A survey of routing techniques for mobile communications networks

S. Ramanathan and Martha Steenstrup

Advanced Networking Department, Systems and Technologies Division, BBN Corporation, Cambridge, MA 02138, USA

Abstract. Mobile wireless networks pose interesting challenges for routing system design. To produce feasible routes in a mobile wireless network, a routing system must be able to accommodate roving users, changing network topology, and fluctuating link quality. We discuss the impact of node mobility and wireless communication on routing system design, and we survey the set of techniques employed in or proposed for routing in mobile wireless networks.

1. Introduction

Mobile wireless networking has enjoyed a dramatic increase in popularity over the last few years, although the technology has existed for more than twenty years and has been commercially available for more than ten years. This recent popularity can be attributed to two factors: (1) advances in hardware design resulting in affordable, portable, low-power, wireless communication and computation devices; and (2) rapid growth in the communications infrastructure resulting in ubiquitous and affordable access to telephone and data networks. Users accustomed to the services available from stationary wireline networks expect to receive the same services from mobile wireless networks, including the ability to execute multimedia applications. Meeting these expectations is a challenging problem. Some services may not be available between all pairs of locations in the network, and hence it may not be possible to provide consistent service for a traffic session as its endpoints move around the network. Moreover, frequent changes in network topology and wireless link quality may make it difficult to maintain a particular level of service, even between two fixed endpoints in the network.

This issue of MONET is devoted to routing, a crucial component of the solution to the problem of providing high-quality communication in mobile wireless networks. We offer this survey of routing systems applicable to such networks, in order to acquaint the reader with the breadth of techniques already available as well as with the open research issues. Included in this survey is an extensive bibliography for further reading.

A mobile network is any network in which some of the constituent nodes (endpoints and switches) change location relative to each other. Endpoints may move among stationary switches (e.g., terrestrial cellular telecommunications networks), switches may move over stationary endpoints (e.g., satellite networks), or switches and endpoints may both move (e.g., mobile packet radio networks). Ideally, the routing system should be able to manage node mobility such that two communicating endpoints are unaware of any mobility in each other or in the network itself.

We define a routing system as not one but a set of several component functions including the following: monitoring network topology and services; distributing this information for use in route construction; locating session endpoints; constructing and selecting routes; and forwarding traffic according to the selected routes. The routing system is responsible for producing routes that meet the service requirements of the traffic sessions within the service constraints imposed by the network. Changes in network or session state may necessitate changes to existing routes in order to maintain their feasibility, and such changes are likely to occur more frequently in mobile wireless networks than in stationary wireline networks. The routing system must be able to quickly detect and respond to such state changes, in order to minimize degradation in services provided to existing traffic sessions, but it must do so using a minimal account of network resources, in order to maximize network throughput.

When tracking and adapting to changes in the location of a mobile node, the routing system consumes network resources in direct proportion to the frequency and speed of the node's movements. Under some circumstances, a node may change locations more rapidly than the routing system can react to such changes, regardless of the amount of network resources available. This situation can occur when an endpoint is moving with high velocity with respect to its network attachment points, either because it is moving with high absolute velocity (e.g., an airplane moving with respect to terrestrial base stations) or because the distance between successive network attachment points is short (e.g., a car moving among microcells). To accommodate highly mobile nodes while consuming a minimal amount of network resources, the routing system must be capable of predicting future node locations, in addition to reacting to current node movements. Prediction is feasible if node trajectory information is available a priori or from a sequence of previously observed node locations. When

accurate trajectory information is not available, the routing system can resort to other means such as distributing the session's traffic to multiple locations within the predicted vicinity of the destination endpoint; in the extreme case, this degenerates into flooding the traffic throughout the network. Use of multiple paths increases the likelihood that the session traffic reaches its intended destination but expends more network resources in the process. In designing a routing system for a particular mobile network, one must compare the costs and benefits of location tracking mechanisms versus those of traffic delivery mechanisms to determine the most suitable combinations.

With the exception of portable computers that attach directly to a wireline infrastructure, most mobile nodes require wireless links to communicate with the rest of a network. Wireless links tend to have lower and less predictable quality than wireless links. Moreover, their quality is highly sensitive to environmental conditions including: the distance between the link's endpoints, resulting in signal attenuation; the terrain between the link's endpoints, resulting in multipath signal propagation or even signal obstruction; externally generated noise, resulting in corrupted transmissions; and interference among multiple transmissions in a particular vicinity, resulting in lost transmissions. While many of these effects can be mitigated by lower-level solutions such as transmission power adjustments, error correction techniques, link-level retransmissions, and efficient channel access procedures, they cannot be eliminated entirely. Thus, it is the responsibility of the routing system to route traffic so as not to create traffic patterns that contribute to the degradation of link quality and so as to minimize the number of low-quality links traversed.

In the following sections, we survey the wide variety of techniques for enabling networks to deal effectively with mobile nodes and wireless links in the context of routing. We address in detail the topics of mobile network organization, node locating tracking, and route selection and traffic forwarding.

2. Network organization

A network consists of *nodes* and the *links* connecting them. Nodes can be functionally classified into two types: *endpoints* (also referred to as *hosts* or *terminals*) that act as sources and sinks of traffic, and *switches* (or *routers*) that forward traffic towards its destination. Depending upon whether or not the endpoints and switches are mobile, we have the four types of networks depicted in Table 1.

The differing requirements of these types of networks give rise to different techniques for network organization. In the remainder of this section, we discuss techniques used for organizing terrestrial cellular networks, packet radio (or ad hoc) networks, and satellite networks.

Table 1

	Stationary endpoints	Mobile endpoints
Stationary switches	Wireline	Cellular
Mobile switches	Satellite	Pkt. radio / ad hoc

2.1. Cellular networks

In cellular networks, the switches are stationary while the endpoints may be mobile and wireless-equipped. Some of the switches may be equipped with both wireline and wireless capabilities and are referred to as *base stations*. The mobile endpoints are partitioned into *cells*, each of which has an associated base station (which may be responsible for one or more cells) used by the endpoints of that cell to connect to the fixed portion of the network. A cellular organization allows frequency reuse among geographically distant cells [1,2], thereby increasing system capacity.

The first widespread realization of a cellular network was developed at Bell Laboratories [2] and resulted in the Advanced Mobile Phone Service (AMPS) [3] supporting analog voice channels in a terrestrial wireless mobile environment. Since then, cellular communications have evolved in three areas: digital microcellular networks for Personal Communications Services (PCS) [4] and related standards (e.g., GSM [26], IS95 [5]); cordless telephony and related standards (e.g., DECT [6], CT2Plus [7]); and paging (e.g., GSC, FLEX [1]). For an overview of the various cellular systems and standards, consult [1,8].

Organization of cellular networks gives rise to two network design problems. The first problem is how to partition the coverage area into cells and involves determining the size of the radius of each cell, the location of the base stations, and the amount of overlap required between cells. Furthermore, as the system grows, cells may need to be subdivided to accommodate additional endpoints. Approaches to such cell *splitting* are addressed in [2]. Microcellular networks [4], as embodied in PCS, have a cell radius of 50 metres to 500 metres and represent a design choice which provides maximal system capacity at the cost of increased handoff rates. A *handoff* (or *handover*) constitutes a change in the cell or base station with which an endpoint is affiliated, resulting from movement of that endpoint. The second problem is how to assign cells to Mobile Switching Centers (MSCs) attached to the fixed portion of the network, or more precisely, how to associate base stations with switching centers. An integer programming formulation of this problem, which considers the cost of handoffs between cells and the cost of cabling between a base sta-

tion and its associated switching center, is given in [9] and a heuristic solution is proposed.

2.2. Packet radio networks

In contrast to urban civilian communications, tactical military communications require survivable adaptive networking and rapid deployment in a variety of potentially hostile environments. In these situations, switch mobility is an important advantage. A network of mobile, untethered switches that employ radio communications is commonly referred to as a packet radio network. Examples include the networks developed for the DARPA PRNET [10,11] and SURAN [12] programs. For an overview of packet radio networks and design issues therein, consult [13,14]. Recently, there has been a growing interest in civilian networks of mobile switches (or mobile hosts with the ability to perform switching functions [15]). Such networks are referred to as *ad hoc* networks, a term adopted by the IEEE 802.11 subcommittee [16], and are conceptually identical to packet radio networks.

Mobility of switches raises organizational problems quite different and rather more challenging than those for cellular networks. In particular, rapid response to switch movement requires autonomous organization mechanisms involving minimal manual intervention. The primary design problem in packet radio networks is that of *clustering* the mobile switches into groups, and is motivated principally by two considerations: controlling channel reuse spatially (in terms of frequency, time, or spreading code) and reducing routing information overhead.

Several clustering techniques have been described for channel control. A local clustering algorithm using *clusterheads* for channel access control is described in [17]. In this distributed algorithm, each node has a distinct numerical identifier. The node with the lowest-numbered identifier in a locality is elected as the clusterhead and may then act as a local controller for channel access. The algorithm described in [18] selects as clusterhead the node with the highest number of neighbors. In [19], a distributed algorithm for forming clusters without clusterheads is given. Here, every cluster has a diameter of at most two hops. Clustering works by forming the first cluster around a highest degree node, the next cluster around a highest degree node in the unclustered part of the network, and so on until all network nodes are in clusters.

Clustering techniques have also been developed, mostly in the context of the DARPA packet radio programs [20,14], to impose hierarchy on large packet radio networks for routing scalability. Hierarchical clustering in packet radio networks is motivated by the fact that the amount of information required to keep each node up-to-date with respect to changes in network topology is proportional to least $N \cdot r$, where N is the number of nodes and r is the rate of topological change. Since r is itself proportional to N, routing information grows as N^2. Network capacity increases only linearly with N, and hence for large N, "flat" routing is not practical. The clustering technique adopted for SURAN allows for arbitrary levels of hierarchy and for cluster reformation in response to mobility. Specifically, cluster formation has two steps: a clusterhead is elected, and nodes join that cluster. A node joins a new cluster if it does not currently belong to any cluster or if the new clusterhead is closer than its current clusterhead. When a cluster grows too large, the node with the lowest-numbered identifier in that cluster elects itself as a new clusterhead. When a cluster becomes too small, the clusterhead resigns and the nodes join other clusters. Special mechanisms (see [14]) handle temporary disruptions in communications caused by cluster (re)formation.

In packet radio networks, typically all of the nodes are switches. As network heterogeneity increases, however, it is likely that one might need to interconnect a mixture of mobile switches and endpoints. Such a network might benefit from a hybrid organization, including features of both packet radio and cellular networks. At the lowest level, endpoints could group themselves into cells with one active switch (or *cellhead*) per cell, each of which would function like a base station in a cellular network, except that it would be dynamically assigned. The cellheads, in turn, would form a multihop packet radio network which could be further structured hierarchically if necessary.

2.3. Satellite networks

In satellite networks, the switching functions are performed by earth-orbiting satellites. Links may be ground-to-satellite or inter-satellite. Some satellite networks consist of a small number of satellites in geostationary orbits and provide extensive ground coverage (except in polar regions), such that a large number of users can be reached over a single hop. Depending on the network, users not covered by the same satellite may communicate through multiple ground-to-satellite hops or may use inter-satellite links. Low-Earth Orbit (LEO) satellite networks consisting of constellations of many satellites are now becoming a commercially viable means for providing global communications (e.g., the Iridium system [21]). Although they have dynamic topologies (i.e., as in a packet radio network, switches move relative to each other and to endpoints), these networks do not require adaptive topology tracking mechanisms. The satellites move in regular, deterministic patterns relative to each other and to the ground, and hence, the network topology (excluding satellite failures) is completely predictable at any time.

Issues related to the organization of satellite networks are somewhat similar to those for cellular systems in that the organization is more or less preassigned. The

problem is to design the best topological layout, including the number of satellites comprising the network; their positions, orbits, grazing angles, and coverage areas; and the inter-satellite connectivity. Satellite coverage areas may be fixed with respect to the satellites as in the Iridium system [21,22] or may be fixed with respect to the earth as in [23]. The Iridium system comprises 66 satellites using polar orbits and grazing angles of 10° to 15°. In contrast, the system proposed in [24] is based on symmetrical orbits over the oblate globe modelled as a 12-facet polyhedron. This system uses higher orbits resulting in a grazing angle of 45° and higher message delays, but reduces crowding in the polar regions and requires a smaller number of satellites.

3. Location tracking

A network must retain information about the locations of endpoints in the network, in order to route traffic to the correct destinations. *Location tracking* (also referred to as *mobility tracking* or *mobility management*) is the set of mechanisms by which location information is updated in response to endpoint mobility. In location tracking, it is important to differentiate between the *identifier* of an endpoint (i.e., what the endpoint is called) and its *address* (i.e., where the endpoint is located). Mechanisms for location tracking provide a time-varying mapping between the identifier and the address of each endpoint.

3.1. Generic issues

Most location tracking mechanisms may be perceived as updating and querying a distributed database (the *location database*) of endpoint identifier-to-address mappings. In this context, location tracking has two components: (1) determining when and how a change in a location database entry should be initiated; and (2) organizing and maintaining the location database. We discuss each of these issues below [#1].

3.1.1. Updating the location database

In cellular networks, endpoint mobility within a cell is transparent to the network, and hence location tracking is only required when an endpoint moves from one cell to another. Location tracking typically consists of two operations: (1) *updating* (or *registration*), the process by which a mobile endpoint initiates a change in the location database according to its new location; and (2) *finding* (or *paging*), the process by which the network initiates a query for an endpoint's location (which may also result in an update to the location database). Most location tracking techniques use a combination of updating and finding in an effort to select the best tradeoff between update overhead and delay incurred in finding. Specifically, updates are not usually sent every time an endpoint enters a new cell, but rather are sent according to a pre-defined strategy such that the finding operation can be restricted to a specific area. There is also a tradeoff, analyzed formally in [25], between the update and paging costs. Fig. 1 illustrates a classification of possible update strategies which are discussed in more detail below.

Static Strategies. In a static update strategy, there is a predetermined set of cells at which location updates may be generated. Whatever the nature of mobility of an endpoint, location updates may only be generated when, but not necessarily every time, the endpoint enters one of these cells. Two approaches to static updating are as follows.

1. *Location areas* (also referred to as *paging* or *registration areas*) [8]. In this approach, the service area is partitioned into groups of cells with each group as a location area. An endpoint's position is updated if and only if the endpoint changes location areas. When an endpoint needs to be located, paging is done over the most recent location area visited by the endpoint. Location tracking in many second-generation cellular systems, including GSM [26] and IS-41 [27], is based on location areas [28]. Several strategies for location area planning in a city environment are evaluated in [29]. These include strategies that take into account geographical criteria (such as population distribution and highway topology) and user mobility characteristics.
2. *Reporting cells* (or *reporting centers*) [30]. In this approach, a subset of the cells are designated as the only ones from which an endpoint's location may be updated. When an endpoint needs to be located, a search is conducted in the vicinity of the reporting cell from which the most recent update was generated. In [30], the problem of which cells should be designated as reporting cells so as to optimize a cost function is addressed for various cell topologies.

The principal drawback of static update strategies is

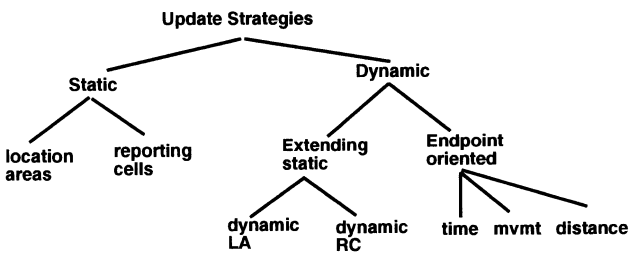

Fig. 1. Classification of update strategies.

[#1] Most of the location tracking literature has been motivated by cellular networks, in particular emerging PCS networks. Thus, the treatment in this section is necessarily of a cellular flavor.

that they do not accurately account for user mobility and frequency of incoming calls. For example, even though a mobile endpoint may remain within a small area, it may cause frequent location updates if that area happens to contain a reporting cell.

Dynamic Strategies. In a dynamic update strategy, an endpoint determines when an update should be generated, based on its movement. Thus, an update may be generated in any cell. A natural approach to dynamic strategies is to extend the static strategies to incorporate call and mobility patterns. The dynamic location area strategy proposed in [31] dynamically determines the size of an endpoint's location area according to the endpoint's incoming call arrival rate and mobility. Analytical results presented in [31] indicate that this strategy is an improvement over static strategies when call arrival rates are user- or time-dependent. The dynamic reporting centers strategy proposed in [32] uses easily-obtainable information to customize the choice of the next set of reporting cells at the time of each location update. In particular, the strategy uses information recorded at the time of the endpoint's last location update, including the direction of motion, to construct an asymmetric distance-based cell boundary and to optimize the cell search order.

In [33], three dynamic strategies are described in which an endpoint generates a location update: (1) every T seconds (*time*-based); (2) after every M cell crossings (*movement*-based); or (3) whenever the distance covered (in terms of number of cells) exceeds D (*distance*-based). Distance-based strategies are inherently the most difficult to implement since the mobile endpoints need information about the topology of the cellular network. It was shown in [33], however, that for memoryless movement patterns on a ring topology, distance-based updating outperforms both time-based and movement-based updating. In [25], a set of dynamic programming equations is derived and used to determine an optimal updating policy for each endpoint, and this optimal policy is in fact distance-based.

Strategies that minimize location tracking costs under specified delay constraints (i.e., the time required to locate an endpoint) have also been proposed. In [34], a paging procedure is described that minimizes the mean number of locations polled with a constraint on polling delay, given a probability distribution for endpoint locations. A distance-based update scheme and a complementary paging scheme that guarantee a pre-defined maximum delay on locating an endpoint are described in [35]. This scheme uses an iterative algorithm to determine the optimal update distance D that results in minimum cost within the delay bound.

3.1.2. Organizing the location database

In organizing the location database, one seeks to minimize both the *latency* and the *overhead*, in terms of the amount of storage and the number of messages required, in accessing location information. These are, in general, counteracting optimization criteria. Most solutions to the location database organization problem select a point which is a three-way tradeoff between overhead, latency, and simplicity. The simplest approach to location database organization is to store all endpoint identifier-to-address mappings in a single central place. For large numbers of reasonably mobile endpoints, however, this approach becomes infeasible in terms of database access time and storage space and also represents a single point of failure.

The next logical step in location database organization is to partition the network into a number of smaller pieces and place a portion of the location database in each piece. Such a distributed approach is well-suited to systems where each subscriber is registered in (and pays bills to) a particular area or "home". With this organization, the location database in an area contains the locations of all endpoints whose home is that area. When the endpoint moves out of its home area, it updates its home location database to reflect the new location. The Home Location Register (HLR) and Visitor Location Register (VLR) schemes of emerging PCS networks [28] are an example of this approach, as are the Mobile IP scheme [36] for the Internet and the Cellular Digital Packet Data (CDPD) scheme [37] for data transport over cellular networks. Studies [38,39] have shown that with predicted levels of PCS users, signalling traffic may exceed acceptable levels. Thus, researchers have considered augmenting this basic scheme to increase its efficiency under certain circumstances. For instance, in [40], *per-user caching* is used to reuse location information about a called user for subsequent calls to that user, and is particularly beneficial for users with high *call-to-mobility ratios* (i.e., the frequency of incoming calls is much larger than the frequency of location updates). In [41], "local anchoring" is used to reduce the message overhead by reporting location changes to a nearby VLR instead of to the HLR, thus increasing the location tracking efficiency when the call-to-mobility ratio is low and the update cost is high.

As with most large organizational problems, a *hierarchical* approach provides the most general and scalable solution. By hierarchically organizing the location database, one can exploit the fact that many movements are local. Specifically, by confining location update propagation to the lowest level (in the hierarchy) containing the moving endpoint, costs can be made proportional to the distance moved. Several papers address this basic theme. In [42], a hierarchy of *regional directories* is prescribed, where each regional directory is based on a decomposition of the network into regions. Here, the purpose of the ith-level regional directory is to enable tracking of any user residing within a distance of 2^i. This strategy guarantees overheads that are polylogarithmic in the size and diameter of the network. In [43], the loca-

tion database is organized so as to minimize the total rate of accesses and updates. This approach takes into account estimates of mobility and calling rates between cells and a budget on access and update rates at each database site. In [44], location database organization takes into account the *user profile* of an endpoint (i.e., the predefined pattern of movement for the endpoint). Partitions of the location database are obtained by grouping the locations among which the endpoint moves frequently and by separating those to which the endpoint relocated infrequently. Each partition is further partitioned in a recursive fashion, along the same lines, to obtain a location database hierarchy.

In the above strategies, the emphasis is on reducing update overhead, but it is equally important to reduce database access latency. One strategy for doing so is *replication*, where identical copies of the database are kept in various parts of the network so that an endpoint location may be obtained using a low-latency query to a nearby server. The problem here is to decide where to store the replications. This is similar to the classical database allocation [45] and file allocation [46] problems, in which databases or files are replicated at sites based on query-update or read-write access patterns. In [47], the best zones for replication are chosen per endpoint location entry, using a minimum-cost maximum-flow algorithm to decide where to replicate the database, based on the calling and mobility patterns for that endpoint.

3.2. Location tracking in PCS

There are two PCS standards for location tracking: the North American standard IS-41 [27,28] and the European standard GSM [26,28]. Both employ a partitioning of the service area into location areas, and both are based on a two-level hierarchy. When a user subscribes to a PCS service provider, an entry is created in the Home Location Register (HLR). When the user moves to a new location area, a temporary record is created in the Visitor Location Register (VLR) which sends a registration message to the HLR. HLRs and VLRs may be integral parts of the Mobile Switching Centers (MSCs), or they may be separate entities such that a single HLR or VLR serves multiple MSCs. We briefly describe location tracking in IS-41; location tracking in GSM is very similar and hence not described here.

In IS-41, once an endpoint enters a new location area, it sends a registration request to the MSC for that area, which is a central switching point for base stations in the location area. The MSC sends a authentication request message to its VLR which in turn forwards the request to the HLR for the endpoint. The HLR's response is delivered to the MSC. If the endpoint is authenticated, the MSC sends a registration notification message to its VLR which in turn forwards the message to the HLR. The HLR updates the location entry corresponding to the endpoint so that the entry points to the new serving MSC/VLR. The HLR sends back to the VLR a response which may contain relevant parts of the user's service profile. If the endpoint was registered previously in a different location area, the HLR sends a registration cancellation message to the previously visited VLR. Upon receiving this message, that VLR erases all entries for the endpoint and sends a cancellation message to the MSC which does likewise.

3.3. Location tracking in the Internet

The Internet Protocol (IP) [48] itself does not support node mobility. To address this need, the Internet Engineering Task Force (IETF) is standardizing a protocol, called Mobile IP [36], which provides support for mobile hosts in the Internet. For a comprehensive overview of mobile IP, see [49].

In Mobile IP, mobile endpoints are allocated permanent IP addresses on a "home" network. When an endpoint moves to a new location outside of the home network, it obtains a temporary forwarding address (called the *care-of address*) in the "foreign" network. The direct way of obtaining the care-of address is through a *foreign agent* in the visited network (whose existence is ascertained through an agent discovery protocol). The endpoint registers with the foreign agent, and the IP address of the foreign agent is used as the care-of address. Another way of obtaining the care-of address is through an address discovery protocol, such as the Dynamic Host Configuration Protocol (DHCP) [50]. The use of DHCP in supporting portable and mobile computing is discussed in detail in [51].

Every mobile endpoint must have a *home agent* on its home network that keeps track of the endpoint's current care-of address (called the *mobility binding*). Each time the endpoint establishes a new care-of address, it must *register* with its home agent so that the home agent always knows the current binding of each endpoint it serves. The home and foreign agents cooperate to provide the illusion that the endpoint is still in the home network. An endpoint wishing to send a message to a mobile endpoint sends the message to the permanent (home) address of the endpoint, where the home agent encapsulates and tunnels the message to the mobile endpoint. The basic Mobile IP protocol does not currently provide any support for direct communication with a mobile endpoint. These features are being developed within the IETF as a separate set of extensions to Mobile IP [52].

3.4. Location tracking in packet radio networks

In a "flat" packet radio network, all of the nodes are visible to each other with respect to routing and, therefore, no location tracking is necessary. As we have previously discussed, however, a flat network organization is not scalable, and hence some form of hierarchical clus-

tering should be employed to hide information. The SURAN packet radio network is a prime example of a hierarchically clustered packet radio network, and we shall describe location tracking techniques in that context. Each node has a *hierarchical address* used in routing packets to that node. This hierarchical address is the sequence of enclosing clusters, starting with the highest level and ending with the lowest level, in which the node resides. *Address servers* located in each bottom-level cluster keep track of a node's hierarchical address. When a node wants to send traffic to another node, it queries the address server present in its cluster for the address of the destination node. To answer the query, the address server in turn may query other address servers. Each address server maintains a cache of responses from other address servers, which can be used to respond to future requests. For a detailed discussion on location tracking in hierarchical packet radio networks and the SURAN address server design, refer to [53].

Some packet radio networks may be able to route messages using position (e.g., latitude and longitude) rather than topologically-derived information about nodes. In [54], two schemes are presented for maintaining and disseminating position information used for position-based routing. These schemes assume that every packet radio determines its position using the NAVSTAR Global Positioning System (GPS) [55].

4. Route selection and forwarding

In any network, the procedures for route selection and traffic forwarding require accurate information about the current state of the network (e.g., node interconnectivity and link quality) and the session (e.g., traffic rate, endpoint locations), in order to direct traffic along paths that are consistent with the service requirements of the session and the service restrictions of the network. Traffic sessions in mobile wireless networks may require frequent rerouting because of network and session state changes. The degree of dynamism in route selection depends on several factors, including the type and frequency of changes in network and session state; the limitations on response delay imposed in assembling, propagating, and acting upon this state information; the amount of network resources available for these functions; and the expected performance degradation resulting from a mismatch between selected routes and the actual network and session state.

As with location tracking, the quantity of network resources required to select routes that reflect rapid changes in network and session state may make such an approach impractical. Moreover, if the interval of time between successive state changes is shorter than the minimum possible response delay of the routing system, better performance may actually be achieved by not attempting to reroute for every state change. Note that many state changes are apt to be small (e.g., slight variations in delay) and hence unlikely to have much effect on the quality of service provided along the route selected for a session. Moreover, the routing system can decrease its sensitivity to small state changes while continuing to select feasible routes, by capturing statistical characterizations of the session and network state and by selecting routes according to these characterizations. If a state change is large enough to significantly affect the quality of service provided along the route for a session, the routing system should attempt to adapt its route to account for this change, in order to minimize the degradation in service to that session.

As in stationary networks, the types of route selection and forwarding procedures employed in mobile networks depend in part upon whether the underlying switching technology is circuit-based or packet-based, and in part on whether the switches themselves are stationary or mobile. In most cellular networks, routes are computed by an off-line procedure, and calls are forwarded along circuits set up along these routes. Handoff procedures enable a call to continue as a mobile endpoint moves from cell to cell. In most packet radio networks, routes are computed by the switches themselves, and traffic is forwarded hop-by-hop at each switch along the route. The switches individually adjust routes according to perceived changes in network topology resulting from switch movement.

4.1. Stationary infrastructure

A network with a stationary switching infrastructure, which may be either wireline or wireless, interconnects a set of base stations with radio interfaces through which mobile endpoints access the network. Such networks include cellular telecommunication networks, wireless ATM networks, and internetworks with mobile endpoints. In these networks, movement of endpoints triggers changes in routing, but such changes are usually performed by entities in the stationary infrastructure. These routing changes might not only affect traffic forwarding from the stationary infrastructure to the mobile endpoint but also traffic forwarding in the stationary infrastructure itself. To increase the probability of uninterrupted service as a mobile endpoint moves around the network, one can employ location prediction to determine the most likely future locations of the endpoint and then establish connections to those locations accordingly. In this special issue, the paper "A class of mobile motion prediction algorithms for wireless mobile computing and communications" by G. Liu and G. Maguire, introduces a set of motion prediction techniques. A model of endpoint movement is built, using information about an endpoint's past movements and the physical constraints imposed by the environment, which attempts to capture regular movement patterns when they exist. Such a model, in combination with the

endpoint's recent movement history, may then be used to predict the endpoint's next location.

4.1.1. Cellular telecommunications networks

Cellular telecommunications standards such as IS-41 [27] and GSM [26,56] are similar procedures for establishing calls to and from mobile endpoints. When an endpoint roams to a service area outside of its home, its current location information contained in its HLR should be updated so that it can continue to receive calls. An MSC may learn of a roaming endpoint in its service area through a registration request or call origination from that endpoint, or by some other means. In any case, the MSC notifies its associated VLR which in turn notifies the HLR of the endpoint's new location. When a endpoint X wishes to communicate with a mobile endpoint Y, the call is routed through the fixed network to an MSC in the home area, according to Y's home number. This MSC then consults Y's HLR to discover the location of the VLR for Y and extends the call setup to the VLR. If the exact location of Y's current base station is unknown, the VLR initiates paging through a local MSC which broadcasts the page to all base stations in the area. The base station in whose vicinity Y currently resides responds to the page, and the local MSC completes establishment of the call to Y.

When Y moves away from its current base station, it must quickly become affiliated with a new base station so as not to interrupt the call in progress with X. Selecting the new base station and determining when to initiate this handoff depends upon perceived signal quality (e.g. power, bit error rate) measured at the mobile endpoint and the nearby base stations. Call handoff may be initiated by the mobile endpoint, by the base station, by the mobile switching center, or by a combination of these, and it may occur within a service area or between adjacent service areas. Handoffs between cells under control of the same base station are called internal handoffs, while handoffs between cells controlled by separate base stations are called external handoffs. The following are the principal types of external handoffs.

Mobile-controlled handoff. The mobile endpoint constantly monitors the quality of the signal from its current base station and from other base stations in its vicinity. It chooses as its new base station the one producing the best signal, with some hysteresis built in to the selection process in order to prevent frequent handoffs when a mobile endpoint crosses back and forth between two cells. This technique is used in both DECT [6] and WACS [57].

Network-controlled handoff. The endpoint's current base station constantly monitors the quality of the signal from the mobile endpoint. When the signal quality falls below a specified threshold, that base station sends a handoff request to the MSC. The MSC then asks other base stations in the endpoint's vicinity to monitor the quality of the signal from the mobile endpoint, and these base stations respond with measurements of signal quality. From among the base stations with sufficiently high signal quality, the MSC selects a new base station. This technique is used in AMPS [3].

Mobile-assisted handoff. The mobile endpoint's current base station asks it to constantly monitor the quality of the signals received from a specified set of neighboring base stations. These measurements are returned to the base station which in turn delivers them to the MSC. The base station may also provide the MSC with its own measurements of the quality of the signal from the mobile endpoint, as in network-controlled handoff. Using all of these measurements, the MSC determines when to initiate handoff and which base station will become the endpoint's new base station. This technique results in lower delays than those encountered in network-controlled handoff and is used by GSM [26] and IS-54 [58]. In [59], enhancements to the basic IS-41 [27] handoff procedure are described, which provide sequential and/or single-copy delivery guarantees to support data services in a cellular network with an IEEE 802.6 MAN infrastructure.

Soft handoff. The mobile endpoint may become simultaneously affiliated with multiple base stations of approximately equal signal quality. In the incoming direction, the mobile endpoint can combine the signals received from these base stations, thus increasing the chances of correctly decoding the signal. Moreover, if one of the signals fades, the mobile endpoint is still likely to receive a signal from at least one of its remaining base stations. In the outgoing direction, the base stations handling the call with the mobile endpoint each send their traffic from the mobile endpoint, along with signal quality information, to the MSC which then selects the highest quality traffic stream to transmit onward. This technique significantly reduces the signal degradation and call dropping associated with the previous handoff techniques and is used in IS-95 digital CDMA [5].

During handoff, a newly-selected base station must have a channel available to receive the endpoint's call, in order for the handoff to be successful. Several techniques have been employed (see [60] for an overview) to decrease the probability that handoff will fail because of a lack of available channels. One approach is to reserve a set of channels that are only available for handoff calls and are not available for new calls. Another approach is to queue handoff requests at the new base station [61,62]. These requests might be served in FIFO order or in priority order (e.g., according to signal degradation from the old base station). To accommodate emergency calls (e.g., 911), base stations may reduce the capacity of

an existing call so that the handoff call and the existing call can coexist on a single channel [57,63].

In this special issue, the paper "A simple and efficient routing protocol for the UMTS Access network" by H. Mitts and H. Hansén describes a protocol to update forwarding information in a rooted-tree access network topology for UMTS [64], in order to enable efficient and seamless handoffs. This protocol minimizes the necessary modifications to the forwarding information stored in the tree and correctly reroutes traffic temporarily misrouted during handoff.

4.1.2. Wireless ATM networks

Recently, there have been several proposals for using a broadband infrastructure supporting ATM to interconnect mobile endpoints, in order to provide quality of service guarantees for a variety of applications. For an overview of the issues involved in supporting multimedia applications to mobile wireless endpoints over a broadband infrastructure, see [65]. These proposals for wireless ATM vary in the amount of mobility management that lies within the ATM network. At one end are approaches that advocate placing mobility management outside of the ATM network in special servers devoted to these tasks [66,67] so that one does not need to modify ATM nor ATM switches in order to support mobile endpoints. At the other end are approaches in which most of mobility management resides in the ATM switches themselves [68]. In [68], a wireless LAN is formed with a collection of "portable base stations". Virtual path/virtual circuit translation has been confined to the edges of the network, by using destination-oriented virtual path labels, in order to keep the base stations simple. All connections to a specified destination bear the same virtual path label and are distinguished by the virtual circuit identifier at the destination. A similar but less general approach to destination-based forwarding in ATM networks has been proposed in [69].

Much of the work on ATM support for mobile endpoints has focussed on techniques for enabling seamless handoffs of virtual circuits to mobile endpoints as those endpoints move around the network. These techniques may be summarized as follows.

Connection reestablishment. With this approach, the network establishes an entirely new virtual circuit each time a mobile endpoint moves to a new base station. While enabling the call processor to determine the current "best" route to that mobile endpoint, this approach requires a large amount of network resources to select the new route and to establish the new connection. Hence, it is likely to be impractical for large networks with many mobile endpoints.

Connection modification. With this approach, the network seeks to modify the original connection to reach the new base station, so as to incur the minimum amount of overhead in connection adjustment. One variant is to graft a new connection to the new base station onto the original connection. The most straightforward graft is to simply extend the original connection to reach the new base station [70,66]. While easy to implement, this approach results in long routes and hence inefficient use of network resources. By moving the grafting point back toward the source to an intermediate switch (or *crossover node* [71]) along the old connection, one may achieve shorter routes at the expense of searching for good crossover points for those routes. In this special issue, the paper "Crossover switch discovery for wireless ATM LANs" by C-K Toh describes and compares several approaches for selecting these crossover points.

Connection prediction. With this approach, multiple connections (called *virtual connection trees* [72]) are established between a selected fixed point in the stationary part of the network and a set of base stations corresponding to an area in which a group of endpoints is expected to move. This approach is best suited to an environment in which the movements of endpoints are known a priori; otherwise, many virtual circuits may be established that will never be used. When an endpoint establishes a connection to a mobile endpoint, that connection consists of two pieces: one from the calling endpoint to the root of the connection tree and one through the connection tree. This technique makes implicit the majority of handoffs, in that handoffs are accomplished by merely switching to different connections in the tree. Note that the portion of the connection from the calling endpoint to the tree's root does not change with handoff.

A connection tree architecture is proposed in [72]. Here, each path through the tree carries a different virtual circuit number. When a mobile endpoint moves between base stations in the same connection tree, it begins transmitting its traffic using the virtual circuit identifier associated with the path between its new base station and the root of the tree. The root detects this handoff by recognizing that the traffic from this mobile endpoint is now using a new virtual circuit number, and it accordingly modifies the virtual circuit number to use when sending traffic to the mobile endpoint, so that the traffic takes the correct path. When the mobile endpoint reaches the boundary of the connection tree, it attempts to gain access to a new tree. Referred to as *virtual connection tree handoff*, this procedure requires participation of a call processor.

In this special issue, the paper "Connection architecture and protocols to support efficient handoffs over an ATM/B-ISDN personal communications network" by O.T.W. Yu and V.C.M. Leung proposes a connection tree approach in which each mobile endpoint has a connection tree anchored at a *tether point* and with branches to the mobile endpoint's current as well as neighboring base stations. All branches carry the same "group virtual

channel identifier". The call processor may adjust the location of the tether point over time so as to minimize the number of links in the connection tree in order to make more efficient use of network resources. In [68], a similar approach is proposed, using connection "homes" for mobile endpoints, where the location of a connection home may be adjusted over time to be close to the mobile endpoint. In each case, these anchor points also serve to coordinate traffic streams during handoff, so that session message order is preserved.

In a connection tree, resources to support quality of service guarantees might be reserved prior to the connection's use, or they might be reserved "on demand" when the mobile endpoint affiliates with a different connection in the tree. The problem of quality of service provision in connection trees is addressed in [73]. In [74], the notion of a "shadow cluster" is proposed, which defines the area of influence of a mobile endpoint (i.e., the base stations to which the mobile endpoint is likely to handoff in the near future). Using information about expected movement of the mobile endpoint, each base station in the shadow cluster can determine the expected amount of resources to reserve in anticipation of a handoff from that mobile endpoint. Information about predicted resource use can also be used to determine whether to accept or reject new calls.

4.1.3. Internetworks

In internetworks, the proposals for mobility management can be divided into two categories: those that support direct routing to a mobile endpoint and those that require all traffic to a mobile endpoint to be routed via its home agent (sometimes referred to as "triangle" routing). The advantages of using a home agent include limiting the number of switches that need to be updated with the new location when a mobile endpoint moves; minimizing the amount of implementation in the network for mobility management; and the ability to keep private the current location of mobile endpoints. The disadvantage of this approach, however, is the potential for long routes and hence inefficient use of network resources. Currently, the home agent approach is the official proposal of the Internet Engineering Task Force for supporting mobile endpoints in the Internet [36], and is also the approach used in Cellular Digital Packet Data (CDPD) [37]. In this special issue, the paper "Analysis of a mobile-assisted adaptive location management strategy" by R. Yates, C. Rose, S. Rajagopalan, and B.R. Badrinath describes and analyzes an optimal routing policy that switches between direct and triangle routing according to the costs of the two routes, the source traffic rate, and the rate of location updating.

Approaches to direct routing [49,75] use location caches for mobile endpoints, enabling direct communication with mobile endpoints without using home agents. Both of these approaches still retain the notion of home network, so that each endpoint may be reached via its home address and its current address, but they differ in the way in which the location caches are maintained in the switches. In [75], each message from an endpoint carries both addresses, and switches in the network learn the current location of mobile endpoints by "snooping" the addresses contained in messages they forward. In [49], distribution of location information is controlled by the home agent for security reasons. A switch with a location cache entry for a mobile endpoint can determine, from addresses contained in a received message, whether the source of the message also has a location cache entry for that endpoint. That switch can then alert the source to request a cache update from the mobile's endpoint home agent, or if the switch is the home agent, it may update the source itself.

Two approaches to supporting seamless session handoff as a mobile endpoint moves throughout an internetwork are as follows: (1) alerting the mobile endpoint's previous foreign agent of the current foreign agent's location [49]; and (2) distributing multiple copies of messages to multiple base stations in the vicinity of the mobile endpoint [76].

Multicast. Multicast distribution in mobile wireless networks is a topic that has only recently begun to be explored. In this special issue, we include two separate papers on multicast. The paper "A framework for delivering multicast messages in networks with mobile hosts" by A. Acharya and B.R. Badrinath provides algorithms for guaranteeing at least once, at most once, and exactly once delivery of a message to members of a multicast group. In this case, a base station joins a multicast group if at least one of the endpoints associated with it is a member of that multicast group. This helps to reduce the dynamism of a multicast group as perceived by the rest of the network. The paper "Efficient solutions to multicast routing in communications networks" by K. Makki, N. Pissinou, and O. Frieder proposes an efficient algorithm to find near-minimum-cost Steiner trees [77]. Such an approach to multicasting is desirable for low-power wireless networks with wireless infrastructure, where transmission capacity is a scarce resource. An efficient off-line algorithm for joining new multicast group members to existing multicast trees, in networks with limited transmission capacity, is described in [78]. This algorithm permits tradeoffs between minimum-hop paths and paths that minimize the amount of traffic over the most congested link, depending upon current network load. For each new subscription to a multicast group, the algorithm produces the lowest-cost path (where cost is a traffic-dependent pricing function defined for each link) from the new member to an existing member of the group.

4.2. Mobile infrastructure

In mobile networks with stationary infrastructure,

the main component of route selection for mobile endpoints is handoff. In mobile networks with mobile infrastructure (i.e., packet radio networks), the switches must not only keep track of the locations of mobile endpoints but also must keep track of each others' locations and interconnectivity as they move. A considerable body of research has been amassed on route selection in packet radio networks. Route selection requires information about the interconnectivity and services provided by the switches as well as information about the service requirements for the session and the locations of the session endpoints. In a packet radio network, where network topology changes frequently and where transmission capacity is scarce, the procedures for distributing routing information and selecting routes must be designed to consume a minimum amount of network resources yet must be able to quickly adapt to changes in network topology.

Most of the proposed approaches to route selection in packet radio networks have centered on distributed adaptive procedures which take into account local or global information about network state in selecting the next hop to a destination. When only local network state information is available, a switch selects the next hop which is the "best" choice within its neighborhood. When global network state information is available, a switch selects the next hop which is on the "best" route to the destination. Centralized approaches have also been proposed for selecting the best routes in packet radio networks. In [79], routing is posed as an optimization problem, where minimizing congestion is the objective, and solutions are obtained through neural network techniques.

4.2.1. Minimum-cost routing

The majority of the approaches to selecting the minimum-cost routes in packet radio networks [80,11,81,18] have used asynchronous, distributed variants of the Bellman-Ford [82] (or *distance-vector*) algorithm. These types of algorithms have been favored over *link-state* algorithms, many of which require each switch to inform every other switch about all its links, because they usually require a smaller amount of routing information to be distributed and because they spread the cost of selecting a route among the switches along the route. The disadvantages of distance-vector algorithms are that they are slow to converge; they may form temporary but lengthy routing loops; and they may discard some useful routes. Not all link-state algorithms require global distribution of routing information, however, and hence such algorithms may be viable candidates for minimum-cost routing in packet radio networks. In this special issue, the paper "An efficient routing protocol for wireless networks" by S. Murthy and J.J. Garcia-Luna-Aceves describes a new distance-vector variant, in which switches exchange information about the distance to a destination and their second-to-last hop on the shortest route to the destination. Using this information, a switch can determine its shortest routes to the destination. To eliminate certain looping problems, a switch that receives a routing update from a neighbor checks all other neighbor's second-to-last hops to determine whether the route implied by the neighbor includes the updating neighbor, and if so, the distance information for those routes is updated accordingly.

4.2.2. Metrics

Several different metrics have been proposed for selecting routes and next hops in packet radio networks, but all of them share in common the goal of promoting efficient use of transmission capacity. Metrics used in minimum-cost routing include hop count, capacity, and various measures of interference. *Tier routing* [80,12] selects minimum-hop routes, in order to minimize the number of network resources used by a session and hence increase network throughput. Maximum-Minimum Residual Capacity Routing (MMRCR) [83] assigns a link cost that depends upon the traffic using the link and indicates the probability of successful transmission and interference. *Subclass routing* [14] assumes adaptive gain control at the transmitter and uses link gain as the link cost. In this case, the objective is to minimize the maximum link gain used along a route, in order to minimize interference.

Direct measures of interference have also been proposed as metrics for minimum-cost routing, in one case to minimize the amount of interference caused by a transmission and in another case to minimize the amount of interference encountered by a transmission. In Least Interference Routing (LIR) [84], the minimum-cost routes are those which cause the least destructive interference. The algorithm assumes adaptive gain control for the transmitter, and different neighbors may be reached through different gains. Each switch computes the potential destructive interference that could result from a transmission to each neighbor. The interference metric is equal to the number of neighbors that can potentially hear the transmission, other than the neighbor to which the transmission is directed, and is independent of traffic levels. In Least Resistance Routing (LRR) [81], the minimum-cost routes are those which encounter the least interference. Interference may be caused by multiple access, jamming, or radios transmitting from outside of the network, and the algorithm does not attempt to determine the cause. Each switch computes an interference measure based on predetection information (e.g., signal energy levels) or postdetection information (e.g., signal quality information from the demodulator). After measuring interference, a switch sends its interference metric to each of its neighbors so that they can determine the interference on their outgoing links.

Several algorithms have been proposed to select next hops based only on knowledge of the position of the cur-

rent switch, the position of the destination, and the state of the network in the vicinity of the switch. These algorithms rely on alternate means, such as GPS [55], to determine position information. All of these algorithms enable a switch to choose a next hop that allows a message to make forward progress toward its destination. The "random" routing approach proposed in [85] selects a next hop at random from the set of its neighbors that are in the direction of the destination. Other approaches, which assume that the transmitter can adjust its power, include the following. Nearest with Forward Progress (NFP) routing [86] chooses as next hop the nearest neighbor with smallest transmission range in the direction of the destination, in order to minimize the potential for interference with other transmission in the area. Most forward with Fixed Range (MFR) routing selects as next hop the neighbor closest to the destination within the specified transmission range, in order to minimize the number of hops along the route. Most forward with Variable Range (MVR) routing [86] is similar to MFR routing except that the transmission range is readjusted to be no more than the distance to the selected neighbor, in order to reduce interference. For a comparison of these next-hop selection strategies, see [87].

4.2.3. Route discovery

Several approaches to routing in packet radio networks have provisions for discovering routes [90,91,14]. In [90], a source wishing to acquire a route to a destination broadcasts a query throughout the network. Any switch with a route to that destination stops propagating the query and responds with a reply, which is flooded over links that do not yet have a route to the destination. The reply may contain route cost information as well, so that recipients can select routes based on costs. Each recipient of the reply may update its routes to the destination using as next hop the switch from which the reply was obtained. A switch may maintain multiple routes to the destination, but route acceptance is done in such a way as to guarantee freedom from loops. Thus, the source as well as other switches may learn about routes to the destination. A similar flooding search algorithm has been proposed for use in tactical circuit-switched networks [92].

An alternative to flooding a route query is for a source, lacking a route to a destination, to forward messages for that destination to a randomly selected next hop. In the approach proposed in [91], a message propagates in this manner until reaching a node that does have a route to the destination, which in turn forwards the message to that destination. Each message, control or data, carries the entire route so that any switch handling the message can obtain a route to the sender by reversing the route contained in the message. As each data message requires an end-to-end acknowledgement, the source and any other switch on the acknowledgement's return path can thus learn a route to the destination. In fact, a switch may learn of and maintain information about multiple routes to a destination. There exist other approaches that also rely on gathering information from passing messages in order to learn about routes. In SURAN, each data message carries the source, destination, number of hops travelled, and number of hops remaining to the destination. Hence, intermediate switches can augment their routes with information obtained from overheard messages, even messages not meant for them [14].

4.2.4. Alternate routes

When network connectivity changes, switches must adapt their routing accordingly, in order to cause minimal disruption to traffic sessions in progress. In some packet radio networks, switches may actively seek out alternate routes upon detecting a failure in the primary route, while in others, switches may maintain multiple routes to each destination so that they can quickly shift to an alternate route if necessary.

The approach described in [90] enables a switch to send a request for an alternate route, upon detection of a break in the current route to a destination. This request propagates in a directed flood upstream from the break. Each recipient of the request determines whether it has a route to the destination. If it has a route through the switch from which the request was received, the recipient erases that route and propagates the request upstream. If it has an alternate route, the recipient responds by broadcasting a reply indicating the route. The propagation of the alternate route request terminates upon reaching a node with an alternate route or when all upstream nodes have been visited. With this approach, not only are failed routes removed quickly but new routes are also found quickly. A route repair technique yielding a similar effect appears in [93]. Here, the routing for a destination is viewed as an acyclic directed graph. When a route breaks, some node (other than the destination) will no longer have any outgoing links. The algorithm iteratively reverses link directions for nodes with no outgoing links, which in turn may cause there to be upstream nodes with no outgoing links. Provided an alternate route exists, the algorithm will terminate with a new acyclic directed graph for the destination.

In the approach proposed in [91], switches may maintain multiple routes to a destination. When a switch fails to receive a link-level acknowledgement for a message from the next hop to the destination, after a specified number of transmissions, it automatically selects an alternate route for that message from its set of candidates. SURAN also uses failure to receive link-level acknowledgements as the cue to seek alternate routes to the destination [14]. In this case, the switch broadcasts the message to all of its neighbors. Any neighbor that receives the message and is closer to the destination than the sending switch forwards the message toward the destination but does not broadcast the message to its neigh-

bors. This approach is very reliable but can potentially consume a large number of network resources as multiple copies of the message are likely to be transmitted to the destination. SURAN also includes mechanisms for quickly eliminating failed routes and obtaining new routes. When a switch detects a break in a route to a destination, it informs its neighbors. That switch may then include, in any message sent to its neighbors, a request for a new route to that destination. Any neighbor with a good route responds to such a request.

For switches that actively maintain multiple routes to a destination, other techniques for selecting alternate routes exist. The following approaches are described and compared in detail in [94]. With *primary n/m* forwarding, a switch selects a primary and a secondary outgoing link for the destination. When transmitting a message, the switch makes m attempts on the primary link, m attempts on the secondary link, and so on, until either the transmission is successful or the number of transmission exceeds n, in which case the message is discarded. *Good-link n/m* forwarding is similar to primary n/m forwarding, except that the switch first makes m attempts on the link used by the last message successfully forwarded to the destination, which is not necessarily the primary link. *EE-based* forwarding is similar to good-link n/m forwarding, except that the switch first makes m attempts on the link with the smaller *resistance* value (as computed in least resistance routing). *EEA-based* forwarding is similar to EE-based forwarding, except that resistance is also a function of whether an acknowledgement is received for a transmission. For each consecutive failure to receive an acknowledgement, the switch increases the resistance value to the neighbor by one. Receipt of a new resistance value from the neighbor replaces the one computed based on failure to receive acknowledgements. Resistance for EEA-based forwarding may also be computed as the constant-with-acknowledgements metric. In this case, the resistance value is initially set to one, incremented for each consecutive failure to receive an acknowledgement, and reset upon receipt of an acknowledgement.

4.2.5. Hierarchical routing

As we have already discussed, hierarchical network structure is an effective way to organize a network comprising a large number of nodes. Much of the work on hierarchically organized packet radio networks was performed in the context of SURAN, and hence we describe the approaches from this perspective. SURAN included two hierarchical routing algorithms to be used in conjunction with its hierarchical clusters (see [20,14] for more details).

Quasi hierarchical routing. This approach, based on [95], enhances tier routing by including the minimum distance to other radios in the cluster, to other clusters in the supercluster, and to other superclusters, in the routing information exchanged between switches. Initialization begins when a *border* packet radio (i.e., a radio on the boundary of a cluster) receives routing information from a radio in another (super)cluster. In one variant, the neighboring (super)cluster is considered to be one hop away, and the radio's route to the (super)cluster includes the shortest route to the border of the cluster. In another variant, the neighboring (super)cluster is considered to be n hops away, where n is the average number of hops from the border packet radio to the members of the (super)cluster. The length of the radio's route to the (super)cluster is the sum of the length of the minimum-hop route to the cluster border plus the average distance from the border to the members of the cluster.

Strict hierarchical routing. This approach uses tier routing within a cluster and link-state routing among clusters. (Super)clusterheads flood routing information about themselves to other (super)clusterheads. Using the link-state information, they then compute (super)-cluster routes and pass this information to the next lower level. Thus, clusterheads learn which next cluster to use to reach which superclusters, and radios learn which next cluster to use to reach a destination cluster in their supercluster and which next cluster to use to reach another supercluster. A similar algorithm is proposed in [96]. In this case, a link-state algorithm is used for both intra-cluster and inter-cluster routing, and border switches play the same role as SURAN clusterheads in distributing routing information.

In SURAN, packet radios are allowed to participate simultaneously in multiple clusters, for the purposes of routing, during a short time interval after changing cluster affiliations. Thus, when a radio moves to a new cluster, for a time it can still receive messages routed to its old cluster, thus limiting session interruption as an endpoint moves.

References

[1] T.S. Rappaport, *Wireless Communications: Principles and Practice* (Prentice-Hall, Englewood Cliff, NJ, 1996).

[2] V.H. MacDonald, The cellular concept, The Bell Syst. Tech. J. 44 (1965) 547–588.

[3] W.R. Young, Advanced mobile phone service: introduction, background, and objectives, The Bell Syst. Tech. J. 58 (1979) 1–14.

[4] D. Cox, Wireless network access for personal communications, IEEE Commun. (December 1992) 96–115.

[5] Telecommunications Industry Association, Mobile station–base station compatibility standard for dual-mode wideband spread spectrum cellular system, TIA/EIA IS-95 (July 1993).

[6] ETSI, Digital european cordless telephone common interface, version 05.03 (May 1991).

[7] Radio Advisory Board of Canada, CT2Plus class 2: specification for the canadian common air interface for digital cordless telephony, including public access services, annex 1 to radio standards specification 120 (1992).

[8] J.W. Ketchum, Routing in cellular mobile radio communication networks, in: *Routing in Communication Networks*, ed. M. Steenstrup (Prentice-Hall, Englewood Cliffs, NJ, 1995).
[9] A. Merchant and B. Sengupta, Assignment of cells to switches in PCS networks, IEEE/ACM Trans. Networking 3(5) (1995) 521–526.
[10] R.S. Kahn, J. Gronemeyer, J. Burchfiel and R. Kunzelman, Advances in packet radio technology, Proc. IEEE 66(11) (1978) 1468–1496.
[11] J. Jubin and J.D. Tornow, The DARPA packet radio network protocols, Proc. IEEE 75(1) (1987) 21–32.
[12] N. Shacham and J. Westcott, Future directions in packet radio architectures and protocols, Proc. IEEE 75(1) (1987) 83–99.
[13] B. Leiner, D. Nielson and F. Tobagi, Issues in packet radio network design, Proc. IEEE 75(1) (1987) 6–20.
[14] G. Lauer, Packet-radio routing, in: *Routing in Communication Networks*, ed. M. Steenstrup (Prentice-Hall, Englewood Cliffs, NJ, 1995).
[15] C.E. Perkins and P. Bhagwat, Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers, *Proc. ACM SIGCOMM*, London, UK (1994) pp. 234–244.
[16] W. Diepstraten, G. Ennis and P. Berlanger, DFWMAC: distributed foundation wireless medium access control, IEEE Document P802.11-93/190 (November 1993).
[17] A. Ephremedis, J.E. Wieselthier and D.J. Baker, A design concept for reliable mobile radio networks with frequency hopping signalling, Proc. IEEE 75(1) (1987) 56–73.
[18] M. Gerla and J.T.-C. Tsai, Multicluster, mobile, multimedia radio network, Wireless Networks 1 (1995) 255–266.
[19] C.R Lin and M. Gerla, Multimedia transport in multihop dynamic packet radio networks, *Proc. IEEE GLOBECOM* (1995) pp. 209–216.
[20] G. Lauer, Hierarchical routing design for SURAN, *Proc. ICC* (1986) pp. 93–101.
[21] J.L. Grubb, The traveller's dream come true, IEEE Commun. 29(11) (1991) 48–51.
[22] R.J. Leopold, Low-earth orbit global cellular communications network, *Mobile Satellite Comm. Conf.*, Adelaide, Australia (1990).
[23] R. Binder et al., Crosslink architectures for a multiple satellite system, *Proc. IEEE* 75(1) (1987) 74–82.
[24] J. Kaniyil et al., "A global message network employing low earth-oribiting datellites, IEEE J. Select. Areas Commun. 10(2) (1992) 418–427.
[25] U. Madhow, M.L. Honig and K. Steiglitz, Optimization of wireless resources for personal communications mobility tracking, IEEE/ACM Trans. Networking 3(6) (1995) 698–706.
[26] M. Mouly and M.B. Pautet, The GSM system for mobile communications, M. Mouly, 49 rue Louise Bruneau, Palaiseu, France (1992).
[27] Telecommunications Industry Association, Cellular radiotelecommunication intersystem operation, TIA/EIA IS-41B (1991).
[28] S. Mohan and R. Jain, Two user location strategies for personal communications services, IEEE Personal Commun. (First Quarter 1994) 42–50.
[29] J.G. Markoulidakis, G.L. Lyberopoulos, D.F. Tsirkas and E.D. Sykas, Evaluation of location area planning scenarios in future mobile telecommunication systems, Wireless Networks 1 (1995) 17–30.
[30] A. Bar-Noy and I. Kessler, Tracking mobile users in wireless communication networks, IEEE Trans. Inform. Theory 39(6) (1993) 1877–1886.
[31] H. Xie, S. Tabbane and D.J. Goodman, Dynamic location area management and performance analysis, *Proc. 43rd IEEE Vehicular Tech. Conf.* (1993) pp. 536–539.
[32] Y. Birk and Y. Nachman, Using direction and elapsed-time information to reduce the wireless cost of locating mobile units in cellular networks, Wireless Networks 1 (1995) 403–412.
[33] A. Bar-Noy, I. Kessler and M. Sidi, Mobile users: to update or not to update?, Wireless Networks 1 (1995) 175–186.
[34] C. Rose and R. Yates, Minimizing the average cost of paging under delay constraints, Wireless Networks 1 (1995) 211–219.
[35] I.F. Akyildiz and J.S.M. Ho, A mobile user location update and paging mechanism under delay constraints, *Proc. ACM SIGCOMM*, Cambridge, MA, (1995) pp. 244–255.
[36] IETF Mobile-IP Working Group, IPv4 mobility support, working draft (1995).
[37] Ameritech Mobile Communications, Inc., Bell Atlantic Mobile Systems, Contel Cellular, Inc., GTE Mobile Communications, Inc., McCaw Cellular Communications, Inc., NYNEX Mobile Communications, Inc., PacTel Cellular, and Southwestern Bell Mobile Systems, Cellular digital packet data system specification (1993).
[38] K. Meier-Hellstern and E. Alonso, The use of SS7 and GSM to support high density personal communications, *Proc. ICC* (1992) pp. 1698–1702.
[39] V.N. Lo, R.S. Wolff and R.C. Bernhardt, Expected network database transaction volume to support personal communications services, *1st Int. Conf. Universal Personal Communications Services*, Dallas, TX (1992).
[40] R. Jain, Y.-B. Lin, C. Lo and S. Mohan, A caching strategy to reduce network impacts of PCS, IEEE J. Select. Areas Commun. 12(8) (1994) 1434–1444.
[41] J.S.M. Ho and I.F. Akyildiz, Local anchor scheme for reducing location tracking costs in PCNs, *Proc. ACM MOBICOM*, Berkeley, CA (1995) pp. 181–194.
[42] B. Awerbuch and D. Peleg, Concurrent online tracking of mobile users, *Proc. ACM SIGCOMM*, Zurich, Switzerland (1991) pp. 221–234.
[43] V. Anantharam, M.L. Honig, U. Madhow and V.K. Wei, Optimization of a database hierarchy for mobility tracking in a personal communications network, Perform. Eval. 20 (1994) 287–300.
[44] B.R. Badrinath, T. Imielinski and A. Virmani, Locating strategies for personal communication networks, *Proc. Workshop on Networking of Personal Communications Applications* (1992).
[45] M.T. Ozsu and P. Valduriez, *Principles of Distributed Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1991).
[46] L.W. Dowdy and D.V. Foster, Comparative models of the file allocation problem, ACM Computing Surveys 14(2) (1982) 287–313.
[47] N. Shivakumar and J. Widom, User profile replication for faster location lookup in mobile environments, *Proc. ACM MOBICOM*, Berkeley, CA (1995) pp. 161–169.
[48] J. Postel, Internet protocol, Internet RFC 791 (1981).
[49] D.B. Johnson, Scalable support for transparent mobile host internetworking, Wireless Networks 1 (1995) 311–322.
[50] R. Droms, Dynamic host configuration protocol, Internet RFC 1541 (1993).
[51] C.E. Perkins and K. Luo, Using DHCP with computers that move, Wireless Networks 1 (1995) 341–354.
[52] D.B. Johnson and C. Perkins, Route optimization in Mobile IP, working draft (1996).
[53] G. Lauer, Address servers in hierarchical networks, *Proc. ICC* (1988) pp. 443–451.
[54] T.-C. Hou and V.O.K. Li, Position updates and sensitivity analysis for routing protocols in mobile packet radio networks, *Proc. IEEE GLOBECOM* (1985) pp. 243–249..
[55] B.W. Parkinson and S.W. Gilbert, NAVSTAR: global positioning system – ten years later, Proc. IEEE 71(10) (1983) 1177–1186.

[56] M. Rahnema, Overview of the GSM system and protocol architecture, IEEE Commun. (April 1993) 92–100.
[57] Bellcore, Generic criteria for version 1.0 wireless access communications systems (WACS) and supplement, Bellcore Technical Reference TR-INS-001313 1 (1993).
[58] Telecommunications Industry Association, Cellular system dual mode mobile station–base station compatibility standard, TIA/EIA IS-54B (1992).
[59] G.P. Pollini and K.S. Meier-Hellstern, Efficient routing of information between interconnected cellular mobile switching centers, IEEE/ACM Trans. Networking 3(6) (1995) 765–774.
[60] Y.B. Lin, S. Mohan and A. Noerpel, PCS channel assignment strategies for hand-off and initial access, IEEE Personal Commun. (Third Quarter 1994) 47–56.
[61] S. Tekinary and B. Jabbari, Handover policies and channel assignment strategies in mobile cellular networks, IEEE Commun. 9(11) (1991) 42–46.
[62] S. Tekinary and B. Jabbari, A measurement based prioritization scheme for handovers in cellular and microcellular networks, IEEE J. Select. Areas Commun. 10(8) (1992) 1343–1350.
[63] Y.-B. Lin, A. Noerpel and D.A. Harasty, Non-blocking channel assignment strategy for handoffs, *Proc. IEEE 3rd Int. Conf. on Universal Personal Communications Services*, San Diego, CA (1994).
[64] E. Buitenwerf et al., UMTS: fixed network issues and design options, IEEE Personal Commun. (February 1995) 30–37.
[65] M. Schwartz, Network management and control issues in multimedia wireless networks, IEEE Personal Commun. (June 1995) 8–16.
[66] B. Rajagopalan, Mobility management in integrated wireless-ATM networks, *Proc. ACM MOBICOM*, Berkeley, CA (1995) pp. 127–141.
[67] M. Veeraraghavan, T.F. La Porta and R. Ramjee, A distributed control strategy for wireless ATM networks, Wireless Networks 1 (1995) 323–340.
[68] K.Y. Eng et al., A wireless broadband ad-hoc ATM local-area network, Wireless Networks 1 (1995) 161–174.
[69] R. Cohen et al., The sink tree paradigm: connectionless traffic support on ATM LANs, *Proc. IEEE INFOCOM* (1994) pp. 821–828.
[70] R. Mishra and M. Srivastava, Call establishment and rerouting in mobile computing networks, AT&T Technical Memo 11384-940906-13TM (1994).
[71] K. Lee, Adaptive network support for mobile multimedia, *Proc. ACM MOBICOM*, Berkeley, CA (1995) pp. 62–74.
[72] A.S. Acampora and M. Naghshineh, An architecture and methodology for mobile-executed handoff in cellular ATM networks, IEEE J. Select. Areas Commun. 12(8) (1994) 1365–1375.
[73] A. Acampora and M. Nagshineh, Control and quality-of-service provisioning in high-speed microcellular networks, IEEE Personal Commun. (Second Quarter 1994) 36–43.
[74] D.A. Levine, I.F. Akyildiz and M. Naghshineh, The shadow cluster concept for resource allocation and call admission in ATM-based wireless networks, *Proc. ACM MOBICOM*, Berkeley, CA (1995) pp. 142–150.
[75] F. Teraoka, Y. Yokote and M. Tokoro, A network architecture providing host migration transparency, *Proc. ACM SIGCOMM*, Zurich, Switzerland (1991) pp. 209–220.
[76] H. Balakrishnan, S. Seshan and R.H. Katz, Improving reliable transport and handoff performance in cellular wireless networks, Wireless Networks 1 (1995) 469–482.
[77] E.N. Gilbert and H.O. Pollack, Steiner minimal trees, SIAM J. Appl. Math. 16 (1968) 1–29.
[78] B. Awerbuch and Y. Azar, Competetive multicast routing, Wireless Networks 1 (1995) 107–114.
[79] J.E. Wieselthier, C.M. Barnhart and A. Ephremides, A neural network approach to routing without interference in multihop radio networks, IEEE Trans. Commun. 42(1) (1994) 166–177.
[80] A. Belghith and L. Kleinrock, A distributed routing scheme with mobility handling in stationless multi-hop packet radio networks, *Proc. ACM SIGCOMM*, Austin, TX (1983) pp. 101–118.
[81] M.B. Pursley and H.B. Russell, Routing in frequency-hop packet radio networks with partialband jamming, IEEE Trans. Commun. 41(7) (1993) 1117–1124.
[82] L.R. Ford Jr. and D.R. Fulkerson, *Flows in Networks* (Princeton University Press, Princeton, NJ, 1962).
[83] D. Beyer et al., Packet radio network research, development and application, *Proc. SHAPE Packet Radio Symposium* (1989).
[84] J. Stevens, Spatial reuse through dynamic power and routing control in common-channel random-access packet radio networks, SURAN Program Technical Note (SRNTN) 59 (1988). Available from the Defense Technical Information Center.
[85] R. Nelson and L. Kleinrock, The spatial capacity of a slotted ALOHA multihop packet radio network with capture, IEEE Trans. Commun. COM-32(6) (1984) 684–694.
[86] T.C. Hou and V.O.K. Li, Performance analysis of routing strategies in multihop packet radio networks, *Proc. IEEE GLOBECOM* (1984) pp. 487–492.
[87] T.C. Hou and V.O.K. Li, Transmission range control in multihop packet radio networks, IEEE Trans. Commun. COM-34(1) (1986) 38–44.
[88] L. Kleinrock and J.A. Silvester, Optimum transmission radii for packet radio networks of why six is a magic number, *Proc. National Telecommunications Conference* (1978) pp. 4.3.1.–4.3.5.
[89] H. Takagi and L. Kleinrock, Optimal transmission ranges for randomly distributed packet radio terminals, IEEE Trans. Commun. 32(3) (1984) 246–257.
[90] M.S. Corson and A. Ephremides, A distributed routing algorithm for mobile wireless networks, Wireless Networks 1 (1995) 61–82.
[91] K. Brayer, Packet switching for mobile earth stations via low-orbit satellite network, Proc. IEEE 72(11) (1994) 1627–1636.
[92] R.P. Lippmann, New routing and preemption algorithms for circuit-switched mixed media networks, *Proc. of IEEE MILCOM* (1985) pp. 660–666.
[93] E.M. Gafni and D.P. Bertsekas, Distributed algorithms for generating loop-free routes in networks with frequently changing topology, IEEE Trans. Commun. COM-29(1) (1981) 11–18.
[94] M.B. Pursley and H.B. Russell, Network protocols for frequency-hop packet radios with decoder sided information, IEEE J. Select. Areas Commun. 12(4) (1994) 612–621.
[95] I. Kleinrock and F. Kamoun, Hierarchical routing for large networks, Computer Networks 1 (1977) 155–174.
[96] W.T. Tsai, C.V. Ramamoorthy, W.K. Tsai and O. Nishiguchi, An adaptive hierarchical routing protocol, IEEE Trans. Commun. 38(8) (1989) 1059–1075.

S. Ramanathan received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Madras, India, and the Ph.D. degree in computer science from the University of Delaware. He is presently a scientist in the Advanced Networking Research group at BBN Corporation. His research interests are in the design and analysis of network algorithms, including internetwork routing, channel access and routing in packet radio networks, and in personal communications networks. He was the recipient of the Outstanding Student Paper Award from ACM SIGCOMM, 1992, and the Outstanding Paper Award from IEEE INFOCOM, 1996.
E-mail: ramanath@bbn.com

Martha Steenstrup received the A.B. degree in mathematics and music from Smith College, the M.A. degree in mathematics from Columbia University, and the Ph.D. degree in computer and information science from the University of Massachusetts at Amherst. Currently, she is a division scientist in the Advanced Networking Research group at BBN Corporation. Her research interests include routing, flow control, and service provision in wireline and wireless networks. She is a member of ACM and IEEE.
E-mail: msteenst@bbn.com