Integrated Digital Services for Cable Networks

by

David Gingold


Submitted to the Technology and Policy Program
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Technology and Policy

at the

Massachusetts Institute of Technology

September 1996

© 1996 Massachusetts Institute of Technology
All rights reserved
Integrated Digital Services for Cable Networks

by

David Gingold
A.B. Physics, Harvard College, 1989

Submitted to the Technology and Policy Program
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Technology and Policy

ABSTRACT

The telephone and cable television networks that provide local-access communication to our homes, schools, and businesses were each designed with a specific application in mind. Modern communication technology, however, allows different applications to converge onto an underlying digital medium. We investigate how such a convergence can happen on cable television networks.

An integrated services digital cable architecture, one that is designed to deliver a variety of communication services and to support both current and future applications, will offer economic benefits both to cable operators and to cable consumers. Modern hybrid fiber/coax cable networks provide a high-capacity communication medium that is suitable for a wide range of communication applications. The economic problem of delivering new services over these networks, however, is characterized by substantial infrastructure investments that must be recovered by selling into competitive, uncertain markets. An integrated services architecture will allow cable operators to respond to changing demand and technology without requiring new investment at every step. But the need for flexibility and efficiency requires addressing specific technical challenges, particularly supporting quality of service in a system with multiple services and heterogeneous communication needs. Systems that solve these problems stand to serve both cable operators, by providing strategic advantages in selling services, and cable users, by promoting the development of future communication applications.

Thesis Supervisor: Dr. Lee W. McKnight, Lecturer, Technology and Policy Program.
ACKNOWLEDGMENTS

Many people provided ideas, advice, insight, and support that contributed to this work. I am especially thankful:

- To my advisor, Lee McKnight, for many times over clearing a path to the finish line;

- To the researchers at the MIT Research Program on Communications Policy: David Carver, Branko Gerovac, Petros Kavassalis, Suzanne Neil, and Richard Solomon, who are full of ideas and inspiration, and who have been a joy to work with;

- To Dave Clark and John Wroclawski at the MIT Laboratory for Computer Science, for sponsoring this work, for lending their expertise, and for arguing with me until I was sufficiently inspired to write something;

- To David Tennenhouse at LCS and the Sloan School, for his reassurance and advice;

- To Jim Mills at MIT LIDS, for keeping me in touch with the industry and for all of those 802.14 documents;

- To Sharon Gillett, who has provided footsteps for me to follow in too many times to count;

- To Chris Avery at the Kennedy School, who probably does not realize how entirely helpful he was;

- To the RPCP graduate students: Joe Bailey, Paul Bosco, Bruce Haimowitz, Tom Lee, Brett Leida, Doug Melcher, Joe Reagle, Jae Roh, Russell Rothstein, and Andrew Sears, who have contributed top-quality expertise, who have been great fun to work with, and to whom I wish future success; and

- To the RPCP administrators, Gill Cable-Murphy and Julia Malik, who keep us in touch with reality when they can, and nonetheless cheerfully tolerate us when they cannot.

______________________________

Above all, I am thankful to my friends and to my family for their persistent love and support, without which nothing would be possible.

______________________________

This research was supported by the Defense Advanced Research Projects Agency of the Department of Defense under contract no. DABT63-94-C-0072, administered by Ft. Huachuca; by DARPA through the grant, “Interface Technologies for Networked Computing,” no. N00174-93-0036; and by the National Science Foundation through the grant, “Networked Multimedia Information Services,” no. NCR-9307548. This material does not reflect the position or policy of the U.S. government, and no official endorsement should be inferred.
# Contents

LIST OF ILLUSTRATIONS .................................................................................................................. 11

LIST OF TABLES .................................................................................................................................. 11

1. INTRODUCTION .............................................................................................................................. 13

   1.1 METHODS ................................................................................................................................... 13

   1.2 THESIS OVERVIEW .................................................................................................................. 14

   1.3 RELATED WORK ........................................................................................................................ 15

      1.3.1 Cable Technology and Digital Services ............................................................................. 15

      1.3.2 Economics ........................................................................................................................ 17

      1.3.3 Integrated Services ........................................................................................................... 17

      1.3.4 Current Standardization Efforts ......................................................................................... 18

   1.4 THE SCOPE OF OUR WORK ....................................................................................................... 18

   1.5 THE INTEGRATED SERVICES APPROACH ............................................................................. 20

2. DIGITAL CABLE TECHNOLOGY AND APPLICATIONS .................................................................. 23

   2.1 THE CABLE MEDIUM ............................................................................................................... 23

   2.2 CABLE TELEVISION TRANSMISSION .................................................................................... 26

   2.3 THE CABLE INFRASTRUCTURE ............................................................................................... 27

   2.4 HYBRID FIBER/COAX SYSTEMS ............................................................................................. 30

   2.5 TWO-WAY TRANSMISSION ..................................................................................................... 32

   2.6 CARRYING DIGITAL SIGNALS ................................................................................................ 34

   2.7 THE CABLE NETWORK AS A COMMUNICATION MEDIUM .................................................. 35

      2.7.1 Topology .......................................................................................................................... 35

      2.7.2 Capacity ........................................................................................................................... 36

      2.7.3 Access .............................................................................................................................. 37

      2.7.4 Latency ............................................................................................................................. 37

      2.7.5 Reliability ......................................................................................................................... 38

   2.8 CABLE APPLICATIONS AND SERVICES .............................................................................. 39

      2.8.1 Applications ..................................................................................................................... 41

   2.9 CONCLUSION ............................................................................................................................ 45

3. CABLE ECONOMICS IN AN INTEGRATED SERVICES ENVIRONMENT ...................................... 47

   3.1 THE ECONOMIC POSITION OF CABLE TELEVISION OPERATORS TODAY ....................... 47

      3.1.1 Cable Television Monopoly Power ................................................................................... 48

   3.2 CABLE REGULATION IN THE UNITED STATES .................................................................. 50

7
3.3 THE PROMISE OF NEW SERVICES ................................................................. 53
3.4 THE COSTS OF DELIVERING NEW SERVICES ............................................. 54
  3.4.1 The Two-Way Hybrid Fiber/Coax Network Architecture ...................... 55
  3.4.2 Categorizing the Costs of New Services ........................................ 57
  3.4.3 Estimating the Value of Bandwidth on a Cable System ..................... 60
3.5 COMPETITION IN A MULTIPLE SERVICES ENVIRONMENT ..................... 64
  3.5.1 Making the Transition to Multiple Services ................................... 65
  3.5.2 Bertrand Competition ................................................................. 66
  3.5.3 Strategies for Competitive Services ........................................... 68
  3.5.4 The Limits of Modeling Competitive Strategies ............................. 72
3.6 THE ECONOMIC IMPLICATIONS OF INTEGRATED SERVICES FOR CABLE ........ 73
  3.6.1 The Benefits of Integrated Services ........................................... 73
  3.6.2 Design Principles for a Digital Cable Architecture ........................ 74
3.7 CONCLUSION ....................................................................................... 76

4. DIGITAL CABLE ARCHITECTURAL CONSIDERATIONS ............................... 77

4.1 WHY DEFINE AN ARCHITECTURE? ....................................................... 77
4.2 CENTRALIZED VS. DISTRIBUTED CONTROL .......................................... 79
4.3 ARCHITECTURE ISSUES FOR THE PHYSICAL LAYER ......................... 82
  4.3.1 Upstream vs. Downstream Transmission ....................................... 82
  4.3.2 Cable Spectrum Use .................................................................. 83
  4.3.3 Multiplexing ................................................................. 84
  4.3.4 Modulation ...................................................................... 85
4.4 ARCHITECTURE ISSUES FOR THE MEDIA ACCESS CONTROL LAYER .......... 87
4.5 SYSTEM-WIDE ARCHITECTURAL CHALLENGES ................................. 90
  4.5.1 Security and Privacy ............................................................. 90
  4.5.2 Plug and Play .................................................................... 94
  4.5.3 Reliability .................................................................. 95
4.6 CONCLUSION ....................................................................................... 96

5. QUALITY OF SERVICE ........................................................................... 99

5.1 SERVICE QUALITY EXPECTATIONS IN THE REAL WORLD: THE EXAMPLE OF TELEPHONY 101
5.2 AN ABSTRACT MODEL OF QoS ............................................................ 104
  5.2.1 Demand When Time Does Not Matter ....................................... 105
  5.2.2 The Temporal Problem .......................................................... 107
  5.2.3 Expectations and Commitments ............................................. 109
  5.2.4 Uncertain Demands ............................................................. 112
5.3 SOLUTIONS FOR QUALITY OF SERVICE .......................................... 113
5.3.1 Integrated Services for the Internet
5.3.2 ATM Services
5.4 CONCLUSION

6. CONCLUSION
6.1 TECHNOLOGY
6.2 ECONOMICS
6.3 ARCHITECTURE
6.4 QUALITY OF SERVICE
6.5 FUTURE WORK
   6.5.1 The Need for Standards
   6.5.2 How Will Cable Benefit Users?

BIBLIOGRAPHY
List of Illustrations

FIGURE 1. COAXIAL CABLE ................................................................. 24
FIGURE 2. CABLE TELEVISION NETWORK INFRASTRUCTURE ....................... 28
FIGURE 3. HYBRID FIBER/COAX NETWORK ......................................... 31
FIGURE 4. TWO-WAY AMPLIFIER ....................................................... 33
FIGURE 5. CABLE SUBSCRIPTIONS IN THE UNITED STATES ......................... 48
FIGURE 6. ESTIMATED CURRENT MARKET SIZES FOR SERVICES ................ 54
FIGURE 7. DISTRIBUTED ARCHITECTURE SIMILAR TO THE ZENITH CABLE MODEM SYSTEM .......................... 80
FIGURE 8. CENTRALLY CONTROLLED ARCHITECTURE WITH HEAD END CONTROLLER ......................... 81

List of Tables

TABLE 1. TECHNICAL CHARACTERISTICS OF APPLICATIONS FOR CABLE ........................................... 45
TABLE 2. U.S. CABLE REGULATION MILESTONES ........................................... 52
TABLE 3. REED'S COST ESTIMATES FOR CONSTRUCTING A HYBRID FIBER/COAX TELEVISION NETWORK ............ 56
TABLE 4. EXAMPLE OF TIME DEPENDENT UTILITY WITH CORRELATED DEMANDS ........................................ 108
TABLE 5. EXAMPLE OF UTILITY AS A FUNCTION OF DECISIONS AND ALLOCATION ...................................... 110
TABLE 6. EXAMPLE OF UTILITY WITH PRICES ............................................ 111
1. Introduction

Communication systems that provide local access—connecting homes, schools, and small businesses—have been designed historically each with a specific application in mind. Telephone networks were built to provide telephone service, and cable television networks were built to provide television service. Modern communication technology, however, promises that these different applications can converge onto an underlying digital medium that does not distinguish one from another. We explore how such convergence can happen on cable television networks. In particular, we consider how a communication architecture can support a variety of communication services, for both current and future applications, carried digitally on modern cable television infrastructure.

The challenge of developing systems that will deliver integrated services over cable is significant. The state of digital cable technology today is such that we cannot fully predict how systems will be built, how they will perform, and how much they will cost, any more than we can predict with certainty the market for applications and services that do not yet exist. Solving just the engineering aspects—designing a system that could deliver integrated services—is a substantial challenge that is being taken on by manufacturers of digital cable systems. But a solution which demonstrates that this can be done does not itself answer the question of how it best should be done. We argue that the best approach is to design systems intended to flexibly and efficiently deliver multiple services over cable networks, and that these systems should in particular provide quality of service for communication in the face of changing applications and demands for services. In the absence of a concrete technical model it is impossible to reach a definitive conclusion about what sort of system would be best. Nonetheless, a careful analysis that ties together the technology, economics, and policy of digital cable systems should provide direction for engineering these systems and insight into associated regulatory and policy problems.

Our thesis seeks to validate this hypothesis: A digital cable architecture that delivers integrated services while satisfying quality-of-service constraints will provide economic benefits to cable operators and users.

1.1 Methods

Our argument that an integrated services approach is beneficial to cable operators and users rests on an economic analysis of the cable communication business, seen in the
light of the opportunities that digital cable technology provides. Cable companies who wish to deliver new services over their networks face a competitive environment in which neither the future demands for services nor the directions of technologies are well known. In the economy of operating a cable system, the burden of recovering sunk infrastructure investments rests entirely on the ability to sell services to consumers. The result, we argue, is that the ability to flexibly meet consumer demands for services, demands that could change dramatically over the next few years, can benefit both cable operators and consumers. This has specific implications for how a digital cable system should be designed.

To assess the technical challenges of building an integrated services digital cable system, we look both generally at the technology and specifically at a particular technical problem. Digital cable systems are not a new technology, but they have not yet found widespread use on cable networks despite many trial systems and much media hype. The path to making systems that can deliver mass-market services requires confronting a number of technical challenges, from the low-level problem of dealing with noise characteristics of these systems to the high-level problem of making equipment that is easily installed and user-friendly.

A particular technical problem that arises in the design of these systems is the need to efficiently share the network resources between users while nonetheless providing for specific connections to maintain a quality of service that is appropriate for their applications. We explore this problem in some detail. Even simple applications that we expect the cable network to be able to provide can have very complicated quality-of-service demands. We model quality-of-service demands in an economic sense, and we will examine how different technical systems attempt to define quality-of-service characteristics for digital networks. We also consider how the way a cable system defines quality of service can affect interoperability with other networks.

1.2 Thesis Overview

Chapter 2 presents an overview of digital cable technology and applications. Two-way digital communication on cable networks is not a new notion, but modern hybrid fiber/coax technology enhances the reliability and flexibility of these systems. Cable as a general-purpose communication medium can be understood in terms of a few basic properties of modern cable systems. We also discuss some of the services that might be
delivered over cable systems and consider how their technical needs are suited to cable's capabilities.

Chapter 3 explores the economics of delivering digital services over cable networks. Cable operators traditionally have enjoyed a degree of monopoly power in delivering a single-service product, but this situation changes as new competition arises for television service and as cable operators deploy services that face direct competition from other providers. By examining strategies that cable operators can pursue, we show generally how an integrated services architecture can provide economic advantages for cable operators and specifically how equipment and operation costs, service bundling, and product differentiation can be critical factors in a competitive environment.

Chapter 4 investigates the technical problems of designing an integrated services digital cable architecture. We identify key functionality requirements, particularly the ability to dynamically use resources on the cable system to adapt to changing demands and the ability to ensure quality for different services. We examine briefly how systems presently being developed and standardized can meet these needs.

Chapter 5 focuses in detail on the technical problem of defining quality-of-service commitments for a shared network. We both consider this as an abstract problem of economics and investigate how actual integrated services solutions address the problem.

Chapter 6 brings together conclusions from our work and discusses the importance of standards in designing integrated services cable systems.

1.3 Related Work

Our work draws on a broad range of ideas, particularly from the fields of engineering and economics, that include theoretical work as well as practical implementations and analyses. Here we outline some of the prior work that our research builds upon.

1.3.1 Cable Technology and Digital Services

Two works provide excellent overviews of traditional cable television technology and applications: Baldwin and McVoy's *Cable Communication*,¹ and Ciciora's "Cable Television in the United States: An Overview," published by CableLabs.²

---

²
Reed's *Residential Fiber Optic Networks* provides very detailed engineering and cost analyses of various modern network architectures, including hybrid fiber/coax systems.\(^3\) We rely on this work as a source of much of the cost data used in our analysis.

Using cable systems to deliver digital two-way services is by no means a new idea. A review of older digital cable systems and their technology is given by Karshmer and Thomas.\(^4\)

Three previous theses at MIT analyzed digital cable systems in different ways. Estrin in 1982 (long before the Internet became mainstream technology) investigated the prospects for carrying digital services on cable and the implication of digital services for cable regulation.\(^5\) Feldmeyer in 1986 performed a deeper technical analysis of the problem of transmitting data over cable and proposed mechanisms for transmission and access.\(^6\) Gillett in 1995 compared cable and ISDN as platforms for delivering residential Internet service, showing that cable can provide a superior and more cost-efficient service.\(^7\)

The technical problem of upstream transmission must be solved to use cable networks for two-way communication, and is thus critical for digital cable systems. Eldering, Himayat, and Gardner provide a solid overview of the sources of the upstream communication problem.\(^8\) CableLabs has also published an in-depth study of this problem.\(^9\)

---


\(^5\) Deborah Lynn Estrin, "Data Communications via Cable Television Networks: Technology and Policy Considerations" (Master's thesis, Massachusetts Institute of Technology, 1982).

\(^6\) David Charles Feldmeyer, "A CATV-Based High-Speed Packet-Switching Network Design" (Master's thesis, Massachusetts Institute of Technology, 1986).

\(^7\) Sharon Eisner Gillett, "Connecting Homes to the Internet: An Engineering Cost Model of Cable vs. ISDN" (Master's thesis, Massachusetts Institute of Technology, 1995).


1.3.2 Economics

The economic problem of operating traditional cable television systems is treated in depth by Webb's *The Economics of Cable Television*. Johnson addresses the more recent situation of providing television service in the face of competition (and its associated regulatory issues) in *Toward Competition in Cable Television*. Baumol and Sidak's *Toward Competition in Local Telephony* is similarly relevant to the situation of cable operators entering into that market.

In addition to the business economics of cable, we look closely at the economic problem of satisfying demands for constrained network resources, which is central to the quality-of-service issues discussed in chapter 5. Much of the recent research to solve similar problems for the Internet is pertinent, and the forthcoming *Internet Economics* edited by McKnight and Bailey provides an overview of that field. MacKie-Mason and Varian explore the problems of network economics in two key papers, and several recent works have critiqued and augmented this, such as that by Shenker, Clark, Estrin, and Herzog.

1.3.3 Integrated Services

Integrated services for communication networks is an idea that has long been proposed and promoted but seldom truly realized in actual systems. The National Research Council's *Realizing the Information Future: The Internet and Beyond*, although not an explicit treatment of this topic, spells out many of the technical, economic, and social implications of integrated services. The work of the Internet Engineering Task Force's Integrated Services Working Group is addressing the practical problem of supporting integrated services in the Internet, and their approach is summarized by Braden, Clark, and

---

Shenker. A different approach to accomplishing integrated services exists within Asynchronous Transfer Mode (ATM) technology; de Prycker provides an overview of these systems in *Asynchronous Transfer Mode: Solution for Broadband ISDN*. These works address the problem of providing integrated services over all sorts of wide-area networks and not specifically cable networks. Application of these solutions to cable networks highlights some difficulties in their approaches, and we argue that integrated services for cable may require quality-of-service mechanisms that are more general and that better address the underlying economic problems of delivering multiple services.

### 1.3.4 Current Standardization Efforts

Two efforts to develop standards for integrated services digital cable architectures provide a wealth of information about the directions and possibilities of this technology. Both of these are attempts at engineering architectures more than standardizations of existing systems. The IEEE Project 802.14 Working Group, with hundreds of submissions contributed by industry and academic participants, is a valuable source of technical details of architectural problems; these efforts are summarized in the group’s “Requirements Document.” Participation in 802.14 is open. Cable Television Laboratories (CableLabs) has issued several requests for proposals for systems that deliver new services over cable including one that specifically addresses the problems of integrated services. CableLabs is a trade organization of cable television operators, including all of the largest U.S. cable companies. The CableLabs work is being pursued privately.

### 1.4 The Scope of Our Work

The technology that is the subject of this thesis is rapidly changing as manufacturers of communication systems develop new products for cable communication.

---


19 Cable Television Laboratories, “Request for Information & Industry Development Proposals (RFI/IDP) for a Integrated Multiple Services Communications Network (MCN)” (Louisville, CO: Cable Television Laboratories, 7 October 1994); “Request for Proposals for a Telecommunications Delivery System over a Hybrid Fiber/Coax (HFC) Architecture” (Louisville, CO: Cable Television Laboratories, July 1994); and “Request for Proposals: High-Speed Cable Data Service (HSCDS)” (Louisville, CO: Cable Television Laboratories, 26 April 1995).
As such, addressing a question as specific as, "Which architecture is the best?" requires capturing the state of this technology at one moment, and can give an answer that may be invalidated by future developments. Case studies of specific systems (such as those by Reed and Gillett) do provide valuable insights with their engineering cost models. However, we seek to demonstrate not the immediate cost benefits of one system over another but the longer term strategic benefits of an architectural approach, particularly in the face of changing technology and changing uses of digital communication. We therefore do not propose a specific solution for carrying communication services, nor do we evaluate one implementation over another. Instead we rely on a general model of cable technology and economics to show both the benefits of the integrated services approach for cable systems and the technological problems that must be addressed to make these systems successful. The arguments we make should provide both direction to engineers developing and standardizing cable communication systems and insight for cable operators who might use this technology.

Considered broadly, the technology involved in delivering communication services over cable networks extends from the televisions, telephones, and computers in our homes to the wide area telephone, computer, and video networks to which cable systems ultimately connect. The business of providing communication services on cable requires consideration of these technologies and particularly of how cable systems will interconnect with other systems and what that will cost.

The technology we are concerned with in this thesis, however, is limited to the problem of communicating on the cable network itself. This includes the devices at the cable head end that initiate and receive communication on the cable system, devices in homes that connect to the cable network, and everything in between. It is this technology that is most critical to the problem of providing local-access communication to homes. The bulk of the cost of cable systems lies in these parts of the infrastructure, and it is often these parts that offer the least flexibility for change once systems are in place.

Our economic analysis of delivering communication services over cable relies on a rough picture of the costs of building, modifying, and operating cable systems, which we draw from published data. To make complete arguments about how to engineer communication systems for cable we would also need to estimate the costs of new technology, the demands for future services, and the future capabilities of competing
communication systems. Such estimates would likely be very uncertain. Instead of relying on this sort of analysis, we develop a general framework for thinking about cable economics in an integrated services environment, relating strategies for cable operators to the costs and flexibility in cable communication technology. The conclusions we draw are ultimately qualitative. Further work could build on our arguments, adding cost and demand estimates to make more specific engineering decisions.

Our work makes some assumptions that are particular to residential cable service in the United States. Serving businesses and government organizations with cable is not a fundamentally different problem, but would change the way services are demanded. Cable systems in other countries use similar technology but may be in very different economic circumstances resulting from different demand, regulation, and competition.

1.5 The Integrated Services Approach

Many existing systems and others under development use cable technology to deliver digital services, but few of these are what we call integrated services systems. For example, Caroll describes a combined cable television and telephony system employed in the U.K. that shares physical infrastructure but uses separate wiring and communication equipment for each service, not even using the cable network for telephony transmission.\textsuperscript{20} Time-Warner, in proposing mechanisms that would control how multiple communication systems would share bandwidth on a cable network, describes an approach that appears closer to integrated services, but in fact uses separate communication systems for different services.\textsuperscript{21}

By integrated services we mean systems that go beyond just sharing the cable infrastructure among different services and applications. An integrated services system should provide a general-purpose communication model that works independently of the applications it supports. With an integrated services system, it should be possible to share network infrastructure, bandwidth, and even the same network transceivers and protocols to support different communication applications. Of course, this does not preclude building equipment that would only be capable of delivering specific services.


We develop integrated services notions further in other parts of this thesis. Chapter 2 discusses the differences between applications and services (and how this distinction can be blurred). Chapters 4 and 5 explore the technical challenges of building integrated services cable systems.
2. Digital Cable Technology and Applications

This chapter presents an overview of cable television network technology and explores the technical characteristics of current and future applications that might be carried on cable networks. Understanding the capabilities and limitations of cable technology is not sufficient to answer the larger question of whether carrying particular services is economically feasible, but is an essential step to addressing this question.

Cable television networks are engineered primarily to deliver a specific service using specific technologies. As we would like to understand how other services might be delivered, we are not concerned so much with the details of how cable television works today as much as we are with the ability of the technology to do other things. Our technical analysis of cable therefore focuses on the basic capabilities of the cable medium. In chapter 3 we will consider the costs of changing the existing infrastructure to deliver new services.

Here we explore cable technology from the ground up, by considering the physical limitations of the cable medium and the technical and topological characteristics of cable networks as they are built today. From this we derive a few general and important characteristics of cable that affect its ability to serve as a general-purpose communication medium.

Understanding what types of future applications might be supported by a digital cable system is a speculative exercise. We identify several communication applications that are important today, but predicting how these applications will be used even a few years in the future is difficult. We recognize that the future of applications for digital cable systems is fundamentally uncertain.

2.1 The Cable Medium

Almost since the first systems were built, cable television networks have been built from two essential components: coaxial cable (often referred to simply as coax) and
broadband amplifiers. The television signals that are delivered over a traditional cable network are not fundamentally different from the signals delivered over the air, but a cable system can often deliver better quality signals and more channels, and there are technological reasons why this is so.

A coaxial cable is simply a wire that is surrounded by a cylinder of insulation which is in turn surrounded by a conducting outer sheath. A signal propagates in coax the same way it propagates along any other wire, but relative to simpler wiring (such as twisted pairs used in telephone systems) coax has two important characteristics that allow it to carry television signals more efficiently. First, coax tends to keep its signals separate from signals in the air, so that signals on the cable do not radiate into the air, and signals in the air do not radiate into the cable. Second, coax can carry signals efficiently over a wide range of frequencies with less attenuation. But a fundamental limitation of coax, and of any wireline medium, is that it attenuates signals more at higher frequencies; specifically, the signal power loss, expressed in decibels per unit length, increases as the square root of the frequency of a signal. A signal of four times the frequency is attenuated at twice the rate.

![Coaxial Cable Diagram](image)

**Figure 1. Coaxial Cable.**

It is incorrect to say that a wireline medium such as coax or twisted pair is limited to carrying a certain bandwidth, since the properties of these media only gradually degrade as we demand more bandwidth from them by transmitting at higher frequencies. Any wire can carry high-bandwidth signals over a sufficiently short distance, but when a signal is attenuated sufficiently it becomes indiscernible from noise in the system. This noise may be inherent in the way signals are transmitted. The point at which that signal is no

---

22 Before broadband amplifiers, the earliest cable networks used one amplifier per channel at every point where an amplifier was needed. See: Baldwin and McVoy, 27-28.

longer useful depends on how the signal is encoded; different ways of transmitting a signal can be less susceptible to noise. So bandwidth itself is not a sufficient measure of how useful a wireline technology is, because it does not take into account signal and noise levels to describe how much information can be carried.

A more useful measure might be *capacity*, which expresses how much information can be carried in a channel, in bits per unit time. In practical terms, a channel’s capacity depends on how signals are encoded, which in turn is limited by the channel’s characteristics, particularly bandwidth and noise. But we know, from Shannon and Weaver,\(^{24}\) that a fundamental relationship limits the capacity that can be achieved in any noisy channel: where \(W\) is the bandwidth in cycles per unit time and \(P/N\) is the signal to noise power ratio, the achievable capacity \(C\) can be at most:

\[
C = W \log_2(1 + P/N)
\]

The *efficiency* of an encoding scheme is the capacity that it achieves relative to this limit.

This relationship already hints at the relative importance of bandwidth and noise in a communication system. If we are already operating in an arena where the signal-to-noise ratio is significant, say 40 dB (a factor of 10,000), then squeezing more capacity out of a channel is difficult without increasing the bandwidth. Increasing the signal-to-noise ratio by 100% (or 3 dB) would increase available capacity by only about 7%. In practice, most wireline systems operate with these high signal-to-noise ratios, although encoding schemes often do not come close to achieving the maximum capacity.

Maintaining a high signal-to-noise ratio over an entire cable system (cable standards specify this should be 48 to 50 dB\(^{25}\)) is difficult. Signal levels are limited at the high end by undesirable transmission properties in the coax and by leakage out of the cable system which can interfere with wireless communication. Within these limits there would be too much attenuation to transmit signals over an entire cable system without amplification. Amplifiers are therefore used throughout the cable system. Because an amplifier itself introduces more noise (and more cost) into the system, designing a cable network for a particular installation requires finding an optimal balance of the lengths and types of coax segments, the number and placement of amplifiers, and the topology of the network.

\(^{25}\) Ciciora, 12.
In principle, a cable system that is designed to support a wider transmission
bandwidth meets these constraints differently. For example, doubling the top frequency
supported, all other things being equal, would result in about 40% more attenuation at the
highest frequencies, and should require that distances between amplifiers be shortened by
40% to make up for that. But in practice re-engineering an existing cable network to
support more bandwidth may not be that difficult if there are other ways to eliminate noise
sources.

2.2 Cable Television Transmission

Television signals that are broadcast over the air in the United States are transmitted
in 6 MHz channels that are allocated to broadcasters by the Federal Communications
Commission. The FCC regulates the location, power, and frequencies used by television
stations, ensuring that stations that use the same channel are sufficiently far apart that they
do not interfere with one another in any of the areas where they may be received. But
preventing interference may also require that two stations do not use adjacent channels in
the same area for a more subtle reason. In the air, a broadcast signal strength falls off
rapidly with the distance from the transmitter, and as a result the signal from a nearby
transmitter can be several orders of magnitude stronger than that from a distant transmitter.
Because transmitters cannot contain their signal perfectly within their designated bandwidth
and because receivers cannot perfectly discriminate between signals from adjacent
channels, a strong nearby signal can interfere with a weak distant signal. This is known as
the near-far problem, and the result is that it is not always practical to use adjacent
television channels in one area.

Television signals delivered over traditional cable television networks are sent the
same way they are over the air: by dividing up the cable spectrum into 6 MHz channels of
bandwidth, and modulating each television signal into one channel. (This is an example of
frequency division multiplexing.) But these cable systems can carry many more channels
than broadcast television for two reasons. First, the near-far problem of broadcast
television is not a problem on cable systems because all channels can be transmitted at the
same power level throughout the network, enabling adjacent channels to be used on the
cable. Second, cable systems are not limited to the bandwidth that is designated by the
FCC for broadcast television; a cable can carry as many channels as the infrastructure will
permit, which in modern systems can be a hundred channels or more.
Because the transmission, or encoding, of analog television signals is done in the same manner as it is for broadcast television, receiving these signals is straightforward. But a television that is not built specifically to receive cable will require an external receiver because a wider range of frequencies is used on cable systems, and sometimes because a television receiver cannot properly discriminate between adjacent channels. This external receiver (a set-top box) re-transmits a selected channel to one that the television can receive. Modern televisions, however, can often tune cable channels directly.

2.3 The Cable Infrastructure

The coaxial cable and broadband amplifier technologies define the essential capabilities of cable networks. But we also need to understand how real cable systems are actually built out of these and other components such as optical fiber, since this introduces both technological and economic constraints on using cable to support other communication applications. Although cable systems in the United States were built independently by a variety of cable companies and equipment providers, the infrastructure is similar enough from system to system so that we can talk about cable systems in general terms. Much of what we need to understand is a matter of terminology used to describe different parts of the network.

We begin by identifying four basic components of a cable network: the head end, the trunk network, the distribution network, and drops.26 Cable networks are topologically organized as a tree, with the head end at the base of the tree, and the trunk, distribution, and drop parts of the network forming successively smaller branches.

---

26 Although cable and telephone networks are structurally similar, they use these terms differently. This can be confusing. The distribution network in a cable plant is sometimes also called the feeder network (we avoid this term in reference to cable systems), but in telephone terminology, feeders connect the central office (structurally similar to the cable head end) to the distribution network, and trunks connect between central offices.
Figure 2. Cable television network infrastructure.

The head end is a central facility that broadcasts signals onto the cable network. A typical head end might serve a city or town of thousands or tens of thousands of subscribers, and is physically a building often located on the outskirts of town, as its location is affected by real estate costs. To support cable television, a head end gathers television signals from a number of sources: over-the-air broadcast television transmission, satellite transmission, land-based microwave transmission, and more recently through optical-fiber distribution networks. The head end can also be the origin of local programming (for example through community access channels), may be where local commercials are inserted into non-local programs carried on the cable system, and is where monitoring and control of the network will typically take place. But the primary function
of the head end is simply to gather all of the television content, modulate the signals onto
the designated cable channels, and feed the resulting signal onto the trunk network.

The trunk network traditionally consists of relatively long coax lines that connect
distribution networks in individual neighborhoods to the head end. Trunks employ high-
quality coax cable and amplifiers along their length; a trunk may be up to 10 miles long
with amplifiers spaced every 2000 feet.\(^\text{27}\)

The trunk network in turn drives distribution networks, which serve neighborhoods
of residences. It is the distribution network that actually passes houses on a street, and
distribution networks are typically shorter (up to a mile) and have more branches than
trunks.

Subscribing residences are connected with drops, which are coax lines that run
from a residence to the distribution network. Drops are connected on a telephone pole or
curb-side pedestal, and are typically disconnected at this point when a subscriber
discontinues service. A drop is made of lower quality (and less expensive) coax cable and
has typical length of 150 to 200 feet.\(^\text{28}\)

Although the path between the head end and a subscriber is mostly in the trunk
network and very little in the drop, the total amount of cable in the network, serving all
subscribers, falls the other way. Ciciora estimates that 50% of the total wiring (by length)
in a typical cable network is in the drops, 38% is in the feeder network, and 12% is in the
trunk network.\(^\text{29}\)

Amplifiers of different varieties are used in different parts of the trunk and
distribution networks. Trunk amplifiers are often higher-quality devices needed to
maintain signal qualities over a long trunk. Bridger amplifiers connect trunks to
distribution networks. Line extenders are amplifiers used within distribution networks.
Passive devices, called splitters and taps, are also used within the system; a splitter divides
the signal energy into several lines, and a tap removes only a small portion of the signal,
typically to connect a drop into the distribution network.

The network infrastructure may be either aerial (on utility poles) or underground.
An underground plant is more expensive, but franchising agreements may require a cable
operator to build the network this way.

\(^{27}\) Ciciora, 47-48.
\(^{28}\) Ibid.
\(^{29}\) Ibid.
Amplifiers within the cable network require power, and this power is usually delivered over the cable system itself by an AC current, similar to ordinary electrical line current and well below the high-frequency television signals, carried directly on the coax. Power is supplied to the cable network in two ways: at the head end, and within the network through power inserters, which are devices that can be mounted on telephone poles to supply power from the electrical power grid into the cable network. Power is not carried beyond the distribution networks; drops are isolated so that power is not carried into subscriber homes.

The requirement for power leaves cable systems vulnerable to power failures, and cable engineers are particularly concerned with this problem when contemplating delivering services such as telephony that may require higher network reliability than cable television. The power failure problem can be addressed by adding back-up power systems at the head end and by using power inserters in the network that have a back-up power capability. Options for back-up power include batteries, redundant power sources supplied by the electric utility network, natural gas generators, and perhaps flywheel power storage systems.\(^{30}\)

2.4 Hybrid Fiber/Coax Systems

Our description of cable systems so far is of traditional systems that use only coaxial cable in their wiring infrastructure. In fact, modern cable systems use both coax and fiber optic cables for transmission. These are called hybrid fiber/coax systems.

As a transmission medium, optical fiber has several advantages over coaxial cable. Foremost are that fiber has the potential of carrying far more bandwidth, is much more immune to noise, and attenuates signals much less. Furthermore, fiber cable itself is not significantly more expensive than coax cable. Why not build communication networks entirely out of fiber? The connections and end points of broadband fiber optic networks are far more expensive than they are with coax. Optical sources and receivers that send and receive electrical signals on a fiber network cost hundreds or thousands of dollars.\(^{31}\) Making connections on a fiber network requires relatively expensive splicing operations.\(^{32}\) The result is that while fiber can be more cost-effective for long-haul point-to-point

---


\(^{31}\) Ciciora, 48-49.
communication, coax is less expensive where there are many branches and connections in a network.

Hybrid fiber/coax systems use fiber in just this way. The trunk parts of the network, where there are long distances of cable with few branches, are replaced with fiber, and the distribution network within a neighborhood remains as an ordinary coax system. The point at which the fiber meets the coax distribution network is called a *neighborhood node*, and for a two way system involves an optical transmitter and receiver.

![Diagram of hybrid fiber/coax network](Image)

**Figure 3.** Hybrid fiber/coax network.

Fiber trunks have the advantage of eliminating long chains of broadband amplifiers that add noise and the potential for failures on the cable system; over these lengths, fiber requires no amplification. The only amplifiers that remain between the head end and subscribers are those in the distribution network, so that the entire path to a subscriber may

---

32 Reed, 93. Reed's entire work treats these issues in much more detail.

33 Ciciora, 47-53.
have at most four to ten amplifiers.\textsuperscript{34} But a hybrid fiber/coax system is not typically a digital transmission system; the fiber portions are generally carrying analog signals that are simply re-transmitted onto the coax at neighborhood nodes.

Since the fiber itself is inexpensive relative to its supporting infrastructure (sheaths, splices, and transmitters and receivers), adding more strands of fiber does not dramatically drive up the cost of a trunk system. As a result, hybrid fiber/coax systems are typically built with individual fibers connecting each coax distribution network to the head end.

Because hybrid fiber/coax networks are be less expensive to build, can carry more bandwidth, and are more reliable than traditional all-coax networks, they are the preferred architecture for building new or upgrading existing cable television networks regardless of future strategies.\textsuperscript{35}

2.5 Two-Way Transmission

Our description of cable systems so far depicts a network that distributes signals only from the head end to subscribers, and cannot transmit signals back to the head end. For the most part this is accurate, as the engineering of these systems reflects the fact that they are used to carry television service to subscribers and not the other way. But it is possible to carry two-way signals on a cable system, and in fact many cable systems are either built with some two-way capability or designed so that this can be added to the system later on.

The coaxial cable medium is not itself directional, and sending signals on it in one direction does not preclude also sending signals in the other direction. This is true generally for any wireline medium; telephone networks in fact carry voice signals in two directions simultaneously over the same circuit, and it is possible to separate these two signals at the ends of the line. In cable systems, the inherent directionality of the network comes from the amplifiers in the system which amplify signals from an incoming cable onto an outgoing cable. The tree topology along with directional amplifiers in the trunk and distribution network combine to deliver signals only outward from the head end.

Cable systems with two-way capabilities use amplifiers that work in both directions. The trick is that only one portion of the cable spectrum is amplified in each

\textsuperscript{34} Ciciora, 48-49; and Cable Television Laboratories, "Request for Proposals for a Telecommunications Delivery System over a Hybrid Fiber/Coax (HFC) Architecture," 131-39.
direction, so that signals in one frequency range are carried in one direction and signals in another range are carried in the other direction. The direction that returns signals toward the head end is called the \textit{upstream path} or sometimes the \textit{return path}.

![Two-way amplifier diagram]

\textbf{Figure 4. Two-way amplifier.}

Although in principle neither the coax cable medium nor this amplification scheme biases the transmission capabilities in one direction or another, the topology of cable systems puts the upstream path at an inherent disadvantage. Branching in the cable tree happens through passive devices—splitters and taps. Downstream, a signal passed through a splitter is attenuated on the splitter outputs, but the noise carried downstream is also attenuated, and the splitter does not change the signal-to-noise ratio. But upstream, a splitter’s outputs become its inputs, and the splitter simply combines the incoming signals and noise. If a signal is present on only one input but noise is present on both, the resulting combination adds the noise together, resulting in a reduced signal-to-noise ratio as the signal is passed upstream. In effect, the head end receives an upstream signal from only a single source but noise from \textit{all sources} connected to upstream channel (all amplifiers and all end points), whereas a subscriber receives a downstream signal from one source and sees noise only from sources along the path between himself and the head end. The problem is called \textit{noise funneling}.\textsuperscript{35} We expect the upstream path to be inherently more noisy than the downstream path, but how much so is a function of the size and shape of the cable system, the sources of noise, and how well noise is controlled.

\textsuperscript{35} Cable Television Laboratories, “Request for Proposals for a Telecommunications Delivery System over a Hybrid Fiber/Coax (HFC) Architecture,” 131-39.

\textsuperscript{36} Eldering, Himayat, and Gardner, 62.
In practice the noise on upstream channels comes mostly from external sources of noise (such as noise generated from radio transmitters, electric motors, computers) that leak into the system through imperfections in the network and home wiring. Such noise is called ingress, and it is difficult to characterize exactly because it depends on the nature of particular sources. Different parts of the cable spectrum can have very different upstream noise characteristics simply because the sources of noise change at different frequencies.

A hybrid fiber/coax architecture can provide some relief to the noise funnelling problem. Because neighborhood distribution networks are connected individually to the head end by fiber trunks, the sources of combined upstream noise can be limited to one neighborhood, and can be reduced as neighborhood nodes are made to serve a smaller number of subscribers. If we expect that the total upstream noise rises in proportion to the number of subscribers, then splitting a 5000-subscriber into ten 500-subscriber nodes might improve the upstream signal-to-noise ratio by 10 dB.

The amount of bandwidth available on the upstream path is a matter of what amplifiers are selected when the system is built, and enabling more upstream spectrum precludes using that spectrum for downstream transmission. In practice, most cable systems either have some upstream capability or use amplifiers that can be upgraded (by plugging in upstream components) to enable upstream transmission. The most common allocation, called a low-split system, allocates four 6-MHz channels at the low-frequency end of the spectrum for upstream transmission. This is little bandwidth relative to the number of downstream channels (which can be over a hundred 6-MHz channels), but that is ultimately a reflection of the business model that the systems have been built to support.

2.6 Carrying Digital Signals

Although the cable technology we have so far described is entirely analog, a cable system can without modification carry properly encoded, or modulated, digital signals. Nothing within the cable network is so specifically oriented towards carrying television that other signals cannot also be carried. As long as a modulated signal fits within the bandwidth and power constraints that the cable system will carry, and as long as the modulation scheme is sufficiently immune to the noise properties of a cable system, cable

\[37\] Ibid., 63. Eldering, Himayat, and Gardner propose that this noise funnelling factor would apply for uncorrelated noise sources, and that a factor that rises with the square of the number of
will carry it successfully. Digital communication can even co-exist with analog television signals, if, for example, we contain the digital signals within their own channels.

Is there anything fundamentally lost when carrying digital signals on an analog cable system? The simple answer is that there is no difference; all signals are analog at heart. But we could imagine designing a cable system differently if we knew that it were carrying only digital signals. For example, instead of using analog amplifiers within the system we can imagine constructing a cable system that uses digital repeaters that decode and re-construct the digital signals and then re-encode and re-transmit them. Such an approach could virtually eliminate the effects of compounded noise in the system, due both to chains of amplifiers and to multiple noise sources along a path, although it does not alleviate noise originating in the transmitters themselves. A system of repeaters could be costly (particularly if it must carry enough information to support a hundred television channels), and in fact only offers advantages in as much as the noise properties of an analog system hinder digital transmission.

Our discussion of carrying digital signals on a cable networks will for the most part assume that we are using a standard analog cable infrastructure that may be a hybrid fiber/coax system but which otherwise has as few structural modifications as possible. Later in this chapter we re-visit the question of whether a different transmission system might be more appropriate for digital communication applications.

2.7 The Cable Network as a Communication Medium

2.7.1 Topology

Cable networks have a tree structure, with the head end located at the root of the tree, subscriber drops as the final branches in the tree, and terminal devices such as set-top boxes and cable modems that are attached to the subscriber drops as the end nodes in the tree. Communication can happen in either an upstream channel or a downstream channel; downstream signals are carried from the head end toward all of the nodes, and upstream signals are carried from a node toward the head end.

Because of the directionality of the network (imposed by its amplifiers), it is not possible, generally, for two nodes in the network to communicate directly with one subscriber would apply for correlated sources (yielding an even larger improvement from subdivision).
another. For example, if two nodes are on two sibling branches in the system, communication from one to the other must pass both upstream and downstream. The cable network topology distinguishes the head end in that it is the only point in the network that receives upstream transmissions from all nodes, and is the only point from which a downstream transmission can be sent to all nodes.

2.7.2 Capacity

Cable networks are built to support a large number of television channels, usually more than fifty and sometimes more than a hundred. For downstream transmission, this means that cable has the potential of carrying tremendous capacity of data. Cable television standards dictate that cable channels should maintain a 48 to 50 dB signal-to-noise ratio in each 6 MHz television channel.\textsuperscript{38} As a result, the theoretical capacity limit of a single downstream channel is about 100 Mb/s. Demonstrated encoding systems have in practice achieved 43 Mb/s capacity for a 6 MHz channel.\textsuperscript{39} For a hundred channel cable system, the aggregate achievable downstream capacity would be over 4 Gb/s.

Upstream capacity is a bit more difficult to characterize because of uncertainty over the noise characteristics of upstream channels; remember that the upstream path is inherently susceptible to noise. But because the theoretical capacity increases as the log of the signal-to-noise ratio, even substantially more noise than downstream channels carry will not preclude high-capacity transmission. If we can maintain a 20 dB signal-to-noise ratio upstream (which is 1000 times as much noise as what we expect downstream), then the theoretical capacity limit of a 6 MHz channel is 40 Mb/s. Research suggests that signal-to-noise ratios might be even better than this.\textsuperscript{40} In practice we can expect to be able to do half that well with actual encoding schemes, but the selection of an encoding scheme might further compromise efficiency in favor of cost (as we discuss in chapter 3). But the fundamental technical disadvantage for upstream communication comes down to increased noise in the upstream channel, and this results in a difference of perhaps a factor of two in the data rate that can be achieved in a given amount of bandwidth.

\textsuperscript{38} Ciciora, 12.
\textsuperscript{40} Eldering, Himayat, and Gardner estimate and measures upstream signal-to-noise ratios in a 500 home hybrid fiber/coax system of about 28 dB, which would correspond to a 56 Mb/s theoretical capacity in a 6 MHz channel.
On a cable system that has four upstream channels, we estimate that the aggregate available capacity, assuming an encoding that achieves 20 Mb/s in a 6-MHz channel, is 80 Mb/s. This is far less than what is available downstream, but almost all of the difference is an artifact of how cable builders have allocated bandwidth on their system: four upstream channels and a hundred downstream channels. This begs the question of what it would cost cable operators to increase the available upstream bandwidth, which we consider in chapter 3.

2.7.3 Access

The cable medium is shared by all subscribers connected to it. This makes cable, considered as a communication system, fundamentally different from, for example, a local telephone network that dedicates a twisted pair line between the central office and each subscriber. Although it is possible to simply divide the available resources on a cable network among subscribers, a scheme that dynamically allocates resources as needed should make much better use of resources if different users have different needs that vary over time. Variable needs for resources are the norm for many communication applications: our telephones and televisions are idle much of the time, and computer network traffic is often bursty, in that short periods of traffic punctuate longer idle periods. Any system that allows multiple users to access a cable network requires a solution for allocating network resources. Such an access scheme can introduce complexity into cable communication systems, and it will be judged both in terms of the costs it imposes and the efficiency with which it operates.

Sharing access among multiple users also gives rise to security and privacy problems. One user connected to a cable network can potentially receive transmissions intended for another, or make transmissions pretending to be another user. Cable systems may need to employ cryptographic techniques to protect against such misuse.

We consider technical solutions to these problems in chapter 4.

2.7.4 Latency

Cable networks impose fundamental transmission latencies that are due to the time it takes signals to propagate through the network. We can estimate this latency based on the longest expected path between a subscriber and the head end, noting that the bulk of this distance is in the trunk network. A long trunk might be ten miles long, and assuming
signals propagate at roughly half the speed of light, the resulting propagation delay is about 100 μs. If hybrid fiber/coax systems consolidate head ends to build larger, regional systems, then the fiber trunks would become longer; a 50 mile trunk would impose a 500 μs delay. Delays of this length are not long enough to be noticeable in interactive applications, but they can have an effect on how multi-access communication schemes work.

Propagation delay imposes problems for multi-access communication in particular when the delay time exceeds typical transmission times. The difficulty is that the system cannot detect collisions (the case of two simultaneous transmissions) within the time that a transmission is occurring, and as a result a transmitting node cannot immediately know if its transmission was successful. For packetized communication, the transmission time depends both on the size of packets and packet transmission rates. Given some flexibility (such as the ability to send data in multiple narrow-band channels) it might be possible to use small enough data rates so that collisions could be detected within the latencies on a cable network. But it seems more likely that packet transmission times will often be smaller than delay times on cable network. For example, a 56-byte ATM cell transmitted in a 10 Mb/s channel would take about 45 μs, which is smaller than the propagation delay for a large cable network. Given such a situation, cable systems may need to adopt multi-access mechanisms that can work efficiently when propagation times exceed transmission times.

2.7.5 Reliability

As they are constructed for television service, cable networks are not engineered to meet the strict reliability standards that telephone networks provide. But the reliability of cable networks might be improved through several mechanisms: by converting to hybrid fiber/coax systems (eliminating trunk amplifiers that can cause failures), by adding monitoring devices in the network to detect failures sooner, by adding back-up power sources, and by using amplifiers that can be automatically bypassed when they fail.

Nonetheless, some sources of reliability problems are fundamental to the cable architecture. Cable networks employ active electronics (amplifiers in the coax networks and optical interfaces at hybrid fiber/coax neighborhood nodes), and these devices rely on electrical power and are often more prone to failures than the cable medium itself. Because multiple users are connected to the same passive coax network, one user's failing
equipment (or the actions of a malicious subscriber) may affect another user’s ability to communicate. Upstream transmission paths in a cable network are prone to ingress, and in the worst case a single source of ingress could cause failures for all the users sharing the coax network.

2.8 Cable Applications and Services

Designing a digital cable system requires some notion of the applications and services that the network might support, for without knowing this we have no way of ultimately judging how well the system performs. But a notion that pervades our work is that the communication applications for cable will evolve over time, and will likely do so in ways that we would not predict today. If factors external to cable networks—the progress of computer technology, the changing uses of electronic communication—are responsible for driving the evolution of network applications, then we should be concerned that cable technology can appropriately keep up with this evolution. If it is the cable network technology itself that drives changes in applications, then the reasons for designing for flexibility may be even stronger. But with either model of evolution, predicting the direction of applications ultimately comes down to speculation.

Recognizing the difficulty of this task, we attempt here to list some of the applications that might be supported by cable networks and to roughly quantify the technical demands these applications put on a communication system. In chapter 3 we consider the economic problems of supporting different applications, which is essential to an actual engineering process. But here we are concerned with the range of technical demands that applications might require, with the extent to which cable technology can support these demands, and with the degree of commonality between the demands of different applications.

The question of common technical demands has bearing on the relationship between applications and services. We use the term application to mean something that users actually do that makes use of the network, and the term service to mean what the network itself performs. Voice and fax communication, for example, are two applications that are supported by standard telephone service. Internet service (specifically

\[\text{\footnotesize 41} \] Within the terminology of the Internet, applications and services have more specific and somewhat different meanings: an application might be a program that runs on a host computer, and a service is a standard way in which one host can initiate communication with another.
the IP transport that is provided by the network) supports the applications of electronic mail, remote login, World-Wide Web browsing, and others. In the case of telephone service, the network technology was engineered around one application but other applications were able to adapt to what the service provided. With the Internet, the network technology was specifically designed to support a range of applications. In both cases, the service limits the applications that can be supported: standard telephone service cannot support applications that demand higher bandwidth than it provides, and Internet service cannot today support applications that need predictable network performance.

The applications that we can easily list for cable are for the most part existing applications that would be carried on a new medium. This is conceptually similar to recent efforts use the Internet to carry voice communication (so called Internet telephony) and fax communication. It is also just what cable television systems did in their original role as distributors of broadcast television. But some important communication applications we use today grew out of (and were not foreseen by) existing services: fax transmissions now account for a significant amount of telephone traffic, and the tremendously popular World-Wide Web application is only a recent use of much older Internet technology.

Several entertainment applications for cable fall into the category of playback applications that involve delivering a data stream, usually for video or audio, over time to a device that plays it to users. Implementations often use data compression schemes, such as the MPEG standards for audio and video, that can dramatically reduce the data rate required for transmission. Playback applications are usually not sensitive to fixed transmission delays even up to a few seconds (since users do not notice the delay), but variation in transmission delay, called jitter, as well as transmission losses, can affect the quality of the application. With the flexibility to introduce more fixed delay, it is possible to build systems that are less sensitive to jitter by buffering data at the receiving end. Error correction schemes can help control transmission losses, also at the cost of increased fixed delays and sometimes increased data rates. These solutions might add cost or complexity to equipment, but there are no fundamental limitations that would prevent transmitting playback applications as long as sufficient capacity is available to support the required data rate.
2.8.1 Applications

The following list presents a sampling of applications that might be carried on cable systems and discusses the sorts of demands these applications might make from a communication network. The list is not exhaustive (a broader list is given by CableLabs\textsuperscript{42}), and the technical demands, which ultimately depend on exactly how applications are implemented, are not exact.

Broadcast Video: Although cable systems already carry broadcast video, using a digital cable system to deliver this might offer a few advantages. A digitally compressed video signal can be transmitted using about two to three Mb/s of network capacity,\textsuperscript{43} and with efficient downstream encoding this could allow cable operators to carry even ten times as many broadcast channels as they do now, or could free up network capacity for other uses. Digital broadcast transmission might naturally provide other flexibility for cable operators, such as the ability to control access to pay-per-view programs and the opportunity to dynamically supply broadcast programming to the network as demanded.

Video-on-Demand: As a substitute for video cassette rentals, video-on-demand may be an immediately marketable service for cable operators. This application requires special video server equipment at the head end (or perhaps at a regional head end), but the technical requirements for actual transmission on the network are essentially those of broadcast television, except that a tiny amount of upstream capacity is needed to interactively request content. However, a video-on-demand system might provide the ability to interactively start and stop playback (mimicking the functions of a video cassette player), which becomes more complicated in the face of substantial fixed delays.

Advanced Television: Proposed standards for high-definition television (HDTV) systems would provide much higher quality video than existing television applications but would require more network capacity. A 1240 by 720 pixel HDTV picture uncompressed requires three times the data rate of an ordinary uncompressed video stream.

Digital Audio: Some cable systems already provide broadcast digital audio services. An uncompressed CD-quality digital audio stream requires 1.4 Mb/s, and

\textsuperscript{42} Cable Television Laboratories, "Request for Information & Industry Development Proposals (RFI/IDP) for a Integrated Multiple Services Communications Network (MCN)," 21-23.
compression can reduce this to 384 Kb/s. Like video, digital audio could also be delivered as an on-demand application.

*Telephony*: Supporting telephone service does not require a tremendous amount of bandwidth from cable networks, but it may impose severe requirements on other aspects of network performance. A telephone connection can be carried digitally in two 64 kb/s channels (one in each direction) without data compression. But telephone service is sensitive to fixed transmission delays for two reasons. First, since it is an interactive service, users will notice very long delays (hundreds of milliseconds) that a system imposes. However, delays of this length are not likely to be a problem for cable systems. Second, much shorter fixed delays, on the order of 5 to 10 milliseconds, can result in audible echoes on voice calls. Echo cancellation techniques can correct this problem, and this is typically done in completing long distance calls where transmission delays are unavoidable. But the need for echo cancellation could add cost to equipment. Delays that can require echo cancellation can come from *packetization delay*, the time spent to fill a packet of data before the packet is transmitted on a network, but such delays can be avoided by using sufficiently small packets or by (inefficiently) not using all of the data in a packet.

The more challenging requirement that telephony poses for cable systems is the need for reliable service. Telephone standards specify that failures should result in down times of less than 53 minutes per line per year, on average. Even with back-up power systems in place, cable systems may require some careful engineering to meet reliability requirements that are this stringent. Cable operators are aware that their cable television customers are sensitive to outages, and given that consumers are accustomed to highly reliable telephone service, the stringent standards may reflect real customer sensitivity.

*Video Conferencing*: Modern video conferencing applications use data compression to allow two-way video connections to work over limited bandwidth channels. The quality of the video connection, seen in terms of the image resolution and

---

44 Ibid.
46 Lee proposes that to avoid the need for echo cancellation equipment, cable systems should limit round-trip delays for voice traffic to 3.4 milliseconds. (Ibid.) To achieve this, voice packets would have to carry no more than 13 bytes of data each.
47 "Residential Broadband Quality of Service," p. 2-4.
48 "Residential Broadband Quality of Service."
update rate, can be traded off against the data rate that the connection requires. Some video conferencing applications can work with a fast modem (28.8 kb/s) and an Internet connection, but perform marginally at those speeds. Cable systems could support much better quality video connections by providing moderately high data rates, on the order of 100 kb/s or 1 Mb/s.

*Computer Networking Applications:* Cable offers an attractive medium for providing network access for home computers, particularly because it can provide high-bandwidth connections and because as a shared medium it can allow many users with bursty network traffic to share the communication resources. But the notion of how to provide network access has changed over the last few years; although once the prospect of providing access to specialized network services seemed attractive, now the Internet appears to be the common and popular medium on top of which everything else must work. Applications such as home shopping and home banking today seem most likely to be provided over the Internet. Many cable companies are already making efforts to provide Internet access through their systems.

The technical demands of providing network connections for computers are difficult to characterize, in part because they change over time. Popular network interfaces today have a wide performance range (28.8 kb/s for telephone modems, 128 kb/s for ISDN access, 10 Mb/s for Ethernet), and some newer systems offer more speed. The usefulness of faster network performance is fundamentally only limited by the internal bus bandwidths in computers and the rate at which software can accommodate data, and both of these change as computer technology improves. As users do different things with computer networks, the characteristics of network traffic change.

Providing computer network access over cable begins to blur our distinction between services and applications. The Internet today can be used to carry many of the applications discussed above, but the performance of these applications can suffer because the Internet lacks mechanisms to support real-time traffic. But even so, the Internet successfully supports multiple applications: electronic mail, web browsing, file transfer, and so on, and these different applications can have different performance needs.

*Electronic Games:* A few services that deliver software to electronic game machines already exist on some cable systems. These services do not require two-way communication, as they broadcast game content that any subscriber can selectively
download to game machines. A more advanced system (with two-way communication capabilities) could provide much more flexibility in delivering game software, and could allow games to be played interactively and between different users on the network.

*Telemetry Applications:* Cable networks might be used to support a handful of applications that we put into the category of *telemetry*, although these applications involve both remote sensing and remote monitoring. A cable network might be used to remotely monitor gas, electric, and water utility meters, immediately saving utility companies the cost of manually reading these devices every month. A full-time connection might offer more flexibility for these systems, allowing utilities to monitor usage more frequently, to charge different rates during the day, to control usage during peak times, and to remotely initiate and disconnect service. Companies that provide burglar and fire alarm services could use cable networks to provide remote sensing of alarms, and could use higher bandwidth connections to perform remote video monitoring. Cable operators themselves may have similar needs to control and monitor their own networks. Most of these applications would make only minimal performance demands from cable systems; the advantage that cable might provide is simply low-cost full-time connectivity.
Table 1. Technical Characteristics of Applications for Cable

<table>
<thead>
<tr>
<th>Application</th>
<th>Network Traffic:</th>
<th>Other Issues:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcast Video</td>
<td>2 x 10^6 b/s downstream (compressed), variable but limited rates.</td>
<td>Providers need security mechanisms for pay-per-view applications.</td>
</tr>
<tr>
<td>Video-on-Demand</td>
<td>3 x 10^6 b/s downstream (compressed), variable but limited rate. Small (10^3 b/s) upstream rate.</td>
<td>Providers need security mechanisms for pay-per-view applications.</td>
</tr>
<tr>
<td>Advanced Television</td>
<td>10^7 b/s downstream (compressed).</td>
<td>Similar to television applications.</td>
</tr>
<tr>
<td>Digital Audio</td>
<td>10^6 b/s downstream.</td>
<td>Playback demands similar to video delivery.</td>
</tr>
<tr>
<td>Telephony</td>
<td>6 x 10^5 b/s two-way (uncompressed) fixed rate.</td>
<td>Packetization delay may be a problem. Users demand privacy. Standards demand very high reliability.</td>
</tr>
<tr>
<td>Video Conferencing</td>
<td>10^5 b/s two-way (compressed), highly variable data rates.</td>
<td>Latency is a problem for interactivity. Privacy is important to users.</td>
</tr>
<tr>
<td>Computer Networks</td>
<td>10^5 to 10^8 b/s or more, two-way traffic. May be very bursty.</td>
<td>Traffic characteristics and future needs depend on what applications will be used.</td>
</tr>
<tr>
<td>Electronic Games</td>
<td>Depends on application.</td>
<td>Interactive multi-player games demand low-latency two-way communication.</td>
</tr>
<tr>
<td>Telemetry</td>
<td>10^3 b/s, very bursty.</td>
<td>Security is important to prevent fraudulent use. Reliability may be essential for some applications.</td>
</tr>
</tbody>
</table>

The table above summarizes the technical demands that these communication applications would place on cable systems. Because the precise nature of these demands depends on how applications would be implemented and used, these requirements cannot be taken as fixed even for these applications. The striking aspect of these requirements, however, is their diversity. The data rates that these applications require vary over several orders of magnitude. Reliability is critical enough to some applications to require careful engineering of cable systems. Some applications can translate increased network performance into better quality, but others have limited needs.

2.9 Conclusion

Although cable networks have been engineered primarily to deliver a broadcast television service, modern hybrid fiber/coax networks have the potential to provide a
powerful communication medium that could support a variety of applications. Because
cable networks are built to transmit high-bandwidth signals, they can support aggregate
data rates on the order of several gigabits per second. The transmission characteristics on
cable are not symmetric—communication upstream is inherently more difficult than
downstream—but the largest bias against upstream traffic is simply the small bandwidth
allocation that is typically given to upstream communication when cable networks are built.

Cable networks share a coax transmission medium among multiple users, and
while this on one hand gives cable the flexibility to dynamically use resources as needed, it
on the other hand requires technical means of allocating resources. Furthermore, the
shared medium can create privacy, security, and reliability problems that cable
communication systems may need to address.

The sorts of applications we envision being supported by cable systems have
diverse communication needs, some taking advantage of the high data rates that cable can
support, some demanding high reliability from the network, and so on. One vision of how
these applications might be supported on cable networks would devote specific services to
specific applications, but it is equally conceivable that many of these applications might be
integrated into a single service. The Internet, or something like it, may in the future
become a common platform for supporting many applications. This may blur our
distinction between services and applications, but regardless of this the technical demands
of applications must ultimately be satisfied by the underlying network technology.
3. **Cable Economics in an Integrated Services Environment**

This chapter explores the economics of delivering multiple digital services over cable networks. The economic challenge for traditional cable television networks involved managing the delivery of a single application, facing limited competition, and using technology that was not changing rapidly. Delivering multiple services over cable changes this situation dramatically. The key challenges cable operators face in this new environment are understanding how to make the transition to the business of delivering multiple services and how best to operate networks that can deliver multiple services.

Making the transition to delivering multiple digital services can require significant infrastructure investment for cable operators. We identify the sources of these costs and the ways in which a communication architecture can affect costs. We furthermore show how the cost structure of delivering new services can affect a cable operator's ability to succeed in the face of competing providers of these services. In such a competitive environment, the ability to differentiate and bundle services may provide a way for cable operators to avoid price wars that could limit their ability to profitably deliver services. Choices made in designing a digital cable architecture can affect all of these factors.

### 3.1 The Economic Position of Cable Television Operators Today

Although cable television service in the United States is provided by a great variety of different companies, we can make a few generalizations that describe the industry as a whole fairly well. Cable television is available to over 95% of American households,\(^4^9\) in almost all cases from a single provider. Most cable systems can carry upwards of 30 broadcast television channels.\(^5^0\) About 65% of homes passed by cable subscribe to the service.\(^5^1\) Cable operators receive annual revenue averaging about $390 per subscriber.\(^5^2\)

---

\(^5^0\) Ibid., appendix B, p. 3.
\(^5^1\) Ibid., 5.
\(^5^2\) Ibid., 12.
Figure 5. Cable subscribeship in the United States.

Exactly how content is sold over cable television networks is rather complicated. Television programming comes from over-the-air broadcasts, satellite networks, and locally originated content. Cable operators in some cases re-distribute programs for free, in others pay for programming, and in others are paid to distribute programs on their networks. Payments can be flat-rate or based on the number of subscribers or viewers. For some programs cable operators sell commercials that they insert into the broadcasts. Pricing of cable services for consumers typically involves bundling that may be tailored to the demands of the community served by a system and may be governed to some extent by franchise agreements. A cable operator might, for example, recognize that an affluent community will be less price-sensitive to premium channels and be willing to deliver low-cost basic-rate service to that community while charging more for premium services.

3.1.1 Cable Television Monopoly Power

Cable television service can be seen as a natural monopoly in two senses. The first sense is that cable television exhibits economies of scale, in that a single operator’s average costs can decrease as his service area increases. A structural savings in a larger service area

---

53 This is explored in depth in: Webb, *Cable Economics.*
comes from sharing equipment between cable head ends, and there may also be an
operational savings in providing service with a larger organization. These economies of
scale provide incentives for cable operators to combine systems together, and in fact we see
this in the form of nearby systems being bought up by large cable operators. Today the
four largest cable operators provide service to over half of the subscribers in the U.S.\textsuperscript{54}

The second sense in which cable television is seen as a monopoly is that two cable
providers competing in a single service area would operate at a much higher cost than a
single cable provider does, since the two providers would require two overlapping
infrastructures. This is generally true for many wireline communication systems where
much of the cost of the infrastructure is in constructing the network to reach residences. If
subscribers buy service from just one provider or the other, then two competitors are
splitting the revenue that a single provider might receive, but each bears the cost of his
infrastructure investment regardless of subscribership. This is particularly true with cable
television systems because once the head end, trunk network, and distribution network are
in place, those parts of the system need not be modified as more subscribers are brought
on line.

In practice, cable television operators rarely face direct competition from a second
cable service. Only about 1.5\% of U.S. homes are passed by multiple systems.\textsuperscript{55} But
cable television has always faced competition in the form of substitutes, first from over-
the-air broadcast television and later from video rentals, that limit the monopoly power of
cable operators. Other multi-channel video services, chiefly satellite services, are beginning
to provide competition that more closely resembles what cable offers. Cable television
operators wield monopoly power only to the extent that their service offers something that
the substitutes do not, such as better picture quality than broadcast television or more local
content than satellite service, and ultimately only to the extent that consumers are interested
in watching television. The business of providing competing video services is growing,
but as it stands cable television is overwhelmingly popular: cable has ten times as many
subscribers as all other multi-channel video services combined.\textsuperscript{56} The simple observation
that a cable franchise typically sells for substantially more than it cost to build ($1800 to

\textsuperscript{54} "Annual Assessment of the Status of Competition," appendix B, table 7.
\textsuperscript{55} Johnson, 17.
\textsuperscript{56} "Annual Assessment of the Status of Competition," 6.
$2500 per subscriber\textsuperscript{57} versus roughly $700 construction cost per subscriber\textsuperscript{58}) is evidence of persistent cable television monopoly power.

3.2 Cable Regulation in the United States

The history of cable television in the United States traces a service that began as a method of bringing broadcast television to areas of poor reception but quickly evolved into a system that provides much more programming than is available on over-the-air broadcasts. The demand for cable services increased in the 1970's and 1980's to the point where consumers have pressured municipal and Federal regulators to impose the sort of controls that are placed on essential services. Since then, cable regulation has swung back and forth between price controls to limit monopoly power and deregulation to encourage growth.

In the early 1960's, the first cable television networks were called \textit{community antenna television}, or CATV, as they were seen not as broadcasters but rather as collective extensions to television receivers. It was not until 1965 that the FCC began to regulate cable television at all, when it became involved in a dispute in San Diego where the cable television company, by carrying Los Angeles stations, was drawing viewers away from local programming. By 1972 the FCC adopted rules that defined what stations cable operators could carry, what public services they were required to carry, technical standards, and a division of responsibility among Federal, state, and local regulatory authorities. But the bulk of control of cable companies remained with municipalities, who had the power to grant exclusive franchises, seek franchising fees, and regulate prices.\textsuperscript{59}

As cable systems grew through the 1970s, regulation began to focus on controlling the natural monopoly power of cable operators. Recommendations focused on promoting the public interest and providing a diversity of content by granting access via public, educational, and government channels.\textsuperscript{60} The conception of cable television, by the late 1970s, had changed from that of a re-transmitter of broadcasts to that of a publisher of content.\textsuperscript{61}

\begin{itemize}
\item \textsuperscript{57} Ciciora, 17
\item \textsuperscript{58} Johnson, 43; and Reed, 316-17.
\item \textsuperscript{60} Johnson, 2-3.
\item \textsuperscript{61} Pool, 158.
\end{itemize}
Still, cable companies felt pressured both by competition from both a growing broadcast industry and by municipal regulation. The Cable Communications Policy Act of 1984, in response to these pressures and to encourage the future growth of cable, effectively deregulated the industry by removing the power of franchising authorities to control prices, except in cases where there were not competing broadcasting channels. After the act took effect in 1986, cable rates rose rapidly.\textsuperscript{62}

In 1990 the FCC prepared a report assessing the impact of the 1984 Act.\textsuperscript{63} The FCC's findings documented the substantial monopoly power that cable companies had amassed in local distribution of television programming, noting that both horizontal and vertical integration had increased in the industry since the 1984 Act. The FCC recommended, as a remedy, introducing more competition in the industry.\textsuperscript{64}

Responding to this report and to public pressure, Congress passed the Cable Television Consumer Protection and Competition Act of 1992, enacted over a Presidential veto, which re-subjected cable companies to price regulation except where effective competition in the form of another cable provider existed. The act furthermore attempted to limit both municipal authority to grant exclusive franchises, and vertical integration within the cable industry.\textsuperscript{65}

The price structures that the 1992 Act mandated were based on a four-tier scheme that differentiated between premium and non-premium channels. Non-premium rates were set by the FCC (using competitive cable systems as a guide for setting the rates), and premium channels were left unregulated. But the ability of cable operators to undermine this regulation by moving channels out of their basic rate packages led to more consumer complaints and ultimately to further FCC rate rollback in 1994.\textsuperscript{66}

With the Telecommunications Act of 1996, Congress once again deregulated cable television rates, favoring competition from new competing services as a means of controlling cable monopoly power. The rate deregulation takes effect immediately for small cable operators and for cable systems that face competition from other video programming services, and within three years for all other cable systems. The more far-

\textsuperscript{62} Johnson, 3-6.
\textsuperscript{64} Ibid., 5003-11.
\textsuperscript{65} Johnson, 3-6.
\textsuperscript{66} Ibid., 10.
reaching measures of the 1996 Act, however, are those that promote competitive entry by telecommunication providers into each other's business. These measures in particular lift several regulatory barriers to providing competitive local telephone service, which could become a lucrative business for cable operators.\footnote{\textit{The Telecommunications Act of 1996: Frequently-Asked Questions} (Atlanta: Deloitte & Touche Telecommunications Industry Services, 1996).}

Table 2. U.S. Cable Regulation Milestones.

<table>
<thead>
<tr>
<th>Date</th>
<th>Policy</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1972</td>
<td>FCC adopts cable rules.</td>
<td>Cable operators must carry some stations, cannot carry others.</td>
</tr>
<tr>
<td>1994</td>
<td>FCC responds to rate complaints.</td>
<td>Price controls are tightened.</td>
</tr>
<tr>
<td>1996</td>
<td>Telecommunications Act of 1996</td>
<td>Price controls removed over next three years, deregulating cable again.</td>
</tr>
</tbody>
</table>

The story of cable regulation, in a nutshell, is of an industry that began as an unobtrusive and competitive service, was encouraged to grow through deregulation, and ultimately amassed monopoly power that the government found difficult to control. Cable television has always faced competition from broadcast television, but by building on its technical advantage of being able to carry more channels (as well as content that broadcasters could not carry) the cable industry has created consumer demand sufficient to become as big as the broadcast industry itself, all the while securing its monopolies by manipulating prices, controlling content, preventing competitive entry, and integrating both vertically and horizontally.\footnote{These are strong accusations that are substantiated in the FCC report. Among the report's findings are that "vertically integrated MSOs have the ability to limit competition to particular programming services," that "most cable companies have the ability to deny or unfairly place conditions on a programming service's access to the cable communities they serve, and...some have done so," and that the local broadcast industry is threatened by the "emergence of strong national cable television companies with rights to serve as exclusive local providers..." ("Competition, Rate Deregulation, and the Commission's Policies Relating to the Provision of Cable Television Service," 5031, 5041.)} Consumers love the service that cable television brings, but they hate the monopoly power that the cable companies wield.
3.3 The Promise of New Services

A proper market analysis for services that might be carried on cable should consider how these services might be sold, how consumers will respond to cable offerings, how competitors will react, and how changing technology and regulation can affect this. The complexity of such analysis is beyond the scope of our work. The market for services that have not yet achieved popularity, such as high-definition television, hinges on factors that are difficult for us to predict. But we can name a few services that would compete directly with existing services which have identifiable markets. To roughly gauge the size of these markets, we consider the following services.69

Broadcast television service: Of course, cable operators are already in the business of delivering broadcast television service. Cable television revenue for all services is roughly $390 per subscriber per year, and with a 65% penetration rate, cable companies yield roughly $250 per year per household passed.

Telephone service: Local telephone companies yield about $44 billion per year for local service, plus another $26 billion in access charges assessed to long distance carriers. Divided over 147 million telephone lines, local telephone revenue is about $475 per year per line. About 94% of households have telephone service (and a few have more than one line), so local telephone service (including access charges) yields roughly $475 per household.

Video-on-Demand: Very little video-on-demand service actually exists today, but this service is almost a direct substitute for videocassette rentals and purchases. Americans spend about $69 per year per person, or $172 per year per household, on prerecorded home video.

Internet Service: Estimates of the size of the market for Internet service vary widely. Just defining what constitutes Internet service is difficult, and the strong growth rate of the Internet quickly invalidates statistics. A recent CommerceNet/Nielsen survey70 estimates that 24% of persons 16 years or older in the U.S. and Canada have access to the Internet—about 45 million people in the U.S., and that this has grown by 50% in the past six months. Other estimates put the figure at 20 to 30 million. Access can come from

---

homes, schools, or workplaces. If we assume that 15% of U.S. homes might pay $20 per month for access, then the market for residential service would be just $36 per year per household. If the popularity of the Internet continues to grow this fast, however, demand might be much higher.

Figure 6. Estimated Current Market Sizes for Services

![Graph showing estimated market sizes for services: Revenue per home passed per year.]

The chart above summarizes these rough near-term market size estimates for the four different services. Although these estimates are not precise, the striking aspect is their relative sizes. The market for local telephone service is substantially bigger than all of cable television's business today. Video-on-demand service, if cable operators can attract business away from the pre-recorded videocassette industry, might also add substantial revenue. Internet service today would be a relatively small business next to any of the others.

3.4 The Costs of Delivering New Services

The technical potential for using cable networks to deliver services other than television is discussed in chapter 2. The economic reality for cable operators, however, is that the infrastructure that they have in place is often built to support only cable television service alone, and their systems require equipment upgrades and other investment to bring new services on-line. Even if we cannot precisely estimate what these costs will be, their structure is important to the problem of delivering new services over cable.

---

3.4.1 The Two-Way Hybrid Fiber/Coax Network Architecture

Two-way transmission capabilities are necessary to support many of the services that we can imagine delivering over cable, notably telephone services and Internet services.\textsuperscript{71} As we discuss in chapter 2, hybrid fiber/coax systems address some of the difficulties in making upstream transmission work well, and are generally thought to be a cost-effective replacement for a coax trunk network even for delivering television service alone.

The cost of upgrading a strictly coax cable system to a hybrid fiber/coax system depends on how much re-building of the system is done with the upgrade. In a hybrid fiber/coax upgrade the trunk network would be mostly or entirely replaced with fiber. A cable operator might additionally choose to extend extra fiber deep into the distribution network (anticipating being able to use more fiber in the future), to upgrade the electronics in the distribution adding more bandwidth or two-way capabilities, or even to entirely replace the distribution network.

Extensive and detailed cost modeling by Reed provides data from which we draw a very rough estimate of the costs of a hybrid/fiber coax upgrade.\textsuperscript{72} In Reed’s model, the cost of building the head end, trunk, and neighborhood node portions of a hybrid/fiber coax network (the parts that would certainly be replaced in an upgrade) is $57 per home passed. The distribution network accounts for another $187 per home passed, much of which is the cost of cable installation. The electronics and power for the distribution network account for $37 of the $187 (including their installation costs). So an upgrade that replaced the electronics and power in the distribution network might cost about $94 per home passed.\textsuperscript{73} Other estimates for upgrade costs are higher; Johnson refers to an estimate of $150 to $180 per home passed, for example.\textsuperscript{74}

\textsuperscript{71} A few cable modem products today implement two-way service on a one-way cable plant by using a telephone line as a return path. This may be acceptable for some Internet applications, but it limits any application that demands higher bandwidth away from homes.

\textsuperscript{72} Reed, \textit{Residential Fiber Optic Networks}.

\textsuperscript{73} Ibid., 316-19.

\textsuperscript{74} Johnson, 32. (We adjust Johnson’s per subscriber figures using a 60% penetration level.)
Table 3. Reed's Cost Estimates for Constructing a Hybrid Fiber/Coax Television Network.

<table>
<thead>
<tr>
<th>Component:</th>
<th>Cost per home passed:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head end equipment</td>
<td>$12</td>
</tr>
<tr>
<td>Fiber trunk network</td>
<td>$26</td>
</tr>
<tr>
<td>Optical network interface (neighborhood node)</td>
<td>$19</td>
</tr>
<tr>
<td>Distribution Network</td>
<td>$182</td>
</tr>
<tr>
<td>Drop</td>
<td>$82</td>
</tr>
<tr>
<td>Subscriber premises</td>
<td>$103</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>$424</strong></td>
</tr>
</tbody>
</table>

Source: Reed, 318-19. Reed's figures are based on an architecture with 500 subscribers per node and 60 percent penetration.

An important result of Reed's model is that, constructing a network from scratch, a hybrid/fiber coax system does not cost more to build than does a traditional system with a coax trunk network. Knowing this, it is not surprising that cable operators feel that the hybrid fiber/coax systems are worthwhile investments regardless of their uncertainty in what might happen with future services.\(^{75}\)

Recognizing that a hybrid fiber/coax network is a basic element of most of the strategies that cable companies are pursuing and that its cost may be justified even without delivering new services, in the context of thinking about multiple services it sometimes makes sense to assume that such an infrastructure with two-way capabilities is already in place in cable systems. This nonetheless leaves open some important infrastructure decisions, such as how power will be supplied in the network, how much bandwidth will be carried, and how large the neighborhood nodes will be. Those decisions may be more closely driven by the particular applications that are carried on the network than the decision to upgrade to a hybrid fiber/coax architecture.

\(^{75}\) See, for example: Fred Dawson, "Investigating before Investing: Filling the Potholes on the Road to Digital," *Communications Engineering & Design*, February 1996.
3.4.2 Categorizing the Costs of New Services

The costs to cable operators of delivering new services to their subscribers come not just from network infrastructure upgrades. Even for delivering television services, cable operators provide subscriber equipment, pay for access to content, and incur the costs of maintaining their networks. New services and the technology used to implement them can affect all of these costs.

We propose that the costs of delivering a new service on a cable system can be thought of in three categories: costs that are fixed for the service, costs that vary with subscribership, and costs that vary with use. This model is useful for analyzing strategic decisions that cable operators can make when bringing new services on-line. But the model is not perfect. Some costs do not fit neatly into one category or another, and sometimes costs may be shared between multiple services in such a way that they cannot be thought of as belonging to one service or another. Different communication architectures can change not only the sizes of these costs but also their structure; a per-subscriber cost in one system might become a per-use cost in another. We illustrate these cost categories below.

3.4.2.1 Costs That Are Fixed for a Service

Modifications to the cable plant are the most prominent fixed costs that cable operators face when bringing new services on-line. These costs include major changes such as upgrading the cable system to a hybrid fiber/coax architecture, adding power sources that can withstand power failures, and adding two-way transmission capabilities to the network. This might also include relatively minor modifications, such as tuning the cable plant to properly support a new communication system or eliminating specific sources of ingress on the upstream channels that prevent devices from working properly.

Other fixed costs include equipment that is added in the head end to support new services and infrastructure that connects the head end with other communication networks. Adding telephone service, for example, might require installing telephone circuits at the head end. We call these bridging costs because they bridge the cable infrastructure to other systems. (Note that this does not refer to a more specific meaning of bridging in computer network terminology).
Some per-service costs are ongoing costs. The costs of operating a billing system, maintaining infrastructure, or leasing interconnection with other networks might not vary with the number of subscribers to a service, but these costs are different from infrastructure costs because they are incurred only as long as the service is offered. (In section 3.5.3.3 we discuss the strategic implications of such costs.)

3.4.2.2 Costs That Vary with Subscribership

Equipment, such as a set-top box or a cable modem, that is provided by cable operators to each subscriber represents a cost that varies with the number of subscribers to a service. The cost of this equipment may be a substantial part of a cable operator's investment; a modern set-top box may cost on the order of $100,76 and cable modems today cost several hundred dollars each, a price that could drop as these devices become more popular.

The drop that connects subscribers to the distribution network is also a per-subscriber cost that is substantial (Reed estimates a cost of $82 per home77). To the degree that drops that were built for cable services can be used for other services, the drop costs for new services are diminished by the number of homes already subscribing to cable services as well as by homes that have drops in place from previous subscriptions. But technical factors may render drops built for television services inadequate for delivering other services.78

We can distinguish between costs that are incurred when bringing on new subscribers to a service, such as the cost of installing drops, and recurring per-subscriber costs, such as the cost of operating a billing service. But in the long run subscribers may come and go on the system, an installation cost in light of a constant churn rate might be seen as a recurring cost.

Costs such as installing a drop or making service calls to subscribers are sunk costs—there is no way to recover the investment. But the cost of equipment provided to subscribers is not a sunk cost if, when service is canceled, the equipment can be taken out and used with another subscriber.

76 Reed, 316-19.
77 Ibid.
3.4.2.3 Costs That Vary with Use

Some of the costs that vary with how much subscribers use services come from costs that are external to the cable network. For example, if a cable operator delivering telephone service pays a per-minute access fee to the public switched telephone network, then calls that the subscribers to this service make incur a cost to the cable operator, a cost that might either be passed on directly to the subscriber or subsumed in a subscription plan. Cable operators might similarly pay upstream for content that is delivered over their network.

Another category of use-variable costs comes from services that, on demand, use limited transmission resources on the cable network itself. To cable operators who control their network resources, this sort of cost is not a direct cost but an opportunity cost. If, for example, a video-on-demand subscriber uses bandwidth that might otherwise be available for delivering a broadcast television channel, then there is an opportunity cost to the cable operator which is the lost value of carrying that extra television channel. This example might seem unrealistic because cable operators plan their broadcast offerings long in advance. But in fact the long-run problem of providing the video-on-demand service involves reserving sufficient bandwidth for the anticipated viewing audience, bandwidth that could otherwise be used for broadcasts, and the size of this resource must grow with the use of the service.

3.4.2.4 Difficulties in Categorizing Costs

The categories above provide a useful simple model for thinking about the costs of new services. But some real costs that cable operators face might be very difficult to fit into these categories. For example, does the cost of maintaining the physical plant depend chiefly on what services are carried, on the number of subscribers, or on how much subscribers use the services? There might be some of each; new services could require that the trunk and distribution networks perform better, but the number of failures in subscriber drops would also increase with the number of subscribers.

Infrastructure costs may also vary with use. For example, a cable operator that provides telephone service needs enough telephone line capacity at the head end to service the maximum number of calls that would likely be placed on the system at one time. This
depends not only on the number of subscribers but also on how often they use their telephones.

The opportunity cost of using bandwidth on a cable system is likely to fall into one category or another depending on how services make use of bandwidth. A new service might require a dedicated but fixed amount of bandwidth throughout the system, in which case this is a per-service cost. But if more channels must be allocated to the service as subscribers use it more, then the cost in the long run begins to look like a per-use cost.

3.4.3 Estimating the Value of Bandwidth on a Cable System

New services may require dedicated bandwidth, or may use bandwidth on demand on a cable system. We argue above that setting aside bandwidth has a cost to cable operators if doing so displaces other things that could be carried; there is an opportunity cost of using bandwidth for new services. What exactly is bandwidth worth to a cable operator? It is difficult to discern this from available data on the cable industry, but we can make rough estimates of the value of bandwidth by looking at choices that cable operators face in building and operating their systems.

Having more channels available on a cable television system gives an operator of a traditional broadcast cable business an increased potential to profit, either by offering more programming that will attract more subscribers or by offering programming that will make subscribers willing to pay more for existing services. But since people only watch a limited (if large) amount of television, new broadcast content only increases demand to the extent that viewers prefer it over everything else available. Given that viewer preferences are somewhat similar among subscribers (people tend to watch popular shows), we expect that as more channels are added to a system each additional channel has a smaller effect on aggregate demand for broadcast cable. Put more precisely, we expect that the marginal value of additional channel capacity on a cable television system diminishes as the total number of channels increases.

If there were no such diminishing marginal value of channels and all channels had equal value, then we could say that the value of each channel was just a fraction of the value of the cable system as a whole. For a cable television system valued at $1200 per home passed that carried 50 channels, the average value of one channel would be $24 per home
but since we expect that there is diminishing value as channels are added, we argue that the value of the last channel’s bandwidth must be lower than this. (The average value serves as an upper bound for any estimate of the marginal value.) We intuitively expect the marginal value to be much lower, since the least-watched channels are far less popular than the most-watched channels.

A better estimate of marginal value of channel capacity is made by examining how cable television companies make decisions about how much channel capacity to build into a cable system. If we assume that cable operators have built their systems to approximately the right proportions, then they have built enough channel capacity so that the marginal cost of adding more capacity has met the diminishing marginal value of the extra channels.

A typical modern cable system might carry about sixty television channels, using all the capacity of the system for television services. Building a system with more channel capacity affects the cost and perhaps the number of amplifiers in the system, but little else: not the amount of cable and passive components needed, and not the labor costs in construction. Ciciora estimates that the electronics in a cable system account for about a quarter of its cost.\textsuperscript{80} If building a cable system with half again as much capacity would have doubled the cost of the electronics, than a cost of $424 per home passed for a 64-channel system\textsuperscript{81} would rise to $536 for 32 more channels. This gives us a very rough but conservative estimate that building additional capacity would have cost $3.30 per 6 MHz channel per home passed. If cable operators have built capacity on their systems that appropriately matches the value of carrying more television channels, then this estimate also represents the marginal value of bandwidth on the system; cable operators should be willing to give up the least- valuable television channels for this price.

\textit{3.4.3.1 Per-Service Bandwidth Costs}

If a new service that a cable operator might carry requires its own 6-MHz channels on the network, is this cost of bandwidth high enough that it precludes carrying a low-penetration service? Suppose that a cable operator is considering deploying a service that

\textsuperscript{79} The cost of building a cable system would be significantly less than $1200 per home passed. But we use an estimate of the value of the system here rather than the cost because we are in effect asking what we would have to pay a cable operator to give up a fraction of his service.

\textsuperscript{80} Ciciora, 17.

\textsuperscript{81} The system cost is taken from Reed’s hybrid fiber/coax model. (Reed, 318-19.)
will require two channels, will serve only 1% of the homes passed by the network, and is sold for $30 per month per subscriber, half of which is an operating profit. The value of carrying this service is $1.80 per home passed per year. If the cable operator expects a 10% annual return on investments, then giving up two channels with a fixed opportunity cost of $3.30 per home passed each should be equivalent to a per-year cost of $0.66 per home passed, just over a third of the operating profits from this hypothetical service. The cost is not high enough to preclude carrying the service, but the example suggests that the opportunity cost of dedicating bandwidth to services could be a barrier to deploying low-penetration services.

3.4.3.2 Per-Use Bandwidth Costs

Another way of thinking of the cost of using bandwidth is its cost in an environment where the bandwidth resources are dynamically shared among services on the cable network. For example, a video-on-demand system could operate on a cable network by dynamically allocating channels to subscribers as they pay to watch movies over the system. The basic problem is the same: as long as there is something else that could be done with the bandwidth resources, there is an opportunity cost of using the bandwidth for the video-on-demand service. But in this case the cost of using bandwidth might be better thought of as a per-use cost since the amount of bandwidth we need is a function of how much the service is used. The analysis in this case is complicated by two factors: first, the opportunity cost might be much higher during peak usage times, and second, the cost depends on how many users are sharing individually controlled pieces of the network.

Suppose, for example, that a video-on-demand system delivers uncompressed video in a 6-MHz channel, and a subscriber needs to use one channel for two hours to watch a movie. What is the economic cost of displacing this bandwidth? If we assume (as above) that a cable operator expects a 10% annual return and values bandwidth at $3.30 per home passed, then the cost of a channel works out to 0.0038 cents per hour per home passed (assuming demand does not vary over time). The critical factor is the number of homes passed, or more precisely, the number of homes that give up the potential to use this channel while one subscriber uses it. In a hybrid fiber/coax system this can be limited to the number of homes connected to one neighborhood node. If there are 500 homes per node, the bandwidth opportunity cost is 1.9 cents for two hours (and if there are 250 homes per node the cost is half that much). Even if cable operators need to over-provision
the allocation over average usage by a factor of ten to accommodate peak usage, the cost is just 19 cents for two hours.

The above analysis suggests that a hybrid fiber/coax architecture opens up a different possibility for relieving bandwidth shortages. Unlike broadcast television, many new services that we might want to carry on cable system—such as video-on-demand or telephone service—allocate bandwidth to individual users on the system. Shortages arise when the demand from all the users sharing the coax network exceeds the capacity of the network. In a hybrid fiber/coax architecture, the cable system is segmented into coax networks that are individually connected to the head end that serve perhaps 500 subscribers. By dividing those segments into smaller networks, a cable operator can increase the average amount of bandwidth available per subscriber. If the system is built with sufficient fiber capacity in the trunk network to do this (and doing so adds little cost), then the cost of sub-dividing is simply the cost of adding more neighborhood nodes that interconnect the fiber and coax networks. In Reed's analysis, the cost of these neighborhood nodes is $9500 each, or $19 per home passed in a system with 500 subscribers per node.\(^{92}\) For roughly that cost of $19 per home passed, a 500 subscriber-per-node system could be modified so that only 250 subscribers compete for bandwidth on a single coax network, doubling the average amount of bandwidth available per subscriber.

Following this method of increasing per-subscriber bandwidth, as the number of homes per neighborhood nodes becomes smaller, the per-home burden for the cost of the neighborhood node equipment would increase to the point where it would be less expensive to instead upgrade the electronics in the coax network to support more bandwidth. But that crossover point may not happen until neighborhood nodes are very small. In Reed's model, a 500 subscriber-per-node system has $169 per home invested in electronics, of which $19 is neighborhood node electronics. Neighborhood nodes in this model might be as small as 60 subscribing homes before per-node cost begins to dominate the cost of electronics.

3.4.3.3 Caveats in Estimating Bandwidth Costs

Our $3.30 per channel per home passed estimate is an oversimplification for several reasons. As discussed above, it does not take into account demands that vary

\(^{92}\) Reed, 318.
during the day, which could magnify the costs of using bandwidth during peak times. It also does not differentiate between using upstream and downstream bandwidth.

Cable systems that have upstream capabilities are commonly built with just four channels that transmit upstream. Cable television service does not require any upstream capability, although some cable television applications can make use of it. Although there may be no shortage of upstream bandwidth now, in the presence of other applications that need upstream capacity, particularly those that dedicate bandwidth to specific applications, this capacity could become scarce. When a cable system is constructed there is an opportunity to choose how much bandwidth is dedicated to upstream transmission (and this is traded off against downstream bandwidth), but changing that allocation later on could require a significant re-working of the system, replacing all of the amplifiers in the trunk and distribution networks. In the long run as systems are re-built it is possible to change how much upstream and downstream capacity is available. In the short run, the options of cable operators may be limited.

An opportunity cost of using bandwidth on a cable system exists only inasmuch as there is an opportunity to do something else with the bandwidth, and only to the extent that there is a shortage of bandwidth on the system. Many factors affect this: not only the demand for services, but also how those services are delivered and how the cable network is constructed. If cable operators switched to a system that delivered television service using compressed video and delivered all programming to viewers on-demand, the notion that more bandwidth can always be used to deliver more broadcast channels might no longer be true.

In the long run the cost of using bandwidth on a cable system is a reflection of the cost of building more capacity into the system, and it therefore does not seem likely that bandwidth would ever come for free. The cost of bandwidth on cable nonetheless is low enough to promise the ability to support high-bandwidth services delivered to homes, particularly when the bandwidth can be dynamically shared between users and perhaps between applications.

3.5 Competition in a Multiple Services Environment

Many of the communication services that cable operators might provide with their networks would be offered in direct competition with services provided by other telecommunication carriers. This is especially true with telephone service, which, if cable
companies can successfully compete with telephone companies, could bring a significant amount of revenue to cable operators.\textsuperscript{83} It is also true with Internet service, which would compete with that provided by a myriad of Internet service providers who use dial-up telephone connections and sometimes other telephone technologies to deliver service today.

Even cable television service itself faces increasing competition. In recent years technology that can provide video over telephone wireline infrastructure (for example by using ADSL systems to deliver compressed video) has threatened to provide telephone companies with a way of directly competing with cable television service, although as yet these services only have been sold to a tiny number of subscribers.\textsuperscript{84} The most widespread competition that cable faces today is from satellite services.

Competition presents a basic problem for cable operators. Much of the investment that cable operators have made is sunk cost in building their network, and some of the additional costs in bringing new services on line are also incurred on a network-wide basis. Cable operators only recover these costs to the extent that they make operating profits selling their services to subscribers. If subscriberhip falls off due to competition, cable operators are simply out of luck.

This section examines the strategic decisions that cable operators must make in competitively providing multiple services on their network. We look both at the problems of entering the market for delivering new services and of pricing and selling services competitively.

\textbf{3.5.1 Making the Transition to Multiple Services}

What stands in the way of cable operators selling new services now? Their most immediate barrier is that much of the technology required is not yet available as mature, mass-produced products. But a fundamental difficulty is that cable operators face \textit{uncertainty} in planning their strategies that comes from several sources. Some uncertainty is technological: not knowing how well systems will perform, what they will cost to own and operate, and which will become mass-produced and standardized. Some uncertainty has to do with demand: not knowing what sorts of services consumers will want and how

\textsuperscript{83} In the U.K., cable companies have had some success delivering competitive residential telephone services, although not using integrated services technology. See: Chuck Carroll, "Development of Integrated Cable/Telephony in the United Kingdom," \textit{IEEE Communications} 33, No. 8 (August 1995).

\textsuperscript{84} "Annual Assessment of the Status of Competition," 6-7.
much they will be willing to pay. Some uncertainty comes from competition: not knowing how other telecommunication providers will offer services and compete with cable. And some uncertainty comes from regulation: not knowing how telecommunication providers will be allowed to price and deliver services.

Rather than making large-scale infrastructure investments in the face of this uncertainty, cable companies can follow a variety of strategies to resolve some of these questions. Trial systems can give cable operators valuable information about the costs and difficulties of installing and operating new technology, and even when that technology does not reflect what would actually be used, a trial can tell cable operators how consumers might respond (in the short run) to new services and prices. Trials cannot resolve some of the fundamental technology and standardization issues, but a legitimate way to cope with that is simply to wait and see what technology becomes successful as other companies build systems. These test, wait, and see strategies do not help cable operators understand how regulatory issues will be resolved, of course. But the larger tragedy for the industry could be that anti-innovative strategies mean missed opportunities to create new services that generate new demand.

The transition to digital services on cable, even if cable operators may someday carry all digital services, will likely come slowly. Digital cable technologies available today are designed to co-exist with ordinary broadcast television services. Since co-existence requires only that systems be able to operate within allocated bandwidth, it makes sense to design integrated systems so that they can operate in the presence of other analog and digital services.

3.5.2 Bertrand Competition

Even in the absence of the technological and regulatory uncertainties, cable operators face difficult problems when entering into the competitive provision of services. In the worst case, a price war among competing service providers could prevent any of the providers from making a profit.

To illustrate how competition might be played out in a telecommunication environment where there are multiple service providers, we imagine an oversimplified analog of the cable and telephone industries. The situation is this: there are two players, \( P_1 \) and \( P_2 \), and two services, \( S_1 \) and \( S_2 \). Each player today is the sole provider of one service and is considering whether to enter the business of providing the second service. To do
that, a player must sink a substantial amount of capital to begin provision, \( C_{P1S2} \) and \( C_{P2S1} \), and then incur a per-month marginal cost of supporting each subscriber to the service, \( MC_{P1S2} \) and \( MC_{P2S1} \). The two competitors may have different costs for providing the same service, but the per-subscriber costs do not change as more subscribers are added. (For the moment we assume that these are the only costs in providing the service: there is no cost in initiating or terminating service to a subscriber, and there are no operating costs other than the per-subscriber costs.)

How does a player decide to enter the business of providing a second service? The decision rests on whether the expected operating profit (the revenue from selling the service less the operating costs) would provide a sufficient return on the required up-front investment. But even if all the costs are known in advance, the expected revenue depends critically on the price charged for the service and the number of subscribers. The number of subscribers in turn depends on how competing services are sold and priced.

If the same service is indistinguishably provided by two competitors and customers base their decision to purchase from one provider or the other solely on price, then the situation becomes a game of Bertrand competition. In this game, once \( P_1 \) has made the investment to enter the market for \( S_2 \), the competitors cannot find an equilibrium price to charge for this service that is above both of their marginal costs. If \( P_2 \) sets a price that is above \( P_1 \)'s marginal cost \( MC_{P1S2} \), then \( P_1 \) is willing to offer service at a slightly lower cost and capture all of \( P_2 \)'s customers. The same is true if \( P_1 \) sets a prices above \( P_2 \)'s marginal cost. Neither competitor is willing to sell below his own marginal cost, but if the two players have different costs then one can capture all of the customers simply by pricing just below the other's marginal cost. In any case, at the equilibrium price it is not possible for both players to profit.

In this Bertrand competition situation, the decision for \( P_1 \) to enter the business of providing \( S_2 \) becomes a question of whether he believes that, pricing this service just below \( P_2 \)'s marginal cost and capturing the entire market, he can yield an operating profit that offsets his fixed costs.\(^85\) This is different from the decision of staying in the business; once sunk costs are expended a player is willing to provide service even if those costs are not

\(^{85}\) This is a stronger condition than there being economies of scope in \( P_1 \) providing both services. We would say that economies of scope exist if, building all infrastructure from scratch, it were cheaper to provide two services with one provider's infrastructure. The condition for \( P_1 \)'s entry is, in effect, that even with \( P_2 \)'s infrastructure already in place it is cheaper to throw that away and provide two services on one network.
being recovered. It is possible that both $P_1$ and $P_2$ would both seek to enter into each
others' markets, but the implication would be that the two players' existing infrastructures
were each better suited to carry the other's services, something that seems unlikely.

The Bertrand competition situation can hold even if we relax some of the
assumptions in our scenario. If there are fixed costs associated with providing each service
and not just per-subscriber costs, then a player would back out of providing the service at
the point where the operating costs are not covered by revenue. That point might be above
the player's per-subscriber marginal cost, but it does not change the situation of being able
to gain subscribers by undercutting a competitor slightly. Adding other per-subscriber
costs, such as the cost of initiating service, changes the decision to add a subscriber
(although pricing might just reflect this cost) but in the long run does not preclude
subscribers from selecting the lower-cost service, particularly if there is gradual turnover in
subscriptions.

If the marginal cost of adding subscribers rises with the number of subscribers
then there might be an equilibrium that would support splitting the subscriber base between
two subscribers. But there is little that could cause an increasing per-subscriber cost for
cable services; a capacity shortage might do this, but there are long-run remedies for
capacity shortages on cable networks (discussed in section 3.4.3). If the per-subscriber
costs are different for different subscribers then there might be a stable allocation of
subscribers between two providers that are allowed to charge different prices to different
subscribers, but in practice regulation of telecommunication services usually prohibits this.

3.5.3 Strategies for Competitive Services

The hypothetical example above shows how cable operators, in the position of
competitively providing services that are differentiated only by their price, could find
themselves in a precarious position where their ability to recover the cost of sunk capital
investments relies solely on their ability to price below what their competition could afford
to do. We argue that the systems that deliver services over cable networks could affect
how competition in these situations is played out. Given the opportunity to choose
between different cable communication systems that have different cost structures and
capabilities, cable operators might pursue several strategies to competitively deliver
services.
3.5.3.1 Incurring Sunk Costs

In the game of Bertrand competition above, an incumbent service provider, faced with a potential competitor who stands to steal his entire market, can pursue a simple strategy to protect his monopoly profits: invest to change his infrastructure so that his marginal costs of providing service are lowered. This might take the form of installing extra capacity, replacing older equipment, or changing the technology used to build the infrastructure. (A telephone company might upgrade its wireline infrastructure with extra capacity; a cable company might upgrade to a hybrid fiber/coax system.) Once the cost of doing this is sunk, the incumbent has demonstrated that he would be willing to continue providing service to customers at a very low cost even if he were not recovering his upgrade costs. This threat is sufficient to keep the other player from entering into competition without the incumbent actually lowering his prices.

The same strategy might also be pursued by a potential entrant against an incumbent service provider, but the argument in this case is less compelling. The incumbent already has sunk capital, and the entrant cannot undo that. Investing sunk costs in this case could influence an incumbent’s decision to make further infrastructure investments. The demonstrated ability to cut prices might also incite competitors to find ways to compete that allow them to avoid a price war; that might come from product differentiation or even from collusion.

The model of Bertrand competition discussed above assumes that sunk costs required to provide different services are independent of one another. This overlooks one of the most important aspects of an integrated services system, which is that such a system can allow new services to be brought on-line without building new infrastructure. Once the infrastructure can support the first service, others come more easily. This is an example of economies of scope in providing multiple services with one system. The strategic effect of such economies of scope in sunk investments can give an operator an advantage in deterring competition in multiple services once the first service is supported. But the deterrent effect ultimately comes from a lower cost of delivering services that is facilitated by the investment, and (as we discuss below) the effect is mitigated by a solution that inflates the cost of delivering services in any way.
3.5.3.2 Avoiding Sunk Costs

Is a cable operator at an advantage if he can find ways to deliver new services without incurring sunk infrastructure investments? To the extent that this reduces his total costs, it is of course helpful. But decisions in the design of a digital cable architecture might be able to trade off sunk costs against per-subscriber or operational costs: more sophisticated and expensive equipment might be able to provide service using existing infrastructure.

To a cable operator who is confident of his strategy to provide a new service in the face of competition, an even trade-off between sunk costs and other (recoverable) costs should not matter; the relevant question is how much profit can be made taking into account all costs. In the face of any uncertainty in providing a new service, however, a strategy that avoids sunk costs has the advantage of leaving open options in the future. If providing a service turns out to be unprofitable, a sunk investment becomes a stranded (unrecoverable) asset.

As discussed in section 3.5.1, some of the uncertainty that cable operators face in delivering services can be resolved in trial systems. Trials demonstrate how consumers react to products, and to some extent might also demonstrate how competitors react. The prices in a trial need not reflect costs, and so the systems used in a trial might be different from those that would be used in a commercial system. But trial systems cannot resolve some uncertainties, and this leaves open an advantage in the strategy of avoiding sunk costs. If the technology available to cable operators changes over time in a way that would make near-term sunk investments obsolete, then cable operators can be better off waiting for new technology. If the regulatory or competitive environments change in ways not anticipated, then cable operators could be better off pursuing strategies that do not sink costs with the risk of stranded assets.

3.5.3.3 Lowering Operating Costs

The model of Bertrand competition discussed in section 3.5.2 assumes that there are no fixed operating costs associated with operating a service, only per-subscriber costs. But real systems will have per-service operating costs that are not just sunk costs. The cost of allocating dedicated bandwidth to a service, discussed in section 3.4.3.1, is an example of such a cost, and other examples include fixed interconnection costs with other networks.
and the fixed components of billing and support costs. With fixed per-service costs, the decision to enter into the business of providing a new service hinges on the same question of whether the operating profit would provide a sufficient return on investment, and here the operating profit must take into account both fixed operating costs and per-subscriber costs. But fixed per-service operating costs can change how a provider chooses to price a service: without such fixed cost, a provider will compete until the price is as low as his per-subscriber costs, but with fixed per-service costs a provider would discontinue service when those costs are not covered.

Trading off these per-service fixed costs against per-service sunk costs could create a strategic dilemma for cable operators. On one hand, operators would like to avoid sunk costs when the prospects for providing service are risky, and would prefer recurring fixed costs that provide a less expensive way out of an unsuccessful strategy. On the other hand, having made sunk investments to achieve low operating costs, an operator would demonstrate a willingness to cut prices that could deter competitive entry.

In principle, operators should be indifferent about trading off per-service fixed costs against per-subscriber costs, since ultimately only the total operating costs for a service matter. But this sort of trade-off might have implications for services for which demand develops over time if high fixed operating costs precluded operators from supporting low-demand services, particularly if the success of “early adopter” users would be responsible for higher demand later on.

3.5.3.4 Service Differentiation and Bundling

The Bertrand competition problems that in the worst case could prevent service providers from recovering the costs of their investments might be avoided if providers can find ways of holding onto subscribers in the face of competitors who offer lower prices. Providers who differentiate their services from the competition can do this if the difference is something that attracts subscribers, and even the ability to attract only a subset of the total subscriber base with some tenacity might be enough to prevent a price war between competitors.

Service differences that could easily be matched by a competitor do not provide a real advantage. But competitors who have different communication infrastructure might find methods of making their service more attractive, in ways that their competitors cannot match. A cable operator, for example, might be able to use his infrastructure to provide
faster Internet service than a competing telephone company can provide at a similar cost, and the telephone company might be able to provide an Internet service that is more reliable than the cable company can provide at a similar cost. If some customers prefer the cable service and others prefer the telephone service, then the competitors can break away from competition strictly based on price.

One form of differentiation that might be particularly applicable to cable operators is the ability to bundle multiple services together as a single package sold to subscriber. Bundling as a pricing strategy is already well understood by cable television operators, who can package different premium television channels together as a way of increasing revenue, particularly when the demand for different channels is negatively correlated (customers who prefer one more tend to prefer the other less). Bundling multiple digital services in this way might also be a good pricing strategy, but here we are proposing something different. If bundling reflects an actual lower cost of providing multiple services together, i.e. if there are economies of scope in delivering multiple services to customers, and if competitors cannot bundle services with similarly low costs, then this sort of bundling can be a form of product differentiation and not just a pricing strategy. A cable operator might be able to attract the customers who want to buy several services together even if a competitor can attract other customers who want to buy only one service by offering individual services at a lower price.

3.5.4 The Limits of Modeling Competitive Strategies

The above discussions illuminate some of the strategic options for cable operators, but as actual models for how operators would do business in a multiple service environment they are very rough at best. The cost structure of cable systems is much more complex than what we discuss; modeling costs with a fixed cost and a flat marginal cost is a rough approximation, and in reality it may be difficult to distinguish sunk costs from recoverable costs. It is difficult to fully account the flexibility of competitors to differentiate their services, to collude formally or informally to set prices above the competitive equilibrium, to bundle products, or to selectively provide services to control costs. We also do not show how regulation can change this picture of competition.

The actual number of services that cable operators wish to carry (at least in the near term) might be small. If Internet service or something like it subsumes other applications
in the long run, then the critical competitive factor might be the ability to support better service rather than the ability to bring on new services.

3.6 The Economic Implications of Integrated Services for Cable

3.6.1 The Benefits of Integrated Services

The sections above hint at some of the reasons why a unified digital cable architecture, one designed to carry a variety of services, might be an attractive platform for a cable operator to use to deliver services. The advantages over systems that are built to deliver only specific services fall broadly into two categories: economies of scope that allow multiple services to be carried more cheaply than separate systems would, and strategic advantages that allow cable operators more flexibility in their business.

Economies of scope offered by a digital cable architecture come from several sources:

- **Plant qualification:** Large-scale modifications to a cable plant such as adding upstream amplifiers, increasing channel capacity, and eliminating noise sources can be very expensive. An integrated services cable system could reduce or eliminate the need for plant qualification work as more services are brought on-line, either by allowing new services to share already-used channels, or by ensuring that an already-qualified plant will continue to work as new capacity is demanded. This could significantly reduce sunk costs that might otherwise have to be incurred as services are added.

- **Terminal equipment:** A digital architecture could allow physical components of terminal equipment (such as transceivers) to be shared among multiple services. This could reduce the cost of providing bundled services, which might be critical to cable operators selling services in competition with other providers. But it is not clear that such savings would be significant, particularly next to savings that might come just from large equipment production scales economies.

- **Bandwidth sharing:** By potentially allowing multiple services to operate within the same channels on a cable system, a multiple-service architecture stands to make more efficient use of cable resources. We estimate that the long-run cost of allocating bandwidth on a cable system is low, but short-term constraints such as limited upstream allocations might put bandwidth at a premium. If overcoming those
problems would require system rebuilds, then a system that shares existing bandwidth efficiently among multiple services could look very attractive to cable operators.

Strategic benefits that cable operators could derive from an integrated services digital cable architecture are more difficult to predict, particularly without detailed knowledge of how cost trade-offs are made in designing a system. But we have identified some potential strategic advantages that include the following:

- **Sunk costs incurred for the first service benefit other services.** The economies of scope that are in the sunk costs that cable operators make in infrastructure can deter competition. The demonstrated ability to deliver more services inexpensively might be enough to keep competing telecommunication providers from investing in capital to deliver competing services.

- **Lower entry costs can reduce the risks of carrying new services.** A system that has lower fixed costs for providing services could help cable operators to deploy services even in the face of uncertain demand, competition, or regulation. If early use of services contributes to later larger scale demand, a system with low entry costs could help cable operators find successful services.

- **Product differentiation can happen through bundling or superior performance.** An integrated services system could provide opportunities to reduce the cost of delivering services by bundling several services together, or to provide services that perform better than what competitors can deliver. Even if such differentiation adds cost, it might allow cable operators to avoid price wars with competitors.

### 3.6.2 Design Principles for a Digital Cable Architecture

Our examination of the economics of operating a digital cable system gives us insight not only into how such a system can benefit cable operators but also into how we can design a better cable architecture. Engineering trade-offs that can be made in designing an architecture will certainly affect both costs and functionality of systems that deliver services. We do not know precisely how to make those trade-offs, but we can identify some important principles:

- **Economies of scope are important.** A chief advantage of an integrated services digital cable system may be providing economies of scope in delivering multiple services, by reducing investment and fixed operating costs and also by allowing several services to be bundled together when delivered to subscribers. Designing an architecture that
allows head-end equipment and terminal equipment to be shared between multiple services will further this goal.

- **Terminal equipment cost is critical.** In an environment where communication services are provided competitively, low-cost terminal equipment can give a cable operator a strategic advantage of a lower marginal cost. Anything in an integrated services system that would drive up terminal equipment costs above what a specialized system might be able to achieve risks pushing cable operators toward specialized systems.

- **Operating costs are critical.** The competitive and strategic advantages that come from being able to deliver services at a low cost depend both on the cost of equipment and on the cost of operating services. The need for reliability for some services, particularly telephony, could drive up operating costs. Finding ways that a cable architecture can reduce operating costs (such as through effective monitoring of the network) might be valuable.

- **Efficient use of bandwidth is not necessarily critical.** We can argue that bandwidth comes cheaply on a cable system if it can be allocated when a system is re-built. However, most existing systems are constrained by limited upstream bandwidth, and cable operators may have difficulty freeing up downstream channels. For services that do not demand much bandwidth it may be quite reasonable to trade off efficient use of a channel for lower equipment cost.

- **Flexible resource use is important.** In the long run cable operators derive advantages from a cable architecture when they can re-allocate resources as demand dictates and when the architecture supports new services and capabilities that might become popular in the future. In the short run, operators derive a strategic advantage if they can use their plant to provide capabilities that their competitors cannot provide.

- **System scalability is important.** Designers of a digital cable architecture should consider how the system can adapt as the demand for services increases and as future applications demand more network performance. More bandwidth on a cable system might come either by dedicating more channels to a digital system or by splitting hybrid fiber/coax systems into smaller coax segments. An architecture that limits these possibilities might impose costs on cable operators in the future.

- **Different cable systems need to co-exist.** The obvious example of this is a multi-service digital platform co-existing with a traditional analog platform, but it is also conceivable
that multiple digital platforms would exist on one system. This can introduce constraints that would preclude assuming that specific channels can be used for a system.

3.7 Conclusion

The prospect of providing new services over cable networks is made attractive by the promise of new revenue and the capabilities of new technologies. But making the transition to a multiple services network requires substantial infrastructure investment, and the ability to yield a return on this investment depends on cable operators selling new services in a competitive and uncertain market. An integrated services cable architecture can help cable operators succeed in this environment by keeping costs low, by providing flexibility that gives operators strategic advantages over their competitors, and by providing opportunities to differentiate their services from competing services. We have outlined several general principles for the design of digital cable systems that could help further such opportunities. With this in mind, the next chapter considers the key design problems for digital cable architectures.
4. Digital Cable Architectural Considerations

Designing systems that can deliver multiple, integrated services over cable networks is a formidable engineering task. The technologies that must come together in these systems range from the broadband communication hardware to the software that implements applications, and include components in between that solve critical problems even if they are ultimately transparent to users.

This chapter presents an overview of the technical problems of designing an integrated services digital cable architecture. We do not attempt to develop solutions for these problems, as doing so is an engineering task that is beyond the scope of our work. Instead we explore the roots of the problems that an architecture will need to address and discuss design alternatives that engineers should consider in solving these problems. Our goal is to map out both the key technical components of a digital cable system and the relevance of these components to the needs of cable system operators and users.

Our analysis draws on ongoing efforts to build systems that deliver digital services on cable systems, and particularly on two standardization efforts that embrace the notion of integrated services: one by Cable Television Laboratories (CableLabs), and the other by the IEEE Project 802.14 Working Group. The CableLabs effort is being pursued privately, but they have published a request for proposals\(^6\) that outlines many of the technical issues that they expect systems to address. The IEEE 802.14 work is available to participants, and is outlined in the group’s draft Requirements document as well in many submissions from participants.\(^7\)

4.1 Why Define an Architecture?

One reason to develop a well-defined architecture for integrated services digital cable systems is the desire to develop standards for these systems. But the need for standards is not as clear-cut as it would be if it directly affected the utility of these systems. As long as cable operator provides access to standard services, such as Internet or telephone connections, and does so through standard interfaces, then there is little reason for users to care how a cable system delivers those services. But cable operators, who buy

\(^6\) Cable Television Laboratories, “Request for Information & Industry Development Proposals (RFI/IDP) for a Integrated Multiple Services Communications Network (MCN),”

\(^7\) Ulm, “IEEE P 802.14 Cable-TV Functional Requirements and Evaluation Criteria.”
equipment and build infrastructure to support these systems, have an incentive to seek standardized architectures. Standard systems can give a cable operator more options for buying equipment that will work with the system he already has, but proprietary systems might limit a cable operator to either buying from a single manufacturer or re-investing to convert to a different system. And although equipment manufacturers might prefer to be able to dominate proprietary markets for these systems, the ability to be able to sell products into a widely standardized market could lower the risks of developing new products for these systems.

A second reason to develop an architecture for digital cable systems has less to do with standards and more to do with the notion of integrated services. One of the benefits of an integrated services system is the ability to deploy systems that can adapt to changing technology and services, even when the direction of those changes is uncertain. We use the term architecture to imply a design that is broader than its near-term implementations might be, one that specifically has provisions for future changes in technology and applications.

Both of these reasons point to interoperability as a goal of an architecture, in one case allowing equipment made by different manufacturers to work together in one system, and in the other case allowing near-term systems to interoperate with future technology and applications.

How much of a system should an architecture specify? The need for interoperability suggests that, at a minimum, an architecture should specify enough to allow manufacturers to build equipment that could be connected to a cable network and implement standard services, so that one manufacturer's terminal device could work with another's head end equipment. In terms of the OSI Reference Model, a digital cable architecture must specify the Physical layer and the Media Access Control (MAC) sub-layer, both of which use mechanisms specifically tailored to the cable system. Other standard protocols might be used above these layers, but to support interoperability we may also need to specify standard ways of delivering specific services above the MAC

---

layer, so that any terminal device built to implement a standard service such as telephony is compatible with how telephony service is delivered on a network.

4.2 Centralized vs. Distributed Control

Should a digital cable architecture employ centralized control over its cable network? Some network technologies, notably Ethernet, are designed without a single point of control, giving every device connected to the network equal access and accomplishing the sharing of network resources through a distributed mechanism. Some digital cable systems employ similar mechanisms to implement an Ethernet-like protocol over cable networks; one example of such a system is the Zenith ChannelMiser system. On the Zenith system, all traffic that is transmitted upstream on the cable network is re-broadcast downstream (through an analog repeater at the head end), and all stations listen to the downstream traffic when transmitting in order to perform collision detection. As a result, every transmission on the system uses both upstream and downstream transmission resources. In Figure 7 below, communication to the user’s terminal device (shown in gray) must travel on both the upstream and downstream paths.

---

89 Edward J. Zylka, “Advanced Data Communications Applications for the CATV Industry,” Proceedings of CABLE '93, June 8, 1993; and Gillett, 70.
Figure 7. Distributed Architecture Similar to the Zenith Cable Modem System.

We argue that such a scheme of re-broadcasting all upstream traffic downstream is inefficient for the uses of integrated services digital cable systems that we envision. The key to this argument is recognizing that we expect most communications to be between a user connected to the cable network and something not on the cable network: a user is most likely to communicate with a telephone subscriber outside of the neighborhood, an Internet node somewhere else in the country, or a video source received at the cable head end. If this is true, then we should be able to feed communication to the user only on the downstream channel and receive communication from the user only on the upstream channel. This can be done only if the cable network interconnects with other networks through the head end, which as the root of the tree topology (see section 2.7.1) is the single point that can both broadcast downstream to and receive upstream from all users. In Figure 8 below, communication to the user’s terminal device travels only on the downstream path.

---

90 We might in fact find some locality of communications within a large cable system, but the critical question here is actually whether locality exists within one shared coax network segment. In a modern hybrid fiber/coax system, the shared coax portion is typically a neighborhood with a few hundred subscribers, with each such neighborhood individually connected to the head end.
In practice, even communication in one direction may require acknowledgments in the other, but the data itself need not be re-sent in the acknowledgment. If upstream bandwidth is scarce, the re-broadcasting scheme makes downstream transmissions also compete for upstream resources, but if we instead gateway the network through the head end without re-broadcasting all traffic then the upstream and downstream resources become independent problems.

The same argument holds for applications where a service provider would cache information that is delivered to subscribers (such as the caching of World-Wide Web content or video-on-demand programming): if there is an opportunity to localize frequently used data, then we are best off placing that data at the cable head end, because putting it anywhere else would require upstream transmission in order to be sent back downstream to users.

So far this is not an argument for centralized control over the cable network but rather an argument for interconnecting the cable network to other communication resources through the head end. But doing so naturally puts a distinguished node at the head end which is the only node that transmits on the downstream channel and the only node that listens on the upstream channel. Since this node must assist in resolving contention for upstream traffic by feeding back information to other nodes connected to the network, it is only a small step to declare the head end a central controller over the network and give it the
responsibility of allocating both upstream and downstream resources. A more distributed resource allocation scheme might be possible, but could only be more complicated. The architectures being developed by both the IEEE 802.14 and CableLabs efforts appear to recognize this, and propose mechanisms that use a head-end controller.

This argument for centralized control challenges the notion, proposed by Carver and Gerovac,\(^9\) that users are inherently better off with a peer-to-peer (i.e. distributed control) local-access architecture than with a centrally controlled architecture. The topology of cable networks provides a natural central point of control, and using this control point to allocate resources is done on the grounds of efficiency and technical simplicity alone. Even if users were cooperatively operating the network, a centrally controlled architecture would be the best solution for them.

4.3 Architecture Issues for the Physical Layer

Although the physical layer issues for digital cable systems are not as complex as those of higher layers, architectural decisions at this level are particularly important because they directly impact equipment costs, operational costs, and system flexibility. The physical interface for digital cable systems will likely be the ordinary coax network that already exists in homes, although it is possible that the poor quality of existing drop and in-home network components will often require upgrades to support digital systems.\(^9\) But this leaves open several remaining questions, particularly: how data is modulated and transmitted on the network, which frequencies are used for transmission, and how transmission schemes dynamically adapt to network irregularities.

4.3.1 Upstream vs. Downstream Transmission

The physical characteristics of upstream channels on cable networks make upstream transmission naturally more difficult than downstream transmission. As discussed in section 2.5, cable networks are particularly susceptible to noise ingress on upstream channels due to the noise funnelling problem, and higher noise levels upstream reduce the capacity of these channels for digital transmission. Compounding this are two practical problems: many cable systems have only a few upstream channels to use, and


\(^9\) John Grothendick, "Is the Weakest Link Ready to Handle Digital?"
these channels are in the low end of the cable spectrum (below 55 MHz) where over-the-air radio transmission (such as from ham radio) and electrical interference can be sources of ingress.\footnote{Eldering, Himayat, and Gardner, 64.} As discussed in section 3.4.3, changing the upstream allocation in a cable plant is possible but not trivial for cable operators.

To some extent it is technically possible to compensate for poor channel characteristics with more sophisticated encoding and transmission technology. But once again upstream transmission is disfavorably biased: a digital cable architecture will likely use just one downstream transmitter at the head end but will use many upstream transmitters, one at each terminal device. Transmitters tend to be more expensive than comparable receivers because they need to more carefully control their signals. As a result, imposing more sophisticated transmission techniques is particularly costly for the upstream channel, where there are more transmitters.

With the above outlined differences between upstream and downstream transmission, we expect that the physical layer solutions that cable systems adopt for upstream and downstream may be very different.

4.3.2 Cable Spectrum Use

Should a digital cable architecture dictate how spectrum is used on a cable system? In particular, should an architecture make assumptions about which frequencies are used for upstream transmission and which are used for downstream? One approach that an architecture can take is to dictate that equipment be \textit{frequency agile}, so that terminal equipment can tune transmitters and receivers to whatever frequency that the cable system uses for upstream and downstream transmission. A benefit of requiring frequency agility is that it could give cable operators the ability to change the upstream and downstream bandwidth allocations in their system without having to change the terminal equipment. Frequency agility might also give a cable system the ability to dynamically respond to frequency-specific problems in a cable network (as might happen with ingress on upstream channels). But frequency agility adds significant cost to transmitters, and as a result we are better off if we can define an architecture that limits the amount of frequency agility for the upstream channel that must be built into equipment.
Short of adopting a model of complete frequency agility for upstream transmission, a cable architecture has several options. Existing cable systems most commonly have upstream capacity in the lowest part of the spectrum, below 55 MHz. A cable architecture could assume that this is the only range available, but this would limit the practical ability of a cable operator to make more upstream bandwidth available on the system through a re-build. An architecture could adopt a wider range, anticipating the possibility of making more bandwidth available, but adding unnecessary cost to equipment in systems that do not take advantage of a wider range. An architecture could use a different part of the cable spectrum altogether for upstream transmission (such as the spectrum above that which is normally used for cable television), but this would preclude using the system on existing plants without modification.

The spectrum use question is further complicated by the need for digital systems to co-exist with traditional cable television service. This should not be a problem if each system occupies its own range of frequencies, but existing usage might make it difficult, for example, for cable operators to re-allocate downstream bandwidth to upstream, and existing upstream use might further constrain what upstream bandwidth is available for digital systems.

Finding a spectrum use plan for a digital cable architecture means facing a dilemma: adding more flexibility to the system (in the form of frequency agility) adds cost to its components. The argument for adding the ability to use more upstream bandwidth than is currently available on cable systems now relies in part on the belief that future services will need this flexibility. We have argued in section 3.4.3 that the cost of allocating more upstream bandwidth on cable systems is not large when cable infrastructure is already being rebuilt. But without available systems that can make use of such bandwidth, cable operators may have little incentive to build more upstream capacity into their networks, and without cable systems that support more upstream bandwidth equipment manufacturers may have little incentive to build equipment that could use it.

4.3.3 Multiplexing

Should the bandwidth that a digital cable system uses be divided into individual channels for individual users, or should the largest possible channel be divided time-wise among users? Neither of the extreme approaches, complete frequency division multiplexing (FDM) and complete time division multiplexing (TDM), is the right solution.
Dividing the spectrum among users by frequency fails to take advantage of the bursty nature of some network traffic to allow users to statistically share network resources. Dividing very wide channels by time, on the other hand, can make transmitters and receivers unnecessarily expensive since each would need to transmit and receive on high-capacity channels; some applications need only small amounts of network capacity and would be over-burdened with a requirement for high-capacity transmitters and receivers.

In practice, digital cable systems often adopt a compromise that performs time division multiplexing within a limited bandwidth channel, and may use several such channels. The spectrum given allocated for a digital cable system will likely come in the 6 MHz increments that are used for broadcast television (because of the need to share cable spectrum with traditional cable television service). Sometimes this is exactly the subdivision that digital cable systems use, but from the point of view of a cable architecture, the 6 MHz division (or any other) is arbitrary. The critical considerations in deciding how to divide cable spectrum frequency-wise and time-wise have to do with both technology costs and applications. Some applications require a fixed and well-defined channel capacity (such as for telephone service), and a pure FDM approach might efficiently support this. Other applications have traffic characteristics that are bursty, and would benefit from a TDM approach. A TDM solution might not significantly drive up equipment costs if it does not use excessively wide channels, and a limited capacity TDM channel might not unnecessarily limit application performance if it is possible to use multiple channels. But an analysis of actual multiplexing schemes cannot properly be done in without knowing how multiplexing is actually accomplished with modulation and media access control systems.

4.3.4 Modulation

The most basic question for the physical layer in digital cable systems is what modulation schemes should be used to encode data on the network. The options available include well-understood techniques such as Quadrature Phase Shift Keying (QPSK) and Quadrature Amplitude Modulation (QAM) as well as more recent techniques such as

---

spread spectrum and discrete multi-tone (DMT) modulation. The various solutions impose different equipment costs, achieve different coding efficiencies (the amount of capacity encoded in a given amount of spectrum), and have different sensitivities to noise and distortion in the cable network. The choice of modulation schemes may depend on the frequency allocation for upstream and downstream communication (discussed in section 4.3.2) because different frequency ranges have different noise characteristics.

An analysis of the merits and performance of these various schemes is beyond the scope of this work. The choice of modulation schemes, particularly for upstream transmission, is nonetheless an extremely important decision in a digital cable architecture. To support an integrated services system, we would like in particular a degree of flexibility in modulation schemes. Future applications may demand much more capacity from the network, and encoding schemes should be able to scale to those needs, recognizing that the cost of technology used for high-performance encoding will fall over time. But some uses of the network require little bandwidth, and an encoding scheme that requires sophisticated equipment for all devices connected to the network could make such applications impractical.

Some modulation technologies specifically address these sorts of scalability issues. Amati Communications has proposed to the IEEE 802.14 group that DMT modulation could be used for upstream transmission in a way that would allow the complexity of each transmitter to scale with the bandwidth that it requires. The DMT technology has the further advantage of being able to adapt to frequency-specific imperfections in the upstream channel and may as a result be able to achieve a higher coding efficiency than other simpler mechanisms. IBM has proposed that more traditional single-carrier coding technology could also be used in a scaleable way by supporting several different data rates and coding efficiencies in one system.

The modulation schemes that digital cable systems will use will likely be implemented with special-purpose chips, not in software with general-purpose digital

---


signal processing (DSP) devices, as DSP technology is not yet practical for the multi-megabit per second data rates that some applications would demand. The high cost of developing special-purpose chip sets complicates the problem of finding solutions for modulation systems. Recovering the development costs requires selling many chip sets, and so developers may be reluctant to build this hardware when its market value is uncertain. The modulation technology decisions that are cut in silicon are in turn coupled with frequency use and noise characteristics decisions that are built into cable networks, and neither the hardware developers nor the cable operators may want to move forward to implement new systems when they are uncertain about what to build. Standardization, either formally or through widespread practice, could help resolve these issues even without completely spelling out a digital cable architecture.

4.4 Architecture Issues for the Media Access Control Layer

In a digital cable architecture that places a centralized node at the head end (as discussed in section 4.2), many of the challenges in devising a media access control (MAC) mechanism have to do with upstream communication, since downstream there is only one transmitting node, but upstream many terminal nodes can transmit on a communication channel. A fundamental part of any MAC solution is a mechanism that accomplishes random access to the network, resolving the contention for resources when more than one node would like to transmit at once. But the need to support real-time services, and more generally the need to control quality of service on the network, puts additional requirements on MAC systems: the network resources that the MAC controls may be allocated to specific connections, and a MAC layer may need mechanisms both to support and to enforce such allocations. Although the random access problem exists only on the upstream channel of the network where multiple transmitters have access, the quality-of-service issues apply to both upstream and downstream transmission.

Solutions for the random access problem are well known both in theoretical models and in implementations of network systems. Some local area network technologies, such as Ethernet, are based entirely on random access systems, but as a result have no provision

"Ibid.

for controlling access to the network in a way that can provide connections with performance guarantees. But a MAC protocol can combine both random access and dedicated access into a single system, for example by allocating only a subset of time slots to random access transmissions. This can allow upstream transmitters both to take advantage of available network capacity to send data immediately and to reserve network capacity for long-term connections. Many of the MAC protocols proposed to the IEEE Project 802.14 Working Group adopt just such a strategy.  

Engineering an actual MAC system begins with this conceptual framework and adds many details. The multiplexing issues discussed in section 4.3.3 may be accomplished at the MAC layer by adopting a system of variable-sized packets or fixed-sized cells. Multiple access requires error detection and control, and may require that transmissions be properly framed within a cell structure so that they occur at exactly the right time, particularly as the round-trip propagation time for a cell can easily exceed its transmission time. A MAC system requires a mechanism for addressing the various nodes on the network, and perhaps way of addressing multiple nodes for multi-cast transmission.

Performance of MAC systems has long been studied, particularly in the context of local area networks. Some performance metrics are concerned with network behavior under unloaded conditions, for example measuring the access delay and the efficiency of the protocol. Other metrics judge the network’s response to congestion, particularly how efficiency is lost to congestion control, whether the performance is stable under congestion, and whether fairness is maintained to allow equal access to network resources for nodes at different locations on the network. Characterizing the performance of a MAC system is difficult to do in an abstract sense because it depends very much on the particular pattern of traffic that the network must carry, but measuring a network’s performance in terms of particular traffic models can provide insight into the relative merits of different systems. The IEEE Project 802.14 Working Group adopts just such a strategy to evaluate the

---


101 See, for example: Henry J. Fowler and Will E. Leland, "Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management," IEEE Journal on Selected Areas in Communications 9, no. 7 (September 1991); and John F. Shoch and Hon A. Hupp,
performance of different MAC proposals for cable networks, based on particular (if arbitrary) traffic models.\textsuperscript{102}

In the layered model of network architectures that the OSI Reference Model envisions, the MAC sub-layer is bounded above and below by interoperability constraints. Below, a MAC system must interact with mechanisms at the physical layer, including not only the transmission of data but also the adjustment of timing, signal levels, and equalization of transmissions based on feedback from the head end. Above, a MAC requires both standard mechanisms that implement standard services (as discussed in section 4.1) and sufficiently general and flexible mechanisms for controlling the quality of service delivered by the network. Laubach et al. propose that the interface between the MAC and physical layers on a cable network be sufficiently general so that in the future, other physical layer implementations can be used with the same MAC mechanism.\textsuperscript{103}

The quality-of-service problem, which we explore in detail in chapter 5, is at its heart both a technical and an economic problem. To support different users, and particularly different applications, that compete for scarce network resources, we need technical mechanisms that determine how the network responds to congestion. The appropriate response may be determined by the type of traffic being carried, by commitments that are made to support specific transmissions, or by a user’s willingness to pay for transmission. The MAC layer is the level at which congestion on the cable network becomes apparent and ultimately must be managed. Decisions of whether and when to transmit data are made at the MAC layer, and this means that the demands and commitments for carrying each packet must somehow be exposed to the MAC interface. Two conclusions of chapter 5 are important to the development of MAC systems for integrated services cable networks. First, the problem of supporting communications with quality of service suggests that the ultimate evaluation of a MAC layer might not be the technical efficiency with which it can carry traffic, but rather the economic efficiency with which it can meet the demands of users. Second, the demands of users for network performance can be complex and difficult to characterize, and we should be concerned that

the mechanisms that implement quality of service are sufficiently general to meet the needs of both current and future users and applications.

4.5 System-Wide Architectural Challenges

   Considered more broadly, designing an integrated services digital cable architecture requires addressing several technical problems that are not specifically confined to the physical or MAC communication layers but can have specific implications for those layers. These problems include providing security and privacy mechanisms for communication on the network, ensuring that terminal devices can be easily connected and operated, and designing systems that will help cable operators achieve network reliability. Some of these needs come from specific applications for cable networks, while other needs are more general. We would like to avoid solutions that would unduly increase the cost of delivering all applications as much as we should avoid precluding carrying specific applications by ignoring their needs.

4.5.1 Security and Privacy

   Because cable networks use a shared, passive medium to connect users, the network itself cannot distinguish where downstream transmissions go (they are broadcast to all terminal devices) or from where upstream transmissions come (they all arrive through the same upstream path). This opens up two problematic possibilities: one user might be able to receive transmissions intended for another user, and one user might make transmissions pretending to be another user. Without a way to control this, cable operators would face several problems in delivering communication services, chiefly: theft of service, where malicious users obtain illicit access to services, and the inability to provide privacy, where malicious users eavesdrop on private communication.

   The privacy and security problems are different for wireline telephone networks because telephone service is provided through a dedicated line between each subscriber and the telephone central office. Eavesdropping on a telephone line cannot so simply be done from a different line coming to another’s home; it may require physical access to the actual line being compromised. Similarly, obtaining telephone service illicitly is difficult because

---

telephone companies know which lines go to which homes. Eavesdropping and theft of service are not impossible on a telephone network, particularly since either might be accomplished by manipulating systems that control the network. But while computer system security is a problem that any modern telecommunication service provider faces, the wireline security problem is severe in the case of cable communication because many users already have physical access to the same wire.

Other communication systems have privacy and security situations that are similar to the problem of cable networks. Ethernet local area networks (in their original form) connect computers with a network that is a passive bus, so that each node can potentially see all of the traffic on the network. Ethernet is sometimes used in situations where eavesdropping and theft of service are not strong concerns because the network is shared among cooperative and trusted users (such as people who work together in an office). But in situations where local area network traffic is neither trusted nor secure, computers can use security systems that are specifically designed to provide both authentication and privacy through encryption techniques. Similar security and privacy problems also show up in wireless communication systems such as cellular telephone service, where providers have responded with technical means of combating the theft of service problem but have done less to prevent eavesdropping.

We are wary of incomplete solutions to these security and privacy problems. One should not assume, for example, that the sophistication of equipment that communicates on the network will be sufficient to keep users from constructing systems that would violate security or privacy. Such assumptions have been discredited by the experiences of telecommunication providers: long distance telephone systems that used in-band signaling were manipulated in the 1970’s by hackers who figured out how to generate signaling tones,\(^{104}\) cellular telephone providers have learned that unsophisticated authentication can be violated by ill:cit “cloning” of legitimate phones,\(^{105}\) and cable television companies have long fought efforts to unscramble premium television services. In all of these cases, it turned out that systems that were thought to be too much trouble to violate were in fact violated, sometimes at a substantial cost to providers.

---


\(^{105}\) Neumann estimates such cellular fraud losses at $0.5 billion per year. (Ibid., 136.)
Modern cryptosystem technology offers mechanisms that are designed to solve exactly the sorts of privacy and security problems that arise as on cable networks. A technique called public key cryptography is particularly useful for cable communication because it simplifies the problem of key exchange (providing a mechanism to communicate securely without an advance exchange of secret keys) and because it can be used to perform authentication, guaranteeing that a transmission is in fact coming from the possessor of a secret key. In practice, public key cryptography and more traditional symmetric cryptography can be integrated into a single hybrid cryptosystem to take advantage of the speed of symmetric cryptography while preserving the convenience of public key methods. Such systems are widely understood, believed to be secure, and can be implemented in software. But an actual implementation of hybrid cryptosystem is complex and requires careful attention to that other software vulnerabilities in the system do not undermine the security provided by encryption.

Should a digital cable architecture incorporate cryptography to address privacy and security needs? One way this could be done would be to implement link-layer security, incorporating cryptography into the MAC system so that the communication between a user's terminal equipment and the cable head end could be private and authenticated. Done properly, this could address the eavesdropping and false identity problems that arise on the cable network, but there are several arguments that might be made against incorporating such a system into a cable architecture. We briefly consider these arguments:

*Does the need for privacy demand such sophisticated systems?* We have motivated this discussion by suggesting that telephone networks provide more security than cable networks because their wires are not shared among users. But in practice, telephone users often use cordless phones and cellular phones that allow eavesdropping without physical access. Privacy might not be a requirement in all situations, but on the other hand, an inability to deliver the level of privacy that other systems can provide might put cable operators at a disadvantage in competitively providing services.

*Does the cost of implementing link-level security outweigh its benefits?* Because these systems can be implemented in software, they might add little cost to equipment that

---

107 Schneier, 177-8.
already has sufficient computing power to support the software. But it should in any case be possible to create an architecture that would allow cryptosystems to be used only as needed, so that terminal devices that did not require the technology would not have to endure an added cost. The strongest need for this technology, however, might come from the need to prevent theft of service and unauthorized access, and adding privacy systems where support for security and authentication already exists might add little cost.

*Is link-level security a complete solution for the privacy needs of users?* Link-level security would only provide protection between terminal equipment and the head end, and not for other parts of the network. As such, it protects against untrusted users of the cable network, but not against untrusted cable operators and not against untrusted communication over wider networks. In practice, users might be willing to trust cable operators with their communication for the same reasons they trust telephone companies: these companies have much at stake if they violate their customers' privacy. But to achieve privacy over a wider area network such as the Internet, security that covers just the cable network is insufficient.109

*Is application-level security a better solution?* An alternative to building cryptography into the MAC layer of cable systems is implementing security and privacy mechanisms at higher levels in the network. Individual services, such as telephone service, that are carried on the network might incorporate their own mechanisms to achieve security and privacy. Applications that run on the Internet sometimes already use cryptography to achieve security over a wider network, not just on the cable network itself. But security at higher layers may be insufficient to prevent low-level eavesdropping and manipulation on the cable network; without link-level protection it may be possible to eavesdrop on the time, amount, and destination of traffic if not its content, and this alone could reveal private information. Incorporating link-layer security does not preclude also using application security at higher layers, and there may be situations when both are called for. Simply pushing the burden of security out of the cable architecture, however, forces applications to cope with a problem that the cable network introduces, and at worse might make users worry about adopting security measures only because the cable system itself does not provide the security they expect.

---

109 Ibid.
Both the IEEE Project 802.14 Working Group and CableLabs have identified security and privacy as considerations for an integrated services digital cable architecture,\textsuperscript{110} but neither group has yet fully addressed solutions to these problems. The best reasons for adopting link level security into a cable architecture are that it solves problems that arise specifically because the cable network uses a shared medium and that addressing these problems once at a low level in the architecture might be simpler than engineering multiple solutions for different services and applications. But we also recognize that link level security is not a complete solution for the privacy needs of users.

4.5.2 Plug and Play

Some business models for selling communication services over cable systems would have cable companies lease and install terminal devices that are connected to the cable network, but in the future, cable operators might adopt (or might be required to adopt) the customer premises equipment model that telephone companies now use, allowing consumers to purchase their own equipment and connect it to the network. In either case, we are concerned that an integrated services cable architecture allows terminal equipment to be easily connected to the network, since complexity in installing this equipment would be a burden to cable operators in one model and a barrier to consumers in another. The ideal situation would be to allow any device connected to the cable network to deliver service immediately, a situation that we call plug and play and that is more-or-less achieved with standard telephone systems today. But achieving plug and play for cable networks is more difficult than it might seem.

Because the cable infrastructure does not run individual lines to individual subscribers, it is not possible to automatically associate terminal devices with particular subscribers. We could, for example, require that all terminal devices be built with unique identifiers that can be read over the network, but this alone does not tell a cable operator which subscriber is served by a terminal device. This might be solved with a phone call to the cable operator, although the device that we wish to connect might be just the device that delivers telephone service to a home. The subscriber identity problem also might be solved by allowing subscribers to key a subscriber identifier into their equipment, but this sort of

\textsuperscript{110} Ulm, "IEEE P 802.14 Cable-TV Functional Requirements and Evaluation Criteria," 36; and Cable Television Laboratories, "Request for Information & Industry Development Proposals (RFI/IDP) for a Integrated Multiple Services Communications Network (MCN)," 61.
solution can increase the complexity of the equipment and starts down the path of having users program their equipment, which we would like to avoid. But in either case, the subscriber identity problem seems solvable without unreasonably impeding plug-and-play.\footnote{Once a terminal device is associated with a particular subscriber, moving that device to another home might not be detectable by the cable operator, who would continue to deliver the same subscriber’s services to that device. This seems odd in the context of telephone networks, but it is not an unreasonable behavior.}

When cable systems employ cryptography to address the privacy and security problems discussed in section 4.5.1, the plug and play problems may become more complicated. If terminal devices require that cryptographic keys be set up to perform secure communication, then the process of exchanging these keys between users and cable providers becomes a vulnerability if revealing those keys would compromise security. Public key cryptography addresses just this problem, but still requires at a trusted way of publicizing the public keys; a malicious user might in principle be able to subvert the system by causing inaccurate public keys to be distributed. The theft of service problems that these mechanisms attempt to address go hand in hand with more practical problems of authentication: if a subscriber initiates service for a new terminal device by phoning her cable provider, the cable operator ultimately needs to validate the personal identity of the subscriber and not just the identity of the equipment.

In addition to mechanisms that allow equipment to work immediately, plug-and-play solutions should address what happens when equipment does not work. The potential sources of failures in these systems include the cable network itself, the drop and inside wiring, power failures, failures in the terminal equipment itself, and failures of equipment connected to the terminal device. In the worst case, a failure shows up simply as a service that does not work—the television has no picture, or the computer has no network connection. Better solutions might help cable operators avoid making service calls and users install equipment more easily.

\subsection*{4.5.3 Reliability}

Some applications that might be carried over cable require a high degree of reliability from the network. In order to match the performance of telephone networks for telephone service, for example, cable networks would need to meet tough standards that
call for only minutes of down-time per year.\textsuperscript{112} This would require that cable operators provide back-up power for equipment within their networks (as discussed in section 2.3) as well as back-up power for terminal devices that need to work when the power fails; as the average unavailability for electric power in the United States is 370 minutes per year.\textsuperscript{113} It also might require that cable operators adopt more effective ways of monitoring their networks and isolating failures.

As much as an integrated services architecture might put more demands on the network, it can also give cable operators better tools to maintain the network. With two-way communication available, cable operators can place monitoring equipment in the network that measures and reports the quality of signals received from the head end.\textsuperscript{114} If terminal devices support simple status and signal quality monitoring, cable operators can collect dynamic performance information without even using special monitoring equipment. This sort of pro-active monitoring could give cable operators the ability to detect poor network performance even before failures occur.\textsuperscript{115}

4.6 Conclusion

Designing an integrated services digital cable architecture is a tremendously complicated task that calls for both assembling existing technologies, such as the modulation schemes used for the physical layer, and developing new ones, such as media access control protocols for the network. Many of the decisions that are made in an architecture can significantly affect the cost of building and maintaining cable networks and the opportunities to deliver new services in the future.

We have explored some of the critical engineering problems that arise in an architecture design, asking in particular how design decisions could affect our ability to expand cable systems in the future to deliver new services. Thinking ahead calls for scalability in two senses: the ability to support applications that demand more communication resources from the network, and the ability to support applications that use

\textsuperscript{112} For example, CableLabs suggests that telephony should meet a Bellcore standard that would require, among things, lines to operate with less than 28 minutes of down time per year. (Cable Television Laboratories. "Request for Proposals for a Telecommunications Delivery System over a Hybrid Fiber/Coax (HFC) Architecture," 33.) A different Bellcore document proposes a 53 minute per line per year standard. ("Residential Broadband Quality of Service," p. 1-2.)

\textsuperscript{113} "Residential Broadband Quality of Service," p. 1-2.

\textsuperscript{114} Baldwin and McVoy, 67.

\textsuperscript{115} Shira McCarthy, "Keeping an Eye on the Network," \textit{Telephony}, 1 July 1996.
the network differently than ones that we envision today. These needs call for flexibility in particular at the physical layer, particularly in how bandwidth in the network is used and in how signals are encoded, and above the physical layer, particularly in how quality of service is supported within MAC protocols (a topic explored in more detail in chapter 5).

The efforts by the IEEE Project 802.14 Working Group and CableLabs to develop standards for integrated services cable systems face the issues that we raise here. Standardization can promote the deployment of new communication services on cable systems by helping cable operators construct their infrastructure in ways that will be compatible with new systems and by allowing multiple vendors to competitively sell interoperable equipment to cable operators. But these standards direct sunk costs in the construction of cable systems and engineering decisions that are etched into silicon when equipment is built. Designing standards with a long-term vision is important.
5. Quality of Service

Many of the applications that an integrated services digital cable system must be able to support require specific performance characteristics from the network in order to operate properly. The need is most pronounced with real-time applications, such as telephony or video conferences, where a failure to deliver data within short time constraints can result in delay or interference that is perceptible to users and can render a service unusable. The problem of providing the necessary quality of service (QoS)\(^{116}\) over digital networks is one that has been studied extensively, notably in the context of delivering real-time services over the Internet and over ATM networks. Some of this research has its roots in much older studies of telephone network performance that advanced the field of queuing theory. But while the telephone problem is largely limited to users of a single application and a single network, modern quality-of-service problems are more complex because they involve sharing network resources among multiple applications that have very different performance needs, and often across multiple networks.

ATM and Internet QoS research is immediately relevant to the problem of designing an integrated services digital cable system. All of these problems are at their heart determining how to share congestible network resources among multiple, often real-time applications. The applications that we wish to support for cable networks are the same, and ATM or IP, or both, could even be the medium through which such applications are delivered on cable networks.

In one sense, the problem of engineering support for real-time services on a cable network is a much simpler problem than doing that over a larger network. In a multi-point network such as the Internet, a single connection can involve many network resources, and it may not be possible to fully understand the state of those resources at any one point in the network. But on a cable system, we are concerned with sharing just a single resource: the coax communication link that connects subscriber terminal devices to the head end, where the network can be centrally controlled.

On the other hand, inasmuch as a digital cable system is a part of these larger networks, it must be able to interoperate with the way QoS is supported in other systems,

\(^{116}\) Quality of service has a specific meaning in the context of ATM networks. We use the term more broadly to mean any sort of network performance characteristics that are not supported by best effort service.
and achieving interoperability can be complicated by the QoS problem. For example, the IP mechanisms that Internet uses today can use almost any sort of underlying network infrastructure: Ethernet, FDDI, modems, ATM, frame relay, and so on. This works because IP offers no performance guarantee other than "best effort," and as such needs little from the underlying network systems other than that they somehow carry the traffic. If we were to add to IP a notion of guaranteed performance for a connection, this would place more demands on the underlying network: in terms of interoperability, we would need a way of translating the service needs of IP into what its underlying network can provide, and an underlying network's capabilities would need to be rich enough to support what IP would like to provide.\textsuperscript{117}

Much of the QoS research for ATM and Internet systems focuses on two problems: how to price network services, and how to efficiently deliver them. If efficiency is measured only in terms of how much information is transmitted within the constraints of a system, what we call network efficiency, then being efficient is largely a technical problem of finding good mechanisms to carry traffic. But a metric that is often more relevant is how much utility is gained by the network's performance, what we call economic efficiency.\textsuperscript{118} Economic efficiency takes into account both the value of transmissions to users and the cost of transmission to network operators (which often is negligible). Network efficiency can be a useful technical measure of, for example, how much of a network's resources are lost to a communication protocol. But for making decisions about how to use the network when there is contention for resources, it is economic efficiency that is the appropriate metric, and this ties the technical problem of delivering network services inseparably to the economic problem of pricing those services.

With the technical and economic problems of delivering real-time services over these networks unresolved, we nonetheless need to make architectural decisions in designing digital cable systems that could affect our ability to interoperate with other networks. Such decisions may be difficult to undo in the future if cable operators make

\textsuperscript{117} Even this notion, that QoS should be delivered through explicit service support in the network, may be controversial. A much different model for supporting real-time services on the Internet that would require only minimal cooperation within the network is proposed in: David Clark, "A Model For Cost Allocation and Pricing In the Internet," in Lee W. McKnight and Joseph P. Bailey, ed., \textit{Internet Economics} (Cambridge: MIT Press, forthcoming).

\textsuperscript{118} The terms are used similarly in: Liam Murphy, John Murphy, and Jeffrey K. Mackie-Mason, "Feedback and Efficiency in ATM Networks" (Ann Arbor: University of Michigan Department of Economics, 1996).
investments in equipment based on a particular architecture or if standardized systems stand as barriers to new innovation. Both new applications for cable systems and new methods of allocating network resources and pricing services will certainly come along in the future. Recognizing this, we would like a digital cable architecture to be flexible enough so that cable operators have a migratory path that allows them to take advantage of new mechanisms without discarding their investment in existing equipment and infrastructure. Ways to achieve such flexibility include both engineering mechanisms that are more general than immediate needs might require and leaving some problems undefined by the architecture and solved only in implementations.

Mechanisms that deliver real-time services over networks are often tied to specific definitions of what sorts of services are supported. ATM networks, for example, are built with the notion of a few general service classes: constant bit-rate (CBR), variable bit-rate (VBR), unspecified bit-rate (UBR), and available bit-rate (ABR) traffic. To the extent that such service classes are designed around specific notions of applications or specific mechanisms that deliver services, imbedding these notions into a digital cable architecture may work against the goal of building a general-purpose system.

This section considers the problem of defining classes of service for real-time applications. Seeking to understand whether the services classes that are defined for existing network solutions might be too limiting for future applications, we consider some very general questions. What does it mean to have a quality-of-service expectation? How does the need for committed service quality arise? How do users communicate their demands for service commitments? To frame these questions, we consider a practical example: the service expectations that exist in telephone networks. We show that the generality of service expectations is quite complex, but we recognize that a perfectly optimal way of specifying expectations is neither practical nor necessary. Nonetheless, considering this general problem gives us insight into the limitations of existing solutions to defining service classes for networks.

5.1 Service Quality Expectations in the Real World: The Example of Telephony

The desire to meet a certain quality-of-service level in a communication application ultimately comes from the application's users. If a carrier degrades the quality of a voice telephone call, or introduces delays in a telnet session on the Internet, it is the users who feel the consequences of an application that performs poorly. Conversely if the provider
changes the service quality in a way that users do not notice through their applications, then users are indifferent to this change; the service is just as valuable to the users as it was before.

Exactly what users expect from communication applications, and the arrangements they make to secure commitments to meet these expectations, are more complicated than we might think. As an illustration, consider telephone service:

I buy telephone service from NYNEX for my home, a service that has not changed significantly in price or quality over the past five or ten years. A simplified description of my relationship with NYNEX is that I expect my phone to work as long as I pay my bill, and if it does not then I am out of luck since NYNEX provides the only wireline phone service I can buy. But if I think harder about this, I realize that I am demanding some specific performance from my phone service. I expect that:

- when I pick up my phone, I should always get a dial tone within a second or two;
- when I place a call I should usually be able to get through to a ring or busy signal, only occasionally encountering a busy circuit problem;
- when I am on a call, I should never be cut off;
- the voice I hear on calls should not be too noisy; and
- I should always be able to make an emergency call to the fire or police department.

In fact, I must clarify what “always” and “never” means here. I am willing to tolerate occasional failures; if a call gets cut off once or twice a year, I will keep paying my bill, and even if my service fails completely I am satisfied if NYNEX fixes the problem promptly and it does not happen again. Because its customers are willing to tolerate some problems, NYNEX can employ phone systems that are not perfect. Occasionally we do get busy circuits on calls because telephone switches cannot handle all possible traffic patterns or even anything approaching what would happen if every subscriber tried to place a call at once. If NYNEX could sell its phone service only if it never blocked calls because of busy circuits, their equipment would have to be far more sophisticated and far more expensive.

How does NYNEX know what quality of service I demand? I have never spelled out the above expectations to them, nor have they spelled out anything as detailed to me (although NYNEX may be regulated to meet technical standards). But I have the potential to do several things if I am dissatisfied: I can call NYNEX and complain, I can call up the
Massachusetts Public Utilities Commission and complain, and ultimately I can cancel my service. Still, I do not expect to have to do any of this because my telephone service has always exceeded the level of performance that would cause me to complain.

The point of this example is to gain insight into how service expectations are manifested in a real communication system. Several points are worth noting:

- **Expectations are ultimately statistical in nature. Meeting expectations relies on this.** Even if our first thought when describing our expectation is “always” or “never,” we may be willing to live with occasional exceptions. This is very important, because it allows our expectations to be satisfied with systems that statistically do a good job using less resources than a perfect job would require.

- **Expectations are complex, and meeting them may require taking this complexity into account.** We know that a telephone call can be carried in a 64 kb/s continuous bit rate channel. But if we are carrying calls on a system that cannot support all possible channels open at once, then to meet expectations we need also to understand the frequency and duration of calls, whether we can drop calls in progress, whether we can block calls before they are made, and so on.

- **Commitments are not always firm.** In our example, the provider and customer have not spelled out the details of their service expectations in advance. Exactly what would happen if expectations are not met is not clear to anyone. Nonetheless we believe that meeting the expectations is important.

- **Communicating expectations is complex and subtle.** Our telephone providers can get information about expectations in many ways: complaints from customers, standards from regulators, canceled service, and so on. Perhaps an increase in quality might prompt customers to buy more services. Perhaps there are important expectations that we do not understand.

- **Over-provisioning simplifies the situation.** Finding out exactly what our expectations are can be more trouble than it is worth. When I place a phone call, the phone company does not ask me how long the call will take; instead they maintain enough capacity so that I can very likely talk for as long as I want without interfering with other customers’ calls.

- **The system works despite imperfect information.** A telephone provider does not completely understand what its customers want at every moment, nor do its customers
completely understand what it is able to provide. However, both have long-term expectations of demand and performance that are developed through experience. Imperfect information between buyers and sellers in theory can be a source of market failure, but in practice we see that this problem can be overcome.

The problem of communicating and meeting expectations can become much more complicated in a network where the applications are not explicitly known. In the telephone example, there is just one basic service provided for a single application (although modern telephone services begin to depart from this notion). In a more sophisticated network, users may have many different applications with different demands, even applications that were never anticipated by providers. The ability to adapt to new services efficiently can benefit both providers and consumers, and can foster the development of new applications.

5.2 An Abstract Model of QoS

In order to understand the roots of expectations and commitments in a communication network, we construct the following imaginary model of a network:

- Our network is controlled by a provider, who allows a number of users connected to the network to access it.
- Access to this network is exclusive and slotted: a slot is a fixed period of time during which only one user can use the network to communicate. Once a slot has passed in time, it is gone and its utility is moot.\textsuperscript{119}
- The provider controls the slots, and so is ultimately responsible for determining who may access them.
- We think about the problem of allocating slots in an economic sense: using the network (accessing its slots) has some value to the users, who ultimately pay the provider to gain access. The provider seeks to allocate access in a way that maximizes his profit.

To simplify this model, we make a few assumptions:

- That there is zero direct cost to the provider of allowing a slot to be used.
- The rate at which slots come is fixed, and the provider has no ability to produce any more, i.e. to increase the capacity of the network.

\textsuperscript{119} Readers should not get caught up in the differences between slots and packets for this example. We use the notion of slots only to emphasize that these are exclusive, temporal resources and to avoid complications in our analysis that might arise with different sized packets.
• Communication between users and the provider (such as requests for slots) happens through some other means, and has no cost.\textsuperscript{120}

The problem as stated is similar to a variety of allocation problems that come up, both within the field of computer networks and in other fields. Varian and MacKie-Mason have presented a good analysis of similar problems in their work on pricing for computer networks,\textsuperscript{121} but their analysis differs in both the assumptions they make and the questions they address. The broad question of how to price network usage, as MacKie-Mason and Varian demonstrate, is important in order to achieve economically efficient use of networks. However, we are concerned with several more specific questions: what sorts of demands might users have, how do they express these demands, how do expectations arise, and how do commitments fulfill expectations?

To this end, we have constructed a model of a network that is substantially different from what Varian and MacKie-Mason analyze. In their network, the consequence of congestion is delay in packet transmission, but in our model congestion only changes how the provider makes admission decisions, and delay happens not in transmission but as a consequence of waiting for slots. While the Varian and MacKie-Mason model of a congestible resource treats each packet sent as an independent economic problem, we are explicitly interested in the case where users value sending groups of packets and not just individual packets.\textsuperscript{122}

5.2.1 Demand When Time Does Not Matter

The fact that the slots are distinguishable, because they come at different points in time, is what makes this model of a network interesting. Suppose this that were not so, and the slots happened all at the same time and so were interchangeable as far as the users were concerned. In this case, the only thing that affects a user's utility is the total number

\begin{footnotesize}
\textsuperscript{120} The inspiration for this model comes from the digital cable paradigm. Some proposed MAC systems for cable do use slotted access controlled by the head end. But the cable problems are technically more complicated, and cannot ignore, for example, how requests for resources are communicated.

\textsuperscript{121} MacKie-Mason and Varian, "Pricing the Internet;" and MacKie-Mason and Varian, "Pricing Congestible Network Resources."

\textsuperscript{122} Shenker et al. point out the need to address this type of demand, and the failure of the MacKie-Mason and Varian model to do that, in. Shenker et al., "Pricing in Computer Networks: Reshaping the Research Agenda," 188.
\end{footnotesize}
of slots available. We can characterize each user $i$'s utility as the price she is willing to pay for a given number of slots plus the money she has left over after paying.$^{123}$  

$$u_i(x_i) + m_i$$

and so a user's demand is simply $u_i(x_i)$, a function on a single integer variable. If the provider charges the same price per slot to all users, then he sees an aggregate demand (the horizontal sum of the individual demand functions):

$$D = \left[ \Sigma_i u_i^{-1}(p) \right]^{-1}$$

and simply sets a price that maximizes his profit, which amounts to maximizing his total revenue since he has no variable costs in allocating slots. The provider's optimization problem happens over just one variable, his price. The solution may not be economically efficient, since the best price can lock out users who would buy at a lower price, but this is the nature of a problem with one provider who has no variable costs.

How does the provider gauge the demands of users? A provider could, for example, set a price for using each slot and see how users react; if this were a repeated game then over time the provider could find a price that maximized his revenue. We can also imagine a system where users communicate their willingness to pay and the provider sets a price based on this information. Just such a "spot market" system is proposed by MacKie-Mason and Varian for pricing transmission on the Internet.$^{124}$

If users are to completely reveal the prices that they are willing to pay for using slots, then this must be expressed as a function of the number of slots used, since we expect that using more slots has diminishing marginal value to users. It is not sufficient for users to specify just a single price in a system that would require users to reveal their preferences.

It is difficult to construct an example of a communication application that reflects the model of user demand in this example, one where time is of absolutely no importance. (If time does not matter, why not put off communication indefinitely?) But some applications care less about short-run time constraints than others. As Shenker et al. point out, when performing file transfers over a network, users care not about the individual packet performance but only about how long it takes the entire transfer to complete.$^{125}$ and

---

$^{123}$ Our notation is similar to that used by MacKie-Mason and Varian. This characterization of demand is a simplification of their congested resource model, since here there is no delay.

$^{124}$ MacKie-Mason and Varian, "Pricing Congestible Network Resources."

so if we look at only a small time scale, it is not when resources are allocated that is important but only that sufficient resources are allocated. A user performing an overnight back-up of a file system might care only that all the data be transmitted by morning, and her demand for resources might be zero for anything less than what is required to accommodate the job. Similarly, users of messaging applications (such as the delivery of electronic mail, faxes, or voice mail messages) care only that the entire message is delivered in a reasonable amount of time, which might not be immediate.

5.2.2 The Temporal Problem

In the exclusive slotted system, slots are in fact distinguishable in time. What happens when users demand particular slots?

Suppose that each user \( i \) demands just one slot, at a particular time \( t_i \), and no other slot will do. We can characterize a user’s demand as a function of which slot she desires and how much she is willing to pay, i.e. her reservation price \( r_i \):

\[
u_i(t) = \delta(t - t_i) r_i
\]

Implicit in this characterization also is that a user’s demand does not change over time and is independent of what happens to other users. This fixed demand makes sense in the context of a truly closed system, since it would take some external event to make a user’s demand change.

So far this is even simpler than the problem above where users demand multiple but indistinguishable slots. In this case, a user’s demand is expressed by just two values and the provider’s problem of maximizing his revenue amounts to, for each time period, identifying the user who is willing to pay the most.

We can easily extend this example to demands for multiple slots if we accept a further constraint that the demands for individual slots are not correlated, or put another way, if a user’s demand for one slot does not depend at all on whether she is allocated some other slot. Such a demand can be characterized as a sequence of reservation prices \( r_{i,t} \):

\[
u_i(t) = \delta(t - t_i) r_{i,t}
\]

Again, the provider’s aim here is to identify in each time period the user with the highest reservation price.

Demands that are not correlated between slots, however, cannot express important notions such as “I would like a slot somewhere in this range of time,” or “I need both of
these slots.” Put another way, these un-correlated demands cannot not express that one slot can be a substitute or a compliment for another.

Consider, for example, a user who would like to use two slots at \( t_1 \) and \( t_2 \), and no other slots. There are four outcomes that the user cares about: she gets neither, she gets the first and not the second, she gets the second and not the first, and she gets both. Her demand is completely characterized by the utilities of these four outcomes (and by which two slots she wants). But we can no longer write the demand as a function of time, because it does not make sense to ask what she wants at one time without knowing what happens at another. Instead, we can express the demand as a function of a binary sequence \( g_{i,t} \) representing whether slot \( t \) was granted to user \( i \). In this case:

\[
    u_i = f(g_{i,t_1}, g_{i,t_2})
\]

Such a demand for three users might look like:

| Slot Allocation | Utility \\n|-----------------|------------|------------|------------|
| \((g_{i,t_1}, g_{i,t_2})\) | \(u_{i,1} \) | \(u_{i,2} \) | \(u_{i,3} \) |
| (0, 0)          | 0          | 0          | 0          |
| (0, 1)          | 0          | 2          | 1          |
| (1, 0)          | 0          | 2          | 2          |
| (1, 1)          | 3          | 2          | 2          |

This expresses, for the first user, “I need both slots, and just one does me no more good than none;” for the second, “I need one or another, but having both does me no more good than just one;” and for the third, “I would rather have one than the other, but either is better than nothing.”

With these sorts of demands, the provider cannot optimize his decisions by considering slots individually. If the demand functions were completely general, then finding an optimal allocation of slots would require searching all possibilities. But in fact the demand function is constrained in that a user’s utility should never decrease when additional slots are granted (since in the worst case, a user has no use for more slots). This might make more specialized optimization techniques possible. But the complexity of even how a user’s demand is expressed increases exponentially with the number of slots a user might demand, so that if we have \( n \) slots in time, then a user’s demand is a function on \( n \) binary variables that we can express with \( 2^n \) values:
\[ u_i = f(g_{i,t_1}, g_{i,t_2}, g_{i,t_3}, \ldots, g_{i,t_m}) \]

All this complexity comes just from allowing for generalized correlation in a user’s demand between different slots. We have not yet introduced the possibility of correlation between what happens to different users or correlation with events external to the system. So far our demand functions are limited to those that remain static over time.

This model of time-dependent, correlated demands for resources comes much closer to specifying the needs of real applications than the time-independent model did. It allows us to specify both the sort of short-term needs for immediate transmission that a real-time application might need and the long-term needs for total throughput that a file transfer might need. For demands that can be completely specified in advance and do not change over time, this model is completely general. But that generality comes at the cost of tremendous complexity in the specification of demand. Any real system where users specified needs for communication services would need a much simpler model of user demands.

### 5.2.3 Expectations and Commitments

Why would a user want a commitment to use a certain slot? By a commitment, we mean some form of assurance, arranged in advance of when a slot arrives, that the user will be able to use that slot. It seems intuitive that knowing in advance that a slot is available for use is something that users would value. But in the scenarios presented so far, there is no basis for valuing a commitment in advance. Consider the following:

A user desires to use the slot at \( t = 2 \). Her reservation price for using this slot is \( r_2 \), and for using any other slot (or for using none) is 0. It is now \( t = 0 \). Suppose the provider offers at \( t = 1 \) a firm commitment to use the slot at \( t = 2 \). This commitment has some value to the user; let us say the value is \( u_i \). Can \( u_i \) possibly be greater than \( r_2 \)? As stated, the user’s utility is only a function of whether she ultimately is granted the slot at \( t = 2 \). The best she can ever do is to achieve a utility of \( r_2 \). Even if at \( t = 1 \) we say that if at \( t = 1 \) she declines the offer of commitment then she will definitely not be granted the slot, the most she is willing to pay is \( r_2 \), otherwise she does not maximize her utility. If the value of the commitment cannot exceed \( r_2 \), then in this case a commitment in advance adds no value.

Despite this argument, we intuitively believe that it is useful to know about things in advance. What is missing from our model? A simple way to give value to an advance
commitment is to introduce a mechanism that can change demand over time. Suppose our user at $t = 1$ must make a decision. The decision, whose outcome we denote as $d_1$, does something that affects the user’s desire to use the slot at $t = 2$. We can characterize the user’s demand as a function of the decision $d_1$ and whether the slot is granted at $t = 2$, denoted by $g_2$. For example, a user’s utility function could have the values:

<table>
<thead>
<tr>
<th>Decision $d_1$</th>
<th>Allocation $g_2$</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 5. Example of Utility as a Function of Decisions and Allocation

In effect, the user has an option at $t = 1$ that will leave her happier if she does not get the slot at $t = 2$, but if she knows she can get the slot, she is better off declining the option. Put simply, the option is a contingency that must be executed in advance. In this situation an advance commitment can have value. But that value depends on what the user’s expectations are about whether the slot will be granted.

Suppose that the user expects that the slot at $t = 2$ will be granted with probability $1/2$. The expected value of her utility is 2 if she chooses $d_1 = 0$, and 2.5 if she chooses $d_1 = 1$, so she would choose $d_1 = 1$. However, if she can just know the outcome of $g_2$ at $t = 1$, then the situation is different: she can choose $d_1$ based on that outcome, and assuming the outcome is resolved with the same probability, her expected value is 3. Resolving the uncertainty earlier, even though it is not resolved any differently, has a value of 0.5 to the user.

We can make a similar argument in the context of a utility that is correlated between two slots. Suppose our user would like to use two slots, one at $t = 1$ and one at $t = 2$, and that just getting one of these slots has no utility to her. Suppose also that the user can request a grant for a slot as that slot arrives, and if it is granted must pay a fixed cost to use the slot, so that the user ends up with a net utility that depends on whether the slots were granted:
Table 6. Example of Utility with Prices

<table>
<thead>
<tr>
<th>Allocation</th>
<th>Demand</th>
<th>Cost</th>
<th>Net Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0, 1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1, 0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1, 1</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

The situation is similar to the previous example because it involves decisions: here the user decides at \( t = 1 \) and \( t = 2 \) whether to request a slot. If the expected probability of a grant is 1/2, then (working backward through the decisions) she will ask for \( g_2 \) if \( g_1 \) was granted, with an expected net utility of 1, and will not ask for \( g_2 \) if \( g_1 \) was not granted, leaving her outcome as 0. Thus before \( g_1 \) is decided, her expected net utility is 1/2.

If this is the only way a user can request slots (just before they are granted), then this is the best our user can do. But if we allow the user to make a request at \( t = 1 \) that says, "If you can give me both \( g_1 \) and \( g_2 \), I will pay you 2, otherwise I do not want either," then the situation is different. Even if we still expect \( g_1 \) and \( g_2 \) each to be resolved with the same probability of 1/2, the user does better. Before \( t = 1 \), she expects an outcome of 3 with probability 1/4, and zero with probability 3/4, so her expected surplus is 3/4. Just as in the previous example, the ability to resolve uncertainty (through a commitment) has a value to the user. Furthermore, in the case were we can make commitments, the expected number of slots used is diminished. But in this example the incentive for making commitments comes not from external decisions but instead from the inability to completely specify a correlated demand in advance. With full advance knowledge of the user's needs, the provider might have been able to find an efficient solution without the need for commitments.

We argue that a definitive need for commitments in our system comes from sources external to the system, particularly from a user's ability to exercise contingencies that affect her demand. Thinking about real user needs in communication systems, this makes sense. A telephone user who places a phone call does not want to be cut off in the middle of the call, and might even prefer not placing a call at all over losing the connection in the middle on the grounds that she has something better to do with her time. One mechanism that addresses this need for commitments is admission control, which exists in telephone networks and in ATM networks but not, currently, in the Internet. But, as we
argue below, admission control might be too simple a solution to a very complicated problem of making commitments for these networks.

5.2.4 Uncertain Demands

The examples above demonstrate that a commitment can have value in a situation where a user has an option, external to our system, that would change her demand. But user demands can change for other reasons, and do so enough that any static model of user demand might be fundamentally unreasonable. Consider a user who would like to use network resources to place a phone call. Before the moment she decided to make the call, her demand for resources was very different, and when she decides to terminate the call, her demand will again change. We have argued that, in the situation where a user is not making a call, our user would like some sort of reasonable expectation of being able to make a call if she should decide to do that. But the characterizations of user demand that we have discussed above do not include any way of distinguishing between a certain need for resources and an uncertain need.

Suppose, for example, that fifty users would each like the opportunity to use one of ten slots sometime in the future, and each knows that she is likely to actually use a slot with a probability of 0.1. As long as the provider believes that the chances of more than ten of the users actually needing a slot is unlikely, it may be reasonable to make commitments to all of these users. The sort of commitment made needs to have a provision for the event in which the demand cannot actually be met; it is not a definite commitment of resources, but users might be willing to live with that.

Why is it important to be able to make such commitments? Suppose, in the same system, there is one more user who is competing for the same resources, but who would like to use all of these slots with certainty. If the provider has not allocated these slots for uncertain use, then the fifty uncertain users may be locked out. The willingness to pay of the five or so users who we would expect to need those resources might be higher than that of the one user with a certain demand, in which case a failure to reserve resources for these uncertain demands would result in an economically inefficient allocation of resources.

It is simple to demonstrate a need for commitments in the face of uncertain demands, but solving this problem would be much more challenging. We have not addressed how users could describe uncertain demands to a provider, nor how users might be made to have the proper incentives to honestly reveal their uncertainty, nor how
providers might specify commitments to users that they might not always be able to fulfill. But this problem shows up in real network systems today (the telephone examples are provided to demonstrate just that), and it only becomes more complicated in integrated services networks that share resources among users who have heterogeneous demands.

5.3 Solutions for Quality of Service

Our abstract model of quality-of-service demands is useful as a demonstration of how the need for commitments arises and why even specifying demands for service is a difficult problem. Actual solutions that implement QoS commitments for integrated services networks take a much more practical approach to the problems we discuss. Here we briefly examine two such solutions, one proposed for the Internet and the other for ATM networks. Both are relevant to the problem of delivering QoS commitments on an integrated services cable network: ATM because its mechanisms are used as the basis for communication mechanisms in some digital cable proposals, and the Internet because it could become a ubiquitous platform for delivering real-time services.

5.3.1 Integrated Services for the Internet

Mechanisms that would add real-time services support to the Internet are presently being debated, developed, and standardized, and as such we can only critique these mechanisms as works in progress. We consider the work of the Internet Engineering Task Force (IETF), particularly the Integrated Services Working Group (ISWG). An overview of mechanisms for supporting real-time services is given by Braden, Clark, and Shenker.\textsuperscript{126}

The problem of engineering support for real-time services on the Internet is in many senses far more difficult than the localized problem of real-time support for digital cable networks. The Internet is composed of many network elements that conspire to deliver traffic in a way that involves little centralized control, and so information about the characteristics of one portion of the network is not generally known to another portion. Furthermore, the Internet protocols work without maintaining within the network information about individual connections, and this is impractical for real-time services.

\textsuperscript{126} Braden, Clark, and Shenker, "Integrated Services in the Internet Architecture: An Overview."
With a digital cable network, it is feasible and straightforward to maintain central control and even connection state at the head end.

The proposed Resource Reservation Protocol (RSVP) addresses the difficult problems of establishing routes through the Internet that will support real-time traffic and maintaining state for connections within the network.\(^{127}\) RSVP presents an interface through which applications can request connections that support real-time services, but the actual sorts of services that can be requested are not specified by RSVP itself. Rather, RSVP acts as a generalized mechanism that brokers requests for services without being tied to a particular model of services.

As a mechanism that works independently of the notion of services that would be provided on the Internet, RSVP is just the sort of architectural solution that we advocate. But RSVP is not the complete picture of real-time services for the Internet. The actual services models that work below the level of RSVP are nonetheless visible to applications (which request specific services through RSVP). Network elements within the Internet, specifically routers, must implement mechanisms that are specific to the service models that they support. This raises the question of how services can be supported when different network elements support different service models; Breslau and Shenker suggest that with some ability to replace one service with another, applications might be able to cope with heterogeneous services within the network.\(^{128}\)

What service models are proposed for the Internet? The ISWG drafts propose a framework for specifying services,\(^{129}\) and within that framework have defined the following services that could be offered in addition to the best effort service that the Internet currently provides:

- **Guaranteed service** delivers packets within bounded delay times, with no congestion losses as long as the traffic conforms to specifications. It is intended for applications that are intolerant of both packet delays and packet loss.\(^{130}\)

• **Predictive service** provides a delay bound that is fairly reliable, but can occasionally impose congestion losses on traffic. The intent of this service is to allow for better network utilization by relaxing the service commitment.  

• **Controlled delay service** allows applications to dynamically adapt their service requests in the face of changing network conditions.  

• **Controlled load service** provides behavior that approximates best-effort service under unloaded conditions.

Each of these service classes is parameterized so that applications can specify the amount of bandwidth and delay bounds that are required. The common model that is used to parameterize traffic requirements is a *token bucket*, which in effect specifies both a sustainable traffic rate (the token rate) and an amount of data that can exceed that rate for a short period of time (the bucket size). In terms of our generalized model of correlated user demands, the token bucket mechanism suggests an important simplification. A token bucket controls the extent to which usage in one time period can affect usage in another; the most short-run capacity a user could ever demand above the sustained level is limited by the bucket depth.

Each of these services performs admission control, so that when the network is too loaded to support the requested connection, service is denied. Once admission is granted, the network has committed to supporting the connection indefinitely, as long as the application keeps the connection open. This admission paradigm is similar to what telephone systems do: we request a call when we dial the number, and once the call is established it is not dropped due to congestion. This simplifies the problem of making admission control decisions, since the only question becomes whether the necessary resources are immediately available. But with this simplification a user can neither request in advance a need for service nor specify when service will end. In some cases, applications may have information available that would allow their needs to be specified in advance: some applications may require regularly scheduled communication, and any file

---

transfer (such as an image or text transfer initiated by World-Wide Web browsers) should be able to specify up front how much data is to be transferred.

Some of the ISWG service models, particularly the controlled load service, are intended to be implemented in a way where a provider might commit to more service than he could deliver under all circumstances, expecting that the average use would not overload his resources. If the controlled load service is meant to deliver network performance that mimics an unloaded best-effort network for applications that work well in this environment, then perhaps what users request with this service is the ability to communicate now and then with the understanding that they are usually idle. This is similar to the sort of commitment that we would expect to arise from user demands that are uncertain, the situation discussed in section 5.2.4. But the controlled load service in fact defines no probability of transmission, and only characterizes traffic in terms of the same token bucket model that the other services use.

Two fundamental differences exist between the abstract user-provider interactions that we discuss and the ISWG model of Internet services. First, the ISWG services limit the way in which users express their demand for services by offering a small set of parameterized services. Second, service requests in the ISWG model include no pricing information: users do not express what they are willing to pay for a service (as they would in the model suggested by MacKie-Mason and Varian), and so cost of service is presumably implied by the services that users request. Neither of these differences necessarily identifies a flaw in the ISWG model. So long as the service models are sufficiently closely aligned to the needs of applications, network efficiency is not lost. And similarly, economic efficiency can be achieved if the pricing models sufficiently reflect user demands.\(^{134}\)

5.3.2 ATM Services

In contrast to the Internet protocols, Asynchronous Transfer Mode (ATM) network technology was designed from the start to support applications, particularly real-time applications, that require specific QoS support. The service model that ATM adopted has been augmented over time and is presently undergoing a standardization process, but it is

\(^{134}\) Shenker makes similar arguments in more detail in: Scott Shenker, "Service Models and Pricing Policies for an Integrated Services Internet" (Palo Alto: Xerox Corporation Palo Alto Research Center, 1995).
well described by specifications produced by the ATM Forum. The ATM model consists of five services:

- **Constant Bit Rate Service (CBR)** is intended for applications that require a continuous data stream at a constant rate and with constrained delay characteristics.

- **Real-Time Variable Bit Rate Service (rt-VBR)** is intended for applications that have variable but well characterized bandwidth needs and can tolerate only limited delay.

- **Non-Real-Time Variable Bit Rate Service (nrt-VBR)** is intended for applications that have similarly variable bandwidth needs but no delay constraints.

- **Unspecified Bit Rate Service (UBR)** is intended for applications that do not have specific bandwidth or delay requirements. The service offers no performance guarantees.

- **Available Bit Rate Service (ABR)** is intended for applications that can respond to changing network performance. The service provides feedback to applications and may dynamically change the performance commitment of a connection.

The VBR services parameterize bandwidth requirements in terms of a sustainable rate and burst characteristics that are very similar to the token bucket scheme that the Internet ISWG proposals adopt.

Like the Internet services, the ATM services involve admission control, make admission decisions at the time the of the request, and provide indefinite commitments. This raises the same concerns about the inability to make advance or fixed-duration requests for service that we discuss in the context of the Internet services. The ABR service does not really address those problems, but it does attack a similar problem of how a network provider can respond to changing demands in the face of existing commitments.

The ATM service model does not embed pricing into service requests, but instead assumes (like the Internet models) that the price of a service is implicit in requests. This may in fact be an efficient way of selling services. But Murphy, Murphy, and MacKie-Mason raise an interesting objection in particular to how ABR service might be supported. They argue that if priority is always given to accepting CBR and VBR traffic on ATM networks at the expense of diminishing service to ABR traffic, this could result in

---


136 Murphy, Murphy, and MacKie-Mason, “Feedback and Efficiency in ATM Networks.”
an inefficient allocation of resources. Network resources are valuable to users even when they cannot characterize their needs in advance, and pricing and allocation mechanisms need to reflect this.

5.4 Conclusion

Delivering quality-of-service communication over digital cable networks, which is required for an integrated services approach, is a difficult problem. The discussions above raise both technical and economic issues that point to the sources of this difficulty. The cable medium is a shared and limited communication resource. Users and applications have different demands for communication resources, and characterizing these needs is inherently complicated, particularly when we understand that real demands can involve advance expectations and uncertainty. A solution for allocating and committing resources addresses an economic problem, not just a technical problem, and we cannot judge the applicability of a system on technical grounds alone without understanding how services are sold. But achieving economic optimality is not the goal; a solution that is simple, flexible, and has the right sort of economic incentives to encourage reasonably efficient use certainly can be preferable to one that achieves a slightly higher degree of efficiency.

In a sense, the problems of providing QoS in an integrated services cable system are similar to the problems of doing this in any integrated services network, and it is likely that solutions such as those adopted for ATM or Internet services will find their way into digital cable architectures. The risk in adopting these solutions (or any others) is that the service models that are supported in a system might be inappropriate or inefficient for the applications that we would like to use. The mechanisms that support QoS are, to an extent, built into cable systems. It would be a shame if the limitations of future digital cable systems provided barriers to using future applications only because we had not anticipated those applications when the systems were designed.
6. Conclusion

6.1 Technology

Our analysis of cable technology begins with considering cable not as a television system incidentally used to provide other services but instead as a general-purpose communication medium. Seen in this light, the technology of modern two-way hybrid fiber/coax cable systems is powerful, flexible, and suitable for delivering a wide variety of communication applications to our homes, including telephony, video on-demand, computer network access, and others. The cable medium provides a substantial amount of bandwidth for communication, but its resources are shared between multiple users. Given the sorts of applications we would like to carry on cable, its shared nature creates fundamental challenges for communication systems that use cable, notably: the need for mechanisms to share network resources among users and applications, the need to ensure security and privacy for communications, and the need for high network reliability.

Although many of the engineering challenges for carrying these applications are confined to the communication systems that use cable networks, supporting new services will put new demands on the cable infrastructure itself. Many cable networks are built to provide only a small amount of upstream bandwidth, and this can constrain the use of cable for applications that can make use of high speed two-way connections. Cable networks may lack the level of reliability required to support some applications (particularly telephony), a problem that could be addressed with the addition of back-up power systems and monitoring equipment. As cable operators upgrade their networks to hybrid fiber/coax systems, they should consider carefully how infrastructure choices can affect their options for future communication services.

6.2 Economics

Although the cable business has long been dominated by television service provided in an environment of limited monopoly power, cable operators today face both the prospect of increased competition in television services and the promise of delivering new communication services with their networks. As much as new communication technology puts cable operators in the desirable position of being able to sell more services, the infrastructure investments that cable operators need to make pose a difficult problem:
the sunk costs of these investments must be recovered by selling into competitive and uncertain markets.

The benefits of integrated services cable systems for cable operators come from both the economies of scope that they can provide in delivering multiple services with one system, and from the strategic advantages of being able to flexibly and efficiently provide services. Sources of economies of scope include reduced plant qualification costs for new services, reduced costs for terminal equipment that implements multiple services, and the ability to dynamically share cable bandwidth between different services. Strategic benefits come from the ability to deter competition by demonstrating an ability to provide service at low cost, and from avoiding competing solely on price by bundling and differentiating service.

Our analysis of cable economics has several implications for designing integrated services systems for cable. These systems should be designed to be able to take advantage of the economies of scope that an integrated services network can provide, for example by sharing equipment between multiple services. Designs that would drive up the per-subscriber costs of providing services, however, will be disadvantageous when those services are sold in a competitive environment, and so systems should allow the costs of equipment to scale with its communication needs. Although many existing cable systems have a shortage of upstream bandwidth, in the long run as cable operators upgrade their systems the costs of bandwidth should be small. The abilities to use cable systems flexibly and to scale systems as new applications demand more resources may be more important than using bandwidth efficiently.

6.3 Architecture

Designing integrated services communication systems for cable networks is a formidable task that must bring together a broad range of advanced technologies to achieve flexible and efficient systems. A well-defined integrated services architecture for cable could both foster the development of communication products for cable by establishing interoperable standards and provide technology that can meet the needs of both current and future applications. Standardization efforts being pursued by Cable Television Laboratories and by the IEEE Project 802.14 Working Group may achieve just this.

The challenges for defining an integrated services architecture begin with finding physical layer solutions for encoding and transmitting data on cable networks and defining
media access control schemes that allow multiple users to efficiently access the cable medium. Media access control solutions must in particular support quality-of-service communication, especially for real-time applications. An architecture must also address several system-wide issues for cable networks: it should provide mechanisms for security and privacy, the ability to install and maintain network equipment without requiring complex configuration, and provisions that will enable cable operators to achieve the high reliability required by some applications.

6.4 Quality of Service

The problem of providing quality of service in communication systems is, at its heart, an economic problem of allocating congestible network resources among different users. The need for quality-of-service support in integrated services systems is especially pronounced because these systems must support applications that have different needs for network resources, including real-time applications that may have especially constrained performance requirements. For cable systems, architectural solutions for quality of service may affect how cable applications can interoperate with other networks and how cable networks can adapt to future applications.

We have shown, in an abstract model of communication, how the need for commitments can arise from time-dependent communication demands combined with alternatives that are external contingencies. Put more simply, users want to know in advance whether their communication will work so that they can do something else if it will not. A mechanism that allocates network resources on a packet-by-packet basis would be inefficient in the presence of such needs. But the general problem of specifying and satisfying needs for commitments appears to be extremely complex. In practice, service commitments in communication systems are made with imperfect information, rely on the statistical nature of demands, and are not always firm.

Solutions for supporting quality of service in the Internet and in ATM networks adopt specific service models that define the sorts of commitments that can be requested from a network. These mechanisms provide for some of the basic needs of existing real-time applications, but they lack the sophistication necessary to address more complicated needs, such as reserving resources in advance of when they are needed. Although we have not proposed better solutions to these problems, our analysis of how expectations and the need for commitments arise in computer networks suggests that a more generalized model
of demands, and one that explicitly takes into account pricing, might better be able to satisfy a variety of communication demands.

6.5 Future Work

The arguments that we have presented advocate the integrated services approach as a long-term solution for communications on cable networks. The benefits of integrated services, however, come not from any such solution but from one done properly. Designing a successful integrated services architecture for cable will require careful consideration of technology, equipment costs, operational costs, and the demands of users and applications. It may also require a better understanding of technical problems, particularly of how to support quality-of-service communication in integrated services networks, than is known today.

6.5.1 The Need for Standards

A standardized communication architecture for cable, particularly one that enables different manufacturers to produce interoperable equipment, could benefit both cable operators and equipment manufacturers. Cable operators do not want to invest in systems that would be made obsolete by competing technology, and would like to avoid being locked into buying proprietary equipment from a single provider. Manufacturers of communication systems do not want to invest in developing products that might be made obsolete by future competing systems. Standards might give both cable operators and cable manufacturers confidence to invest. A standard that achieves interoperability need not entirely define how systems are built, and might leave room for innovation and differentiated products.

A single standard for cable communication may not be necessary. Cable systems operate independently of one another, and if cable operators ultimately own and control the devices that are attached to their networks, different communication systems could be used with different networks. Well-defined applications for cable, such as telephone and Internet service, may be sufficient to insulate users from non-interoperable cable systems that provide these services. However, if consumers purchase the terminal devices that they attach to cable networks (as is done with telephone networks today), the situation is different: consumers may demand a universal standard that would allow them to move their equipment between cable systems.
As cable operators upgrade their networks to modern hybrid/fiber coax systems, they face infrastructure decisions that can affect their ability to support future communication systems: how much bandwidth the network should support, how large neighborhood nodes should be, how power should be provided in the network, and so on. We have argued that the relatively small amount of upstream bandwidth typically provided by cable networks could create unnecessary complications and limitations for cable communication, but the cost of making more upstream bandwidth available when upgrades are performed is small. But communication systems may be engineered making specific assumptions about cable performance, including how upstream and downstream bandwidth is allocated. Therein lies a paradox: without confidence that cable systems will support a better upstream bandwidth allocation, cable equipment manufacturers might not build equipment that would take advantage of it; but without available equipment to use it, cable operators might not find reason to build more upstream bandwidth into their systems. Even before new cable communication systems are developed, standards that define cable infrastructure performance, particularly how more upstream bandwidth should be provided, could significantly promote the development of better communication systems.

6.5.2 How Will Cable Benefit Users?

A presumption of our work is that as long as user demands ultimately drive the services that are sold on cable systems, users will derive benefits from cable systems that let them buy more of what they want. This view, however, is only a starting point to understanding how cable (and other communication systems) can benefit users. Although the services at stake for cable today are very different from what cable could do in its earlier days, the philosophical problem of deriving public good from cable systems still retains its arguments. If cable systems are monopolies, they can exercise control that would raise prices above what would exist in a competitive market and constrain the content and services that they provide. The effects of such monopolization can be both to take economic benefits away from users and (perhaps more disturbing) to limit the market for services in ways that results in economic inefficiency, where the potential net benefits to cable operators and users are not realized.

The situation that changes this picture is increased competition in the telecommunication industry, driven by both technological and political change. Our work
is an example of how digital technology can transform an infrastructure built to provide a single service into one whose offerings would compete with several other telecommunication systems. The Telecommunications Act of 1996 exemplifies the trend of moderating telecommunication monopoly power through competition instead of controlling it through regulation. If competition effectively limits the ability of cable operators to monopolize their services, then users will find benefits in both better services and lower prices.

The lessons of history, however, should not be lost in the movement to promote telecommunication development. Cable television faced competition from the beginning in the form of broadcast television, but over time cable operators were able to transform their television service into one that outperformed the competition and one that many consumers found to be essential. The monopoly power that cable television operators found came from a technological advantage that allowed cable to carry more channels (as well as content that broadcasters could not carry), combined with an economic advantage of being the only provider who could do this. The mere presence of multiple providers of television service was not sufficient to guarantee that one system did not gain market power.

Cable today is once again in the position of being able to provide services that would compete directly with what other telecommunication providers offer. But the true power of cable technology may come not by matching what its competition does but by exceeding it. The benefits of integrated services systems come from both supporting those applications in use today and enabling future applications. The challenges of integrated digital services for cable networks include both understanding how this technology will be built and ensuring that it will serve the public good.
Bibliography


http://rpcp.mit.edu/Pubs/inet/abstract.html


http://info.internet.isi.edu/in-notes/rfc/files/rfc1633.txt


draft-ietf-intserv-hetero-01.txt


_______. "Request for Information & Industry Development Proposals (RFI/IDP) for a Integrated Multiple Services Communications Network (MCN)." Louisville, CO: Cable Television Laboratories, 7 October 1994.


_______. "Request for Proposals: High-Speed Cable Data Service (HSCDS)." Louisville, CO: Cable Television Laboratories, 26 April 1995.


http://rpcp.mit.edu/Pubs/AAI_Local_Access_Paper/abstract.html


http://www.cablelabs.com/CATV.pdf


_______. “True Multimedia Capability: Cable on Track to Get What It Wants in Data.” *Communications Engineering & Design*, March 1996.


ftp://alfred.sims.berkeley.edu/pub/Papers/
Pricing_the_Internet.ps.Z


ftp://alfred.sims.berkeley.edu/pub/Papers/pricing-congestible.ps.Z


http://www.spp.umich.edu/spp/papers/jmm/icc96.pdf


http://xerxes.nas.edu:70/l/nap/online/rtif
http://www.commerce.net/work/pilot/nielsen_96/exec.html


draft-ietf-intserv-control-del-svc-02.txt

draft-ietf-intserv-svc-template-02.txt


ftp://ftp.loc.gov/pub/thomas/c104/s652.enr.txt


http://www.tns.lcs.mit.edu/publications/VI/abstract.html


draft-ietf-intserv-ctrl-load-svc-01.txt


