 1995.

Manasse, M.S. "The Millicent protocols for electronic commerce." First USENIX Workshop on Electronic Commerce, 1995.

Markatos, E., and C. Chronaki. "A top-10 approach to prefetching the web." Internet Society INET'98, Geneva Switzerland, July 1998.

McAfee, R.P., J. McMillan, and M.D. Whinston. "Multiproduct monopoly, commodity bundling, and correlation of values." *Quarterly Journal of Economics* 104 (1989): 371-383.

Metz, P., and P.M. Gherman. "Serial pricing and the role of the electronic journal." *College & Research Libraries* 52 (1991): 315-327.

Michel, S., K. Nyugen, A. Rosenstein, L. Zhang, S. Floyd, and V. Jacobson. "Adaptive web caching: towards a new global caching architecture." Third International WWW Caching Workshop, Manchester England, June 1998.

Mitchell, B.M., and I. Vogelsang. *Telecommunications pricing: theory and practice*. Cambridge University Press, 1991.

Moy, J. "Multicast extensions for OSPF." RFC 1584, 1994.

Nagle, D. et. al. "Active networking for storage: exploiting active networks for network-attached storage." Carnegie Mellon University, Proposal to DARPA BAA 98-03, 1998.

Narayan, S., P. Losleben, and F.-C. Cheong. "A market-based economic model for multi-media object storage and distribution." International Workshop on Multi-Media Data Base Management Systems, August 1995.

Obraczka, K. "Massively replicating services in wide-area internetworks." Ph.D. dissertation, University of Southern California, 1994.

Obraczka, K. "Multicast transport protocols: a survey and taxonomy." *IEEE Communications Magazine* 36, no. 1 (1998): 94-102.

Ohnishi, H., T. Okada, and K. Noguchi. "Flow control schemes and delay/loss tradeoff in ATM networks." *IEEE Journal on Selected Areas in Communications* 6, no. 9 (1988): 1609-1616.

Okerson, A.S., and J.J. O'Donnell, eds. *Scholarly journals at the crossroads: a subversive proposal for electronic publishing*. Washington DC: Association of Research Libraries, 1995.

Ordoover, J.A., and R.D. Willig. "On the optimal provision of journals qua sometimes shared goods." *American Economic Review* 68 (1978): 324-338.

Parekh, A.K. "A generalized processor sharing approach to flow control in integrated services networks." Ph.D. dissertation, Massachusetts Institute of Technology, 1992.

Parekh, A., and R.G. Gallager. "A generalized processor sharing approach to flow control in integrated service network - the single node case." *ACM/IEEE Transactions on Networking* 1, no. 3 (1993): 344-357.

Parekh, A., and R.G. Gallager. "A generalized processor sharing approach to flow control in integrated service network - the multiple node case." *ACM/IEEE Transactions on Networking* 2, no. 2 (1994): 137-150.

Partridge, C., T. Mendez, and W. Milliken. "Host anycasting service." RFC 1546, 1993.

Pejhan, S., M. Schwartz, and D. Anastassiou. "Error control using retransmission schemes in multicast transport protocols for real-time media." *IEEE/ACM Transactions on Networking* 4, no. 3 (1996): 413-427.

Plaxton, C.G., R. Rajaraman, and A.W. Richa. "Accessing nearby copies of replicated objects in a distributed environment." Department of Computer Science, University of Texas at Austin, TR-97-11, 1997.

Rekhter, Y., and T. Li. "An architecture for IP address allocation with CIDR." RFC 1518, 1993.

Rekhter, Y., and T. Li. "Implications of various address allocation policies for Internet routing." RFC 2008, 1996.

Rekhter, Y., P. Resnick, and S. Bellovin. "Financial incentives for route aggregation and efficient utilization in the Internet." Telecommunications Policy Research Conference, Solomons MD, 1996.

Rendleman, J. "Reducing web latency -- Stanford University tries web hosting to boost 'net access." *Communications Week*, June 30 1997, 9-.

Ruth, G.R. "Usage accounting for the Internet." Internet Society INET'97, Kuala Lumpur Malaysia, 1997.

Sairamesh, J., D.F. Ferguson, and Y. Yemini. "An approach to pricing, optimal allocation and quality of service provisioning in high-speed packet networks." IEEE INFOCOM, 1995.

Salama, H.F., D.S. Reeves, and Y. Viniotis. "Evaluation of multicast routing algorithms for real-time communication on high-speed networks." *IEEE Journal on Selected Areas in Communications* 15, no. 3 (1997): 332-345.

Saltzer, J.H., D.P. Reed, and D.D. Clark. "End-to-end arguments in system design." *ACM Transactions on Computer Systems* (1984).

Schelen, O., and S. Pink. "Resource sharing in advance reservation agents." Mimeograph, 1998.

Schill, A. "Migration, caching and replication in distributed object-oriented systems: an integrated framework." *IFIP Transactions C (Communication Systems) C-6* (1992): 309-329.

Schmalensee, R. "Gaussian demand and commodity bundling." *Journal of Business* 57, no. 1 part 2 (1984): S211-230.

Sedgewick, R. *Algorithms in C*. 3rd ed. Reading MA: Addison Wesley, 1998.

Shapiro, C., and H.R. Varian. *Information rules: a strategic guide to the network economy*. Cambridge MA: Harvard Business School Press, 1998.

Shenker, S. "Service models and pricing policies for an integrated services Internet." In *Public Access to the Internet*, ed. B. Kahin and J. Keller, 315-337: MIT Press, 1995.

Shenker, S., D. Clark, D. Estrin, and S. Herzog. "Pricing in computer networks: reshaping the research agenda." *Telecommunications Policy* 20, no. 3 (1996).

Shenker, S., C. Partridge, and R. Guérin. "Specification of guaranteed quality of service." RFC 2212, 1997.

Shenker, S., and J. Wroclawski. "General characterization parameters for integrated service network elements." RFC 2215, 1997.

Sidgmore, J. "The electronic future." Telecommunications Managers Association Conference, Brighton, UK, October 5-7 1998.

Sirbu, M.A., and J.D. Tygar. "NetBill: an electronic commerce system optimized for network delivered services." *IEEE Personal Communications*, August 1995.

Sirbu, M.A. "Credits and debits on the Internet." *IEEE Spectrum* 34, no. 2 (1997): 23-29.

Sollins, K., and L. Masinter. "Functional requirements for uniform resource names." RFC 1737, 1994.

Sollins, K. "Architectural principles of uniform resource name resolution." RFC 2276, 1998.

Songhurst, D., and F. Kelly. "Charging schemes for multiservice networks." International Teletraffic Congress 15, 1997.

Spigai, F. "Information pricing." *Annual Review of Information Science and Technology* 26 (1991): 39-73.

Spragins, J.D., J.L. Hammond, and K. Pawlikowski. *Telecommunications protocol and design*. Addison Wesley, 1991.

Stigler, G.J. "United States v. Loew's Inc.: a note on block booking." *Supreme Court Review* (1963): 152-157.

Stoller, M.A., R. Christopherson, and M. Miranda. "The economics of professional journal pricing." *College & Research Libraries* 57 (1996): 9-21.

Stonebraker, M. et al. "An economic paradigm for query processing and data migration in Mariposa." Third International Conference on Parallel and Distributed Information Systems, September 1994.

Thaler, D., D. Estrin, and D. Meyer. "Border Gateway Multicast Protocol (BGMP): protocol specification." *Work in progress*, 1997.

UUNET. "UUNET announces multicast service for mass Internet broadcasting." Press release, September 23 1997.

van Steen, M., F.J. Hauck, and A.S. Tanenbaum. "A model for worldwide tracking of distributed objects." TINA'96, September 1996.

van Steen, M., F.J. Hauck, P. Homburg, and A.S. Tanenbaum. "Locating objects in wide-area systems." *IEEE Communications Magazine*, January 1998, 104-109.

Varaiya, P.P., R. Edell, and H. Chand. "INDEX: the Internet demand experiment." , 1998.

Varian, H.R. "Pricing information goods." Scholarship in the New Information Environment Symposium, Harvard Law School, 1995.

Waitzman, D., C. Partridge, and S. Deering. "Distance vector multicast routing protocol (DVMRP)." RFC 1075, 1988.

Wang, Z., and J. Crowcroft. "Prefetching in the world wide web." *IEEE Global Internet*, London UK, November 1996.

Wang, Z., and J. Crowcroft. "Cachemesh: A distributed cache system for world wide web." NLANL Web Caching Workshop, Boulder CO, June 1997.

Wang, Q., J. Peha, and M.A. Sirbu. "Optimal pricing for integrated services networks." In *Internet Economics*, ed. L. McKnight and J. Bailey, 353-378: MIT Press, 1997.

Wei, L., and D. Estrin. "The trade-offs of multicast trees and algorithms." ICCCN, 1994.

Wei, L., and D. Estrin. "Multicast routing in dense and sparse modes: simulation study of tradeoffs and dynamics." University of Southern California Computer Science Department, 95-613, 1995.

Wessels, D., and K. Claffy. "Internet Cache Protocol (ICP), version 2." RFC 2186, 1997.

Wessels, D., and K. Claffy. "ICP and the Squid web cache." *IEEE Journal on Selected Areas in Communications* 16, no. 3 (1998): 345-357.

Whinston, M.D. "Tying, foreclosure, and exclusion." *American Economic Review* 80 (1990): 837-859.

Williams, S., M. Abrams, C.R. Standridge, G. Abdulla, and E.A. Fox. "Removal policies in network caches for world-wide web documents." ACM SIGCOMM, 1996.

Williamson, O.E. "The modern corporation: origins, evolution, attributes." *Journal of Economic Literature* (1981): 1537-1568.

Willig, R.D. "Pareto-superior nonlinear outlay schedules." *Bell Journal of Economics* 9 (1978): 56-69.

Wilson, R. *Nonlinear pricing*. Oxford University Press, 1993.

Wolfson, O., S. Jajodia, and Y. Huang. "An adaptive data replication algorithm." *ACM Transactions on Database Systems* 22, no. 2 (1997): 255-314.

Wroclawski, J. "Specification of the controlled-load network element service." RFC 2211, 1997.

Zahray, W.P., and M.A. Sirbu. "The provision of scholarly journals by libraries via electronic technologies: an economic analysis." *Information Economics and Policy* 4 (1990): 127-154.

Zegura, E.W., K. Calvert, and S. Bhattacharjee. "How to model an internetwork." IEEE Infocom, San Francisco CA, 1996.

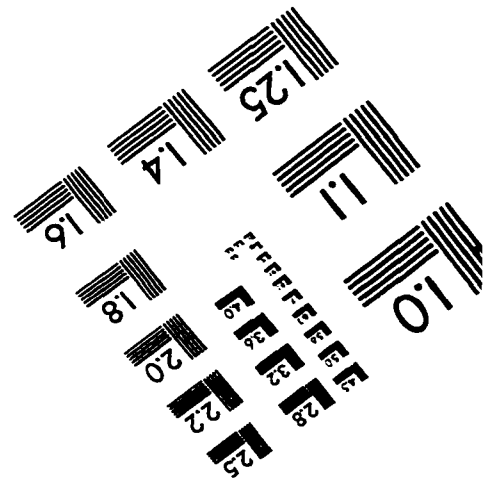
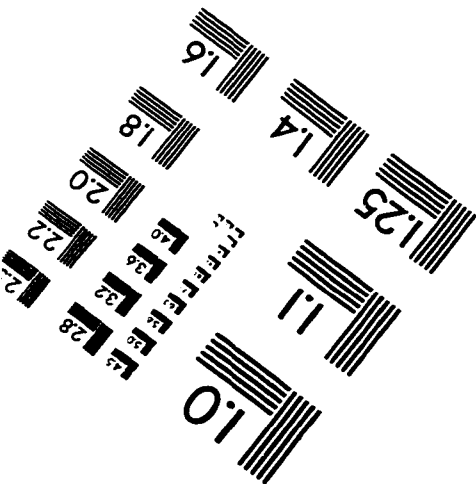
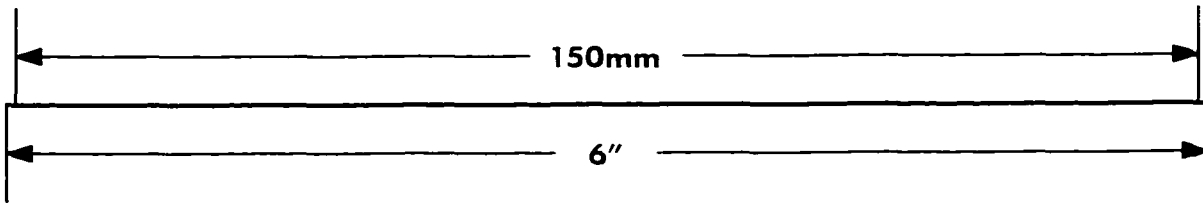
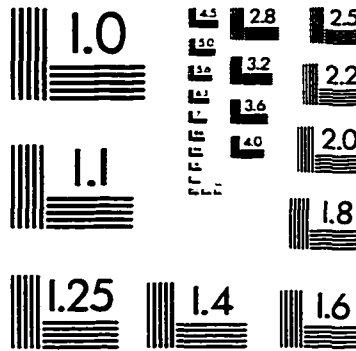
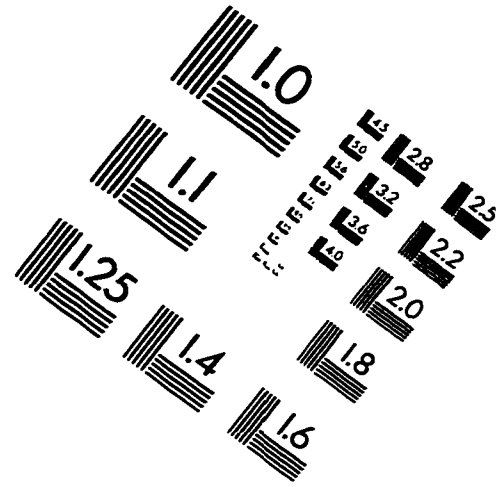
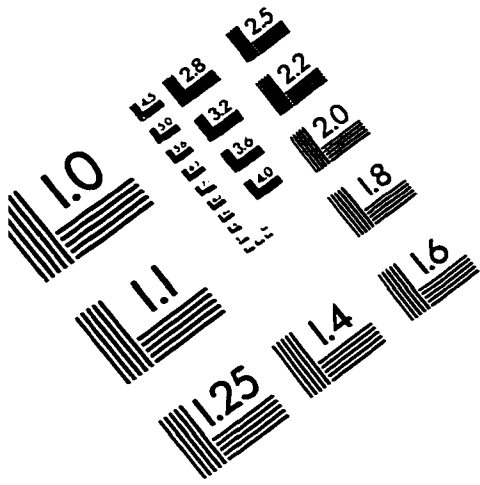
Zhang, H. "Service disciplines for guaranteed performance service in packet-switching networks." *Proceedings of the IEEE* 83, no. 10 (1995): 1374-1399.

Zhang, L., S. Deering, D. Estrin, S. Shenker, and D. Zappala. "RSVP: a new resource ReSerVation Protocol." *IEEE Network* 7, no. 5 (1993): 8-18.

Zhang, L., S. Floyd, and V. Jacobson. "Adaptive Web Caching." Initial proposal, 1997.

Zipf, G.K. *Human behavior and the principle of least effort*. Cambridge MA: Addison-Wesley, 1949.

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc.
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc.. All Rights Reserved