Footprint handover rerouting protocol for low Earth orbit satellite networks

Hüseyin Uzunalioğlu [a], Ian F. Akyildiz [b], Yelena Yesha [c] and Wei Yen [d]

[a] *Bell Labs Advanced Technologies, Lucent Technologies, 101 Crawfords Corner Road, Holmdel, NJ 07701, USA*
[b] *Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[c] *Department of Electrical Engineering and Computer Science, University of Maryland at BC, Baltimore, MD 21250, USA, also with NASA CESDIS, Goddard Space Flight Center, Code 930.5, Greenbelt, MD 20771, USA*
[d] *Mitsubishi Telecommunications Network Division, 3885 Crestwood Parkway, Suite 300, Duluth, GA 30096, USA*

Low Earth Orbit (LEO) satellite networks will be an integral part of the next generation telecommunications infrastructures. In a LEO satellite network, satellites and their individual coverage areas move relative to a fixed observer on Earth. To ensure that ongoing calls are not disrupted as a result of satellite movement, calls should be transferred or *handed over* to new satellites. Since two satellites are involved in a satellite handover, connection route should be modified to include the new satellite into the connection route. The route change can be achieved by augmenting the existing route with the new satellite or by completely rerouting the connection. Route augmentation is simple to implement, however the resulting route is not optimal. Complete rerouting achieves optimal routes at the expense of signaling overhead. In this paper, we introduce a handover rerouting protocol that maintains the optimality of the initial route without performing a routing algorithm after intersatellite handovers. The FHRP makes use of the footprints of the satellites in the initial route as the reference for rerouting. More specifically, after an optimum route has been determined during the call establishment process, the FHRP ensures that the new route due to handover is also optimum. The FHRP demands easy processing, signaling, and storage costs. The performance results show that the FHRP performs similar to a network without any handovers in terms of call blocking probability.

1. Introduction

Terrestrial wireless networks such as cellular and PCS networks provide mobile communication services with limited geographic coverage. To provide global coverage to a more diverse user population, a number of *low Earth orbit* (LEO) satellite systems have been proposed [10,12,17]. The LEO systems can support both the areas with terrestrial wireless networks and those areas that lack any wireless infrastructure. In the former case, a satellite system could interact with the terrestrial wireless network to absorb the instantaneous traffic overload of the terrestrial wireless network. In other words, mobile users would alternatively access a terrestrial or a satellite network through dual-mode handheld terminals. In the latter application area, LEO satellites would cover regions where terrestrial wireless systems are economically infeasible due to rough terrain or insufficient user population.

The term LEO is used to classify satellites with orbiting altitudes between 500 and 1500 km above the Earth's surface. This low altitude provides small end-to-end delays and low power requirements for both satellites and terminals. As figure 1 depicts, users can access LEO satellites with their small handheld phones. Moreover, the satellites can be connected to terrestrial networks via gateways. Another feature of LEO satellites is that intersatellite links (ISL) allow the routing of a connection through the satellite network without requiring any terrestrial resources.

LEO satellites move with respect to a fixed observer on the Earth surface. The velocity of a LEO satellite relative to a fixed observer is very fast ($\approx$ 8 km/s) [9]. Because of this nonstationary characteristic, the coverage area of a LEO satellite changes continuously. The global coverage at any time is still possible if a certain number of orbits and satellites are used. As an example, the Iridium system uses 6 polar orbits with 11 satellites in each orbit [10,11]. Due to nonstationary coverage regions of individual satellites, the source or the destination terminals on the ground may not stay in the coverage region of the initial source or

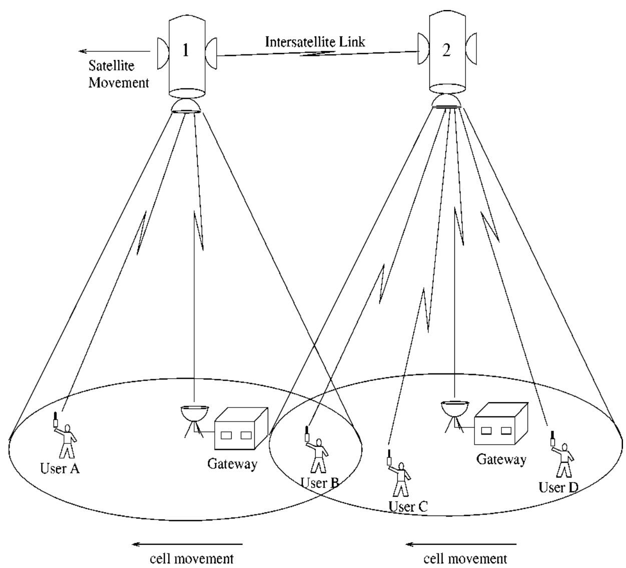


Figure 1. Wireless communication via LEO satellite network.

destination satellites throughout the communication. Thus, the initial source and the destination satellites may need to transfer the ground source and destination terminals to other satellites whose coverage regions contain the ground source and destination terminals. This event is called as *handover*. There are two types of intersatellite handovers in LEO systems: *intraorbit* and *interorbit* handovers. The former refers to handovers between neighbor satellites in the same orbit, while the latter refers to handovers between neighbor satellites in adjacent orbits. In LEO systems, handovers occur frequently because of the relatively high speed of the satellites (19,000–25,000 km/h) [8]. A handover may result in the addition of new satellites in the existing connection route. During a handover, the existing connection route should be updated accordingly.

When a call is accepted to the network, the connection route is determined using a certain *optimality* criterion such as minimum hop [2] or minimum cost [2]. The route established using such criterion is referred to as *optimal route*. Due to their high mobility, satellites are not used to determine the optimal route by themselves. Indeed, in the Iridium system [11], the connection routes are determined by the gateways as shown in figure 1. As an example, assume that user A wants to communicate with user D with a hand-held phone. The connection request is sent to the gateway via satellite 1 (source satellite), which covers user A. The gateway locates user D and computes the connection route. Then, the route information is sent to satellite 1, which handles the signaling to establish the connection. The resulting path, in figure 1, is a 2-hop route. However, in other cases, the route could have more than 2 hops. Routing using gateways causes signaling overhead between the satellites and the gateways. Thus, minimizing the number of routing operations performed by the gateways would enhance the performance of the network. During a handover, connection route should be modified to include the new satellite into the connection route. The route change can be achieved by augmenting the existing route with the new satellite or by completely rerouting the connection. Route augmentation is simple to implement, however the resulting route is not optimal. This may cause inefficient utilization of the satellite resources. The problem can be alleviated by completely rerouting the connection whenever a handover is necessary. Complete rerouting achieves optimal routes at the expense of signaling overhead and route establishment delay. The handover rerouting problem has not been adequately addressed in the context of satellite networks. In this paper, we propose a handover rerouting scheme referred to as *Footprint Handover Rerouting Protocol* (FHRP), which consists of two phases: route augmentation and rerouting. The motivation is to balance the optimality of the complete rerouting with the simplicity of the path augmentation. The FHRP addresses both intra- and interorbit handover problems. Moreover, the rerouting phase in the FHRP can be handled by satellites without any intervention of gateways. Therefore, the signaling and computation overhead can be reduced in the gateways.

The remainder of this paper is structured as follows: In section 2 we briefly review the state-of-the-art on satellite handover schemes. In section 3 we introduce the footprint handover rerouting protocol (FHRP). We also describe how the FHRP can handle intraorbit and interorbit handovers. In section 4 we investigate the performance of the algorithm. Finally, we discuss the future research directions and conclude the paper in section 5.

2. Related work

Satellite handover problems have become an active research area recently [5,6]. An analytical model has been proposed to calculate the handover rate for single-hop satellite connections in [6]. The model only considers intraorbit handovers. Due to the single-hop nature of the connections, no rerouting scheme is proposed. In a more recent study [5], interorbit handovers are addressed in a single-hop network environment. After developing call blocking probabilities for new calls and handovers, authors investigated a handover prioritization strategy based on the queueing of handovers. This study lacked the support for a multihop handover scheme. Although neglected in the existing literature, multihop connection routing is necessary in mobile satellite networks, since, even in the case of a connection between two users near to each other, the source and the destination terminals would be covered by different satellites; hence, necessitating at least two satellites for the connection.

The multihop satellite routing problem has been addressed in [15] with an emphasis on setting up routes between pairs of satellites to minimize the rerouting frequency, i.e., optimization was performed for the routes between two satellites. Realistically, the optimization is needed for the route between two ground terminals. An optimal route between two satellite nodes is not necessarily optimum for a connection between two ground terminals since the handovers between the ground terminals and the satellites result in changing satellite end nodes for the connection. The study did not address the handover rerouting problem.

The handover rerouting problem has been studied in the context of terrestrial wireless networks [1,7,13]. For example, a whole new route is established after a handover in [7]. Although an optimum route is used all the time, complete rerouting would cause handover call blocking because of excessive route re-establishment delay. Partial rerouting algorithms have been proposed in [1,13]. These algorithms basically make use of a tree-based structure for the network. During a handover, the node which is a parent of both nodes involved in the handover are determined, and the route between the parent and the original end node is replaced with a route between the parent and the new end node. Even though partial rerouting algorithms cause much less overhead in the network compared to the new route establishment, the route after the handover is not optimum.

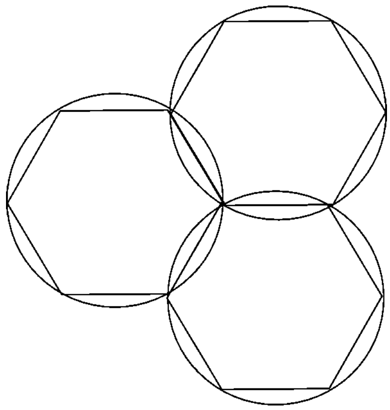


Figure 2. The footprints of the LEO satellites.

3. The Footprint Handover Rerouting Protocol (FHRP)

The service area, i.e., the *footprint*,[1] of a single satellite is a circular area on the Earth's surface in which the satellite can be seen under an elevation angle equal to or greater than the minimum elevation angle determined by the link budget requirement of the system. For a complete coverage of the Earth's surface, some overlapping between the footprints of adjacent satellites is necessary. The largest possible *effective footprint* of a satellite is then equivalent to the largest hexagon inscribed into the footprint as shown in figure 2.

In the LEO system described, the satellites are moving in O circular polar orbits. Each orbit has L satellites, and the total number of satellites is $N = O \cdot L$. The visibility period of a satellite, T_V, is defined as the maximum time duration that a ground terminal resides in the coverage region of a satellite and can directly communicate with that satellite. The visibility period of a typical LEO satellite is around 10 minutes. The period of an orbit T_O, on the other hand, is the minimum time interval required for the location of the satellites sharing a common orbit to repeat itself. If *Loc*(t) is a function that gives the location of the satellites at time t, then *Loc*$(t) =$ *Loc*$(t + T_\mathrm{O})$. If it is assumed that only one satellite is visible to a ground terminal (minimal coverage) at any time, it is trivial to show that the visibility period and the orbit period are identical, i.e., $T_\mathrm{O} = T_\mathrm{V}$. The case in which multiple satellites are visible to a ground terminal is also possible if more than the minimum number of satellites for global coverage is used. In this case, the user can pick any of the visible satellites subject to a certain objective such as the Signal-to-Noise Ratio (SNR) maximization or maximum time to a possible handover [3]. After the selection of a particular satellite, the handover problem is identical to the handover problem in a satellite network with minimal coverage. Thus, for the sake of presentation clarity, the satellite network discussed in this paper is assumed to provide minimal coverage, and $T_\mathrm{O} = T_\mathrm{V}$. The analytical relations between the orbit altitude, satellite speed, total number of orbits and satellites, and the visibility period can be found in [16].

[1] Service area, coverage area, and footprint are used interchangeably in this paper.

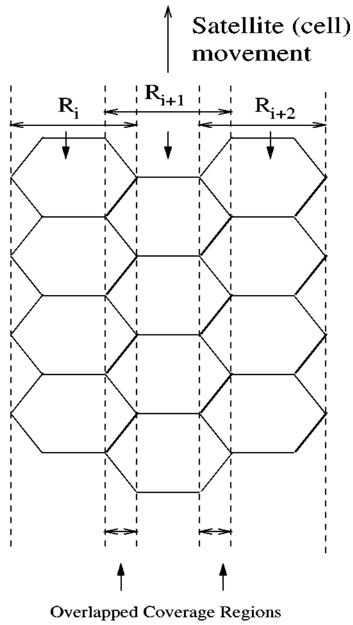


Figure 3. Overlapped coverage regions of adjacent orbits.

A single satellite in the ith polar orbit traces the coverage region of that orbit, R_i, as it circulates the Earth. In other words, all satellites in the ith polar orbit have exactly the same coverage region, R_i. But, at a given time, each satellite in the ith orbit handles traffic from a portion of R_i. In general, the coverage regions of two adjacent orbits may overlap with each other as shown in figure 3. Note that the overlapping coverage regions of adjacent orbits are different from the overlapping coverage of adjacent satellites. The former results from the movement of the satellites along their orbits while the latter is due to the circular footprints of individual satellites.

Each satellite has up and down wireless links for communication with ground terminals and I intersatellite links (ISL) for communication between satellites. The ISLs that connect the adjacent satellites in the same orbit are called *intraorbit* ISLs while the ISLs that connect adjacent satellites in neighbor orbits are called *interorbit* ISLs. Intraorbit ISLs are permanent while interorbit ISLs might be turned off temporarily when the satellites are crossing polar regions [16]. Moreover, left and right neighbor satellites of a satellite crossing polar regions switch their positions, i.e., the left neighbor becomes the right neighbor and vice versa. This results in a dynamic, but deterministic, network topology. Routing strategies that handle dynamic LEO satellite network topology have been investigated in [4,14,15]. Note that the routing problem and the rerouting protocol described in this paper are orthogonal to each other. Hence, for the clarity of the presentation, it is assumed that the LEO satellite network described in this paper has static topology and each satellite has 2 permanent intraorbit and 2 permanent interorbit ISL links as shown in figure 4. In

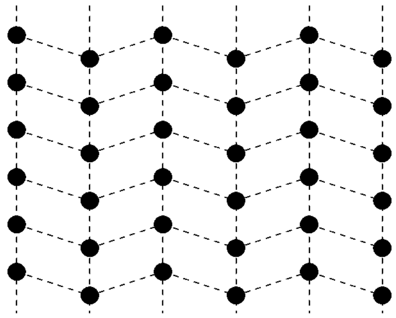


Figure 4. Intersatellite links between adjacent LEO satellites.

the Iridium system, the orbits 1 and 6 are counter-rotating, i.e., the satellites in the neighbor orbits 1 and 6 rotate in opposite directions. In [16], it is concluded that only ISLs between latitudes of approximately 60° north and south would be maintained between counter-rotating orbits in the Iridium system. The area with no interorbit ISLs is referred to as *seam*. Satellites going into seam switch their ISLs to the neighbor orbits off temporarily. Any connection passing through these links requires rerouting. The network depicted in figure 4 is a seamless one, but the protocol described in this paper handles the networks with a seam as explained later in this section.

In the remainder of this paper, it is assumed that a connection is routed through a number of LEO satellites for the source and the destination terminals on the ground. The route is denoted by $S_1 \to S_2 \to \cdots \to S_K$ where K is the number of satellites in the route. S_1 and S_K are referred to as the *source* and *destination* satellites, respectively. For the sake of clarity, S_1 and S_K are also labeled as S_{s} (source satellite) and S_{d} (destination satellite), respectively. The ordered set of satellites involved in the route is referred to as the *routing set,* $\mathcal{A}$, i.e., $\mathcal{A} = \{S_1, S_2, \ldots, S_K\}$. Assume that the connection is set-up at $t = t_{\mathrm{e}}$ using a routing algorithm such as minimum hop or minimum cost. If the route is optimum, we use the notation $\mathcal{A}_{\mathrm{opt}}$, $S_{\mathrm{s_{opt}}}$ and $S_{\mathrm{d_{opt}}}$ for $\mathcal{A}$, S_{s}, and S_{d}, respectively. The optimum routing set and the current routing set are identical if no handover is performed after the optimum route is established.

Given the limited visibility period and the high speed of the satellite, which is much faster than the velocities of mobile terminals, it is realistic to assume that ground terminals are stationary in this specific environment [5,6]. As an example, in the Iridium system, satellites travel with a speed of 26,000 km/h ($\approx$ 8 km/s). The diameter of the footprint of an Iridium satellite is approximately 4,000 km [9]. Thus, mobility of the user terminal, including the Earth's rotation, is negligible, and all handovers are caused by the mobility of the LEO satellite instead of the ground terminal. In the case of intraorbit handovers, the takeover satellite that assumes the responsibility for the ground terminal is referred to as the *successor* of the handover satellite. However, as shown in figure 3, if a ground terminal is located in the overlapping coverage regions of two adjacent orbits, the interorbit handover may occur. In this scenario, the takeover satellite is referred to as the *neighbor successor* of the handover satellite because these two satellites reside in the neighbor (adjacent) orbits. Also note that, due to the overlapping footprints of the adjacent satellites, the ground terminal would communicate with two satellites simultaneously when it is located in the overlapping region.

Definition. The rerouting of a connection passing through satellites $S_1 \to S_2 \to \cdots \to S_K$ to $S'_1 \to S'_2 \to \cdots \to S'_K$ where S'_i is the successor satellite of S_i is referred to as the *Footprint Rerouting* (FR).

Theorem 1. P is a multihop LEO satellite route established at time $t = t_{\mathrm{e}}$, and P' is the footprint reroute of P. Then:

(a) If P is a minimum hop route between S_{s} and S_{d}, then P' is a minimum hop route between S'_{s} and S'_{d}.

(b) If P is a shortest path route between S_{s} and S_{d} and the link cost is a function of the time-homogeneous traffic load, then P' is a shortest path route at time $t = t_{\mathrm{e}} + T_{\mathrm{O}}$.

(c) If P is a shortest path route between S_{s} and S_{d}, then P' is a shortest path route between S'_{s} and S'_{d} under the assumption that link cost is a function of time- and location-homogeneous traffic load.

Proof. The first part of the proof is trivial since there is a one-to-one correspondence between the nodes of the original and the footprint reroute. The proof of the second part is based on an imaginary terrestrial network where a one-to-one correspondence between each terrestrial node and a satellite switch exists. The relation between the satellites and the nodes of the imaginary terrestrial network is determined by two functions, f and g. The function f maps the satellites in the network to the ground points. Each ground point, s_i, in the mapping corresponds to the center of coverage of a unique satellite at time t, i.e., $f(S_i, t) = s_i$ for $i = 1, \ldots, N$, where S_i denotes the ith satellite of the system and N is the number of satellites in the system. The function g maps a ground point s to a satellite at time t if $f(S_i, t) = s$ for any i in the set $\{1, 2, \ldots, N\}$. The value of $g(s, t)$ is undefined if there is no satellite i that results in $f(S_i, t) = s$.

Next, we construct a ground network which has the same topology of the satellite network at time t_{e}. Specifically; if there exists a route between satellites S_i and S_j shown as $\overline{S_i S_j}$, then there exists a route between ground nodes s_i and s_j shown as $\overline{s_i s_j}$, where $s_i = f(S_i, t_{\mathrm{e}})$ and $s_j = f(S_j, t_{\mathrm{e}})$. Also, the cost function associated with the satellite and the ground routes are equal, i.e., $C(\overline{S_i S_j}) = C(\overline{s_i s_j})$, where $C(\cdot)$ is the cost function. Suppose the traffic load in location (x, y) at time t_{e} is $L(t_{\mathrm{e}}, (x, y))$. An optimal routing algorithm A can compute the optimal path P for a source

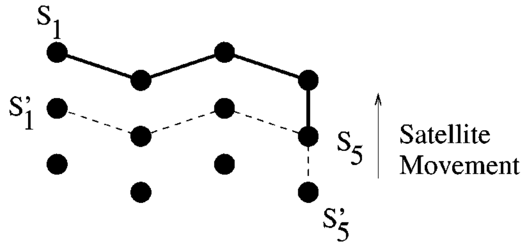


Figure 5. Example use of theorem 1(a).

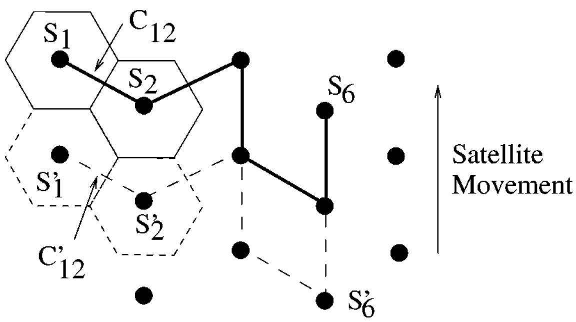


Figure 6. Example use of theorem 1(b).

and a destination pair, i.e., $A(L(t_e,(x,y))) = P$. For time-homogeneous traffic load, the time reference can be ignored. Thus, the path P is optimal along time in the ground network. On the other hand, $f(S_i,t_e) = s_i = f(S_i', t_e + T_O)$. Thus, the optimum route in the ground network corresponds to the satellite route P at time t_e and to the satellite route P' at time $t = t_e + T_O$, where P and P' correspond to the paths $S_1 \to S_2 \to \cdots \to S_K$ and $S_1' \to S_2' \to \cdots \to S_K'$, respectively.

The proof of part (c) is trivial since the shortest path route is also a minimum hop route if the traffic is time- and location-homogeneous, i.e., the cost of each ISL link is identical at any time. Thus, the route P' is also a shortest path route all the time. □

Example 1. Figure 5 shows a minimum hop route between satellites S_1 and S_5. Theorem 1(a) proposes that the route between S_1' and S_5', shown as dashed lines, is also a minimum hop route.

Example 2. Figure 6 shows a shortest path route between satellites S_1 and S_6. The cost of each link in the route, denoted as $C_{i,i+1}$ for $1 \leqslant i \leqslant 5$, depends on the traffic load in the footprints of the satellites i and $i+1$. Hence, the route between S_1' and S_6' denoted as dashed lines will be a shortest path route when the footprints of the satellites $S_1', \ldots, S_6'$ coincide with the footprints of the satellites $S_1, \ldots, S_6$. This event, which corresponds to Theorem 1B, occurs at $t = t_e + T_O$.

The goal of the Footprint Rerouting (FR) is to find an optimum route without performing the optimum route finding algorithm after a handover. Theorem 1 always guarantees optimality if the minimum cost algorithm is used to obtain the original route. On the other hand, the optimality is guaranteed at certain time instants if the original route is obtained using time-homogeneous traffic load. Clearly, time homogeneity assumption is not realistic for a global network. However, the change in the traffic load would be unnoticeable in a time interval comparable to the lifetime of a connection. The third part of theorem 1, which requires both time and locational homogeneity, is given only for completeness.

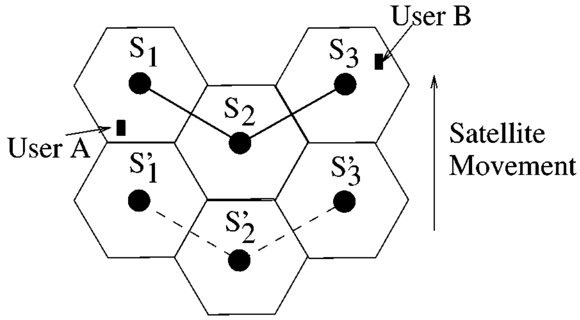


Figure 7. Application time of FR.

A handover is necessary when one of the end satellites, either the source or the destination satellite, goes out of the visibility region of the ground terminals involved in the communication. The connection should be transferred to either the successor or the neighbor successor satellite that is visible to the ground terminal that requires the handover. It is not possible to use the footprint rerouting at this instant since FR replaces both end satellites with their respective successors. In other words, FR is only possible when the new end satellites are the successors of the end satellites in the original route. This situation is depicted in the example shown in figure 7. The original route for users A and B is $S_1 \to S_2 \to S_3$. The FR of the original route is $S_1' \to S_2' \to S_3'$. User A is very close to the border of the footprint of S_1, and hence it is subject to a handover shortly. During the handover instant, FR is not applicable since user B is not yet located in the footprint of S_3'. Hence, there is a need for a mechanism to handle the routing problem until FR becomes applicable. We propose an augmentation algorithm to handle the handovers until FR becomes applicable. In the augmentation phase, a direct link is set up between the new end-satellite and the original route. If no such link exists, the connection is rerouted using the original routing algorithm. The FR algorithm is applied after both of the end nodes become the successors of the original end nodes. During a handover process, ground terminals decide whether the augmentation or FR should be used. The decision depends on the current time, the set-up time of the most recent optimum route, t_e, the routing set $\mathcal{A}_{\text{opt}}$ of the optimum route and the current routing set $\mathcal{A}$. The mobile terminals keep this information during the lifetime of their connections. The handover protocol, which consists of augmentation and FR phases, is referred to as the *Footprint Handover Rerouting Protocol* (FHRP) in the remainder of this paper.

3.1. Augmentation algorithm

In this section, the augmentation algorithm for the source node of a connection is described. It is shown later that the same algorithm is also applicable to the destination node. Assume that the most recent optimum route establishment has been performed at $t = t_e$. At $t = t_1 > t_e$, the ground terminal goes out of the coverage region of the source satellite, S_s. Since the global coverage is guaranteed, a new satellite which is denoted as S' covers the mobile terminal. S' is either the successor (intraorbit handover) or the neighbor successor (interorbit handover) of S_s as shown in figure 8. Assume that the necessary conditions as explained in the next section for FR is not held at this moment, and, thus, the source terminal decides to start the augmentation algorithm. A service request message including the current routing set, $\mathcal{A}$, is sent to S' to initiate the augmentation algorithm. The rest of the algorithm is handled by S' as follows:

1. The satellite S' checks whether the handover connection can be supported by its up-/downlinks. The connection would be blocked due to insufficient capacity in the uplink channels. Moreover, the connection blocking is also possible due to insufficient capacity in the downlink if the destination terminal is also in the coverage region of S'. If there is no sufficient bandwidth in the up and/or downlink(s), the connection is blocked and the source terminal releases the previous route.

2. If there is sufficient capacity in up- and/or downlinks, the new source satellite, S', first checks whether it is already serving the connection, i.e., it is checked whether S' is already in the routing set $\mathcal{A}$. If the result is positive, say $S' = S_i$, where $i = 2, 3, \ldots, K$, the portion of the route up to S' is deleted and the new route becomes $S_i \to S_{i+1} \to \cdots \to S_K$. The new routing set, $\mathcal{A}$, is sent to the ground terminals.

3. If S' is not in $\mathcal{A}$, a direct link to one of the satellites in $\mathcal{A}$ is searched starting with the last member (satellite with the largest index) of $\mathcal{A}$. This is because a link to a satellite with the largest index number results in a route with the shortest length. If a direct link with sufficient capacity to support the connection is found, the link is augmented to the original route. As an example, assume that a link between S' and S_i is found. Then, the new route is $S' \to S_i \to S_{i+1} \to \cdots \to S_K$. The unused portion of the previous route, $S_1 \to S_2 \to \cdots \to S_{i-1}$, is removed. The handover process is completed after the ground terminals are informed about the route changes.

4. If a direct link between S' and the nodes in $\mathcal{A}$ with required capacity is not found, the original routing algorithm is performed. If a route with required capacity is found, the resulting routing set, $\mathcal{A}_{\mathrm{opt}}$ and route establishment time are sent to the ground terminals.

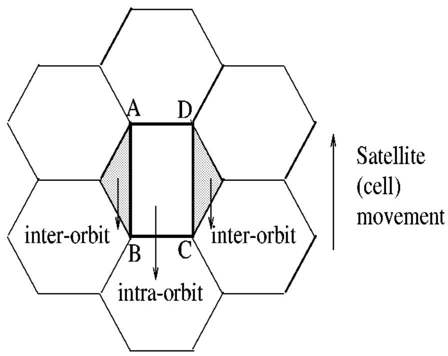


Figure 8. Inter- and intraorbit handover in augmentation algorithm.

Note that step 4 handles the handovers between satellites in counter-rotating orbits that was defined as the seam earlier in this section. Since no ISL is maintained between the satellites in counter-rotating orbits, step 4 enforces the connection to be completely rerouted.

The described augmentation algorithm can be applied to a handover involving the destination ground terminal with minor changes. To avoid simultaneous handovers of the source and the destination satellites of a connection, the destination terminal sends a handover request message to the source terminal. If the source terminal is in the process of a source handover, the destination's request is held until the source handover is completed. The source sends a handover permission message back to the destination that follows a similar augmentation process. Only difference is that, when a direct link is searched between S' and $\mathcal{A}$ in step 3, the satellite with the smallest index is checked first. Upon completion of the augmentation process, the new routing set is sent to the source terminal.

3.2. Footprint rerouting phase

The augmentation algorithm can be performed very fast, since only local changes are needed in the existing route. However, the resulting route is not guaranteed to be optimal. Thus, there is a need to update the route at certain time intervals. The selection of the update interval is important since the frequent rerouting attempts waste the network resources while a large rerouting interval results in the use of a non-optimal route for a long time. Here, we use theorem 1 to solve the routing update interval problem. Theorem 1 states that the optimality of the original route is preserved in the Footprint Reroute at certain times. The time when the Footprint Reroute is optimum would only be known by the end terminals and, thus, FR should be initiated by one of the end terminals. Here, we assume that FR is initiated by the source terminal. The route update time is based on the establishment time of the most recent optimum route t_e, optimum routing set $\mathcal{A}_{\mathrm{opt}}$, and the current routing set $\mathcal{A}$.

The first requirement to apply FR is that the source and the destination satellites of the current route are the successors of the source and the destination satellites of the original route, respectively, i.e., $S_d = S'_{d_{\mathrm{opt}}}$ and $S_s = S'_{s_{\mathrm{opt}}}$, where $S'_{d_{\mathrm{opt}}}$ and $S'_{s_{\mathrm{opt}}}$ are the successors of the source and the destination satellites of the optimum route. Second requirement depends on the nature of the traffic load of the system and the optimum routing criterion based on theorem 1 as follows:

1. If the original route is a minimum hop route, FR can be applied anytime when the successor end nodes are serving the connection, $S_{\rm d} = S'_{\rm d_{opt}}$ and $S_{\rm s} = S'_{\rm s_{opt}}$.
2. If the traffic load is time-homogeneous and the routing criterion depends on the traffic load (minimum cost routing), FR should be performed at $t = t_{\rm e} + t_{\rm O}$.

One of the above conditions is chosen based on the traffic and the routing criterion of the network and is applied for all connections. FR is initiated with a rerouting request, which is sent from the source terminal to the source satellite. The rerouting request includes the optimum route set $\mathcal{A}_{\rm opt}$. The source satellite tries to establish a connection traversing the optimal route. If the connection rerouting is performed successfully, i.e, no blocking occurs, the current route is removed. Upon the completion of the rerouting process, the source and the destination terminals update their routing information.

3.3. Storage requirements

The storage requirement of the route information does not introduce major overhead for the ground terminals, since the longest possible loop-free route in a seamless LEO network as shown in figure 4 is bounded by $N-1$, where N is the number of satellites in the network. This result is trivial since the longest loop-free route from the source to the destination passes through all the nodes in the network only once. The length of such a route is equal to $N-1$. When the minimum hop routing algorithm is used, the bound for the length of the route is smaller as proven in the following theorem.

Theorem 2. Assume that P is a loop-free route in a seamless LEO satellite network and *Length*(P) is the length of the route given in number of links. If P is a minimum hop route, then *Length*$(P) \leqslant \lfloor O/2 \rfloor + \lfloor L/2 \rfloor + 2$, where O is the number of the orbits and L is the number of satellites per orbit.

Proof. The proof is based on the connectivity structure of the network and the properties of the minimum hop routing algorithm. Assume that the orbits are indexed as R_i for $i = 1, \ldots, O$, where O is the number of orbits. The orbits R_i and R_{i+1} are adjacent to each other. Moreover R_1 and $R_{\rm O}$ are also neighbors due to the circular symmetry of the system. Thus, the maximum length of a route between two orbits is equal to $\lfloor O/2 \rfloor$. The satellites in each orbit are indexed similarly as S_i for $i = 1, \ldots, L$, where L is the number of satellites per orbit. Each satellite has direct links with its up/down and left/right neighbors. Thus, S_i can communicate with S_{i+1} with a direct link. Similar to the orbits, S_1 and S_L have a direct link between them due to the circular symmetry of the system. The maximum length of a route between two satellites sharing the same orbit is equal to $\lfloor L/2 \rfloor$. The length of the minimum hop route between two satellites in different orbits is equal to the sum of the maximum distance between the orbits and the maximum distance between the satellites sharing the same orbit, i.e., the length of the minimum hop routes is bounded by $\lfloor O/2 \rfloor + \lfloor L/2 \rfloor$. The constant term 2 in the theorem is due to the augmentation algorithm. As explained in section 3.1, the augmentation algorithm would extend the route at most 1 link. In the worst case, the augmentation algorithm is applied twice (one for each end terminal for interorbit handover). Thus, the length of the worst-case minimum hop route is bounded by $\lfloor O/2 \rfloor + \lfloor L/2 \rfloor + 2$. □

4. Performance evaluation

An event-driven LEO satellite network simulator has been written to evaluate the performance of the FHRP. The connections are voice calls. The simulation variables are the call arrival rate, call holding time, number of ground channels, and number of ISL channels. Both the call interarrival and call holding times are exponentially distributed. In particular, the average call holding time is set to 3 minutes for all experiments. The simulated LEO satellite network has 6 orbits and each orbit has 6 satellites as depicted in figure 4. The simulation time for each experiment is 200 minutes. The performance metrics are the total, new, and handover blocking probabilities. The total blocking probability is computed by dividing the number of blocked calls by the total number of calls during the simulation. The handover call blocking probability is the ratio of the number of blocked calls and the number of handovers, while the new call blocking is the ratio of the number of blocked new calls and the total number of new call arrivals. The handover blocking is more important than the new call blocking, because the interruption of a conversation is more annoying than blocking a new call. Blocking would occur due to insufficient ground or ISL channels. Since the scope of this paper is related to the multihop ISL routing, we intentionally increase the number of ground channels in the simulations so that most of the results presented below do not have any blocking due to insufficient ground channels.

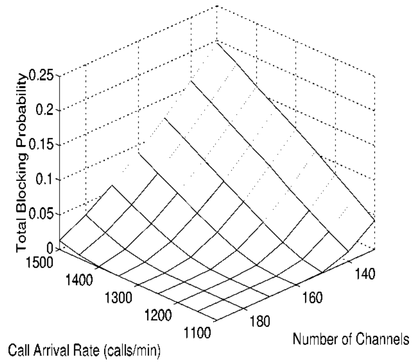
The Dijkstra algorithm [2] is used to find routes for new calls. The cost of a path is equal to the sum of the costs of all links in that path. Two different cost functions are tried in the simulations. In the first, the cost of each ISL link is equal to the number of busy channels. In the second, the cost of each ISL link is equal to one, and, thus, the resulting route corresponds to the minimum hop (minimum delay) route. Note that even in the minimum hop routing, the load on the ISL channel is considered so that the Dijkstra algorithm finds the minimum hop route that does not contain any congested ISL link. However, we found that the simulation results for different cost functions are very close in all cases. We present only the results of the minimum hop routing due to the space limitation.

The performance of the FHRP is compared with two other scenarios. In the first scenario, we have a *static* network with the same topology of the simulated LEO satellite network. The static network consists of switches that do not move. Thus, no handover is necessary. In a static network, the connection routes preserve their optimality at all times. In other words, it represents the best achievable blocking performance with the given switch capacity. In the second scenario, *pure augmentation* approach is used during the satellite handover. The pure augmentation approach results in handover blocking if the handover satellite is not on the existing route or an augmented link cannot be found between the handover satellite and the existing route. Thus, the difference between the performance of the pure augmentation and the FHRP shows the performance gain achieved by the rerouting, especially by the footprint rerouting.

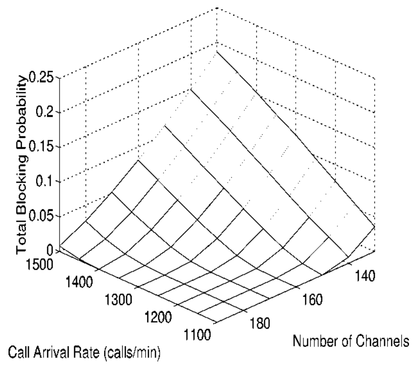
4.1. The case of homogeneous traffic

For homogeneous traffic, the new call arrivals are distributed uniformly in the coverage regions of the satellites. Specifically, the source and the destination satellites are generated uniformly among the satellites in the network. Also, within the footprints of these two satellites, the locations of the source and the destination mobile terminals are generated uniformly. In figure 9, the performances of the FHRP, the static network, and the pure augmentation scenarios are shown. The number of ground channels in these three figures is equal to the number of ISL channels. The number of channels shown in the figures is for each link (ISL or ground) of each satellite.

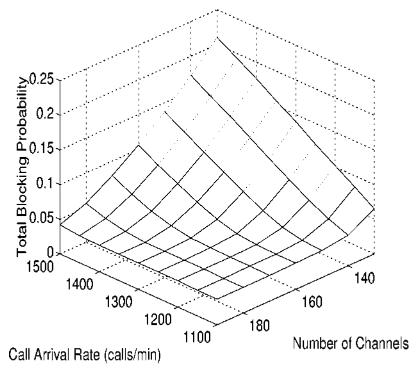
The static network scenario performs better than the FHRP and the pure augmentation. The blocking performance of the FHRP is very similar to that of the static network. The pure augmentation, on the other hand, performs poorly. Especially, for each call arrival rate, the total blocking probability of the pure augmentation approach decreases to around 5% as the number of ISL channels increases to 190. However, the FHRP and the static network scenario have no blocking at all as the number of ISL channels increases. This result shows that pure augmentation is not sufficient by itself to achieve low blocking probabilities. In all scenarios, the total blocking probability for small call arrival rates or large number of channels is almost constant. When the call arrival rate increases (or, the number of channels decreases), the performance starts degrading quickly. We also observed that the blocking due to ISL channel congestion contributes to the total blocking probability much more than the blocking due to ground channel congestion, although the numbers of ground channels and ISL channels are equal. In most cases, the ISL channel blocking contributes more than 90% of the total blocking. The reason for this behavior is because each connection has about 4.8 hops on average while only one uplink and one downlink are required for the ground channels. It is more probable for a multihop route being congested compared to a single hop route.



(a) The FHRP.



(b) Static network.



(c) Pure augmentation.

Figure 9. Total blocking probability for (a) the FHRP, (b) static network, and (c) pure augmentation scenario.

In the second set of experiments, the number of ground channels is kept constant at 190 so that no blocking due to ground channel congestion occurs. The number of ISL channels is equal to 150. Figure 10 shows the new call blocking and handover blocking performances of the FHRP and the pure augmentation as a function of the call arrival rate. The static network case has not been simulated, since no handover occurs in the static network. According to figure 10, the FHRP has a much smaller handover blocking than the pure augmentation. In the pure augmentation almost 4% of the handover calls are blocked. The new call blocking for the FHRP is higher than that of the pure augmentation, since the pure augmentation blocks more han-

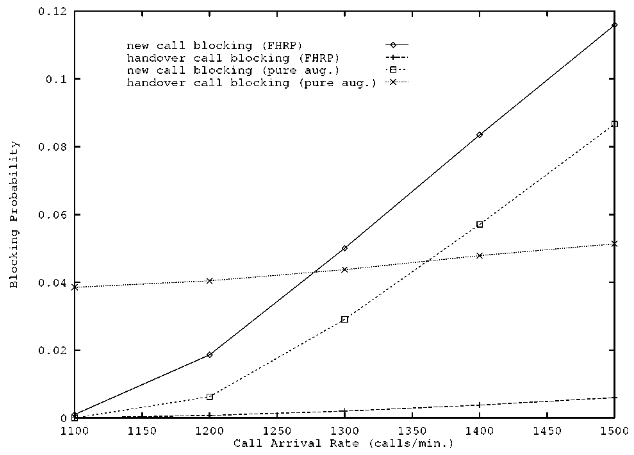


Figure 10. New call and handover blocking probabilities vs. call arrival rate for static, FHRP and pure augmentation networks.

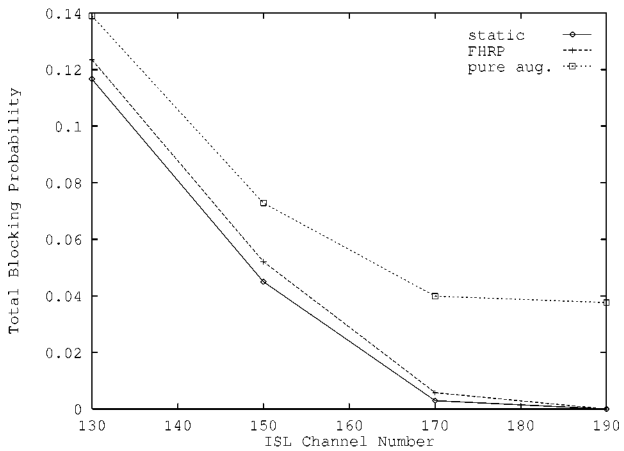


Figure 11. ISL blocking probabilities vs. number of ISL channels.

dover calls and, hence, can accept more new calls. On the other hand, the sum of the handover and the new call blocking probabilities is lower in the FHRP case.

In the last experiment for the homogeneous call arrival case, we kept the call arrival rate at 1300 calls/min and the number of ground channels at 190. The effect of increasing the number of ISL channels is investigated. As shown in figure 11, increasing the number of ISL channels helps to decrease the blocking probability. The FHRP and the static networks perform very similar and are superior to the pure augmentation. The results of the first three sets of experiments show that the FHRP performs very similar to a network without any mobility of the switches or the user terminals, i.e., the handovers do not degrade the performance of the system. On the other hand, pure augmentation results in higher call blocking compared to the other two. Especially, the handover call blocking is comparably high. Thus, it is clear that a handover algorithm that solely consists of route augmentation causes very high blocking probability. We repeated the experiments to see whether rerouting would help when the calls are blocked in pure augmentation. Thus, if a handover call is blocked, it is rerouted. The performance results for this modified pure augmentation algorithm were very similar to those for the FHRP. Thus, the FR phase of the FHRP can be replaced with a rerouting function as used in the call establishment phase. On the other hand, as explained in section 1, FHRP can be performed by the satellites while rerouting without FR needs to be done by the gateways. Furthermore, performing FR in the satellite can reduce the signaling and computation overhead in the gateways. According to our experiments, more than 90% of handovers only require FR instead of complete rerouting.

4.2. The case of heterogeneous traffic

In the heterogeneous traffic, the source satellite is still generated uniformly. However, we use a different approach to select destination satellite. First, two random numbers uniformly distributed between 0 and 1 are generated to determine the magnitude of the movement in the x and y axes from the source to the destination satellite. In the meantime, a probability vector P is used, where $P = (p_0, p_1, p_2, p_3, p_4, p_5, p_6) = (0, 0.1, 0.35, 0.6, 0.8, 0.9, 1)$. If the random numbers fall between p_i and p_{i+1}, then i is assigned to their associated magnitude. Since we are more interested in the multihop connection, p_1 is assigned to 0.1 so that the source and destination satellites will not be the same unless both random numbers are less than 0.1. After the magnitude is decided, two random numbers are generated to decide the direction of the movement. Once we know the direction and magnitude of the movement, the destination satellite can be obtained. Note that the movement in the y direction circulates on the orbit. For example, if the source satellite is 9 and the y movement is -4, then the resulting destination satellite is $9 \to 8 \to 7 \to 12 \to 11 = 11$ because satellites 7–12 are in the same orbit in the simulated network. However, the x movement circulates on different orbits. If the obtained destination satellite is not in the correct range,[2] then the same process repeats until a correct one is selected. Since there is no circulation in the x axis, the satellites located in the central orbits are subject to more load in their ground channels, especially, in the downlinks.

In figures 12 and 13, the number of the ground channels is chosen as 600 to avoid any ground blocking. Figure 12 shows that the static network still has the best blocking performance. The performance of the FHRP is almost identical to that of the static network. So, this experiment confirms that the relative performance of the FHRP is not affected by the heterogeneous traffic pattern. In fact, the relation among the performances of the FHRP and the static network is similar to that in the homogeneous traffic case. However, the blocking probability for the heterogeneous traffic is higher than that for the homogeneous traffic because the call arrival rate is 1100 calls/min in figure 12 while it is 1300 calls/min in figure 11. In figure 13, the

[2] For example, if the source satellite locates in the first orbit and the x movement is -2, then the resulting satellite will not be in the correct range.

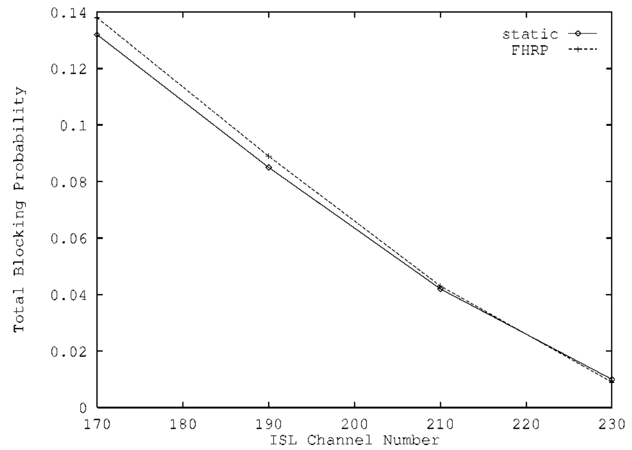


Figure 12. ISL blocking probabilities vs. number of ISL channels for heterogeneous case.

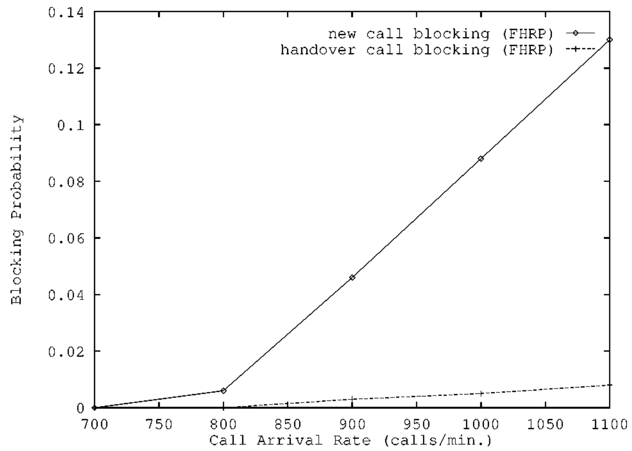


Figure 13. New call and handover blocking probabilities vs. call arrival rate for heterogeneous case.

ratio between the new call and handover blocking is similar to that in the case of homogeneous traffic. The number of ISL channels is 170 for this experiment. In conclusion, figures 12 and 13 confirm that the FHRP performs consistently in the cases of homogeneous and heterogeneous traffic.

5. Conclusions

A handover rerouting algorithm, referred to as Footprint Handover Rerouting Protocol (FHRP), has been proposed for LEO satellite networks. The FHRP is a hybrid algorithm that consists of the augmentation and the footprint rerouting phases. In the augmentation phase, a direct link from the new end satellite to the existing route is found. In case there is no such link with required capacity exists, a new route is found using the optimum routing algorithm. In the footprint rerouting phase of the FHRP, the connection is routed through footprint reroute determined by the original optimum path. The goal of the rerouting is to establish an optimum route without applying the optimum routing algorithm after a number of handovers. This property is significant because, in the ideal case, the routing algorithm computes a single route for each connection. As proven in section 3, the optimality of the original route is maintained during the communication. The performance of the FHRP is compared with a static network and pure augmentation. The results show that FHRP performs very similar to the static network and substantially better than the pure augmentation algorithm in terms of call blocking probability. Moreover, handover calls have less blocking compared to the new calls.

References

[1] B.A. Akyol and D.C. Cox, Rerouting for handoff in a wireless ATM networks, IEEE Personal Communications 3(5) (October 1996) 26–33.

[2] D. Bertsekas and R. Gallager, *Data Networks* (Prentice-Hall, Englewood Cliffs, NJ, 1992).

[3] P. Carter and M.A. Beach, Evaluation of handover mechanisms in shadowed low Earth orbit land mobile satellite systems, International Journal of Satellite Communications 13 (1995) 177–190.

[4] H.S. Chang, B.W. Kim, C.G. Lee, Y.H. Choi, S.L. Min, H.S. Yang and C.S. Kim, Topological design and routing for low-Earth orbit satellite networks, in: *Proc. of IEEE GLOBECOM* (1995) pp. 529–535.

[5] E. Del Re, R. Fantacci and G. Giambene, Handover requests queueing in low Earth orbit mobile satellite systems, in: *Proc. of the 2nd European Workshop on Mobile/Personal Satcoms* (1996) pp. 213–232.

[6] F. Dosiere, T. Zein, G. Maral and J.P. Boutes, A model for the handover traffic in low Earth-orbiting (LEO) satellite networks for personal communications, International Journal of Satellite Communications 11 (1993) 145–149.

[7] K.Y. Eng, M.J. Karol, M. Veeraraghavan, E. Ayanoglu, C.B. Woodworth, P. Pancha and R.A. Valenzuela, A wireless broadband ad-hoc ATM local area network, Wireless Networks 1(2) (1995) 161–174.

[8] A. Ganz, Y. Gong and B. Li, Performance study of low Earth-orbit satellite systems, IEEE Transactions on Communications 42(2/3/4) (February/March/April 1994) 1866–1871.

[9] P.R. Giusto and G. Quaglione, Technical alternatives for satellite mobile networks, in: *Proc. of the 1st European Workshop on Mobile/Personal Satcoms* (1994) pp. 15–27.

[10] J.L. Grubb, IRIDIUM overview, IEEE Communications Magazine 29(11) (1991).

[11] Y.C. Hubbel, A comparison of the IRIDIUM and AMPS systems, IEEE Network Magazine 11(2) (March/April 1997) 52–59.

[12] M.A. Sturza, Architecture of the TELEDESIC satellite system, in: *Proc. of Int. Mobile Satellite Conference* (1995) pp. 212–218.

[13] C. Toh, The design and implementation of a hybrid handover protocol for multimedia wireless LANs, in: *Proc. of MOBICOM '95* (1995) pp. 49–61.

[14] H. Uzunalioğlu, Probabilistic routing protocol for low Earth orbit satellite networks, in: *IEEE ICC '98* (1998) pp. 88–93.

[15] M. Werner, C. Delucchi, H.-J. Vogel, G. Maral and J.-J. De Ridder, ATM-based routing in LEO/MEO satellite networks with intersatellite links, IEEE Journal on Selected Areas in Communications 15(1) (January 1997) 69–82.

[16] M. Werner, A. Jahn, E. Lutz and A. Bottcher, Analysis of system parameters for LEO/ICO-satellite communication networks, IEEE Journal on Selected Areas in Communications 13(2) (February 1995) 371–381.

[17] R.A. Wiedeman and A.J. Viterbi, The Globalstar mobile satellite system for worldwide personal communications, in: *Proc. of Int. Mobile Satellite Conference* (1993) pp. 46–49.

Hüseyin Uzunalioğlu received his B.S. and M.S. degrees from the Middle East Technical University, Ankara, Turkey, in 1990 and 1992, respectively, and the Ph.D. degree from the Georgia Institute of Technology in 1998, all in electrical engineering. From 1992 to 1996 he was a recipient of a research scholarship from Middle East Technical University. He worked as a researcher in Georgia Tech Research Institute from 1995 to 1998. He is currently with Bell Labs Advanced Technologies, Lucent Technologies. His current research interests are in traffic and mobility management for broadband networks. He is a member of IEEE.
E-mail: huseyin@lucent.com

Ian F. Akyildiz received his BS, MS, and PhD degrees in computer engineering from the University of Erlangen-Nuernberg, Germany, in 1978, 1981 and 1984, respectively. Currently, he is a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology and Director of Broadband and Wireless Networking Laboratory. He has held visiting professorships at the Universidad Tecnica Federico Santa Maria, Chile, Universite Pierre et Marie Curie (Paris VI), Ecole Nationale Superieure Telecommunications in Paris, France, and Universidad Politecnico de Cataluna in Barcelona, Spain. He has published over two hundred technical papers in journals and conference proceedings. He is an editor for IEEE/ACM Transactions on Networking, Computer Networks Journal, ACM–Springer Journal for Multimedia Systems, ACM/Baltzer/URSI Journal Wireless Networks, and Cluster Computing Journal. He is a past editor for IEEE Transactions on Computers (1992–1996). He was the program chair of the "9th IEEE Computer Communications" workshop in 1994. He also served as the program chair for ACM/IEEE MOBICOM '96 (Mobile Computing and Networking) conference and for IEEE INFOCOM '98. Dr. Akyildiz is an IEEE Fellow and an ACM Fellow. He served as a National Lecturer for ACM 1989–1998. He received the ACM Outstanding Distinguished Lecturer Award for 1994. Dr. Akyildiz also received the IEEE Leonard G. Abraham Prize award (best original paper published in the IEEE Journal of Selected Areas in Communications in 1996). His current research interests are in satellite networks, wireless networks, Internet and multimedia communication systems.
E-mail: ian@ee.gatech.edu
WWW: http://www.ee.gatech.edu/research/labs/bwn

Yelena Yesha received the B.Sc. degree in computer science from York University, Toronto, Canada, in 1984, and the M.Sc. and Ph.D. degrees in computer and information science from The Ohio State University in 1986 and 1989, respectively. Since 1989 she has been with the Department of Computer Science and Electrical Engineering at the University of Maryland at Baltimore County, where she is presently a Professor. In addition, since December, 1994, Dr. Yesha is serving as the Director of the Center of Excellence in Space Data and Information Sciences at NASA. Her research interests are in the areas of distributed databases, distributed systems, electronic commerce, and trusted information systems. She edited 6 books and authored over 70 refereed articles in these areas. Dr. Yesha was a program chair and general co-chair of the ACM International Conference on Information and Knowledge Management and member of the program committees of many prestigious conferences. She is a member of the editorial board of the Very Large Databases Journal and editor-in-chief of the International Journal of Digital Libraries. During 1994, Dr. Yesha was the Director of the Center for Applied Information Technology at the National Institute of Standards and Technology. She is presently serving as a member of the U.S. delegation to the G7 program on global market place for small and medium enterprizes. Dr. Yesha has extensive experience in consulting for leading computer corporations. Dr. Yesha is a senior member of IEEE, a member of the New York Academy of Science, and a member of ACM.
E-mail: yeyesha@cs.umbc.edu

Wei Yen was born in Kaohsiung, Taiwan. He received his B.S. degree from Tatung Institute of Technology, Taiwan, in 1989, and his M.S. and Ph.D. degrees both in electrical and computer engineering from Georgia Institute of Technology in 1993 and 1997, respectively. He is currently with Mitsubishi Telecommunications Network Division as a Senior Product Engineer. His current research interests include communication protocol design and analysis for multimedia applications, multicast routing, WDM and ATM networks. Wei Yen is a co-winner of the 1997 Leonard G. Abraham Prize Paper Award and a member of IEEE.
E-mail: wyen@mtnd.com