Exploiting User Profiles to Support Differentiated Services in Next-Generation Wireless Networks

Vijoy Pandey, Nortel Networks
Dipak Ghosal and Biswanath Mukherjee, University of California, Davis

Abstract

In the next-generation wireless network, user profiles such as the location, the velocity (both speed and direction), and the resource requirements of the mobile device can be accurately determined and maintained by the network on a per-user basis. We investigate the design of a wireless network architecture that exploits user profiles to maximize network efficiency and provide better Quality-of-Service (QoS) to different classes of users. In this article we provide implementation guidelines of such an architecture for the Third-Generation Partnership Project (3GPP) network. The key underlying primitive of the architecture is the use of both real-time and aggregate user profiles to perform advance resource reservation in the handoff target cells of the wireless cellular network. We identify various factors that can influence the efficiency of the resource reservation scheme, and through a simulation analysis of an example scenario we show the impact of these factors on the QoS that profiled users receive. The example scenario comprises two service classes: a high cost, profiled service with higher QoS; and a lower cost, non-profiled service with best-effort QoS. The results show that high QoS can be guaranteed to users who subscribe to the profiled service.

In order to achieve the goal of providing high-quality multimedia services in next-generation wireless networks [1], it will be necessary to implement new techniques that can guarantee Quality of Service (QoS) while accounting for the limited bandwidth and the delay and error characteristics of the wireless access network. This requires a differentiated-services architecture that can offer multiple service levels, each with a different QoS guarantee.

Implementing a differentiated-services architecture in a wireless network is complex due to two key factors. First, due to user mobility, the network needs to guarantee resources spatially as the user moves and attaches to different points in the network. Second, the resource requirements may change due to changes in the mobile device. For example, a user switching from a palmtop to her laptop may request higher bandwidth from the network.

User mobility in wireless networks has been characterized by random-walk or Brownian-motion-based mathematical models [2]. In reality, however, user mobility has a high degree of predictability due to temporal and spatial locality. Temporal locality arises due to the fact that a mobile user typically takes predictable routes, which implies that a user will typically cross the same set of cells at predictable times in a wireless cellular network. For instance, a mobile user will typically follow the same path to work in the morning, and the reverse route back home in the evening. Spatial locality refers to the fact that user mobility is constrained along pathways and highways, which results in a mobile user crossing the cells in an ordered sequence determined by the manner in which these pathways and highways intersect the cellular coverage area.

Recently, with the Federal Communications Commission (FCC) Enhanced 911 mandate, location-based services have become a top priority for cellular service providers. There is a focused effort to deliver personal, time-critical, and location-dependent information to users, such as driving directions, current traffic condition, tracking other family members, and local facilities-based services. As this is a growing market, different methods have been developed to accurately determine the location of a mobile user. Advances have also been made in velocity-estimation algorithms [3]. Moreover, due to the strong interaction between the different protocol layers — physical, data-link, network, and applications layers — in a wireless network, there is need for and it is possible to accurately determine the resource requirements of the mobile device.

In this article the real-time and aggregate values of a user's mobility and resource requirements are referred to as the "user profile." These user profiles are maintained by the network, and the goal of our study is to investigate the design and implementation of a user-profile-based differentiated-services architecture for the next-generation wireless network. We describe specific implementation details for the Third-Generation Partnership Project (3GPP) [1] cellular network architecture and study the performance benefits of our proposed approach in a network with two types of users:

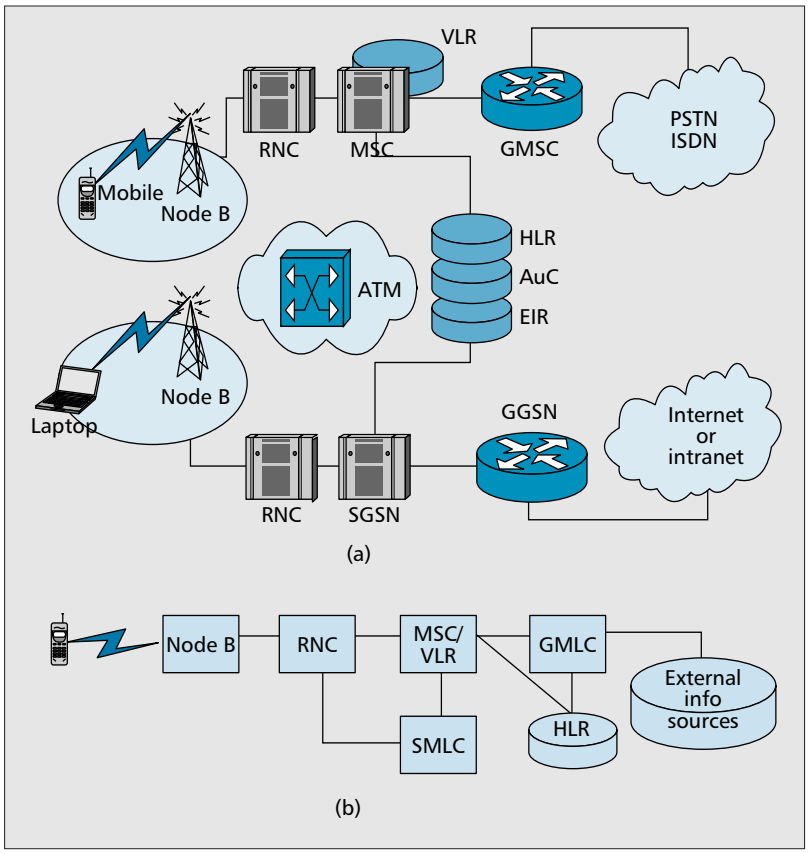
- Profiled users who subscribe to a higher-cost profiled service that guarantees higher QoS.
- Regular (non-profiled) users who receive best-effort service.

We have used dropping probability, the ratio of dropped handoff attempts to the total number of handoff attempts

0890-8044/04/$20.00 © 2004 IEEE

generated, as the metric for QoS in this study. We observe that the network provides improved QoS to profiled users by significantly reducing their dropping probability through advanced reservation of cell resources along the path predicted by the user profile. There are optimal values of the reservation distance (which is the distance prior to a cell crossing when the reservation is attempted) and the reservation granularity (which is related to the frequency of the re-attempts when a reservation attempt fails) which result in the maximal improvement in dropping probability.

The novelty of our work is that it describes the implementation and studies the characteristics of an on-demand resource-reservation algorithm based on user profiles and path prediction for two classes of users. The idea of user profiles has been proposed in the literature in the context of the mobility tracking and path prediction for wireless and mobile networks [4–6]. For a detailed discussion on related work the reader is referred to [7]. While most studies have utilized fixed allocation of a set of resources for profiled users, such approaches would lead to higher QoS at the cost of low resource utilization. Some studies propose a more dynamic approach [8, 9] but do not provide a concrete mechanism to implement such ideas in the real network. Our proposal dynamically and efficiently allocates resources only in those target cells where the profiled user is most likely to handoff to, thus achieving high resource utilization. A key ingredient of this dynamic reservation is an algorithm based on path and service prediction using statistical tables of user profiles generated and stored in network databases.



■ Figure 1. *a) Components of the 3GPP network architecture; b) network elements needed to support profile-based channel reservation.*

The remainder of this article is organized as follows. We describe our profile-based cellular network architecture and our channel-reservation algorithms. We outline the concept of a user profile register (UPR) and discuss its design. We introduce our user-profile-based resource-management algorithm. Implementation details for the 3GPP network architecture are then provided. We then describe our simulation model and illustrate our numerical results. Finally, we conclude the article.

Existing and Emerging Network Infrastructure

Figure 1a shows the components of the 3GPP Universal Mobile Telecommunications System (UMTS) reference architecture[1] [1]. The UMTS network consists of three interacting domains: the core network (CN), UMTS Terrestrial Radio Access Network (UTRAN), and user equipment (UE). UMTS supports Global System for Mobile communications (GSM) radio access as well, for compatibility with older mobiles, but we will only concentrate on the new UMTS elements for clarity.

There are two key entities in UTRAN. The base station is known as the Node B. It is the physical unit for radio transmission and reception within a cell. Depending on sectoring, one or more cells may be served by a Node B. The radio network controller (RNC) manages one or more Node Bs. An RNC is responsible for channel assignment, admission control, and management of handoffs between neighboring Node Bs within its area of control.

The core network has evolved from the GSM core network with General Packet Radio Service (GPRS). It is divided into the circuit-switched and packet-switched domains. Certain components are common to the circuit and packet segments of the UMTS core network. The home location register (HLR) is a database containing user-specific data — such as user identity, subscribed services, and current user location — for all users registered in a particular geographical area, known as the home area. The authentication center (AuC) and the equipment identity register (EIR) are databases that aid in identification and authentication of mobile users.

The circuit-switched part of the UMTS core network consists of the mobile switching center (MSC), visitor location register (VLR), and gateway MSC (GMSC). The MSC constitutes the interface between the radio system and the fixed network. It is responsible for setting up, managing, and clearing calls (connections); routing incoming calls to the appropriate cell; and managing inter-RNC handoffs. The VLR contains information about all users currently "visiting" its particular geographical area. The GMSC is the gateway between the UMTS network and external circuit-switched networks.

The packet-switched elements of the UMTS core network consist of the serving GPRS support node (SGSN) and gateway GPRS support node (GGSN). The GGSN has functionality equivalent to the GMSC, while the SGSN performs duties similar to a combined MSC and VLR, such as routing and visitor management.

Wireless service providers are rapidly shifting focus from simple voice services to mobile LoCation services (LCS) [10]

[1] *We have chosen the 3GPP Universal Mobile Telecommunications System (UMTS) network architecture for our examples, but the concepts outlined in this study are also applicable to other wireless cellular architectures as well.*

that utilize a user's position information to provide localized and personalized services. A few key network elements are required for supporting LCS, and will also be used for our profile-based algorithm described below. The gateway mobile location center (GMLC) is responsible for interfacing with the LCS clients who request the mobile's position. The serving mobile location center (SMLC) and the location measurement unit (LMU) determine the geographical coordinates of the mobile.

User Profile and User Profile Register

We introduce a new architectural component called the user profile register (UPR) which is a database similar to the HLR and VLR. The UPR contains user-profile information that can be queried by the wireless network to provide differentiated services to its customers. A user profile consists of mobility patterns and services accessed by the mobile user, possibly also tabulated against the time of day and day of the week. It contains pointers to network elements that can provide real-time values of the user's location and velocity information. A UPR should have interfaces to external information sources, such as network information databases (described later), to aid in QoS management. The components of the UPR are outlined below.

User Location Interface: This is a logical interface to devices such as the SMLC and the LMUs that can provide real-time user location values to the UPR.

User Velocity Interface: This logical interface will query the network element responsible for real-time velocity estimation of a mobile user [3].

User Path Table (UPT): This table is an ordered list of the most probable paths a mobile user could traverse at any given time on any day of the week. A mobile path is a list of cell-IDs that a mobile user traverses. For example, let $<c_1, c_2, \ldots, c_n>$ represent a mobile user's path, which starts from cell c_1 and ends in cell c_n.

This path could contain a number of *hot-spots*. A hot-spot is defined as a collection of cells within a geographical region where the mobile-user population density is higher than some pre-defined threshold. City downtown regions, train stations, airports, or dense residential areas represent typical hot-spots. Hot-spots could be dynamic in nature in that they can change with the time of day, traffic conditions, and special events. A city downtown can be a hot-spot during the work day, while residential neighborhoods can become hot-spots on weekends. The network can detect a hot-spot near a stadium on Superbowl Sunday, while a snow storm could cause multiple hot-spots along the highways near Lake Tahoe.

An user can terminate her call at any point along the path, $<c_1, c_2, \ldots, c_n>$. If all the cells in the path are contained within a single hot-spot, the path is considered *internal* to the hot-spot, and does not appear in the UPT. Such intra-hot-spot paths are usually short as they are contained within a very limited geographical area, and therefore they do not lend themselves very well to statistical learning algorithms, as there is not enough time during the path traversal to predict or update network databases. Moreover, mobility patterns for calls that are generated and terminated within a microcellular hot-spot environment (e.g., a downtown area) are usually Brownian in nature, and barely show any statistical route preferences. We discuss this issue later.

A User Path Table (UPT) can include two types of paths:

- User-specified: These paths are defined by the user before usage. For example, when signing up for the premium profiled cellular service, a user may define a few *typical* paths she might take during an average workday. Paths can also be defined using mapping services such as MapQuest or using GPS devices by specifying the end points of the route. These pre-calculated routes can then be "pushed" out to the UPR.
- Statistical: Each user's UPT could be initialized to an empty value or to a user-specified list. Subsequently, every time a path is traversed by the user, its ranking is increased in the UPT. Over time the UPT would typically converge to a fixed set of paths, as users statistically take the same routes for the same chores.

Every time the mobile traverses a new path that does not exist in the UPT, an entry is added to the table. Whenever the mobile repeats a particular path the weight of the entry is increased. The table is ordered in decreasing order of weight and, over time, a user profile is developed for the user. The top entries of the UPT would correspond to the most probable path for specific times of day.

User Resource Table (URT): This table is an ordered list of resources (services) used by a mobile user at any given time on any day of the week. The URT contains two types of entries:

- User-specified: The user could supply a list of services to be used for each path that she could take. For example, voice services could be accessed during the drive to work and back, while Web services could be accessed during the lunch break.
- Statistical: The URT could be initialized to an empty value or to the user-specified list above, and subsequently updated by a positive weight every time that service is accessed along the mobile path. The URT will also be constantly updated, and it would converge over time to a virtually static list.

Interfaces to External Information Systems: This entry contains variables obtained from external information systems such as network architecture databases or GPS and Global Information Services (GIS) devices. For example, network architecture databases could provide information about cell layout and sizes. GPS and GIS devices could supply real-time changes in hot-spot definitions since, as discussed earlier, hot-spots can be dynamic in nature and can change depending upon the time of day, traffic conditions, and special events.

User-Profile-Based Differentiated-Services Architecture

In this study we focus on a user-profile-based call QoS control algorithm that utilizes the statistical profiles of mobile users to manage and control the network resources in order to guarantee the negotiated QoS to users. We have considered two classes of service for two types of users:

- Profiled users who subscribe to a premium profiled service and expect better QoS.
- Non-profiled users who pay less and expect a "best-effort" service from the network.

Below we outline our user-profile-based resource-management algorithm.

PARMA: Profile-Assisted Resource-Management Algorithm

Figure 2 shows a part of the trajectory of a user commuting from the Arden Town suburb (not shown) to Richards Boulevard, near downtown Sacramento, California. The procedure to reserve resources for a profiled user that is a core part of the profile-assisted resource-management algorithm (PARMA) is outlined below:

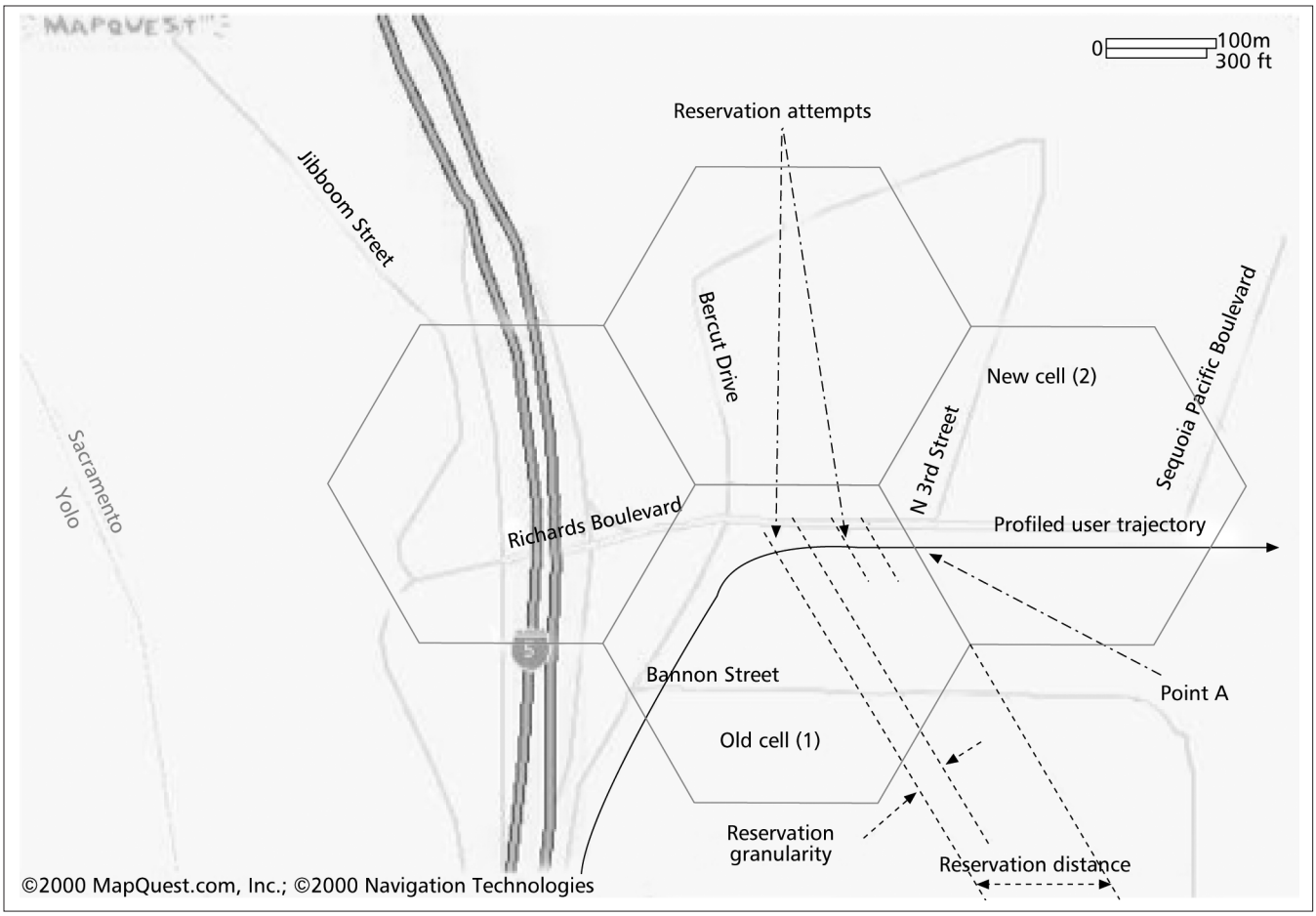


Figure 2. *Illustration of PARMA.*

1 When a mobile user is *close* to a cell boundary the network would query the subscriber database to determine if the user is a profiled customer.
2 If the user is not a subscriber for profiled services then normal handoff procedure would follow, and if there are no channels available in the target cell (Cell 2 in this example) the user's call will be dropped.
3 If the user subscribes to profiled services the network will query and extract the user's resource requirements from the URT. For the purpose of this discussion we will assume that the user has subscribed to voice services only and hence would require a channel to be reserved in the target cell.
4 The network tries to predict the target cell, based on the UPT and the current location and velocity of the mobile.
5 The network would then attempt to reserve a channel in advance for the user in Cell 2 when the user is at a distance r_d from the cell boundary, where r_d is known as the *reservation distance*.
6 If the reservation attempt succeeds, the user is handed off to Cell 2 on the reserved channel at the cell boundary, which is shown as point A in Fig. 2. If the reservation fails, the network re-attempts the reservation every r_g (*reservation granularity*) distance apart, till the reservation succeeds or till a handoff takes place at point A. This re-attempt could be explicit if the intelligence resides in the mobile device, or it could be performed implicitly by the MSC as it already has the reservation request information.
7 If the reservation is unsuccessful till Point A (even after several attempts) then the user session gets dropped. By allowing multiple reservation attempts the dropping probability of a profiled user can be substantially reduced.

As mentioned in Step 1, PARMA is initiated when the mobile user is close to a cell boundary. There are two key approaches to proximity evaluation, and both approaches are conceptually equivalent. In the first approach the mobile device measures the signal strengths from adjacent base stations and reports this information to the Node B. The network could then decide, based on these measurements, whether the user is "close enough" to initiate PARMA. In the second approach the network could keep track of the velocity (speed and direction) of the user. Knowing the trajectory of the user from the user profile, the network can then calculate when handoff would occur. Therefore, based on this calculation the network can decide when to initiate PARMA. In our study we choose the first approach to measure a user's proximity to a cell boundary.

We have conceptually used the parameters r_d and r_g to initiate reservation and to retry the reservation attempts until they succeed or the mobile has had a handoff. PARMA is not specifically tied to these parameters, and any other initiation or retry triggers could be used. For example, if the signal strength at a Node B or the RNC falls below a pre-defined threshold (s_d), the first reservation attempt could be initiated. Henceforth, whenever the signal strength falls lower by pre-defined decrements s_g, a reservation could be re-attempted.

Although we have considered channel reservation to highlight our algorithm, PARMA is much broader in scope. Any user-specific resource, as defined in the URT for the profiled user, can be allocated for in the target cell. Resources such as the browser cache for mobile browsers, session and state information for data connections, application proxy states for thin clients running on mobile devices, etc., can all be reserved in the target cell through PARMA.

PARMA is not limited to only two classes of users, that is, profiled and non-profiled users. The profiled users can be classified into multiple categories, each with different QoS guarantees, and hence, reservation requirements. For example, PARMA can be used to reserve resources in more that one cell, depending on the user's priority class within the profiled user base. As an illustration, emergency vehicles such as ambulances can be given the highest priority, and, for a given source and destination (such as a hospital), resources can be reserved all along the path. This will result in uninterrupted service between the paramedics in the ambulance and the doctors in the hospital.

Design and Implementation Issues

In this section we provide guidelines to implement PARMA in a 3GPP network architecture. First we describe the modifications we need to perform to the existing network components to support PARMA, and then we propose a possible implementation.

Modifications to 3GPP Network Elements

We can leverage the location-measurement infrastructure described earlier for obtaining current updates to the position of a profiled user. Additionally, we must modify some of the network elements to support PARMA, as described below.

- The SMLC should be modified to store a short history of the mobile's position instead of storing only the current position. The size of this history depends on the accuracy of the path-prediction algorithm.
- The UPR database should be implemented to include the profile tables and real-time values for each mobile user in her home area. Each UPR should also be able to accept and incorporate updates to user profiles available from the MSC. To keep storage requirements small, less probable paths and services could be phased out over time. The network could employ a multi-level virtual-memory-like storage mechanism and caching schemes such as Least Recently Used (LRU) to keep a subset of the table space in faster memory while phasing out the rest of the information to permanent storage. Given the recent advances in storage technology, network designers can lean toward enabling longer history for more accurate path prediction.
- Logical interfaces should exist between the UPR and external information databases and systems, such as the HLR, network architecture databases, the LMUs, and the SMLC. The HLR interface would aid the UPR in gathering subscription information about a user. Network architecture databases would provide information on hot-spot definitions and cellular layouts. LMUs and the SMLC would provide current location information and would assist in user-velocity estimation.
- The software in the MSC should be enhanced to statistically update the UPT and the URT as described earlier. The MSC should be able to make corrections to the user profile depending upon the success or failure of the user-profile-based path-prediction process and the services that the user has accessed, by increasing the "rank" of successfully predicted paths and resource requirements in the UPT and URT, respectively. The MSC should then provide this feedback to the UPR database.

Implementation of PARMA

Figure 3 shows the message sequence chart for PARMA. In current networks a mobile device periodically sends a list of Node Bs and their signal strengths to the current Node B for the purpose of a handoff, using a *SignalStrengthList* message. In UMTS the list of Node-Bs to which the mobile is currently connected is known as the Active Set. The set of cells that are not in the Active Set, but are being monitored for possible handoff, form the Monitored Set. The current Node B, with help from the MSC, uses this information to decide the target handoff cell. We can modify the Node B software to trigger a *PredictTargetCell* signal to the MSC whenever the signal strength of another Node B comes within a trigger threshold (S_T) of the signal from the current Node B. We study the impact of this threshold on network performance later.

On receiving the *PredictTargetCell* signal, the MSC sends out a *GetLocation* message to the SMLC requesting the past few coordinates (positional history) of the mobile user. The SMLC functionality can be modified to keep track of the positional history of a mobile user. The length of this history can vary, depending on the accuracy of the path-prediction algorithm. On receipt of *GetLocation*, the SMLC forwards the positional history of the mobile to the MSC. For performing path prediction, the MSC also requires the user's path profile (UPT) and the current velocity of the user, which are stored in the UPR database. It also requires the URT for gauging the resource requirements of the mobile in the target cell. The *GetProfile* and *SendProfileList* messages accomplish this task.

On receiving the current location, velocity, and the UPT, the MSC performs path prediction and informs the most probable target Node B to reserve resources for the mobile depending on the most probable services accessed by the customer. If the trigger threshold S_T is designed accurately, the handoff will occur immediately after the *AllocateResource* message, and the transition to the new cell will be smooth. After the mobile undergoes the handoff, the new Node B sends a status update to the MSC through the *HandoffStatus* signal and includes its own Cell-ID. The MSC checks the Cell-ID to confirm whether this was the cell to which it had sent the *AllocateResource* message. *HandoffStatus* also mentions whether the resource reservation was sufficient or whether the Node B had to allocate fewer or more resources. Finally, the MSC updates the UPT and the URT at the UPR based on the results of the comparison in the previous step.

Design Issues

There are several issues to consider when implementing PARMA in a wireless network. In this section we outline some design issues, and we generalize on the basic scheme presented earlier.

Speed-Modulated Reservation Granularity: When calculating reservation re-attempts to the target cell, PARMA could take the velocity measurements into account to dynamically modify the reservation granularity. The speed at which a user is traveling toward the target cell would determine how small or large the reservation granularity should be.

Direction-Modulated Reservation Granularity: The direction of movement of a profiled user with respect to the cell boundaries also plays an important role in the implementation of PARMA. Consider a user who is moving almost tangentially to the cell boundary. The mobile device's list of strongest Node B signals could change very frequently over time, hence triggering frequent reservation attempts. Therefore, by using direction information coupled with the network's information of cell boundaries, PARMA can employ techniques such as hysteresis and signal thresholds — which are used to avoid the ping-pong effect in handoffs [11] — to reduce unwanted reservation re-attempts.

Hot-Spot Modulated Reservation and Impact of Link Loss Characteristics: The number of permutations of highways and freeways that a user can take to travel between a pair of hot-spots is generally expected to be quite small. Therefore, over time inter hot-spot portions of a UPT entry tend to converge

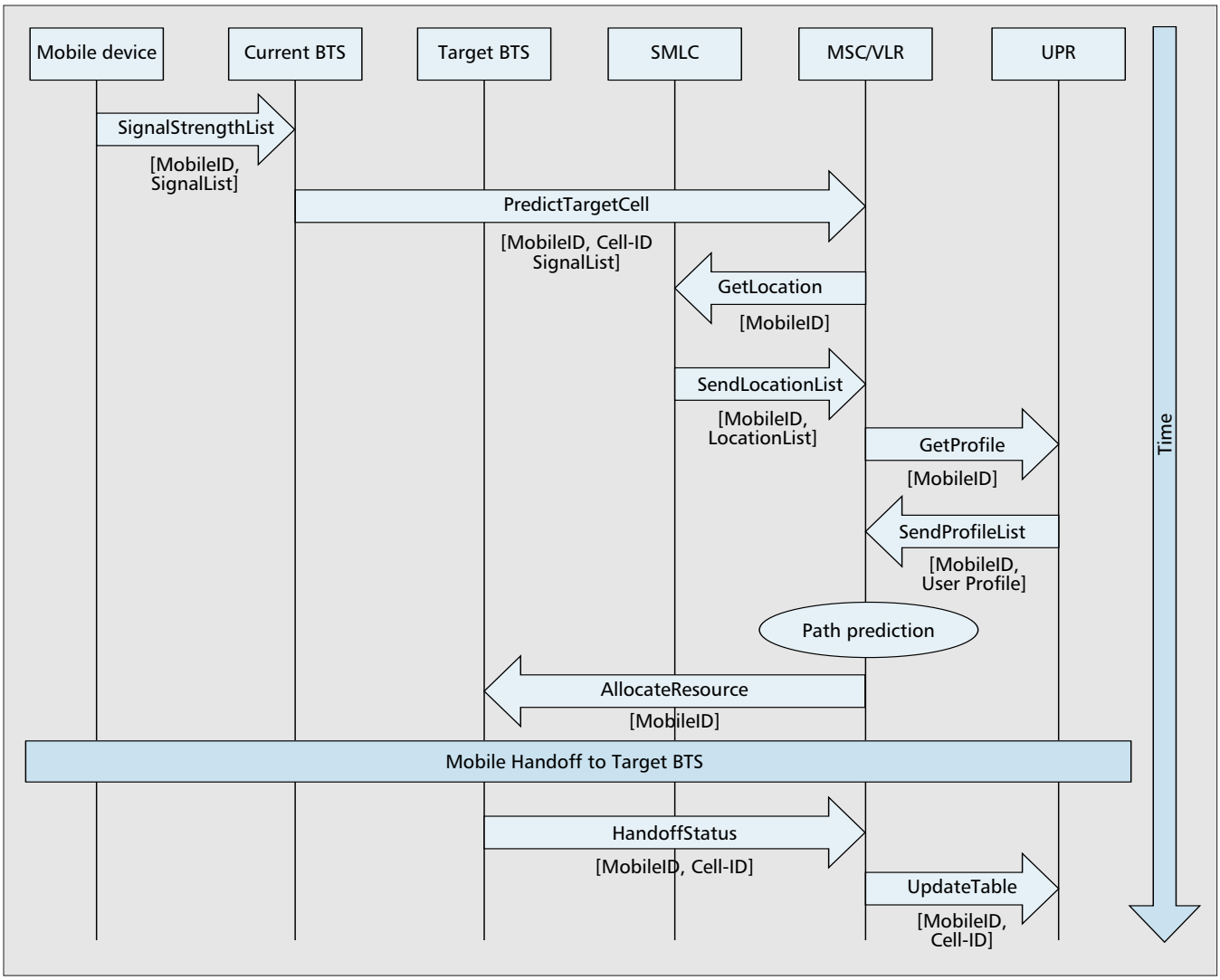


Figure 3. *Message sequence chart for PARMA.*

to a few fixed paths. Hence, path prediction can be performed on inter hot-spot paths with sufficient accuracy by consulting the UPT entries for the user. Moreover, fast table lookups will reduce overheads for resource reservation.

Intra hot-spot paths are more difficult to predict since the number of permutations of streets that can be taken to reach the destination could be significantly larger as compared to the inter hot-spot scenario. Once the network has determined that the user has entered an intra hot-spot region of her mobile path, the network should solicit help from velocity-estimation algorithms and location-measurement technologies to better predict the user's target cell.

For intra-hotspot paths, link loss characteristics will have some adverse impact on path prediction and profile accuracy. PARMA currently reserves resources in exactly one predicted target cell. For areas such as city centers we can expand PARMA to reserve resources in a subset of the Active Set. Typical Active Set sizes are anywhere between two and three cells [12, 13], and these are the cells to which the mobile is most likely to handoff. The positional history of the mobile, coupled with the current speed and direction measurements, will provide better rules for selecting the most probable set of paths, and hence the most probable subset of target cells in the Active Set where PARMA can allocate resources. Once the handoff completes, or after a fine-tuned timeout, the resources in the remaining cells can be released.

Macrocell Reservation: Hot-spots are regions of high mobile-user population density. For such geographical areas the concept of a microcellular/macrocellular multi-tier architecture has been proposed to provide greater channel capacity through microcells while providing better management using the overlayed macrocells.

If all resource-reservation attempts are unsuccessful, a handoff takes place; and if there are no available resources in the target microcell, the overlayed macrocell can be used to temporarily "hold" the session (and resources) of a profiled user. Under these circumstances, PARMA would keep re-attempting the reservation requests in future target microcells. The goal would be to vertically handoff the session to a microcell at the first available opportunity.

Reservation Granularity versus Round Robin: PARMA introduces the concept of reservation granularity. Resource-reservation requests are *re-attempted* after every r_g distance until the request succeeds or a handoff takes place, resulting in a success or a dropout.

Instead of the above strategy, all resource-reservation requests could be queued in the target cell if they were unsuccessful on the first attempt. The network could then use a round-robin scheme to scan through the queue and re-attempt each request. The request would be dequeued on a reservation success or on a successful or unsuccessful handoff.

Signaling Overheads: Though we have shown nine mes-

sages for implementing PARMA (Fig. 3), the overheads are minimal as most messages can be piggybacked on existing signals, as described below. To aid in handoffs, a mobile device periodically sends the list of closest Node Bs to its current Node B. As mentioned earlier, this information is contained within the Active Set and Monitored Set for UMTS. The current Node B can use the same message, coupled with a threshold S_T, to trigger the *PredictTargetCell* message. *GetLocation* and *SendLocationList* will soon be implemented in the 3G networks to provide LCS. *GetProfile* and *SendProfileList* can be piggybacked onto HLR database queries for authentication and authorization most of the time (if UPRs are implemented as a part of the HLR database). Moreover, user profiles can be cached in the VLR database inside the MSC and updated at the UPR once the mobile leaves that location area. The *AllocateResource* signal is sent to a target Node B after a handoff decision has been made. For PARMA, we send the signal just after path prediction, imposing no extra overheads. Similarly, the *HandoffStatus* signal is sent irrespective of whether PARMA has been implemented.

The additional signals in PARMA include the *PredictTargetCell* and *UpdateTable* messages.

Discussion of Storage Space: Let us attempt to approximately gauge the storage requirements for implementing the UPR in a wireless network. For a quick "back-of-the-envelope" calculation, let us assume that there are up to a total of 65,000 cells in the geographical area being served by a single UPR. A cell-ID can then be represented by 2 bytes, either in a flat addressing scheme or in a hierarchical scheme for faster addressing. Over the duration of a call the mobile would typically traverse on the order of 10 cells. This would imply that an average path in the UPT is 20 bytes long. Since the UPT is a sparse graph it can be represented by an adjacency list of size $p \times l$, where p is the number of paths stored, and l is the length of the path. The number of "most likely" paths in the UPR per profiled user would impact the accuracy with which user profiling can be used to predict the mobile's current trajectory. Typically, the UPT of a user would contain two or three "most likely" paths per weekday, and maybe three or four paths for the weekend, giving us a total of approximately 10 paths per user. Hence, the mobile network would expense approximately 200 bytes per user for path information. As for the services accessed by any user, a couple of bits (lets say a byte) would suffice. Savvier ways of path representation can probably pull the storage requirements even lower.

Discussion of Running Time: The above implementation guidelines show that PARMA and path prediction are not resource-heavy algorithms. On the other hand, the duration of the call may be short enough such that no useful path information could be gathered. This section discusses the possible interactions between the duration of the call and the running time of PARMA. Let us consider the following two cases.

- If the call is very short the caller would most likely stay within the same cell and would not generate any handoffs. In this case, when the network verifies the validity of the caller it can also check whether the user is profiled and handle the call with the appropriate priority. Since there are no handoffs, reservations are never attempted and the impact of the running time of PARMA is not under consideration.
- If the call is long enough to generate handoffs, we believe that a profile-based reservation procedure is not much more expensive than the steps required to service a handoff. As can be seen from Fig. 3 and the related description, PARMA requires only two messages over and above what is already used today in handling a handoff. Most of the other information is piggybacked on existing message types. Additionally, the decision and update algorithms in PARMA try to model themselves after currently available user access validation and handoff procedures, and the extra cycles to update and query the UPT are quite lightweight with regard to running time.

Moreover, path prediction is an iterative process that converges over time. The statistical UPT is built over time and will converge to a fairly accurate list of paths traversed by the user. For a new call the network first assumes the most popular path in the UPT. Over the duration of the call, with the help of velocity-prediction schemes and external information sources such as GPS co-ordinates, the network can construct a weighted view (over the path in the UPT and the current GPS values) of what the user's trajectory might look like. Thus, even though there might be slight deviations from the statistical paths in the UPT, these variations will result in over-reservations only in certain cells, which is a slight penalty to pay for better resource utilization in the average case.

A Quantitative Analysis

We have studied the performance of PARMA using detailed simulation experiments. Even though we have tried to create an accurate simulation model, we had to make a few deviations from the practical implementation to better analyze the problem. The following section describes our simulation model and how it relates to the practical implementation.

Simulation Model

We have simulated a single-tier cellular network architecture. The primary resource accessed by cellular users in this network are channels. There are a limited total number of channels, C, available to the network which are allocated to cells using fixed-channel allocation with a static reuse pattern. The reuse distance ratio, i.e. the ratio of the cell radius to the distance from the center of the cell to the next co-channel cell, is denoted as R.

We employ a hexagonal cell structure with a cell radius of c km. We model a limited user population of U users at any given time in the network, out of which a fraction p of the users are profiled. To model the density of users within hot-spots, we use the following algorithm, assuming a circular hot-spot with a radius of h:

1 Choose (or define) the hot-spot radius, h.
2 Choose an angle (δ) as a uniform random number between 0 and 2π.
3 Once δ is fixed, choose a uniform random number ρ between 0 and h. Place a user at an angle δ and at a distance of ρ from the center.
4 Repeat steps 2 and 3 for each user in a hot-spot.

There are H hot-spots in the region covered by the network. The size and location of these hot-spots can be defined using a graphical interface (which we have developed) to the simulation. For our investigation we assume that each user has a different direction of movement (D) and speed (V). For the purpose of this study we assume that we can accurately predict a user's trajectory at every given point in time.

PARMA requires a threshold to trigger the channel-reservation algorithm when a profiled user travels close to the edge of a cell. We have used a reservation distance r_d to trigger PARMA, while in practice a threshold on the received signal strength can be used for triggering the algorithm, as described earlier.

We assume that new-call arrivals into the network follow a Poisson distribution with parameter λ calls/sec. Call-holding time is assumed to follow an exponential distribution with a mean of $1/\mu$ seconds.

Parameter	Description	Value
H	Number of hot-spots	3
h	Hot-spot radius	3 km
A	Number of cells in region	20×20
c	Cell radius	0.5 km
C	Total number of channels in the network	200
R	Reuse distance ratio	2
V	User speed (range)	40-65 km/h
U	Total number of user streams in the network	5,000
p	Fraction of profiled user streams	0.4
r_d	Reservation distance	0.2 km
r_g	Reservation granularity	0.01 km
$1/\mu$	Mean call-holding time	120 sec
λ	New-call arrival rate per user stream	0.025 calls/s

■ Table 1. *Default values of the network parameters.*

We have used *Improvement in Dropping Probability*, denoted by γ, as an important performance metric. This is defined as the reduction in dropping probability of a profiled user as compared to a non-profiled user. Specifically, if P_{dp} is the dropping probability for profiled users and P_{dn} is the dropping probability for non-profiled users, then γ is given by:

$$\gamma = \frac{P_{dn} - P_{dp}}{P_{dn}} * 100\% \quad (1)$$

The default parameter values are shown in Table 1. Assuming hexagonal cell approximations, fixed channel allocation and $R = 2$ leads to a three-cell cluster, that is, the frequency assignment pattern repeats every three cells. Given that $C = 150$ or 200 channels over $A = 20 \times 20$ cells, we have approximately 50 to 67 channels per cell. The total number of user streams in the network is fixed at 5,000, out of which 2000 streams are profiled, on average. This implies that each cell inherits an average of 12.5 user streams. Each user stream generates calls as a Poisson process, with mean in the range of [0.005–0.05] calls/second. This leads to a cumulative Poisson call arrival rate in the range [0.0625–0.625] calls/second for each cell.

Results and Discussion

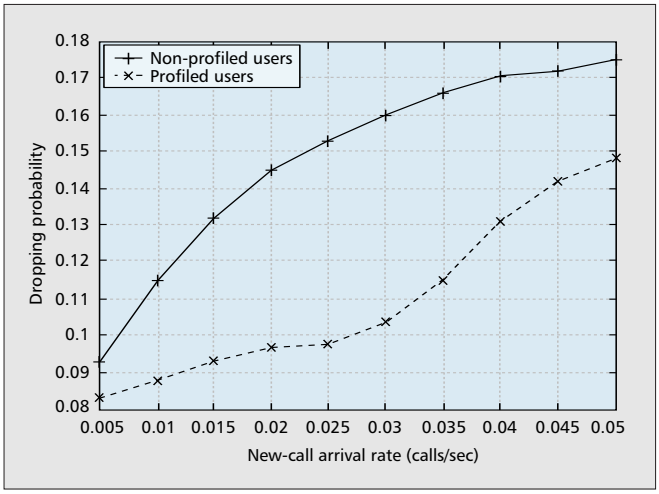
Figure 5 shows the improvement in dropping probability, γ, experienced by profiled users as the new-call arrival rate is increased from 0.005 calls/sec to 0.05 calls/sec for $C = 150$ and $C = 200$ channels. Figure 4 shows the dropping probability for profiled and non-profiled users as a function of the call-arrival rate for $C = 200$ channels. As expected, dropping probability for profiled users is significantly lower than that for non-profiled users. At the left extreme of both figures, when λ is small the load to the network is very light. Hence, very few handoff attempts of both profiled and non-profiled users are dropped. Therefore, reserving channels for profiled users in advance does not result in a large γ. At the other extreme, when λ is large the network load is high. A significant number of the channel-reservation requests for profiled users get blocked, also resulting in a small γ. When the network load is moderate the profiled users obtain the most benefit from channel reservation. As can be observed in Fig. 4, the dropping probability for non-profiled users keeps increasing even at moderate load, while the dropping probability for profiled users starts to flatten out, benefiting from the reservations. This results in a substantial improvement in dropping probability, with a peak occurring at $\lambda = 0.025$ calls/sec (for $C = 200$ channels), when we observe an improvement of 35.8 percent.

Through our experiments (whose results have not been shown for brevity) we also observed that there is an optimal value of r_d (at $r_d = 0.2$ km) which results in the maximum γ. When r_d is small (e.g., at $r_d = 0.05$ km), the reservation attempts are made too close to the cell boundary. Hence, there is not enough time to recover from a failed reservation attempt before the handoff occurs. When r_d is too large (e.g., at 0.25 km and beyond) the channel-holding time of profiled users is inflated by a large amount, which causes the overall load in a cell to increase. Additionally, we observed an optimal value for r_g, which resulted in the maximum γ. When r_g is small the network makes a large number of reservation attempts on behalf of the profiled user, resulting in higher load to a cell, and hence a large dropping probability. When r_g is large there are not enough reservation re-attempts for profiled users. For these and additional results the reader is referred to [7].

Conclusion

With the continuing deployment of intelligent network components, it is becoming easier to collect and maintain accurate real-time data on the location, velocity, and resource requirements of a mobile user. These data can be exploited to develop user profiles, and they can be aggregated to develop mobility and resource-requirement patterns of users in a region. We have made the following contributions in this work:

- We have described the design and implementation of a new scheme, called Profile-Assisted Resource-Management Algorithm (PARMA), which utilizes user profiles to provide better QoS to mobile users in a wireless network. Specific implementation details have been proposed for the 3GPP network architecture, though the concepts would be broadly applicable to most wireless network architectures.



■ Figure 4. *Dropping probability for different new-call arrival rates at C = 200 channels.*

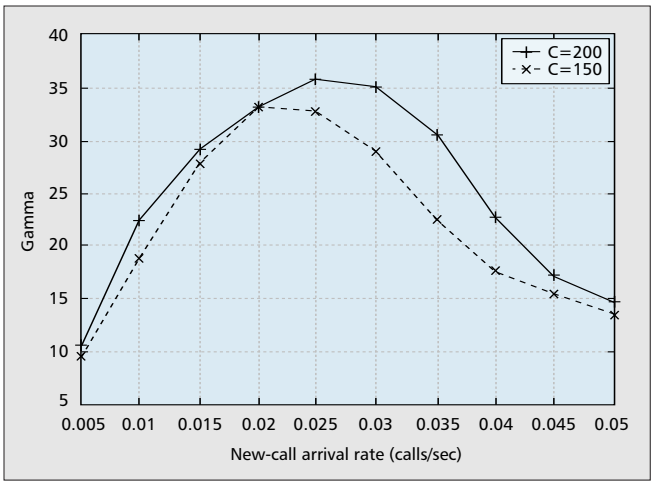


Figure 5. *γ for different new-call arrival rates.*

- There are numerous challenges and design issues in implementing such a scheme for the next-generation wireless networks, and we attempted to resolve some of these design issues.
- Through detailed simulation, we have studied the benefits of user profiles by analyzing a resource-allocation scheme using the concept of reservation distance and reservation granularity. We have shown that this concept can produce significant improvement in dropping probability of profiled users over their non-profiled counterparts. We observed that there are optimal values of the reservation distance and the reservation granularity parameters, which result in maximal improvement in dropping probability.

Acknowledgments

We would like to thank the anonymous reviewers and the editors, Dr. Peter O'Reilly and Dr. Chatschik Bisdikian. Their comments and feedback significantly improved the quality of this article.

References

[1] 3rd Generation Partnership Project, Network Architecture (T1TRQ3GPP 23.002-330), Standards Committee T1 Telecommunications, version 3.3.0 edition, 2000.
[2] J. S. M. Ho and I. F. Akyildiz, "Dynamic Mobile User Location Update for Wireless PCS Networks," *ACM/Baltzer Wireless Networks*, vol. 1, no. 2, Dec. 1995, pp. 187–96.
[3] B. Jabbari and W. Fuhrmann, "Teletraffic Modelling and Analysis of Flexible Hierarchical Cellular Networks with Speed-Sensitive Handoff Strategy," *IEEE JSAC*, vol. 15, no. 8, Oct. 1997, pp. 1539–48.
[4] S. Tabbane, "An Alternative Strategy for Location Tracking," *IEEE JSAC*, vol. 13, no. 5, June 1995, pp. 880–92.
[5] M. H. Chiu and M. A. Bassiouni, "Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning," *IEEE JSAC*, vol. 18, no. 3, Mar. 2000, pp. 510–22.
[6] A. Bhattacharya and S. K. Das, "Lezi-Update: An Information Theoretic Approach for Personal Mobility Tracking in PCS Networks," *ACM/Baltzer Wireless Networks*, vol. 8, no. 2–3, Mar.–May 2002, pp. 121–35.
[7] V. Pandey, D. Ghosal, and B. Mukherjee, "PARMA: A Profile-Assisted Resource Management Algorithm for Improving QoS in Wireless Networks," Tech. Rep., University of California, Davis, 2004.
[8] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks using the Shadow Cluster Concept," *IEEE/ACM Tran. Net.*, vol. 5, no. 1, Feb. 1997, pp. 1–12.
[9] F. Yu and V. C. M. Leung, "Mobility-Based Predictive Call Admission Control and Bandwidth Reservation in Wireless Cellular Networks," *Proc. IEEE INFOCOM*, Anchorage, AK, April 2001, vol. 1, pp. 518–26.
[10] Nokia, http://www.nokia.com, Mobile Location Services, 2001.
[11] G. P. Pollini, "Trends in Handover Design," *IEEE Commun. Mag.*, vol. 34, no. 3, Mar. 1996, pp. 82–90.
[12] 3rd Generation Partnership Project, "3G-TR-25.942, RF System Scenarios," Tech. Rep., 3rd Generation Partnership Project, 2000.
[13] "Evaluation Report for ETSI UMTS Terrestrial Radio Access (UTRA) ITU-R RTT," Tech. Rep., ITU-R, 1998.

Biographies

VIJOY PANDEY (vijoy@nortelnetworks.com) received the B.Tech.(Hons.) degree from the Indian Institute of Technology, Kharagpur, in 1995, and an M.S. degree from the University of California, Davis in 1997. He is currently pursuing his Ph.D. at UC Davis while working in the ethernet switching group at Nortel Networks in Santa Clara, California. At UC Davis he was nominated for the Professors for the Future Fellowship Award in 1999. His research interests include architectures and protocols for next-generation wireless networks, and intelligent packet switching for secure wired and wireless local area networks.

DIPAK GHOSAL (ghosal@cs.ucdavis.edu) received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1983, the M.S. degree in computer science from the Indian Institute of Science, Bangalore, India, in 1985, and the Ph.D in computer science from the University of Louisiana, Lafayette, in 1988. From 1988 to 1990 he was a research associate at the Institute for Advanced Computer Studies at the University of Maryland (UMIACS) at College Park. From 1990 to 1996 he was a member of technical staff at Bell Communications Research (Bellcore) in Red Bank, NJ, USA. Currently he is with the faculty of the computer science department at the University of California, Davis. His research interests are in the areas of IP telephony, peer-to-peer systems, mobile and ad hoc networks, and performance evaluation of communication systems.

BISWANATH MUKHERJEE (mukherje@cs.ucdavis.edu) [S'82 M'87] received the B.Tech. (Hons) degree from the Indian Institute of Technology, Kharagpur (India) in 1980 and the Ph.D. degree from the University of Washington, Seattle, in June 1987. At the University of Washington he held a GTE Teaching Fellowship and a General Electric Foundation Fellowship. In July 1987 he joined the University of California, Davis, where he has been a professor of computer science since July 1995, and where he served as chairman of the computer science department from September 1997 to June 2000. He is a winner of the 2004 Distinguished Graduate Mentoring Award at UC Davis. Two PhD dissertations (by Laxman Sahasrabuddhe and Keyao Zhu) supervised by Professor Mukherjee were winners of the 2000 and 2004 UC Davis College of Engineering Distinguished Dissertation Awards. He is co-winner of paper awards presented at the 1991 and the 1994 National Computer Security Conferences. He serves or has served on the editorial boards of the *IEEE/ACM Transactions on Networking*, *IEEE Network*, *ACM/Baltzer Wireless Information Networks* (WINET), *Journal of High-Speed Networks*, *Photonic Network Communications*, and *Optical Network Magazine*. He also served as Editor-at-Large for optical networking and communications for the IEEE Communications Society. He served as the technical program chair of the IEEE INFOCOM'96 conference. He is author of the textbook *Optical Communication Networks* published by McGraw-Hill in 1997, which received the Association of American Publishers, Inc.'s 1997 Honorable Mention in Computer Science. He is a member of the board of directors of IPLocks, Inc., a Silicon Valley startup company. He has consulted for and served on the technical advisory board of a number of startup companies in optical networking. His research interests include lightwave networks, network security, and wireless networks.