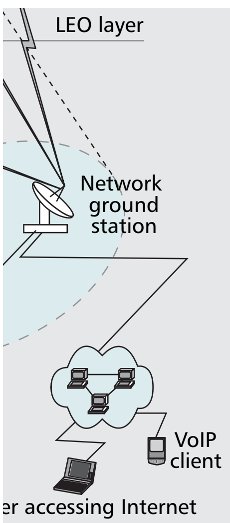
ACCEPTED FROM OPEN CALL

EXPLORING THE ROUTING STRATEGIES IN NEXT-GENERATION SATELLITE NETWORKS

FATIH ALAGÖZ AND ÖMER KORÇAK, SATLAB, BOGAZIÇI UNIVERSITY
ABBAS JAMALIPOUR, UNIVERSITY OF SYDNEY



Employing an efficient routing algorithm in satellite networks is a critical issue. Because of the particular features of satellite networks, new routing approaches should be developed.

ABSTRACT

Employing an efficient routing algorithm in satellite networks is a critical issue, because satellite network resources are costly and must be managed in an optimal and effective manner. Because of the particular features of satellite networks, such as dynamic topology, non-homogeneous traffic distribution, limited power and processing capabilities, and high propagation delays, new routing approaches should be developed. In this article, we first examine these properties with particular emphasis on their effect on the routing objectives. Subsequently, we provide a survey of various routing protocols that aims to address the crucial issues stemming from these properties of satellite systems. We classify these protocols according to their objectives, and discuss their advantages and disadvantages. We also elaborate on relevant technical aspects.

INTRODUCTION

With the rapid globalization of the telecommunications industry, satellites are expected to appear frequently in future telecommunications systems, due to their extensive geographic coverage, inherent broadcast capability, and support for mission-critical applications. Communications over satellites began with the use of individual satellites rotating on the geostationary orbit (GEO). However, focus has shifted toward the development of non-geostationary (NGEO) systems consisting of low earth orbit (LEO) and medium earth orbit (MEO) satellites because of the high propagation delays encountered in GEO systems. The high altitude of GEO satellites leads to high propagation delays that are not suitable for most applications (especially for those running in real-time). Therefore, for better utilization of satellites and to increase the performance of the system, new NGEO systems usually provide on-board processing capabilities including modulation/demodulation, encoding/decoding, transponder/beam switching, and routing. In such systems, any two satellites within line-of-sight are connected to each other via inter-satellite links (ISL). The use of ISL raises the issue of routing in the satellite network. Employing an efficient routing protocol is a critical issue, as satellite network resources are costly and must be managed optimally. The conventional routing techniques proposed for terrestrial networks are not directly applicable to the next-generation satellite networks because of certain features that are particular to satellite systems. In the next section, we provide a brief background on these features. In the rest of the article, we present a survey of various routing techniques proposed for satellite networks. Several technical issues regarding the application of these techniques are discussed. The final section provides a conclusion and identifies some future research directions.

FEATURES OF SATELLITE NETWORKS

A conventional satellite constellation consists of a number of orbits at a certain altitude, a number of satellites per orbit (plane), and ISL between some satellite pairs. Keeping the orbit closer to the earth is more attractive, as it allows for a reduction of the required antenna size as well as the transmission power level and leads to shorter communication delay. However, these advantages come with a price: footprints of satellites at lower altitudes are smaller. Therefore, a higher number of satellites are required for global coverage. In addition, lower-orbit satellites move with higher speeds relative to the earth's surface, resulting in a highly dynamic satellite constellation topology. The mobility of satellites constitutes a major difference between satellite networks and their terrestrial counterparts. Moreover, satellite nodes have different capabilities than terrestrial nodes. Finally, satellite constellation topologies present some specific geometric properties. In the next section, we present a brief overview of these properties of satellite constellation networks. A high-performance routing algorithm should consider the effects of all these features.

EFFECTS OF SATELLITE MOBILITY

Satellite Network Topologies are Dynamic — Although the ISL between satellites in the same plane (intra-plane ISL) are fixed in length, the length of the ISL between satellites from different

1536-1284/07/$20.00 © 2007 IEEE

The long lead times between design, development, launching, and service phases of the satellite systems usually make it difficult to implement state-of-the-art technology. Therefore, satellites should not be designed specifically for use only with current technology.

planes (inter-plane ISL) changes depending on the movement of the satellites. In polar constellations, for example, intra-plane ISL are the longest when satellites are over the equator and the shortest when they are over the polar region boundaries. Moreover, network connectivity can also vary. ISL connectivity between satellites may change based on the distance and the viewing angle between them. ISL passing over a seam also are switched on and off continuously. On the other hand, when a satellite enters a polar region, its adjacent inter-plane ISL are switched off.

Network Topology Changes Are Predictable and Periodic — Although the topology of a satellite network rapidly changes, it is deterministic and can be predicted quite accurately. Moreover, the complete topology dynamic is periodic, that is, it repeats itself within a known period.

Traffic Is Very Dynamic and Non-Homogeneous — Because satellites cover smaller areas in low-orbit systems, traffic requirements are unbalanced due to the varying population density, which is high in cities, low in rural areas, and almost zero over the oceans (which cover about 70 percent of the surface of the earth). As satellites move, traffic received from terrestrial nodes varies continuously depending on the user density in the footprint area.

Handovers Are Required — In connection-oriented satellite network structures, where ISL are switched off due to mobility of the system, a link handover process is required to maintain active connections. Link handover is required when ISL connectivity changes or when the link between a user node and a satellite becomes unavailable.

LIMITATIONS AND CAPABILITIES OF SATELLITES

Power and Onboard Processing Capability Are Limited — As more complex processing is carried out in satellites, they consume more power and their lifetime becomes shorter. There is a trade-off between the lifetime and the processing capability of satellites. Actually, a very long lifetime is not required, because technology improves very quickly. One drawback of a satellite as compared to a terrestrial node is that it is infeasible to upgrade the technology or extend the storage and processing capabilities of a satellite after it is launched.

Implementing State-of-the-Art Technology Is Difficult — The long lead times between design, development, launching, and service phases of the satellite systems usually make it difficult to implement state-of-the-art technology. Therefore, satellites should not be designed specifically for use only with current technology, and interfaces with terrestrial networks also should be designed with a similar approach.

Satellites Have a Broadcast Nature — Satellites offer great potential for multimedia applications because of their capability to broadcast and multicast large amounts of data over a very large area.

FEATURES DUE TO THE NATURE OF SATELLITE CONSTELLATIONS

Higher Propagation Delays — Because of the long distances between satellites and the high altitude of the constellation, propagation delay is considered the most important cost factor in satellite networks. As the altitude of the orbit increases, the effect of propagation delay becomes more evident.

Fixed Number of Nodes — Disregarding satellite failures, usually the number of nodes in a satellite network is fixed unlike most terrestrial networks, where new links or destinations can be added on a daily basis.

Highly Symmetric and Uniform Structure — Since the constellation topology is highly symmetric and uniform, there can be many alternate paths between two satellite nodes. Selection of the most appropriate path can effectively increase the utilization of the system.

Taking these features into consideration, various routing techniques are proposed for satellite networks. The main objectives behind these proposals can be classified as follows:

1. To handle dynamic topology changes with minimum overhead: for this purpose, mainly the periodicity and predictability of the constellation topology are considered. Frequent handovers and limited storage and processing capabilities of satellites are challenges.

2. To prevent the dropping of an outgoing call due to link handover as the satellite topology changes: for this purpose, some of the proposed routing algorithms aim at reducing the probability of link handover occurrence. Mainly periodicity and predictability of topology changes are considered.

3. To minimize the length of the paths in terms of propagation delay and/or number of satellite hops to avoid poor resource utilization as well as high end-to-end delay: constant size and highly symmetric and uniform nature of the constellations constructs are advantageous features that should be considered.

4. To prevent congestion of some ISL while others are idle: for this purpose, various load balancing algorithms are proposed. As a common feature, these algorithms are adaptive to dynamic and non-homogeneous traffic. Load balancing algorithms could also benefit from symmetric and uniform characteristics of constellation topologies and must consider physical restrictions of satellites.

5. To perform traffic-based routing to satisfy quality of service (QoS) requirements: in this context, the main problems are dynamic topology and traffic, frequent handovers, and physical limitations of satellites.

6. To provide better integration of satellite networks with terrestrial networks: in this context, some works try to apply existing routing algorithms to satellite constellations to provide easier integration of the satellite network to the terrestrial network. As an example, some works examine how to adapt IP routing to satellites, so that the constellation can be seamlessly integrated with the Internet. On the other hand, if the

Effect	Features particular to satellite systems	Aim of routing techniques						
		1	2	3	4	5	6	7
Satellite mobility	Dynamic network topology	√	√	√		√	√	√
	Predictable and periodic topology changes	√	√	√			√	√
	Dynamic and non-homogeneous traffic				√	√		√
	Handovers	√	√			√	√	√
Limitations and capabilities of satellites	Limited power and onboard processing capacity	√			√	√	√	√
	Difficulty of implementing latest technology						√	√
	Broadcast nature							√
Nature of satellite constellation	Higher propagation delays			√				√
	Fixed number of nodes			√	√		√	√
	Highly symmetric and uniform structure			√	√			√

Table 1. *The relationship between classes of routing objectives and satellite features.*

satellite network uses its own arbitrary routing protocol, the problem to solve is how to integrate it with the terrestrial networks.

7. To perform efficient multicasting over satellites, regarding the characteristics of satellites in broadcasting.

To achieve each of these seven objectives, some of the aforementioned features of satellite systems should be considered. Table 1 shows the relationship between classes of routing objectives and satellite features. The check marks indicate the existence of an effective relation but are not necessarily restricted to those shown in Table 1.

Routing in Satellite Networks

In this section, we examine how the proposed routing techniques may achieve the objectives stated in the previous section and discuss some of their advantages and disadvantages.

Routing Issues in a Dynamic Topology

Due to the rapid changes in the status of the links and satellite positions, a satellite network can be considered as a dynamic-topology network. The utilization of conventional routing techniques widely used in terrestrial networks is not feasible for satellite networks, because these protocols rely on the exchange of topology information that must be constantly refreshed, which incurs substantial overhead. However, although the topology of a satellite network rapidly changes, these changes are periodic and predictable because of the strict orbital movements of the satellites. Therefore, some routing techniques are proposed utilizing this periodicity feature. In [1], a LEO satellite network is modeled as a finite-state automaton, where the system period (which is the least common multiple of the orbital period of the satellite layer and the earth period) is divided into states. The states are derived from the ISL connectivity data, so that the network has a fixed topology in each state. Due to the periodicity of the constellation topology, there can be only a finite number of states. Then, it is proposed to perform optimal routing on each of these fixed topologies for the best use of ISL in the system. A number of routing tables are stored on-board and retrieved when topology changes. Although the messaging overhead and computational complexity is reduced, large storage capacity is required in the satellites. To reduce the on-board storage requirements, a suitable number of network control centers (NCC), usually located on the ground, can be used. In this context, deciding on the number of NCC to use and their distribution on the globe are left as open issues.

Another concept worth mentioning that is tailored to dynamic satellite constellation is the virtual node (VN) concept. In the VN technique, a fixed virtual topology consisting of VN is superimposed over the physical topology to hide the mobility of satellites from the routing protocols. Routing is performed in the fixed virtual topology by the use of a common routing protocol. This scheme can directly integrate the space network with the terrestrial IP network and may provide good support for IP-multicast and IP-QoS [2]. However, this approach presents some challenging problems such as scalability of routing tables and high computational complexity in space devices. Moreover, [3] states that one-to-one mapping of physical topology to a virtual topology is problematic for several reasons. One reason is the *sun outage* phenomenon, where the satellite serving a fixed footprint is in the same line of sight with the sun. This incurs inefficiency due to highly non-uniform user density and the possibility of better coverage in a region of a non-serving satellite. Therefore, further work is required to improve the VN concept.

The utilization of conventional routing techniques widely used in terrestrial networks is not feasible for satellite networks, because these protocols rely on the exchange of topology information that must be constantly refreshed, which incurs substantial overhead.

When compared to terrestrial networks, the propagation delay is more important in space networks due to the altitude of the satellites orbit and long distances between the nodes. Moreover, as a greater number of satellite hops is traversed, total processing and propagation delays increase.

REDUCING LINK HANDOVERS AND REROUTING ISSUES

When a connection is broken, the issue of (re)connection setup overhead is imperative for satellite networks due to the highly dynamic nature of the network topology. When some ISL are switched off or the up/down link (UDL) between the terrestrial node and the corresponding initial satellite is broken, handover is required to maintain the active connections. Rerouting attempts during link handovers cause delay jitter and signaling overhead. Moreover, because of the possibility of resource unavailability in alternate paths and the delay caused by rerouting, the forced termination probability of ongoing connections is increased. Therefore, it is desirable to minimize the possibility of rerouting due to handovers. It is especially important for real-time multimedia applications to maintain QoS guarantees using a connection-oriented routing protocol. Thus, minimizing the number of connection handovers is a crucial issue.

To reduce link handovers, Jukan *et al.* propose to reduce the handover ratio by favoring ISL with higher lifetimes [4]. For each connection request, request packets are flooded towards the destination. While traversing the route, these packets gather the lifetime information of intermediate satellites, and this information is used by the destination node for deciding on the most appropriate path. Chen *et al.* [5] also consider minimizing handovers in their proposed adaptive routing scheme. Among the set of paths that satisfy the QoS requirements, a path that can minimize the possible number of handovers and that also is not inferior to the best possible path with a predefined degree is selected.

The above-mentioned procedures create optimization between two satellite nodes. Hence, only ISL handovers are considered. However, a connection also should be re-established when the UDL between the source ground station and the ingress satellite — or between the destination ground station and the egress satellite — is broken. This kind of handover is defined as *inter-satellite handover* [6] and should be considered for better optimization. It is preferable to reduce inter-satellite handovers due to the movement of satellites with respect to user stations. In [6], Uzunalioglu, *et al.* introduce a probabilistic routing protocol (PRP) that tries to reduce rerouting between two ground end-users by utilizing LEO satellite topology dynamics and call statistics. Basically, the algorithm tries not to use ISL that would be switched off before the connection is over. Since exact call duration is not known a priori, the probability density function (PDF) of the time duration in which the call uses the established route is utilized. The determined PDF is used to establish the routes of the connection, such that the routes are terminated by a call termination event or an inter-satellite handover instead of a link handover with a target probability p. A distinction is made between route calculation for newly arrived calls and inter-satellite handover calls because dropping an ongoing call that experiences an inter-satellite handover is more annoying than blocking a new call.

PATH MINIMIZATION ALGORITHMS

In a satellite network, the cost of a path is determined by the propagation and processing delays on the satellites. When compared to terrestrial networks, the propagation delay is more important in space networks due to the altitude of the satellites orbit and long distances between the nodes. Moreover, as a greater number of satellite hops is traversed, total processing and propagation delays increase. Reducing the processing delay has beneficial side effects, such as reducing the data blocking probability and yielding better power consumption in satellites. Therefore, path minimization is a crucial task in satellite networks.

In some works, it is assumed that ISL have fixed lengths. Various authors argue that this assumption is reasonable because in most of the constellation topologies, the length variation is low, especially for crowded regions that are closer to the equatorial region. Moreover, it also is claimed that minimizing the hop count is more critical for improving the performance of the system, and therefore it is reasonable to ignore dynamic length variations of ISL. On the other hand, numerous authors think that the dominating factor for performance is the propagation delay, and they aim to find a minimum-propagation-delay path with the minimal hop count among the paths. This complicates the task, because ISL lengths change with time due to the dynamism of the topology. The predictability of topology changes, and the known facts about the nature of satellite networks (e.g., links over the polar regions are shorter than links over the equatorial region) can be used to simplify this task. Nevertheless, extra storage and processing complexity is required to consider the propagation delays.

Sun *et al.* [7] deal with static routing in a regular LEO satellite network that is modeled as a two-dimensional N-ary hypercube. The minimum-hop path is found by the Dijkstra algorithm, and contention resolution schemes are investigated for maximizing the throughput. It is shown analytically and validated by simulation results that a scheduling scheme favoring packets closest to their destinations results in maximum system throughput, at the expense of degraded system fairness. Actually, there could be many minimum-hop paths in a satellite constellation, due to its symmetric and uniform nature. Therefore, it is proposed in some works to favor the one with the minimum-propagation delay. The most trivial way to do this is to store the length information of all links for a system period in each satellite (or in ground stations that perform routing) and to apply a shortest-path algorithm using this information. This requires a high storage capacity. Henderson *et al.* [3] and Ekici *et al.* [8] develop distributed algorithms for minimizing the propagation delay. The geography-based algorithm of Henderson *et al.* is based on the hypothesis that the series of locally optimal forwarding decisions will yield a route that is close to the optimal route. In other words, depending on the geographic information embedded in the addresses, each satellite forwards the packet to its neighbor that most

Load balancing scheme		Advantages	Disadvantages
Source-based	Isolated	Simple	No global view of the network
			Low utilization
	Nonisolated	Global view of the network	Difficult to guarantee up-to-dateness of the traffic information
		Good traffic adaptiveness	High signaling overhead
Central		Global view of the network	Difficult to guarantee up-to-dateness of the traffic information
		Whole information can be used for an overall optimization procedure	High signaling overhead
		Computational complexity is carried from satellites to a central node	Scalability problem
Distributed		Each node uses up-to-date local information	No global view of the traffic load distribution
		Low signaling overhead	Utilization is somewhat low
		No rerouting issues	
		Fast adaptation to changes	
Hierarchical		More routing choices	Physical challenges in providing interorbital satellite communications
		Better adaptation to traffic changes with less computational and signaling cost	Increased system complexity

Table 2. *Comparison of different load balancing schemes.*

reduces the distance to the destination. On the other hand, Ekici *et al.* [8] introduce the datagram routing algorithm (DRA) for an idealized polar constellation. It regards the satellite network as a mesh topology consisting of logical locations (virtual nodes). Data packets are routed in a distributed fashion in this fixed topology. DRA consists of two phases: at a given satellite hop, it first finds all the neighboring satellites that can move the packet one hop closer to the destination. Then, from the candidate next hops, it selects the one that most reduces the remaining distance to the destination.

There are many other algorithms that utilize information on expected traffic characteristics and handover possibilities of ISL while applying a shortest-path algorithm or similarly consider dynamic traffic characteristics while deciding on the most appropriate path among a set of shortest paths. The objective of those algorithms from the latter group is mainly to distribute the traffic load in a more balanced way. They are described in the following section.

Load Balancing Algorithms

Because population distribution on the surface of the earth is highly non-uniform, traffic requirements are unbalanced in a low orbit satellite network. This may lead to congestion in some resources, while others are under-utilized. To overcome this problem, the routing algorithm should distribute the flows in a balanced way over appropriate ISL between the communicating nodes.

We classify these algorithms according to the place where the routing is performed: source-based, central, distributed, and hierarchical load-balancing algorithms, as shown in Table 2.

Source-Based Load Balancing — In source-based load-balancing algorithms, the route to a given destination node is calculated by the ingress node. The ingress node could be a terrestrial node or a satellite. If it is located on the ground, an extra signaling delay is introduced. However, in the latter case, computational and storage requirements to perform route calculation can exceed the capacity limits of a satellite.

Franck *et al.* [9] further classify source-based load balancing algorithms as isolated and non-isolated algorithms. Isolated algorithms use only information local to the node where routing is performed. In non-isolated algorithms, traffic information is gathered from the whole network. Authors suggest a non-isolated algorithm as follows: each node keeps the graph of the whole network. When routing is performed, all nodes and edges that are near to saturation point are pruned. Then a shortest-path algorithm is run, considering the propagation delay as well as the constant transit delay per hop.

Chen and Jamalipour in [5] propose an alternative adaptive routing algorithm that uses the information on both the average and the mini-

Because of the highly symmetric and uniform structure of satellite constellations, there can be many minimum-hop paths between two satellites, and routing probably can be done efficiently.

mum number of occupied channels per route. First, the algorithm finds a set of candidate minimum delay paths that also minimize the handover probability and delay jitter. Then, among these alternate paths, the one with minimum traffic weight, which is determined by a weighted combination of average and minimum number of occupied channels over the route, is selected.

Non-isolated routing technique increases the computational and signaling complexity of the routing architecture, but it is superior to isolated algorithms because it considers traffic adaptivity in the whole network. However, there is a potential drawback of non-isolated algorithms: the gathered traffic information may not reflect the actual condition since the information takes time to be distributed in the constellation (due to high propagation delays) and to avoid excessive signaling, state changes in the network are not always advertised.

Central Load Balancing — In central load-balancing algorithms, routing tables are calculated in a central node and then stored in the satellite nodes. This central node can be a satellite or a terrestrial node. Mainly, we can consider optimal-routing algorithms in this context. Optimal-routing algorithms are network-oriented, aimed at minimizing the mean-blocking probability in the network by providing better load balancing.

The aforementioned finite-state automaton (FSA)-based algorithm of Chang *et al.* [1], which is offline, assigns expected traffic loads to links depending on the statistical information on the potential requirement density for each node and the distance between the nodes. For each state, it aims to maximize the minimum residual capacity in the network. Because this is an NP-complete mixed-integer optimization problem, it uses heuristics to provide optimal routing. Authors conclude through simulation that optimal routing is superior in terms of newly initiated and ongoing call-blocking probability.

Papapetreu *et al.* [10] propose to perform the flow deviation (FD) algorithm, which is a well known optimal routing algorithm that aims to find a routing pattern that minimizes the mean network delay. Depending on the information gathered from the whole network, a designated central node performs the FD algorithm. The FD algorithm splits the load among different paths according to path lengths that are defined as a flow dependent metric. As the lengths of the paths change, the FD algorithm continuously adapts these changes by deviating traffic from one path to another, so that the defined cost metric is minimized. The authors show by simulation that the FD algorithm is superior to the Dijkstra algorithm for finding the path minimizing the propagation delay and to the adaptive Dijkstra algorithm for finding the path minimizing a flow metric.

By performing routing at a central node, better traffic engineering could be maintained using the global view of the network. However, central load-balancing algorithms share the same problems with non-isolated source-based routing algorithms. Computational complexity can be carried to a ground node that does not suffer much from power limits, but the high signaling requirement and the difficulty of accurately transferring traffic information are the most challenging problems. Moreover, the centralized routing approach may present scalability problems due to the capacity limits of the central node and the rapid increase in the computational complexity with the enlargement of the network size.

Distributed Load Balancing — Because of the highly symmetric and uniform structure of satellite constellations, there can be many minimum-hop paths between two satellites, and routing probably can be done efficiently. Establishing a static connection between two nodes may lead to poor utilization of alternate paths. Moreover, as we mentioned before, the connection-oriented approach may suffer in attaining path connectivity by handover mechanisms. Rather, a distributed next-hop routing strategy seems to be simpler. Each satellite independently decides on the best next hop to forward the packet. Ekici *et al.* [8] implement this approach in the aforementioned DRA. The main objective of the algorithm is to minimize the propagation delay, but a satellite may change its decision if the output queue for the ISL over the minimum-propagation-delay path is congested. Taleb *et al.* [11] claim that a better load balancing might be achieved, given that a satellite is aware of the traffic conditions at the next-hop satellite. They propose an explicit load-balancing (ELB) scheme, where a congested satellite sends a signal to its neighboring satellites to decrease their sending rates, and its neighbors search for alternate paths. This method reduces the packet dropping probability, but it is not safe from signaling congestion due to feedback packets (even though signaling packets are sent only when it is necessary, they could be required very frequently in some conditions). Algorithms in [8] and [11] do not take any action for load balancing until some nodes experience a certain level of congestion, that is, they are not proactive. In [12], we show that it is more appropriate to avoid congestion before it happens and provide a priority-based adaptive-routing (PAR) algorithm that aims to balance the traffic before congestion occurs. We compare PAR with other adaptive minimum-hop-path algorithms, such as fixed adaptive routing (FAR) and random adaptive routing (RAR). At each hop, among the ISL that are on a minimum-hop path, FAR first selects the one that is toward a given direction (vertical or horizontal). If that link is congested, the other direction is selected. RAR makes the selection of the initial ISL candidate in a random manner. Simulation results show that using the PAR algorithm not only increases the throughput, but also decreases the queuing delay. Moreover, the PAR algorithm does not incur any signaling overhead due to feedback packets.

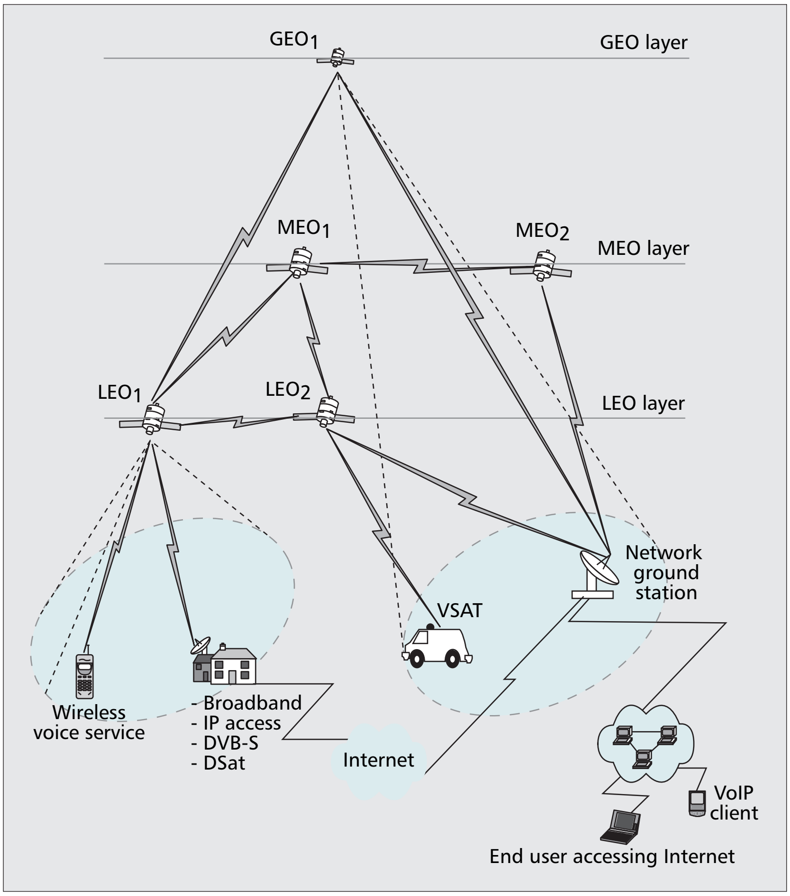
The distributed load-balancing algorithms mentioned previously provide fast reaction to traffic changes when compared with the source-based and centralized load-balancing algorithms. However, they use only the local traffic information, which might not reflect the entire traffic load distribution. Surely, it is possible to distribute the local information to the whole network and use it in local next-hop decisions, but this will cause extensive signaling overhead.

Hierarchical Load Balancing — Hierarchical (multi-layered) satellite architectures with inter-orbital links (IOL) between layers of satellite constellations are of much interest, as they may yield better performance than individual layers. Figure 1 depicts a layered satellite architecture with some applications. The hierarchical approach aims to reduce the computational complexity on the satellites and the communication load on the network, compared to non-isolated algorithms, while enabling better adaptation to traffic changes.

Lee *et al.* [13] propose a satellite-over-satellite (SOS) system, where the satellite architecture is composed of LEO and MEO layers. The lower layer (LEO) satellites send ISL state messages to the upper layer. The upper layer (MEO) satellite uses this state information to derive some local routing information (LRI) about the LEO satellites that are in its coverage area (these change with time due to the relative mobility of LEO satellites with respect to MEO satellites). This information also is exchanged between MEO satellites. In addition, MEO satellites derive global routing information (GRI), including all routing information of the LEO and MEO layer satellites, by using this exchanged state information and send it through IOL to all LEO-layer satellites that are within their coverage area. In the proposed routing algorithm, short-distance-dependent traffic is transmitted through lower layer satellites, but long-distance-dependent traffic is transmitted through IOL up to the MEO layer to minimize the average number of satellite hops and resource consumption.

The satellite grouping and routing protocol (SGRP) proposed in [14] is another hierarchical algorithm where LEO satellites are divided into groups according to the footprint area of the MEO satellites in each snapshot period. Each LEO group is managed by a MEO. Similar to the proposal in [13], a MEO satellite computes the minimum-delay paths for its LEO members, depending on the link-state information arriving from the LEO satellites. The authors provide a detailed analysis of their proposed system.

Dash *et al.* [15] consider a three-layered architecture consisting of GEO, LEO, and high altitude platforms (HAP) for voice over IP (VoIP) applications. GEO act as the backbone routers, LEO as the second layer, and HAP cover special areas with high and sensitive traffic, such as battlefields and disaster areas. The LEO layer satellites are assumed to have earth-fixed footprints, and they are modelled as virtual nodes; hence the GEO satellites cover logically fixed LEO topologies. The LEO satellites exchange their routing tables as their footprint area changes. This architecture enables all the satellites to see other layers as stationary. The LEO satellites and HAP measure residual bandwidth on their outgoing links and send the information to the GEO layer within a given period. Since the GEO satellites have limited on-board processing capacities, this link-state information can be sent to the fixed terrestrial gateways for processing.



■ **Figure 1.** *A layered satellite network with sample applications.*

After forming intra-domain routing tables, the gateways upload these tables to the GEO layer, and the GEO satellites flood them to the LEO satellites and HAP that are in their coverage area. An aggregated routing table for each domain is also formed that includes the maximum-residual-bandwidth paths between different border nodes of the domain. Then, these tables are exchanged between GEOs and transferred to the lower layers.

The advent of hierarchical architectures implies more redundancy and routing choices in satellite systems. A variety of topological design and routing issues should be investigated to enable the best use of satellite technologies in future communication systems.

Traffic-Based Routing

To support the rapidly-growing real-time multimedia services, satellite systems should offer QoS guarantees, which is particularly difficult in connectionless networks due to the difficulty in accounting for the delay aspects of QoS and sequencing. Usually, QoS guarantees are provided through connection orientation. However, due to the high mobility of satellites in non-geostationary systems, it is difficult to attain path connectivity. Therefore, some of the algorithms described previously aim at minimizing the rerouting probability due to handovers while calculating the routes. Nevertheless, it is not easy to offer QoS guarantees without reducing the rerouting probability to very low levels. The VN

There is a significant time gap between the design and the operational stages of a satellite system. That lag usually makes it difficult to provide and to utilize state-of-the-art technology in orbit.

concept can be used to remove topology changes. However, this approach also has its drawbacks as described previously. Therefore, this topic requires further study.

Another issue in providing QoS guarantees is to reduce delay variation (jitter) that occurs due to path rerouting. In a LEO satellite network, the movement of satellites causes changes in relative positions between any two satellites that belong to different orbits. This may result in unacceptable levels of delay jitter. Therefore, a routing algorithm must try to reduce the delay jitter for better QoS conditions while at the same time keeping the delay itself as low as possible. This issue is considered in the context of various proposed algorithms in the literature [5]. Regarding handovers, a routing algorithm is expected to choose a new path that is not much different from the previous one in its length, although sometimes the selected path is not the best (shortest) one.

Jukan *et al.* [4] propose a distributed QoS-routing approach. The source node floods connection request packets towards the destination. At each intermediate node, the quality parameters (delay, lifetime of ISL, etc.) are updated. When connection request packets reach the destination node, the destination eliminates the paths that do not satisfy the QoS requirements. Among all feasible paths, the one with the minimum number of hops and maximum lifetime is selected.

Kandus *et al.* [16] propose a traffic-class-dependent (TCD) routing algorithm. Three classes of traffic are considered: (1) delay-sensitive traffic, (2) throughput-sensitive traffic, and (3) best-effort traffic. The routing algorithm behaves differently for each class of traffic. Each satellite has three separate outgoing queues (one for each traffic class), serving for each outgoing link. A scheduler should be implemented that defines the actual transmission sequence of packets in outgoing queues. Obviously, the selection of the scheduling policy has a large impact on the routing performance of the particular traffic class. Therefore, the authors investigate five different scheduling policies that we do not describe in this article. The TCD algorithm attempts to guarantee QoS for different traffic classes. However, it may assign a single route for a specific class with high data traffic and may heavily overload the chosen path. This may negatively affect the traffic load balancing over the entire constellation.

Routing from a Space-Ground Integration Point of View

The problem of how to integrate space networks with terrestrial networks arises when considering using satellite systems as a part of a global communication system (such as the Internet). Basically, there are two trends in this context: according to the first trend, the goal is to apply the existing algorithms as extensively as possible and provide interfaces with terrestrial networks as easily as possible. Conversely, there is a second trend to use an arbitrary routing protocol in the space segment. In other words, the satellite network can be designed and operated independently of the terrestrial network. The disadvantage of the latter trend is that an address resolution protocol and complex interworking functions are required. However, it is a better approach than the former trend in terms of scalability.

Currently, the Internet Protocol (IP) dominates the end systems attached to the satellite terminals. Therefore, research that investigates how to apply IP routing directly to satellite systems can be considered in the context of the first trend. Wood *et al.* [2] examine a strategy (based on the VN concept) that aims to adopt IP routing at the satellites. This strategy permits direct support for IP-multicast and IP-QoS (integrated and differentiated service models). However, there are challenging problems with this technique. First, the variable-length IP packets must fit into a fixed-length frame structure in the space segment. The authors propose to achieve this by using either explicit IP-level fragmentation or implicit lower-level fragmentation to break packets, such that they can fill in the frames, and by using padding to fill up the frame structures. The second important problem is scalability. When the terrestrial network increases in size, a large amount of routing information must be updated for both the terrestrial and the satellite networks. However, this is not feasible for satellites, because the capacity of satellites is limited and cannot be upgraded after they are launched. Therefore, it is better to separate and isolate satellite and Internet routing updates. Finally, because IP-routing is slow and requires high processing power, it is not suitable for satellites that are equipped with limited on-board processing capacities. The authors argue that the IP-routing performance is continuously improving and that by using some shortcut IP-switching techniques, such as multi-protocol label switching (MPLS), it seems possible to overcome this problem. We briefly discuss MPLS over satellite constellations in the last part of this section.

There is a significant time gap between the design and the operational stages of a satellite system. That lag usually makes it difficult to provide and to utilize state-of-the-art technology in orbit. Moreover, next generation satellite networks are expected to provide and support multiple types of services and to interwork with different terrestrial networks, such as broadband integrated services digital network (B-ISDN), Internet, and so on. Therefore, it is reasonable to isolate the routing in the space segment from that in the terrestrial networks. For this purpose, it is suggested to view the satellite network as an autonomous system (AS) with a different addressing scheme.

In terrestrial networks, the cost metrics between different AS are the dominating factor when establishing a route between nodes from different AS. Generally, this is because the internal paths in a terrestrial AS are much shorter than the inter-AS paths. However, this is not the case for a satellite network, because an internal path easily can be as long as an external path. Thus, in the satellite context, the routing protocol should consider the internal cost metric to be as important as the external one.

Wood *et al.* [2] claim that in the future, satellite constellations can carry IP traffic by using a combination of border routing protocols, tunneling, network address translation, and MPLS. MPLS enables the adopting of new paradigms for conventional IP traffic by decoupling packet forwarding from the information carried in the IP header. This is achieved by distributing the routing information to the core routers and assigning short labels to the packets at the ingress point. MPLS has appealing mechanisms that essentially support the integration of the IP world with QoS and traffic engineering features. Donner *et al.* [17] deal with how to adopt MPLS over a satellite constellation. Constellation topologies that do not have a seam (inclined Walker constellations) are seen as the strongest candidates to host MPLS, due to their permanent nature. Nevertheless, the inherent and frequent handovers between ground and satellite stations and topology dynamics due to varying ISL lengths remain a challenge. At the ingress point of an MPLS network, there are label edge routers (LER) that manage the label distribution and in some cases, perform route computations. The authors propose to locate LER on the ground to avoid

- Expensive and complex on-board processing in the satellites
- Extra signaling overhead required for restarting QoS negotiation or admission control for rerouting of a label switching path (LSP)

Different scenarios for route computation (including "intelligence" in terms of traffic engineering, adaptiveness, and/or optimization) and LSP management (establishing the result of *route computation* in the network) are investigated. Centralized routing approaches are viewed as more promising for use within an MPLS framework. For further information, see [17]. MPLS-based networking in satellite constellations is an appealing approach. However, some important practical problems related to rerouting and maintenance overhead are still unsolved and deserve further study.

Multicast Routing

Given the ability of satellites to broadcast large amounts of data over a very wide area, multicasting over satellites has become a very hot research topic. Ekici *et al.* [18] indicate that none of the existing multicast routing protocols (such as reverse-path multicast (RVM), distance vector multicast routing protocol (DVMRP), or multicast routing extensions for open shortest path first (OSPF), such as multicast OSPF (MOSPF), are suitable for satellite networks, because they employ some type of periodic message exchanges to form or maintain multicast trees, and this is not favorable due to the physical limitations of satellites. To fill the gap, authors developed a multicast-routing protocol for LEO satellite IP networks, where multicast trees are formed based on DRA [8]. The algorithm aims to minimize the number of branches extending from a satellite at each step. Authors conclude that their algorithm outperforms existing multicast-routing algorithms in terms of end-to-end delay. But, multicast-routing algorithms for multi-layered satellite networks still is an appealing research area.

Conclusions

Satellite communication systems have some intrinsic features that significantly affect the performance of their routing algorithms. In particular, adopting traffic and topology dynamics may incur significant overhead in the course of utilizing satellite resources. Various routing algorithms were proposed for satellite networks to overcome this drawback. In this article, we classified these algorithms and described relevant technical issues for use in the next generation satellite networks. Although some algorithms seem to meet performance criteria in certain cases, routing algorithms that support minimum overhead with better resource utilization remain a practical problem for the next generation satellite communication networks. Both QoS and multicast routing algorithms are still formidable tasks in satellite networks. Furthermore, the advent of sophisticated satellite network architectures, such as multi-layered systems and new space technologies such as HAP, continue to broaden potential routing choices [19].

Both QoS and multicast routing algorithms are still formidable tasks in satellite networks. Furthermore, the advent of sophisticated satellite network architectures, such as multi-layered systems and new space technologies such as HAP, continue to broaden potential routing choices.

Acknowledgment

This work is supported by the State Planning Organization of Turkey under the Next Generation Satellite Networks Project, grant no. DPT-03K 120250.

References

[1] H. S. Chang *et al.*, "Performance Comparison of Optimal Routing and Dynamic Routing in Low-Earth Orbit Satellite Networks," *Proc. VTC '96*, Atlanta, GA, 1996.

[2] L. Wood *et al.*, "IP Routing Issues in Satellite Constellation Networks," *Int'l. J. Sat. Networks*, vol. 19, no. 1, Jan./Feb. 2001, pp. 69–92.

[3] T. H. Henderson and R. H. Katz, "On Distributed and Geographic-Based Packet Routing for LEO Satellite Networks," *Proc. GLOBECOM*, San Francisco, CA, Dec. 2000.

[4] A. Jukan, N. H. Nguyen, and H. R. van As, "An Approach to QoS-Based Routing for LEO Satellite Networks," *IEEE Int'l. Conf. Commun. Tech.*, Beijing, China, vol. 1, 2000, pp. 922–29.

[5] J. Chen and A. Jamalipour, "An Adaptive Path Routing Scheme for Satellite IP Networks," *Int'l. J. Commun. Sys.*, vol. 16, no. 1, Feb. 2003, pp. 5–21.

[6] H. Uzunalioglu, M. D. Bender, and I. F. Akyildiz, "A Routing Algorithm for LEO Satellite Networks with Dynamic Connectivity," *ACM-Baltzer J. Wireless Networks*, vol. 6, no. 3, June 2000, pp. 181–90.

[7] J. Sun and E. Modiano, "Routing Strategies for Maximizing Throughput in LEO Satellite Networks," *IEEE JSAC*, vol. 22, 2004, pp. 273–86.

[8] E. Ekici, I. F. Akyildiz, and M. D. Bender, "A Distributed Routing Algorithm for Datagram Traffic in LEO Satellite Networks," *IEEE/ACM Trans. Net.*, vol. 9, no. 2, 2001, pp. 137–47.

[9] L. Franck and G. Maral, "Static and Adaptive Routing in ISL Networks from a Constellation Perspective," *Int'l. J. Sat. Commun.*, vol. 20, no. 6, 2002, pp. 455–75.

[10] I. Gragopoulos, E. Papapetrou, and F.-N.Pavlidou, "Performance Study of Adaptive Routing Algorithms for LEO Satellite Constellations under Self-Similar and Poisson Traffic," *Space Commun.*, vol. 16, 2000, pp. 15–22.

[11] T. Taleb *et al.*, "ELB: An Explicit Load Balancing Routing Protocol for Multi-Hop NGEO Satellite Constellations," *Proc. GLOBECOM '06*, San Francisco, CA, Nov. 2006.

[12] Ö. Korçak, F. Alagöz, and A. Jamalipour, "Priority-Based Adaptive Shortest Path Routing for IP over Satellite Networks," *Int'l. J. Commun. Sys.*, vol. 20, no. 3, 2007, pp. 313–33.

[13] J. Lee and S. Kang, "Satellite over Satellite (SOS) Network: A Novel Architecture for Satellite Network," *Proc. IEEE INFOCOM*, Tel-Aviv, Israel, vol. 1, Mar. 2000, pp. 315–21.

[14] C. Chen and E. Ekici, "A Routing Protocol for Hierarchical LEO/MEO Satellite IP Networks," *Wireless Networks*, vol. 11, no. 4, July 2005, pp. 507–23.

[15] D. S. Dash, A. Durresi, and R. Jain, "Routing of VoIP Traffic in Multilayered Satellite Networks," *Proc. SPIE Performance and Control of Next-Gen. Commun. Networks, ITCOMM 2003*, Orlando, FL, vol. 5244, pp. 65–75.

[16] G. Kandus, A. Svigelj, and M. Mohorcic, "The Impact of Different Scheduling Policies on Traffic Class Dependent Routing in Intersatellite Link Networks," *Int'l. J. Sat. Commun.*, vol. 22, 2004, pp. 533–46.

[17] D. Donner, M. Berioli, and M. Werner, "MPLS-Based Satellite Constellation Networks," *IEEE JSAC*, vol.22, no.3, Apr. 2004, pp. 438–48.

[18] E. Ekici, I. F. Akyildiz, and M. D. Bender, "A Multicast Routing Algorithm for LEO Satellite IP Networks," *IEEE/ACM Trans. Net.*, vol. 10, no. 2, Apr. 2002, pp. 183–92.

[19] S. Bayhan, G. Gur, and F. Alagöz, "VoIP Performance in Multilayered Satellite IP Networks with OBP," DOI: 10.1002/doc.874, 2007.

Biographies

FATIH ALAGÖZ (alagoz@boun.edu.tr) obtained his D.Sc. degree in electrical engineering in 2000 from George Washington University, Washington, DC and his B.Sc. in electrical engineering in 1992 from METU, Turkey. He is an associate professor in the Department of Computer Engineering, Bogaziçi University, Turkey. He is the director of the Satellite Networks Research Laboratory at the university. His research interests include wireless, mobile, and satellite communications. He was with the Department of Electrical Engineering, United Arab Emirates University, during 2001–2003. He has edited two books and published more than 50 scholarly papers in selected journals and conferences.

ÖMER KORÇAK (korcakom@cmpe.boun.edu.tr) is currently working toward a Ph.D. degree in the Department of Computer Engineering, Bogaziçi University. He received his B.Sc. and M.Sc. degrees in computer engineering from Bogaziçi University in 2002 and 2004, respectively. He is a Ph.D. scholar of the Scientific and Technical Research Council of Turkey. His main research interests include networking in non-geostationary satellite systems with a focus on routing and traffic engineering issues and all-optical networking with a focus on wavelength assignment in optical burst-switching networks.

ABBAS JAMALIPOUR [S'86, M'91, SM'00, F'07] (a.jamalipour@ieee.org) has a Ph.D. from Nagoya University, Japan. Currently, he is vice-chair of the Communications Switching and Routing Technical Committee and chair of the Chapters Coordinating Committee, Asia-Pacific Board, all part of IEEE ComSoc. His areas of research are wireless data communications networks, wireless IP networks, next-generation mobile networks, traffic control, network security and management, and satellite systems. He was one of the first researchers to disseminate the fundamental concepts of next-generation mobile networks and broadband convergence networks, as well as the integration of WLAN and cellular networks, some of which are gradually being deployed by industry and are included in ITU-T standards. He is the author of the first book on wireless IP and two other books and has co-authored five books and over 180 technical papers, all in the field of mobile communications networks. He is a Fellow of the Institute of Engineers Australia and an IEEE Distinguished Lecturer, Editor-in-Chief of IEEE Wireless Communications, and a Technical Editor of several scholarly publications, including IEEE Communications Magazine, Wiley International Journal of Communication Systems, and Journal of Communication Networks. He has authored several invited papers and been a keynote speaker at many conferences. He served as chair of the Satellite and Space Communications Technical Committee from 2004 to 2006. He has received several awards, including the 2006 IEEE Distinguished Contribution to Satellite Communications Award, the 2006 IEEE Communications Society Best Tutorial Paper Award, and the 2005 Telstra Award for Excellence in Teaching.