Satellite-Based Internet: A Tutorial

Yurong Hu and Victor O. K. Li, The University of Hong Kong

ABSTRACT

In a satellite-based Internet system, satellites are used to interconnect heterogeneous network segments and to provide ubiquitous direct Internet access to homes and businesses. This article presents satellite-based Internet architectures and discusses multiple access control, routing, satellite transport, and integrating satellite networks into the global Internet.

INTRODUCTION

The Internet has enjoyed explosive growth in the past few years. At the same time, the proliferation of new applications, and expansion in the number of hosts (computers connected to the Internet) and of users impose new technical challenges to Internet development. New Internet infrastructure and technologies capable of providing high-speed and high-quality services are needed to accommodate multimedia applications with diverse quality of service (QoS) requirements. Furthermore, in order to provide ubiquitous Internet access, appropriate mobility support is required.

A satellite communication system, distinguished by its global coverage, inherent broadcast capability, bandwidth-on-demand flexibility, and the ability to support mobility, is an excellent candidate to provide broadband integrated Internet services to globally scattered users. A satellite system, if properly designed, can cover the entire surface of the Earth, making it extremely appealing to aeronautical and maritime users, and to those in remote areas lacking terrestrial communication infrastructure. Even for the densely wired parts of the world, it offers an alternative to the increasingly congested terrestrial links. A satellite network is inherently a broadcast system. It is particularly attractive to point-to-multipoint and multipoint-to-multipoint communications which are experiencing rapid development, especially in broadband multimedia applications. Satellite networks can serve as broadband access networks, high-speed backbone networks connecting heterogeneous networks, or simply as communication links between users with fixed or mobile terminals.

However, the interoperation between a satellite system and the existing terrestrial Internet infrastructure introduces new challenges. This article attempts to survey ongoing research efforts on integrating satellite systems into the global Internet. It clarifies the crucial technical difficulties and provides insights for further research. The rest of the article is organized as follows. In the next section we present the basic background on satellite systems. The article then describes proposed satellite-based Internet architectures. Several technical issues in constructing satellite-based Internet and some suggested solutions are discussed. The final section provides a summary and identifies some future research directions.

SATELLITE COMMUNICATION FUNDAMENTALS

A satellite system consists of a space segment and a ground segment. The ground segment consists of gateway stations (GSs), a network control center (NCC), and operation control centers (OCCs). The NCC and OCCs handle overall network resource management, satellite operation, and orbiting control. The GSs act as network interfaces between various external networks and the satellite network. They also perform protocol, address, and format conversions. The space segment is composed of satellites, which may be classified into geostationary orbit (GSO) and nongeostationary orbit (NGSO) satellite, including medium earth orbit (MEO) and low earth orbit (LEO) satellite, according to the orbit altitude above the Earth's surface.

GSO — The majority of satellites in operation nowadays are placed in GSO orbit. The GSO satellite is 35,786 km above the equator, and its revolution around the Earth is synchronized with the Earth's rotation. Therefore, it appears fixed to an observer on the Earth's surface, and may serve as a repeater in the sky. Its high altitude allows each GSO satellite to cover approximately one third of the Earth's surface, excluding the high latitude areas. The area of coverage of a satellite is called its *footprint*. Three GSO satellites are sufficient for global coverage. However, the cost of launching GSO satellites is high. Due to its high altitude and the inherent signal degra-

0163-6804/01/$10.00 © 2001 IEEE

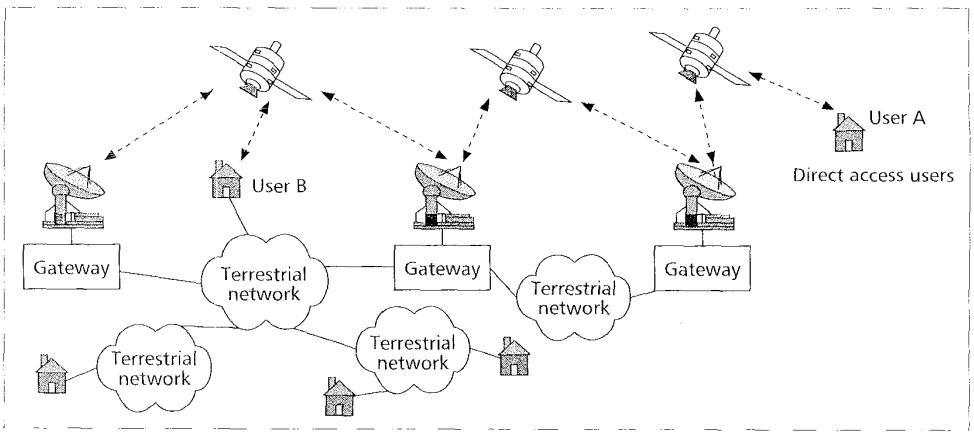

Figure 1. *The satellite-based Internet with the bent-pipe architecture.*

The idea of using satellites as a solution of the last mile problem (i.e., connecting users to the network access points), inspired by the usage of cost-effective VSATs and improvements in satellite technologies, is relatively new.

dation with distance, large antennas and transmission power are required for both the GSO satellite and ground terminals. The most significant problem is the large propagation delay for GSO satellite links. The typical value of round-trip delay is 250–280 ms, which is undesirable for real-time traffic.

MEO and LEO — MEO's distance from the Earth's surface is from 3000 km up to the GSO orbit with a typical round-trip propagation delay of 110–130 ms. LEOs are located 200–3000 km above the Earth's surface. For a LEO satellite the round-trip delay is 20–25 ms, which is comparable to that of a terrestrial link. Since LEO/MEO satellites are closer to the Earth's surface, the necessary antenna size and transmission power level are much smaller; but their footprints are also much smaller. A constellation of a large number of satellites is necessary for global coverage. The lower the orbit altitude, the greater the number of satellites required. In addition, since satellites travel at high speeds relative to the Earth's surface, a user may need to be handed off from satellite to satellite as they pass rapidly overhead. Therefore, steerable antennas are crucial to maintain continuous service.

Satellite Payload — The satellite payload is responsible for the satellite communication functions. Once the satellite is launched, it is very expensive and almost impossible to upgrade or repair. The space environment, with radiation, rain, and space debris, is harsh for satellites. Therefore, the satellite payload is required to be simple and robust. Traditional satellites, especially GSOs, serve as bent pipes. They act as repeaters between two communication points on the ground. There is no onboard processing (OBP). It is simple and easy to implement. Some satellite systems allow OBP, including demodulation/remodulation, decoding/recoding, transponder/beam switching, and routing to provide more efficient channel utilization. OBP can support high-capacity intersatellite links (ISLs) connecting two satellites within line of sight. By using a sophisticated constellation with ISLs, connectivity in space without any terrestrial resource is possible.

Frequency Bands — The most commonly used satellite frequency bands are C band (4–8 GHz), Ku band (10–18 GHz), and Ka band (18–31 GHz). With a higher frequency band and a corresponding shorter wavelength, smaller antennas can be used to receive the signal. Some satellite systems use C band and thus employ large antennas with minimum diameter of 2–3 m. The majority of direct broadcast satellites use Ku band for broadcasting as well as for Internet connections from the server to the users, with a terrestrial return link. A Ku band antenna can be as small as 18 inches in diameter. There are proposals to provide a Ka band return link for these systems. Ka band potentially offers much higher bandwidth than Ku band, and can use very small antennas, but it suffers from environmental impairments such as fading and rain attenuation. There are also plans to use frequencies beyond Ka band, but the technologies for using those frequencies are immature and further investigation is needed.

Satellite-Based Internet Architectures

The satellite-based Internet has several architectural options due to the diverse designs of satellite systems, in orbit types (GSO, MEO, LEO), payload choice (OBP or bent pipe), and ISL designs. There are suggestions that multiple satellite types (i.e., GSOs, MEOs, and LEOs) be included in a hybrid GSO/NGSO network to fully utilize the best characteristics of each orbit type.

A satellite network can serve as part of the Internet backbone, a high-speed access network, or both. Using a satellite system as part of the Internet backbone has a long history that dates back to the Atlantic SATNET interconnecting ARPANET with European research networks [1]. However, the idea of using satellites as a solution of the last mile problem (i.e., connecting users to network access points), inspired by the usage of cost-effective very small aperture terminals (VSAT) and improvements in satellite technologies, is relatively new.

A typical satellite-based Internet scenario with bent-pipe satellites is depicted in Fig. 1.

Multiple access control, defined as a set of rules for controlling access to a shared channel among contending users, plays an important role in efficiently and fairly utilizing the limited satellite system resources.

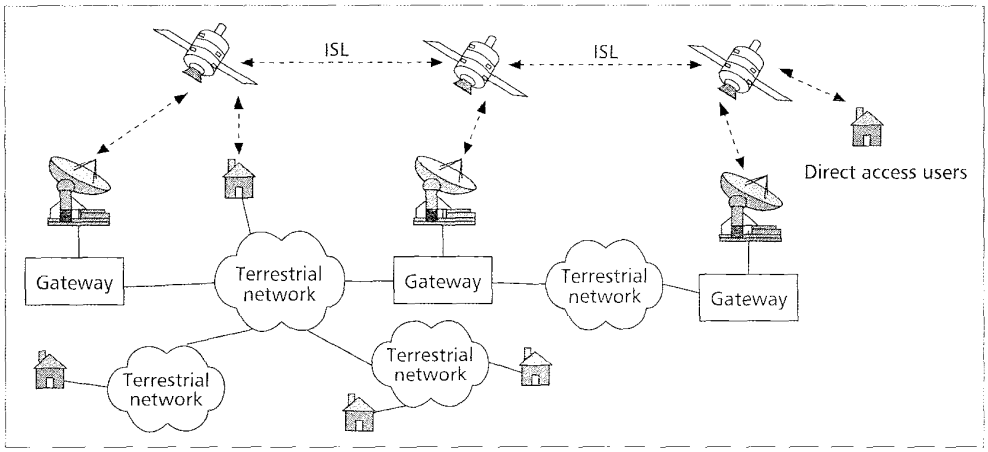

Figure 2. *The satellite-based Internet with the OBP and ISL architecture.*

The satellites adopted can be GSO, MEO, or LEO. It provides Internet access as well as data trunking service. The satellite network interfaces with the ground Internet infrastructure via GSs on the Earth. It may be the only access method for some users (e.g., user A) when no other communication method is available, or a backup connection in addition to an existing terrestrial access network (e.g., user B).

However, the bent-pipe architecture's lack of direct communication paths in space results in low spectrum efficiency and long latency. OBP and ISLs may be used to help construct a network in the sky (Fig. 2). This architecture is again a combination access and backbone network. Teledesic is one such system using a constellation of 288 LEO satellites with ISLs [2]. The rich connectivity in space will provide more flexibility but also bring complex routing issues, which will be discussed later.

In Table 1 we summarize some proposed worldwide broadband satellite systems that aim to provide high-speed Internet service. We have not included Iridium [3], which started operation in 1998, since it is primarily designed for voice and paging services.

In the two aforementioned general architectures, user terminals are assumed to be interactive, which means they can directly transmit data up to the satellite and receive data from the satellite. Although rapid advancements in satellite technology have spawned small user terminals, such as ultra small aperture terminals (USAT) with 60 cm antennas, the interactive terminal is still expensive and thus frustrates direct-to-home implementation. Enlightened by Internet traffic asymmetry where considerably more data is transmitted from the server to the end user than in the reverse direction (e.g., Web browsing), there is a trend to offer Internet access via direct broadcast satellites (DBSs) used for television broadcasting [1]. Each home has a receive-only satellite dish to collect data delivered in the high-speed satellite broadcast channel. The reverse path to the server is provided by a terrestrial link (Fig. 3). Hughes's DirecPC system is an example. In order to make full use of the wide bandwidth of satellite broadcast links, DBS is also extended by using the receive-only terminals as gateways to interconnect remote networks.

The above architecture contradicts the traditional symmetric network assumption of two-way balanced load and identical link characteristics, and causes the so-called unidirectional routing problem elaborated on in the next section.

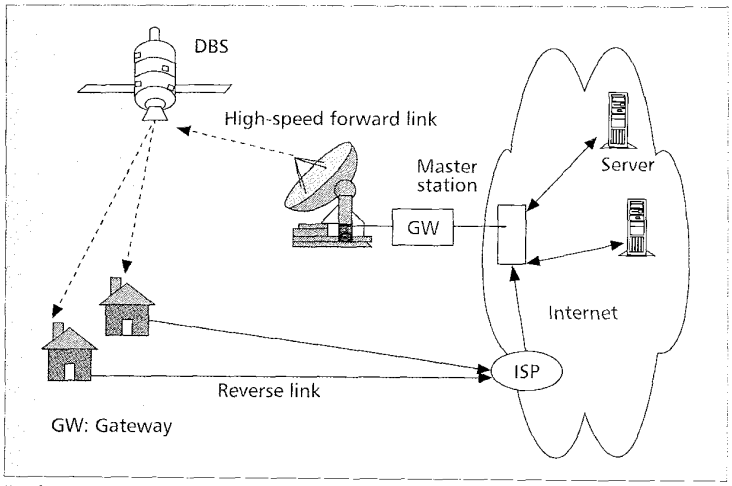

Figure 3. *Internet access via DBS.*

Technical Challenges

In this section we summarize the technical challenges in designing and implementing the satellite-based Internet. We focus on special requirements unique to satellite systems, and leave out general considerations common to terrestrial networks. Multiple access control schemes, essential for satellite systems, are described first. Then we investigate transmitting IP packets in satellite networks. Finally, transport issues based on TCP modification and satellite-specific transport protocols are presented.

Multiple Access Control

In interactive satellite systems, a large number of user terminals widely scattered within the satellite footprint contend for the satellite uplink channel.[1] Multiple access control (MAC), defined as a set of rules for controlling access to a shared channel

System	Major sponsors	Constellation	Satellite payload	Frequency band	Data rate	Service date
Astrolink	Lockheed Martin	Up to 9 GSO satellites	OBP and ISLs	Ka	Up to 200 Mb/s downlink Up to 20 Mb/s uplink	2003
Skybridge	Alcatel Espace, Loral Space	80 LEOsatellites at 1469 km	Bent-pipe	Ku	16 kb/s–20 Mb/s downlink 16 kb/s–2 Mb/s uplink	2002
Spaceway	Hughes	4 GSO satellites (ultimately 21 satellites)	OBP and ISLs	Ka	Up to 92 Mb/s downlink 16 kb/s–6 Mb/s uplink	2002
Teledesic	Motorola, Lockheed Martin	288 LEO satellites at 1375 km	OBP and ISLs	Ka	16 kb/s–64 Mb/s downlink 16 kb/s–2 Mb/s uplink	2004

Table 1. *A summary of proposed worldwide broadband satellite systems.*

among contending users, plays an important role in efficiently and fairly utilizing the limited satellite system resources. MAC protocol performance can significantly affect higher-layer protocols and the QoS provided by the system.

The performance of MAC protocols depends on the characteristics of both the shared communication media and the traffic. The long latency in satellite channels (especially the GSO links) excludes some MAC schemes used in terrestrial local area networks (LANs) such as carrier sense multiple access (CSMA), and the limited power resource in satellites constrains the transponder and computational capacity on board the satellite. Internet traffic turns out to be bursty in nature. Besides the current best-effort service, the Internet is expected to provide diverse QoS guarantees (e.g., on delay, delay jitter, packet loss ratio) for a wide range of traffic types. Thus, a candidate MAC protocol must implement priorities. Real-time traffic with transmission deadlines is usually given higher priority than non-real-time traffic.

Generally speaking, a good MAC scheme for a satellite-based network should be simple to implement, robust, and flexible to accommodate network reconfiguration. The MAC should be able to achieve high throughput, maintain channel stability, and enjoy low protocol overhead and small access delay.

Depending on how bandwidth is allocated among all contenders, candidate MAC schemes for satellite systems can be categorized into three groups: fixed assignment, random access, and demand assignment.

Fixed Assignment — Fixed assignment may be made on a frequency, time, or code basis. Major techniques include frequency-division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access (CDMA). In FDMA and TDMA systems, each station utilizes its own dedicated channel. They are contention-free, and can provide QoS guarantees. However, this is at the expense of inefficient utilization of resources. Their lack of flexibility and scalability makes them only suitable for small-scale networks with stable traffic patterns. FDMA was the first fixed assignment multiple access method used in satellite systems. TDMA is popular mainly because of its compatibility with the nonlinear nature of transponders[2] and is used in the majority of current satellite systems. In a CDMA system, each user is assigned a unique code sequence which is used to spread the data signal over a wider bandwidth than that required to transmit the data. If code sequences are guaranteed to be orthogonal, all other simultaneous transmissions in the same channel act as additive interference to the desired signal and can be removed completely at the receiver side, where a reverse procedure, despread, is taken to recover the original data. Thus, in a CDMA system, the whole bandwidth is used by all users, making it more flexible for system expansion.

Random Access — Due to technological advances, small and inexpensive terminals (i.e., VSATs and USATs) with lower data rates are now widely available, thus stimulating home or personal use of satellite access service. The number of stations within a satellite footprint increases from a few to several hundreds or thousands. In addition, the traffic generated by each user is very bursty. Fixed assignment schemes are replaced by contention-based random access (i.e., Aloha and its variations). In random access schemes, each station transmits data regardless of the transmission status of others. Retransmissions after collision increase the average packet delay, and frequent collisions may cause low throughput.

Demand Assignment — Although random access may better accommodate a large number of terminals with bursty traffic, it provides no QoS guarantees. Demand assignment multiple access (DAMA) protocols attempt to solve this

[1] *This section describes protocols used by those interactive satellite systems which use a satellite link for the reverse link from the users to the Internet servers, and are not applicable to those systems, such as DirecPC, which use a terrestrial reverse link.*

[2] *In order to efficiently utilize the power of transponders, we must drive them into a saturation area where the amplifiers operate as nonlinear devices. In an FDMA system, users' signals may be received simultaneously, and the nonlinear amplifier generates undesired interference. In TDMA, only one user accesses the transponder at any given time interval, and the problem is avoided.*

A hybrid scheme called Round-Robin Reservation is based on fixed TDMA. The number of stations is required to be less than or equal to the number of time slots. Each station obtains a dedicated channel, and extra or unused slots are accessed in a round-robin manner or via slotted Aloha.

problem by dynamically allocating system bandwidth in response to user requests. A resource request must be granted before actual data transmission. The transmission of requests is itself a multiple access problem. However, since the request message is typically much shorter than actual data transmission, we can afford to have reservation requests collide and be retransmitted. After a successful reservation, bandwidth is allocated on an overall FDMA or TDMA architecture, and data transmission is guaranteed to be collision-free. This article focuses on the TDMA architecture in which equal-sized time slots are grouped into frames, repeated periodically.

The reservation may be made under centralized or distributed control. The central controller can be located at an Earth station or at a satellite with OBP. For a ground-based controller, the minimum request delay is two round-trip times before the reservation request is granted. The minimum request delay can be halved for a space-based controller, but the satellite payload capacity limits this implementation. Distributed control, in which each station receives all request information from the satellite broadcast channel and makes a decision on its own, is more robust and reliable. The channel overhead associated with reservation announcements is reduced greatly, and the minimum reservation delay is as small as one round-trip time. Although it may put the processing burden on the stations, distributed control is still preferred, considering its overall advantages.

Resource reservation can be made either explicitly or implicitly. Explicit reservation is on a per-transmission basis, and usually a dedicated reservation channel is shared among all stations. Each station sends a short request via the reservation channel specifying the number of time slots needed. Stations access the reservation channel in fixed assignment mode, such as TDMA reservation, or random access mode, such as Aloha reservation. Data is transmitted in the data channel after successful reservation.

In implicit reservation, there is no explicit reservation message, and a successful data transmission in a slot serves as an indication of reservation for the corresponding time slot in subsequent frames. Packets belonging to a long transmission can repeatedly occupy the same slot in consecutive frames. An empty slot in a frame indicates the end of the transmission, and other users can then contend for this slot starting in the next frame. This scheme is attractive for relatively steady traffic patterns such as voice and video connections. An example of this scheme is Reservation Aloha, which uses the first data packet as an implicit request unit and accesses the available time slots via the slotted Aloha protocol.

Priority-Oriented Demand Assignment (PODA) and first-in first-out (FIFO) Ordered Demand Assignment (FODA) [4] combine implicit and explicit requests. Each PODA TDMA frame consists of a control part and a data part with an adjustable boundary. Explicit requests contend in the control part by slotted Aloha, while implicit requests are piggybacked on data packets. FODA further divides the data part into stream and datagram subframes. One FIFO stream queue and two FIFO datagram queues, for short interactive traffic and bulky traffic, are maintained with decreasing priorities. The latter two types of traffic are transmitted in the datagram subframes.

There are proposals to make use of the unreserved resource after the demand assignment. Combined free/demand assignment multiple access (CFDAMA) [5] freely assigns remaining channels according to some strategy (e.g., round-robin). In combined random access and TDMA-reservation multiple access (CRRMA) [6], remaining resources are open for random access. By randomly accessing the unreserved channel, some bursty interactive traffic may be transmitted immediately without waiting for the two-hop reservation delay. A hybrid scheme called Round-Robin Reservation (RRR) is based on fixed TDMA. The number of stations is required to be less than or equal to the number of time slots. Each station obtains a dedicated channel, and extra or unused slots are accessed in a round-robin manner or via slotted Aloha. In satellite systems, large GSs interfacing with terrestrial networks function as multiplexers for traffic from those directly connected networks. Gateway stations are usually much more heavily loaded than small terminals, and the number of GSs is much smaller than that of small terminals. The RRR mechanism may be suitable for this scenario. For example, GSs may obtain dedicated time slots, while small terminals contend for the remaining slots in each frame. Similar hybrid methods may combine the advantages of different schemes, but further performance analyses and simulations are needed to demonstrate their feasibility.

Routing Issues in Satellite Systems

Routing Issues in a LEO Constellation — The significant advantages of LEO with OBP and ISLs, such as small delay and full connectivity, make it a very attractive approach to the Internet in the sky. In such networks, the major technical issue is the complex dynamic routing issue due to satellite movements.

Dynamic Topology — Due to the relative movement between the LEO satellite and the Earth, a satellite has a very short visible period to motionless users on the ground. To maintain 24-hr continuous coverage, a carefully designed satellite constellation is crucial. At any time there should be at least one satellite within line of sight of a user. When a satellite moves out of a user's visual field and another satellite moves in, intersatellite handover happens. For a satellite with multiple antennas and transponders, the satellite footprint is divided into a number of spotbeams, each covered by an antenna beam. Thus, frequent interbeam handover from spotbeam to spotbeam occurs within a single satellite's visible period.

The ISLs in the constellation form a mesh network topology. Each satellite is typically able to set up 4–8 ISLs. There are two types of ISLs: intraplane ISLs connecting adjacent satellites in the same orbit, and interplane ISLs connecting neighboring satellites in adjoining orbits. Intraplane ISLs are maintained permanently, but some interplane ISLs may be temporarily switched off

when the viewing angle or distance between two satellites changes too fast for the steerable antennas to follow. This may occur between two counter-rotating orbits or when two orbits cross. The routing scheme should be able to handle topological variations. Fortunately, although the constellation topology changes frequently, it is highly periodic and predictable because of the strict orbital movements of the satellites.

Some dynamic routing mechanisms popular in the Internet, such as distance vector (DV) and link state algorithm (LSA), are not directly applicable in satellite constellation routing, because frequent topological changes in satellite constellation will cause large overhead and oscillation if these schemes are used. Two new concepts tailored to dynamic satellite constellation are worth mentioning: discrete-time dynamic virtual topology routing (DT-DVTR) [7] and the virtual node (VN) [8].

- **DT-DVTR** — DT-DVTR makes full use of the periodic nature of satellite constellation and works completely offline. It divides the system period[3] into a set of time intervals so that the topology changes only at the beginning of each time interval and remains constant until the next time interval. In each interval, the routing problem is a static topology routing problem that can be solved easily. A number of consecutive routing tables are then stored onboard and retrieved when the topology changes. With this strategy, online computational complexity is transformed into a large storage requirement on the satellites. In order to minimize the storage needed and the inter-satellite handover attempts when topology changes, an optimization procedure can be used to choose the best path or a small set of paths from the series of instantaneous routes. Although it can significantly reduce the storage size, some links may become congested while others are underutilized.
- **VN** — The objective of this scheme is to hide the topology changes from the routing protocols. A virtual topology is set up with VNs superimposed on the physical topology of the satellite constellation. Even as satellites are moving across the sky, the virtual topology remains unchanged. Each VN keeps state information, including routing tables and information of users within the VN's coverage area. In a certain period, a VN is represented by a certain physical satellite. As this satellite disappears over the horizon, the VN is represented by the next satellite passing overhead. The state information is also transferred from the first satellite to the second. A routing decision is made on the virtual topology, and the protocols are not aware of the dynamic satellite constellation concealed in state transfers.

Based on these two concepts, some routing schemes are proposed for carrying IP packets through the satellite constellation. Some commercial satellite systems (e.g., Teledesic) use proprietary routing techniques that are highly dependent on explicit orbital and constellation knowledge and optimized for specific designs. Due to their lack of generality, they are not covered in this article.

IP Routing at the Satellites — To route IP packets through a satellite constellation, it seems straightforward to adopt IP routing at the satellites. This strategy is addressed in [9], and is based on the VN concept. It can seamlessly integrate the space network with the terrestrial Internet, and permits direct support for IP multicast and IP QoS (integrated and differentiated service models). However, how to deal with variable-length IP packets, the scalability problem of onboard routing tables, and computational and processing capacity limitations in space devices are challenging problems. The scheme is still in its infancy, and some practical problems are unsolved in the implementation of the VN concept.

How to deal with variable length IP packets, the scalability problem of on-board routing tables, and computational and processing capacity limitations in space devices are challenging problems. The scheme is still in its infancy, and some practical problems are unsolved.

ATM Switching at the Satellites — Many proposed systems use ATM as the network protocol for the constellation (i.e., Cyberstar, Astrolink, Spaceway, and Skyway) with a satellite-specific signaling protocol and link layer protocol [10]. An ATM version of DT-DVTR is investigated in [7], where all the virtual channel connections between the same pair of ingress and egress satellites are grouped into a virtual path connection (VPC), and onboard switching is done according to the VPC labels. A modified S-ATM packet is suggested to reduce the overhead without changing the cell size [10]. If such a system is adopted to provide Internet service, IP over ATM or other similar technologies will be used.

External Routing Issues — It is reasonable to assume that the internal routing schemes for satellite constellation will continue to be heterogeneous. Satellite manufacturers and operators will probably select routing methods best suited to their own system designs. The internal protocol should be kept simple. Details of the satellite network should be hidden from the terrestrial Internet as well. Today's Internet achieves this kind of isolation by using the autonomous system (AS) concept. Typically, some external routing schemes are used for inter-AS routing, while internal routing is handled by the AS's own internal routing protocol.

A satellite system can be considered an AS in the Internet (Fig. 4) [9]. A number of border gateways (BGs) running exterior routing protocols (e.g., Border Gateway Protocol, BGP, used by terrestrial ASs) will communicate with terrestrial ASs. Only BGs on the constellation periphery need be aware of the outside addresses and topological information. All packets going through a satellite constellation enter the satellite AS from one entry BG, which is responsible for determining the exit BG of each packet. If necessary, the entry/exit BGs perform encapsulation/decapsulation and address resolution. The BGs can be implemented either onboard the satellites or in ground GSs. If space-based BGs are used, computational and storage requirements may be too much for the satellites. On the other hand, if terrestrial gateways are used,

[3] *The system period is the least common multiple of the orbit period and the Earth's rotation period.*

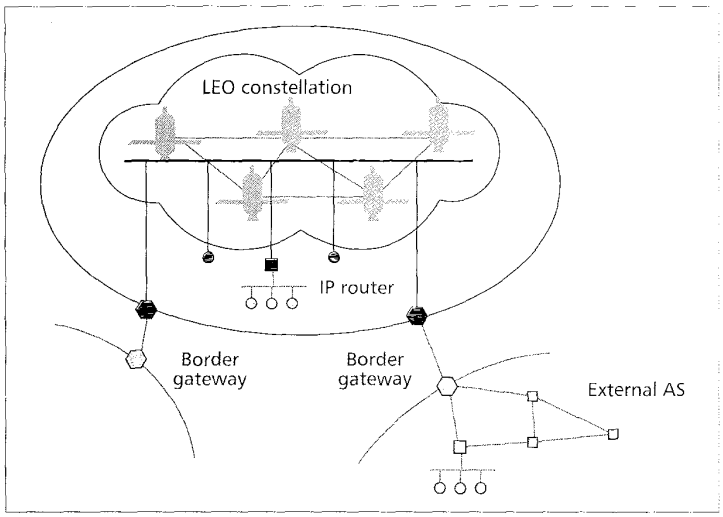

Figure 4. *A LEO constellation's autonomous system.*

packets must be bounced back to the ground for IP routing, introducing an extra round-trip delay. However, ground BGs are more realistic. Furthermore, external routing protocols popular in terrestrial networks cannot simply be reused in satellite constellations. In terrestrial networks, any internal link within an AS always has smaller cost than an inter-AS link. It is generally true because in terrestrial networks an AS is usually limited to a small geographical area, while an inter-AS link travels a much longer way. But a satellite system extends globally, and routing within a satellite constellation may be as expensive as traversing several ASs. Thus, multiple pairs of BGs should be used from a satellite constellation to a destination in some AS.

Unidirectional Routing — As described earlier, Internet access via DBS poses the unidirectional routing problem which cannot be handled by traditional dynamic routing schemes where bidirectional links are assumed. For example, in distance vector routing, a router receiving the distance vector tuple {destination, cost} from its neighbor deduces that it can reach the destination via this neighbor. It is no longer true in the satellite broadcast scenario where the direct reverse link to the satellite does not exist. Multicast routing protocols (e.g., DVMRP) based on the reverse shortest path tree also face such a problem [11].

There are three ways to handle this problem. Instead of dynamic routing, static routing may be an option; but with thousands of users served by a DBS, it is impossible to manually configure all of the routing entries. The other two methods are routing protocol modification and tunneling, proposed in the Internet Engineering Task Force (IETF) Unidirectional Link Routing (UDLR) working group. They are discussed next.

Routing Protocol Modification — In unidirectional routing, the router at one end of a unidirectional link with a send-only interface is referred to as a feeder, while the router at the other end of the unidirectional link with a receive-only interface is called a receiver. The key idea of the modification is twofold. First, the modified protocol should enable a receiver to identify the potential feeders whenever it receives routing updates from them, and to ignore the unusable routing information in those packets while keeping the useful reports to maintain the neighboring connectivity [11]. Second, the receiver periodically delivers its own routing message to all feeders through the terrestrial reverse channel. Thus, when a feeder gets the routing information, it can update the related routing entries for reachable destinations through the unidirectional link passing the receiver. The idea is used by the UDLR working group in the proposals to modify some popular protocols (i.e., RIP, OSPF, and DVMRP).

Tunneling — Tunneling offers a link layer approach to hide the network asymmetry from the routing process. A virtual bidirectional link is set up between a DBS and a user by encapsulation and decapsulation. This virtual link is called a tunnel. Packets destined for the DBS from the user are delivered via the tunnel. The tunnel endpoint at the user side first encapsulates the packet, and then passes it to the routing protocol where it is delivered through the actual terrestrial reverse channel. When the packet arrives at the satellite, the tunnel endpoint captures it, decapsulates it, and forwards it to the routing protocol to which it seems to come from a bidirectional link.

The above two approaches are simple, and since tunneling is transparent to all upper layer protocols, it may be quickly implemented in DBS Internet access architectures. However, the two schemes are designed based on point-to-point unidirectional links, although satellites are point-to-multipoint broadcast systems. Thus, further study is needed to design new approaches optimized for this architecture. The two approaches also focus only on the routing issue within a single AS and fail to address interdomain routing. New interdomain routing schemes that can handle unidirectional links are needed.

Satellite Transport

The TCP/IP and UDP/IP protocol suites form the basis of the Internet. Due to their tremendous legacy, it is unlikely they will be totally discarded in the near future. Therefore, the satellite-based Internet is expected to continue to serve applications based on TCP and UDP. However, the performance of both protocols will be affected by the long latency and error-prone characteristics of satellite links. The impacts on TCP will be much greater, and heated debates have been spawned regarding the feasibility of TCP in a satellite environment. Researchers working with NASA's ACTS satellites are performing research regarding TCP/IP over satellite connections. The IETF TCP over Satellite working group is also dedicated to improvement of TCP performance in satellite systems. In this section we first present the main limitations of TCP over satellite links and then summarize the ongoing research efforts on the satellite transport problem.

TCP Performance over Satellite — TCP uses a positive feedback mechanism to achieve rate control and reliable delivery. The long latency of

satellite links (especially GSO links) increases the TCP end-to-end delay and results in sluggish acknowledgments. The slow feedback will weaken the functionality of rate control and congestion avoidance, and thus affect the throughput. In addition, a potential problem is the large fluctuation of measured round-trip time (RTT) that may be caused by dynamic topology in LEO constellation networks. Large variations in RTT measurements may result in false timeouts and retransmissions.

In the initial slow start stage of TCP transmission, although the sending rate increases exponentially, it is still too slow for the high-bandwidth satellite links. One proposed solution is to increase the initial value of the window. TCP originally allows a window size of 64 kbytes, which also limits the maximum sending rate to 64 kbytes/RTT. The satellite link will be underutilized, and a large window scaling up to the bandwidth-delay product of the satellite link is required for higher throughput. A set of window scaling options to TCP implementation are defined in IETF Request for Comments (RFC) 1323.

Satellite links are subject to various impairments (i.e., interference, fading, shadowing, and rain attenuation). Therefore, a high bit error rate (BER) is expected. Although advanced modulation, coding schemes, and forward error correction (FEC) techniques are used to reduce the BER, in some environments high BER persists. But TCP does not distinguish between corrupted data caused by transmission error and packet loss due to congestion; both are unacknowledged and interpreted as a notification of network congestion. When there is a corrupted packet, the window size is halved even though there is no congestion. Furthermore, transmission errors on a satellite link are bursty in nature, especially under bad weather conditions. Bursty errors in one RTT will dramatically reduce the throughput. The Space Communications Standards-Transport Protocol (SCPS-TP) [12] defined for the general space environment provides two mechanisms to distinguish the sources of loss and responds differently. In addition, network asymmetry can also impair TCP performance. Satellite network asymmetry occurs in two situations. One is in the asymmetric DBS Internet access architecture described earlier. The other is due to bandwidth asymmetry in some interactive satellite terminals. These terminals may be capable of downloading at tens of megabits per second, but with uplink speed of only several hundred kilobits per second. The limited reverse link capacity may cause the acknowledgment starvation problem. The backlogged feedback will slow down the window refresh. In addition, acknowledgment loss due to reverse link congestion may trigger unnecessary retransmissions.

Another problem inherent in TCP is the fairness issue between different TCP connections with various RTTs. When those TCP connections share a bottlenecked link, the TCP connections with longer RTTs will suffer unfair bandwidth allocation.

Performance Enhancements — The IETF TCP over Satellite working group has recently made a number of recommendations to enhance the performance of TCP over satellite links in its RFCs. The last two schemes listed below are non-TCP techniques:

- TCP selective acknowledgment-(SACK) options (RFC2018) allow the receiver to specify the correctly received segments. Thus, the sender needs to retransmit only the lost packets. TCP SACK can recover multiple losses in a transmission window within one RTT.
- TCP for transaction (T/TCP) (RFC1644) attempts to reduce the connection handshaking latency from two RTTs to one RTT, which is a significant improvement for short transmissions.
- Persistent TCP connection, supported in HTTP1.1 (RFC2068), allows multiple small transfers to download in a single persistent TCP connection. It is more efficient.
- The Path maximum transfer unit (MTU) discovery mechanism allows TCP to use the largest possible packet size, thus avoiding IP segmentation. It reduces the overhead, and eliminates fragmentation and defragmentation.
- FEC is employed in link layer protocols to improve the quality of satellite links, but it should not be expected to fix all problems associated with manmade noise, such as military jamming, and some natural noise, such as that caused by rain attenuation. Besides FEC, some other link layer approaches (e.g., bit interleaving, link layer automatic repeat request schemes) can also be used to improve packet error rate in transmissions over satellite links.

TCP extensions can solve some of the limitations of standard TCP over satellite links, but other problems such as long end-to-end latency and asymmetry are not effectively addressed. One way to alleviate the effects of large end-to-end latency is to split the TCP connection into two or more parts at the GSs connecting the satellite network and terrestrial networks. There are three approaches to splitting TCP connections over satellite links:

- **TCP spoofing**: The divided connections are isolated by the GSs, which prematurely send spoofing acknowledgments upon receiving packets. The GSs at split points are also responsible for retransmitting any missing data. The performance of TCP spoofing is examined in [13].
- **TCP splitting**: Instead of spoofing, the connection is fully split. A proprietary transport protocol can be used in a satellite network without interference to standard TCP in terrestrial networks [14]. It is more flexible, and some kind of protocol converter should be implemented at the splitting points.
- **Web caching**: In contrast to the above two schemes, the TCP connection is split by a Web cache in the satellite network. Users in the satellite network connected to this Web cache need not set up TCP connections all the way to servers outside if the required contents are available from the cache. Web caching effectively reduces connection latency and bandwidth consumption.

In [14], a Satellite Transport Protocol (STP) is designed and used in the TCP splitting

The satellite-based Internet is expected to continue to serve applications based on TCP and UDP. However, the performance of both protocols will be affected by the long latency and error-prone characteristics of satellite links.

One way to alleviate the effects of large end-to-end latency is to split the TCP connection into two or more parts at the gateway stations connecting the satellite network and terrestrial networks.

approach as well as for traffic management in a satellite network. STP is based on the basic operation of Service-Specific Connection-Oriented Protocol (SSCOP). The sender periodically requests the receiver to report successful receptions, and retransmission is triggered by explicit selective negative acknowledgment. It uses a hybrid window and rate congestion control mechanism. STP performs well in asymmetric networks since the reverse traffic is significantly reduced, but it does not distinguish between different sources of packet loss and also leaves the fairness problem unsolved.

CONCLUSION

In this article we present an introduction to the satellite-based Internet. Some possible architectures based on bent-pipe and OBP satellites are discussed. Several technical challenges, including multiple access control, IP routing in LEO constellations, unidirectional routing, and satellite transport issues are investigated. In addition to what we elaborate on in this article, some important research issues are identified as follows:

- IP QoS support. There is no lack of research regarding QoS support in satellite systems. However, most of them are based on ATM QoS classes [8, 15], and the mapping of ATM service classes to IP QoS requirements is a nontrivial problem. Moreover, the implementation of TCP/IP over ATM brings much overhead, extra processing time, and protocol complexity. Direct support of the Internet integrated or differentiated service model is desired. In [9], multiprotocol label switching (MPLS) is proposed to support Internet QoS (integrated or differentiated service) in a satellite-based network. The integration of space and terrestrial communication systems, internetworking different satellite networks, and the advent of hybrid satellite systems will bring more redundancy and routing choices. QoS routing in satellite systems will be a very important research problem.
- Traffic and congestion control. To ensure that the satellite network achieves desired performance and fulfills the IP QoS requirements, a set of mechanisms to control traffic and avoid congestion is required. A well-designed MAC protocol will not by itself prevent congestion in the network. Traffic management, traffic shaping, policing, and scheduling are also required. Some preventive congestion control schemes, such as admission control, and efficient congestion notification schemes are important to maintain the specified QoS guarantees.

REFERENCES

[1] H. D. Clausen, H. Linder, and B. Collini-Nocker, "Internet over Direct Broadcast Satellites," *IEEE Commun. Mag.*, June 1999, pp. 146–51.
[2] D. J. Bem, T. W. Wieckowski, and R. J. Zielinski, "Broadband Satellite Systems," *IEEE Commun. Surveys*, vol. 3, no. 1, 2000.
[3] S. R. Pratt *et al.*, "An Operational and Performance Overview of the IRIDIUM Low Earth Orbit Satellite System," *IEEE Commun. Surveys*, vol. 2, no. 2, 1999.
[4] N. Celandroni, and E. Ferro, "The FODA-TDMA Satellite Access Scheme: Presentation, Study of the System, and Results," *IEEE. Trans. Commun.*, vol. 39, no. 12, Dec. 1991, pp. 1823-31.
[5] Le-Ngoc and S. V. Krishnamurthy, "Performance of Combined Free/Demand Assignment Multiple-Access Schemes in Satellite Communications," *Int'l. J. Satellite Commun.*, vol. 14, no. 1, Jan./Feb. 1996, pp. 11–21.
[6] H. W. Lee and J. W. Mark, "Combined Random/Reservation Access for Packet Switched Transmission over a Satellite with Onboard Processing: Part I — Global Beam Satellite," *IEEE Trans. Commun.*, vol. COM-31, Oct. 1983, pp. 1161–71.
[7] Markus Werner, "A Dynamic Routing Concept for ATM-Based Satellite Personal Communication Networks," *IEEE JSAC*, vol. 15, no. 8, Oct. 1997, pp. 1636–48.
[8] R. Manger and C. Rosenberg, "QoS Guarantees for Multimedia Services on a TDMA-Based Satellite Network," *IEEE Commun. Mag.*, July 1997, pp. 56–65.
[9] L. Wood *et al.*, "IP Routing Issues in Satellite Constellation Networks," *Int'l. J. Satellite Commun.*, to appear.
[10] I. Mertzanis *et al.*, "Protocol Architectures for Satellite ATM Broadband Networks," *IEEE Commun. Mag.*, Mar. 1999, pp. 46-54.
[11] W. Dabbous, E. Duros, and T. Ernst, "Dynamic Routing in Networks with Unidirectional Links," *Proc. 2nd Int'l. Wksp. Satellite-Based Information Services (WOSBIS '97)*, Budapest, Hungary, Oct. 1997.
[12] R. C. Durst, G. J. Miller, and E. J. Travis, "TCP Extensions for Space Communications," *ACM MobiComm '96*, Nov. 1996.
[13] M. Allman, H. Kruse, and S. Ostermann, "TCP Performance over Satellite Links," *WOSBIS '96)*, Nov. 1996.
[14] T. R. Henderson and R. H. Hatz, "Transport Protocols for Internet-Compatible Satellite Networks," *IEEE JSAC*, vol. 72, no. 2, Feb. 1999, pp. 326–44.
[15] R. Goyal *et al.*, "Traffic Management for TCP/IP over Satellite ATM networks," *IEEE Commun. Mag.*, Mar. 1999, pp. 56–61.

BIOGRAPHIES

YURONG HU (yrhu@eee.hku.hk) received her B.E. from Tsinghua University, China, in 1999. She is currently an M.Phil. student in the Department of Electrical and Electronic Engineering at the University of Hong Kong, China. Her research interests are in the areas of satellite networks, multimedia communications, and Internet QoS.

VICTOR O.K. LI [F] (vli@eee.hku.hk) received his S.B., S.M., E.E., and Sc.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology in 1977, 1979, 1980, and 1981, respectively. He is Chair Professor of Information Engineering at the University of Hong Kong, and Managing Director of Versitech Ltd., the technology transfer and commercial arm of the university. Previously, he was professor of electrical engineering at the University of Southern California (USC), Los Angeles, and director of the USC Communication Sciences Institute. He has published over 200 technical articles, and has lectured and consulted extensively around the world. His research interest is in information technologies, focusing on the Internet and wireless networks.