Chapter 1

Introduction

The latter half of the 1990s has seen a resurgence of interest in satellite-based data networks. Satellite communication systems have long been one of the hallmarks of advanced communications technology, with their remarkable and distinctive ability to link most of the populated areas of the Earth. Yet, until recently, the satellite communication industry had increasingly begun to look more like a dinosaur, with competition from fiber optic and terrestrial wireless networks steadily eating away at the industry's most profitable markets.

As of this writing, however, the satellite industry is poised for rapid growth, with funding in place for the deployment of technically ambitious, multi-billion dollar systems, and growing competition in both service provision and hardware manufacturing. Figure 1.1 illustrates one analyst's projection of the growing market for satellite broadband data services [137], and other analysts project that the growth in the satellite industry will outpace the growth of the entire communications market over the next ten years, as the satellite sector's market share rises from 2.3% today to 6% a decade from now [94]. What has triggered this rapid turnaround? The answer lies in the confluence of two economic and technological trends:

1. **The Internet boom** The 1990s will likely be remembered as the decade during which the Internet came of age. There is presently an incredible (and increasing) demand for faster and cheaper Internet services, and many companies are scrambling to offer these broadband services. Satellite networks provide a fast way to reach customers because they do not rely on buildout of a high-speed terrestrial network, which may take years to accomplish in many areas of the world. Moreover, with the advent of the World Wide Web [12], broadband Internet access tends to be highly asymmetric in traffic usage, with users downloading (consuming) much more information than they generate. As we shall discuss, this type of traffic pattern matches well with satellite networks, where it is much cheaper to receive data at broadband rates than to transmit at such rates.

2. **Advances in satellite technology** The rapid technological progress that has spurred the growth of the Internet has also helped to significantly advance the state-of-the-art in satellite technology. Most notably, miniaturization of electronics has allowed more and more sophisticated satellite and terminal hardware to be economically deployed. Satellites, which once were mainly repeaters in space, have much more on-board processing functions and have the capability to juggle multiple directional "spot" beams on the Earth's surface while also com-

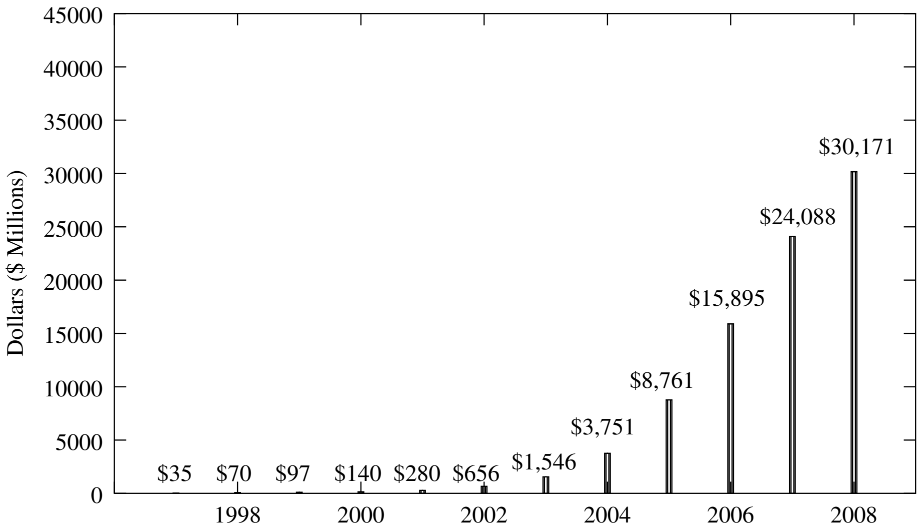

Figure 1.1: Projections of Annual Satellite Broadband Revenue (Source: Merrill Lynch and Co.).

municating with other satellites via high frequency radio links. Sophisticated constellations of low-earth-orbiting satellites, and handsets and terminals that can track the motion of these satellites, are now being designed and deployed. New frequency bands at 20-30 GHz, the use of which was once precluded by the lack of affordable high-frequency hardware, are now being opened to satellite communications, greatly increasing the available bandwidth of newer satellite systems. Probably the most widely-observable evidence of the impact of advances in electronics on this industry can be seen in the growth of direct-to-home (DTH) satellite television services, one of the most rapidly deployed consumer electronics product in history. DTH services, using small, affordable satellite dishes, have brought satellite services into the mainstream in a way that was not possible using technology of a decade ago.

While the use of satellite networks as a part of the Internet backbone dates back almost twenty five years, the use of satellites to provide high-speed network access is relatively new. The success of new satellite networks in delivering high-speed access hinges on the ability of the underlying protocols to function correctly and efficiently in the satellite environment, an environment characterized by (for traditional geostationary (GEO) satellites) much longer propagation delays than are found in terrestrial networks, and (for newer low-earth-orbiting (LEO) satellites) a rapidly time-varying network topology. In this dissertation, we concentrate on the application of satellite systems to provide this “last-mile” access connectivity, and focus in particular on two problems relevant to Internet data networking over these broadband next-generation systems: i) improving the performance of reliable transport protocols over high-latency paths, and ii) routing strategies for networks of low-earth-orbiting satellite systems. Before introducing these problems in more detail, we first digress to describe the fundamental characteristics and technological trends of both satellite communication systems and the present-day Internet.

1.1 Satellites and the Contemporary Internet

1.1.1 A Brief Overview of Satellite Communications

The field of satellite communication systems is a rich, multidisciplinary field involving several areas of electrical, aeronautical, and mechanical engineering. Several books provide overviews of the field as a whole; among the works in wide use today are those by Maral and Bosquet [81], Gordon and Morgan [50], and Pratt and Bostian [113].

The first idea of using a satellite orbiting at a geostationary altitude (35,780 km above the equator) to provide communication services is attributed to author Arthur C. Clarke in 1945. The first artificial communications satellite (SCORE in 1958) did not follow long after the Sputnik launch in 1957, and the first commercial geostationary satellite (INTELSAT 1, or "Early Bird") in 1965 ushered in the era of overseas telephony via satellite. This first INTELSAT satellite had a capacity of 480 telephone channel at an annual cost of $32500 per channel [81].

In the 1970s and 1980s, both the market for satellite communication services and the technology grew rapidly. Besides providing international telephony and data services between large earth stations owned by national carriers, communication satellites were increasingly used for video (television) distribution. The international organization INMARSAT was founded to provide telephony and data services to maritime customers. As we describe later, the first satellite network experiments based on packet switching (the Atlantic Packet Satellite Network, or SATNET) commenced in 1976. Finally, the construction of systems based on Very Small Aperture Terminals (VSATs) for transaction-oriented traffic such as credit card verification and database management was begun in the 1980s.

As mentioned above, in the 1990s the growth of alternative, cheaper technologies such as high speed fiber optic networks has gradually eliminated much of the international telephony service for non-mobile customers. However, technological advances enabled the creation of direct-to-home (DTH) satellite television services that are competitive with cable television systems. And because of the explosion of interest in the Internet, in the latter 1990s satellite channels have begun to be used for trunking between international Internet Service Providers and the US backbone.

In the coming decade, having been revitalized by the demand for broadband Internet access and broadcast television, the satellite market is poised for large growth. The two biggest technological advances are likely to be the emergence of systems at Ka-band (20-30 GHz) and systems composed of tens to hundreds of low-earth-orbiting (LEO) satellites. Ka-band systems are advantageous because they permit satellite terminals smaller than one meter in diameter to be used for two-way communications, and because the amount of spectrum allocated in this new frequency band is larger than previous allocations at lower frequencies. As we explore later in this thesis, LEO systems promise to offer services with much lower latency and terminal power requirements than those offered by geostationary satellites. Figure 1.2 illustrates the current placement of the roughly 200 commercial geostationary communications satellites in orbit around the Earth; with Ka-band systems and directional antennas that allow more than one satellite to occupy an orbital slot, the density could double in the next decade. Table 1.1 summarizes some commercially proposed LEO systems under development. In the long run, satellites are well positioned to offer broadcast services at competitive rates and to provide general communications capacity to points on the globe that are not well served by terrestrial networks. However, it is clear that there is insufficient allocated spectrum capacity for satellites to significantly displace terrestrial wireline or wireless networks, even if

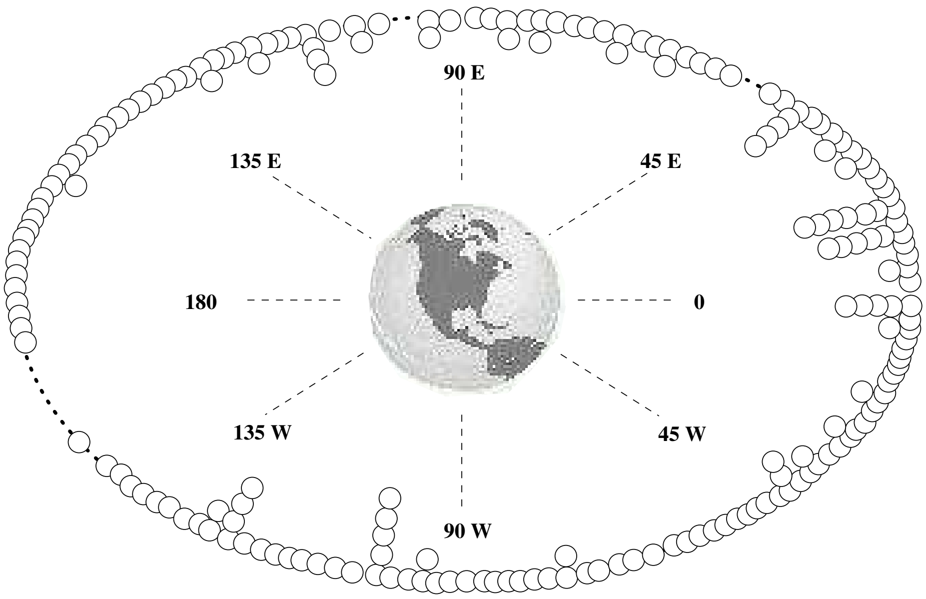

Figure 1.2: Current commercial communications satellites in geostationary orbit (Source: Hughes Space and Communications Company; reproduced with permission).

it were economical to do so [70].

1.1.2 The Internet Protocol Architecture

The term "the Internet" refers to a wide collection of packet switching networks that are tied together through the common use of the Internet Protocol (IP) and its associated routing and addressing conventions. Each network can be thought of as a separate "autonomous system" that takes responsibility for delivering traffic within its own network however it sees fit while conforming to standard protocol mechanisms at exchange points (interfaces) with other participating networks. The most distinguishing characteristic of this network architecture is that it is decentralized and has no single administrator. Another key aspect of the architecture is how the various protocols interrelate. Figure 1.3 illustrates a popular view of the Internet protocol architecture, sometimes called the "hourglass figure," which illustrates how there is one common protocol (IP) used by everyone (at network exchange points) but that protocol layers above and below the IP layer are more heterogeneous [110]. In fact, it was the need to interconnect different networks such as the original ARPAnet, SATNET, and the Mobile Radio Network in the San Francisco Bay area that propelled the usage of the Internet protocol [121]. In Figure 1.3, we have labelled five network "layers" that are commonly associated with the Internet architecture (also sometimes referred to as the "TCP/IP" architecture). In this thesis, we will explore problems involving protocols that lie at the network and transport layers.

The current Internet can be characterized by the following features:

- **Best effort packet delivery** The Internet makes no guarantees about bandwidth, latency,

	Iridium	Teledesic	Skybridge	Globalstar	ICO
Uses/services	voice, messaging fax	broadband access, private networks	broadband access, private networks	voice, messaging fax	voice, messaging position location
Data rates (Kb/s)	2.4	64,000 down 2000 up	20,000 down 2000 up	7.2	2.4
Number of satellites	66	288	80	48	10
Orbital planes	6	12	8	8	2
Altitude (km)	780	1375	1450	1410	10,400
Frequency band	L-band	Ka-band	Ku-band	L/S-band	S-band
Payload type Payload type	circuit switched	packet switched	repeater	repeater	repeater
Satellite crosslinks	yes	yes	no	no	no
System costs ($ billions)	3.7	9	4.2	2.2	2.6

Table 1.1: Summary of major LEO system proposals (data from various sources, but mainly [143]).

sequencing, or even the successful delivery of a packet. Instead, packets are routed among destinations as best as the routing infrastructure can do, and it is up to higher-layer protocols that operate end-to-end between corresponding hosts to provide whatever service guarantees are necessary (such as in-order, reliable delivery). This architectural decision has contributed to the nice scaling properties of current Internet deployment, albeit at the cost of supporting poorly those applications that require strict performance guarantees from the network.

- **Heterogeneity** Any way you look at the Internet, there is a tremendous amount of heterogeneity. The performance characteristics of end-hosts and communications links operating in the Interent vary widely. The success of the network relies on successful deployment and operation of protocol mechanisms that mitigate the problems posed by this heterogeneous environment. Two recent research projects in our research group (one on multimedia-related proxies for heterogeneous networks [7] and one on transport protocol performance over heterogeneous wireless networks [9]) focused specifically on dealing with heterogeneity in the Internet.

- **Huge installed base** Throughout the 1990s, the number of hosts on Internet has grown exponentially. As of July 1999 there were over 56 million hosts, double the number present at the start of 1998 [63]. As a result, it is increasingly hard to make protocol changes that do not interoperate cleanly with existing hosts on the network. New protocols or protocol enhancements that require changes to end hosts are not likely to be quickly adopted unless they either enable a new service not easily supported by the existing protocol base or provide a very large performance enhancement. Therefore, the onus is on the developer to elaborate a clear deployment path for any proposed enhancements and to demonstrate that the proposed

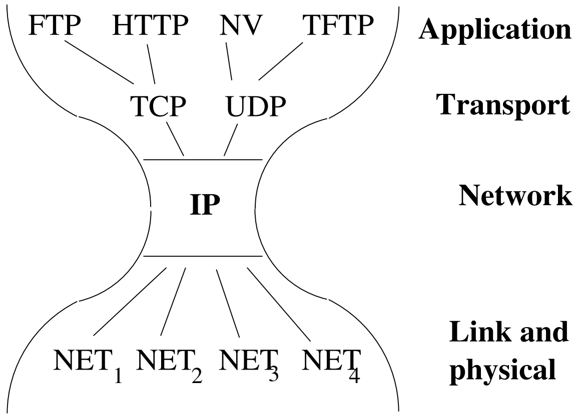

Figure 1.3: The Internet protocol architecture.

changes interoperate well with the installed base.

1.1.3 Satellites and the Internet– A Match Worth Making?

It is already well known and unavoidable that the end-to-end packet delay over GEO satellite links is commonly several hundred milliseconds worse than most high-speed Internet connections. Although future LEO systems are aimed at reducing this absolute delay, the transmission (error) performance of these systems is uncertain at this time and is likely to be worse than what users typically observe with wireline systems. Both these long latencies and (potentially) high error rates can cause performance problems for Internet connections. Often, the performance problems can be overcome by proper protocol design, but experience has proven that it is unrealistic to expect that protocol designers will take the special characteristics of a satellite link into account when designing protocols– satellites are too much of a niche market for optimization. Nevertheless, we believe that the Internet and satellite links can be a very good match for providing broadband access, for the following reasons:

- **Current Internet performance is spotty** Packet loss rates in the Internet can be quite high; it is not uncommon for packet loss rates on portions of the Internet to approach 10 to 20 percent at times due to congestion. Many current Internet users use dial-up access, which (at rates of 30 Kb/s) can add hundreds of milliseconds of delay for large packets. Additionally, Web server response times, while usually fast, can sometimes be very sluggish (i.e., server response times have heavy-tailed distributions [51]). In such an environment, satellite link performance does not look bad in comparison. In fact, for some applications such as low-bit-rate video, it is possible for satellite transmissions of a video stream to be decoded earlier than a terrestrially-routed stream if the terrestrially-based stream must overcome a higher packet loss rate. Because there is no admission control nor fine-grained traffic policing in the Internet, it is not clear whether the performance will improve markedly anytime soon.

- **Elasticity of popular applications** Geostationary satellite connections are considered problematic for "real-time" applications involving human interactions; for example, telephone calls over satellite links encounter enough latency that the timing nuances of normal conversation can be disturbed for many users. These types of applications have been termed "non-elastic" [125]. In contrast, the two most popular uses of the network– Web related activities and email– are "elastic" in that they do not require stringent performance guarantees from the network for users to be satisfied. While it is likely that delay-sensitive wide-area applications (such as IP telephony) will evolve in the near future, LEO systems are being designed specifically to better support such applications.

- **Bandwidth asymmetry matches traffic asymmetry** Satellite networks that are designed to enable low cost user terminals are typically built around a "star" network topology, in which a large hub antenna and large power amplifiers is used to broadcast a high bit rate stream to customers. However, because the most expensive component of a customer terminal is a power amplifier, a low wattage power amplifier is used and consequently the customer cannot use a large return carrier for transmissions back to the hub. Fortunately, the type of asymmetric bandwidth provided by such systems is well suited for Web browsing, by far the most popular Internet application in terms of network usage. Recent traffic traces of client PCs have shown an asymmetry ratio between inbound and outbound bandwidth usage of over 5 to 1. Satellite operators have already capitalized on this traffic asymmetry by offering hybrid satellite systems, such as the DirecPC network which allows users to download data at rates up to 400 Kb/s but which relies on a dial-up return channel (typically around 30 Kb/s) [100].

1.2 Contributions and Thesis Overview

1.2.1 Statement of Research Problems

In this thesis, we concentrate on *the application of satellite systems to provide broadband access to the Internet*. This is in contrast to using such networks for trunking or transit connectivity between carrier networks, as has often been done historically. Because point-to-point satellite connections will continue to be more expensive than fiber-based options (if such options exist), satellite communications will most likely be used i) for broadcasting of information to multiple users and ii) for broadband access where viable terrestrial infrastructure is lacking. In this application context, we focus on the following two problems:

- the performance of reliable transport protocols over GEO satellite links, and
- the design of unicast routing protocols for LEO satellite networks.

The problem of designing and deploying reliable transport protocols that perform adequately over satellite links is well established one, but in our opinion the problem has not been completely solved and in particular we approach the problem from different angles. Specifically, we consider the transport protocol performance that a satellite-based user is likely to encounter when his or her connection traverses a portion of the wired Internet, in contrast to looking at transport connections in isolated satellite environments. We explore changes that can be made in existing transport protocol implementations and specialized protocol gateways that can be deployed within satellite broadband access networks. We especially emphasize studying protocol performance in the context of other competing Internet connections that share the same path as the measured connection; such an emphasis is not often found in the available literature. And rather than focusing exclusively on file transfers, we explore the performance of small data transfers commonly associated with Web browsing.

In contrast, the design of unicast routing protocols for LEO networks is an emerging problem, with much of the previous work on networking for LEO systems focusing on connection-oriented routing rather than packet switching. This is because it has long been thought that Asynchronous Transfer Mode (ATM) networks would form the backbone of all future broadband wide-area networks. Since the future of ATM is no longer clear, and because we believe that a connectionless networking paradigm is better suited for rapidly time-varying network topologies, we instead choose to focus on IP-based routing. By constructing a network simulator that is able to provide detailed packet-level simulations of future LEO networks, we explore not only some fundamental performance properties of such constellations but also more specialized routing techniques tailored specifically for proposed LEO networks.

1.2.2 Contributions

Regarding the two main problems identified above, we were able to make a number of contributions, which we summarize here and discuss in more detail in the following chapters:

- Our study of TCP performance over GEO satellite links is among the first that explores in detail the interactions between satellite TCP connections and other (non-satellite) connections that share part of the same end-to-end path. One long-standing problem in this context has been the fairness performance of TCP's congestion avoidance algorithm when multiple connections with different round trip delays share a bottleneck link. Using simulation models, we provide evidence that, while TCP fairness problems may not be easily solvable in a manner that can be incrementally deployed by making changes to host implementations, small changes to a satellite connection's congestion avoidance algorithm can substantially improve the fairness of the bottleneck link usage without comprimizing link utilization. In particular, in congestion situations we were frequently able to double the throughput of satellite connections by making their congestion avoidance policies slightly more aggressive than normal, but not so aggressive as to unfairly penalize other competing connections. Next, we highlight how imperfect implementations of standardized TCP options relating to loss recovery and congestion avoidance can lead to poor file transfer performance over satellite links and, through experiment and simulation, construct a reference implementation of these options that can achieve good performance in a non-congested satellite environment. We also quantify how much two proposed TCP options (TCP for Transactions and options for increasing TCP's initial window) reduce the latency of short transfers, finding that the use of both options can reduce the user-perceived latency by a factor of two to three. However, since these options are not guaranteed to become widely deployed, and since the file transfer performance of even satellite-optimized connections can be derailed by the fairness problem discussed above, we investigate the performance benefits that can be achieved by using TCP protocol gateways in a satellite network. In particular, we find that well-tuned TCP protocol gateways, which split a single TCP connection into two separate connections, can achieve performance comparable to connections that do not traverse a GEO satellite link.

- Given that TCP protocol gateways are possible in the network architecture, we explore whether protocols other than TCP may be more suitable for the long-delay and bandwidth-asymmetric environment of satellite access networks. We describe the overall design and performance of a satellite-optimized transport protocol (which we have named "STP") that is specifically designed for a broadband satellite network characterized by high degrees of bandwidth asymmetry. Some benefits of this protocol relative to TCP's performance include good performance in a high loss environment, less sensitivity to large variations in the round trip delay experienced by packets, and a reduction of up to an order of magnitude in the amount of bandwidth used on the reverse channel to return acknowledgments.

- Finally, we conduct a detailed investigation of packet routing alternatives in the context of LEO satellite networks. Our study is believed to be among the first that focuses on the connectionless packet routing problem for the highly time-varying network topologies of LEO networks. We describe the construction of a LEO network simulator, suitable not only for packet routing studies but also for other aspects of networking over future satellite constellations. This simulator reveals some interesting fundamental delay performance properties of LEO networks that have not yet been described in the literature. After illustrating some of these properties, we turn our attention to packet routing. Although existing distributed routing protocols can be made to work in this environment, we seek to exploit the specialized topological properties and system constraints of LEO networks in the design of new routing strategies. We first explore the hypothesis, advanced by other researchers, that by making locally optimal packet forwarding decisions that minimize the geographic distance to the destination, one can obtain routes that are close to optimal in terms of delay performance. Although we find that a distributed protocol based on this concept is fundamentally difficult to construct because of the structure of commercially-proposed network topologies, we are able to demonstrate the benefits of geographic-based addresses in centralized routing systems. We develop an addressing strategy for a particular cellular structure on the Earth's surface and prove its optimality from the standpoint of maximizing opportunities for address aggregation of geographically contiguous cells. By taking advantage of temporal consistencies in routing tables and address aggregation possibilities, we propose a centralized routing strategy that, when compred to traditional non-hierarchical routing approaches, leads to a reduction of over an order of magnitude in both the amount of routing traffic that must be conveyed between network nodes and the number of satellite routing table entries used for non-local destinations.

1.2.3 Thesis Overview

The remainder of this dissertation is organized as follows.

In the next chapter we delve deeper into background material related to the problems we address in this thesis, and survey the related work that provides the foundation for the research presented herein.

In Chapter 3 we describe our overall research methodology and provide an overview of the simulation environment that we used to generate numerical results and the experimental testbed used to study the performance working implementations of the protocols we constructed.

Chapters 4 through 6 form the core of the thesis, with Chapters 4 and 5 focusing on the problem of transport protocol performance over GEO satellite links, and Chapter 6 exploring the problem of unicast packet routing in LEO satellite constellations.

Chapter 4 is concerned with the performance of TCP in a satellite environment, focusing first on potential remedies to fairness problems inherent in TCP's congestion avoidance algorithm, then studying the interaction of different TCP implementation options in a satellite environment, and finally exploring the potential gains achievable through the use of TCP protocol gateways in satellite access networks.

In Chapter 5 we introduce our satellite transport protocol (STP). We begin by describing the basic design and operation of the protocol. We then study its performance through the use of simulation models and an implementation.

We begin Chapter 6 with a detailed discussion on the construction of a simulation environment to study the problem of networking in LEO satellite constellations. After presenting some fundamental delay performance results obtained from our simulator, we focus in the remainder of the chapter on the strategy of using geographic-based addressing information to simplify routing.

Finally, in Chapter 7 we conclude by summarizing our our work and discussing directions for future work.