Minimum-Cost Data Delivery in Heterogeneous Wireless Networks

Haining Chen, *Student Member, IEEE*, Hongyi Wu, *Member, IEEE*, Sundara Kumar, *Student Member, IEEE*, and Nian-Feng Tzeng, *Senior Member, IEEE*

***Abstract*—With various wireless technologies developed over the past few years, a ubiquitous and integrated architecture is envisioned for future wireless communication. An important optimization issue in such an integrated system is how to minimize the overall communication cost by intelligently utilizing the available heterogeneous wireless technologies while, at the same time, meeting the quality-of-service requirements of mobile users. In this paper, we first identify the cost-minimization (CM) problem to be NP-hard. We then present an efficient minimum-cost data-delivery algorithm based on linear programming (LP), with various constraints, such as channel bandwidth, link costs, delay budgets, and user mobility, taken into consideration. In case of insufficient bandwidth for communication with the core network, prefetch is employed to fully utilize the wireless-network capacity. If multiple routes are available, a probability-based approach is taken for CM. Extensive simulations are carried out to evaluate the proposed CM scheme. Our results show that the proposed LP approach can effectively reduce the overall communication cost, with small overhead ($< 3\%$) for signaling, computing, and handoff. We expect that minimum-cost data delivery will become imperative for the future heterogeneous wireless networks and the emerging 4G wireless systems.**

***Index Terms*—Cost minimization (CM), heterogeneous wireless networks, linear programming (LP), quality of service (QoS).**

I. Introduction

With VARIOUS network characteristics and commercial concerns, a number of wireless technologies have been developed over the past few years, and they are likely to coexist for many years to come. For example, the cellular systems [1]–[3] have evolved from the first-generation analog system to the second-generation digital system, and they are presently entering the era of 3G that supports not only voice but also data traffic at a speed of up to 2 Mb/s, while the 4G system is under development for achieving a data rate that is ten times higher. On the other hand, a series of complementary IEEE standards, including 802.20 [4], 802.16e [5], 802.16 [6], 802.11 [7], and 802.15 [8], have been developed or are currently under development to effect data communication in mobile and fixed broadband wireless-access networks, local- and metropolitan-area networks, and personal-area networks, respectively. In particular, 802.20 and 802.16e target at mobile broadband wireless-access networks, providing users moving at vehicular speed with a data rate from 1 to 30 Mb/s in a wide area. 802.16 offers fixed broadband wireless-access network with data rate up to 75 Mb/s, which can be allotted to T1-level connections for business customers and/or to the best effort DSL-speed connections for home customers. 802.11 supports low-mobility users in small cells, at the data rates varying from 1 to 54 Mb/s. Recently, this cost-effective technology is being deployed aggressively for establishing metro-scale "cellular WiFi" networks [9] to support seamless Internet access in urban areas. In addition to aforementioned terrestrial communication systems, the satellite [10] is a vital component in the wireless system, providing global coverage and high-speed data transmission.

While most of these wireless technologies are deployed independently for now, the service providers have most interest to own and operate overlaid heterogeneous wireless systems, which integrate multiple wireless technologies with partially overlapped coverage areas and provide ubiquitous network service to mobile users. For example, several mobile carriers such as Verizon, Sprint PCS, and T-Mobile are anxious to include wireless LAN (or WiFi) access among their service offerings. In order to access various wireless networks/technologies, the mobile host (MH) may be equipped with one or multiple programmable wireless-interface card(s) (e.g., based on the programmable radio technology [11], [12] or an approach similar to mobile-access router [13]), resulting in twofold flexibility that may enable the optimization of data delivery: 1) An MH may select one of multiple available wireless-access technologies at a particular location, because one area may be covered by multiple wireless networks with different costs, data rates, and mobility-support capabilities, and 2) an MH may use different access technologies when it travels in the network and arrives at different locations covered by various wireless networks. From the standpoint of the service provider, it is an important issue to minimize the overall communication cost by intelligently using the available heterogeneous wireless technologies.

In this paper, we consider a typical scenario where an MH X is involved in massive data transmission while traveling (or staying, as a special case). For example, MH X may participate in a large peer-to-peer (P2P) network, where the members share resources such as movie files [14], [15]. Given the large data volume and the limited link capacity, a long data-transmission time (e.g., up to hours) may be expected, during which MH X needs to serve as either a receiver or a data

Manuscript received November 24, 2005; revised December 24, 2006 and February 4, 2007. This work was supported in part by the U.S. Department of Energy (DoE) under Award DE-FG02-04ER46136, by the Board of Regents, State of Louisiana, under Contract DOE/LEQSF(2004-07)-ULL, and by the National Science Foundation CAREER Award under Award CNS-0347686. The review of this paper was coordinated by Prof. T. Hou.

The authors are with the Center for Advanced Computer Studies, University of Louisiana, Lafayette, LA 70504 USA (e-mail: hxc5633@cacs.louisiana.edu; wu@cacs.louisiana.edu; sxk6124@cacs.louisiana.edu; tzeng@cacs.louisiana.edu).

Digital Object Identifier 10.1109/TVT.2007.901049

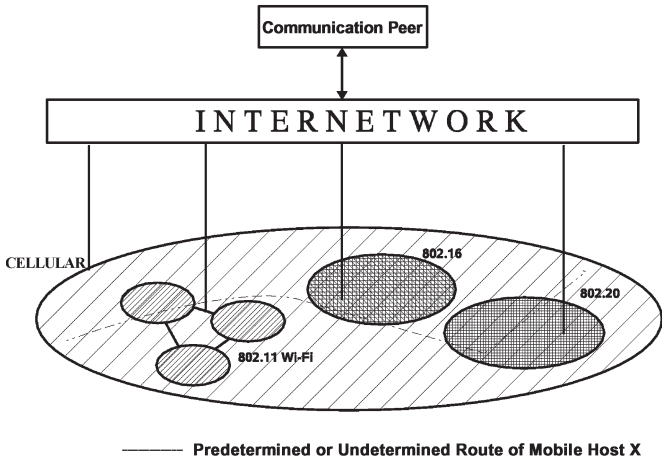

Fig. 1. Scenario of a heterogeneous wireless environment.

source continuously upon its travel or stay. MH X accesses the Internet through available heterogeneous wireless links in order to communicate with its peers. In addition, certain quality-of-service (QoS) requirements (e.g., delay) may be called for, depending on the type of applications. For example, the user may require downloading a movie file and playing it at the same time, and thus, the delay requirement is associated with such an application. The objective is to minimize the communication costs of the data transmission to/from the MH while meeting the QoS requirements.

Intuitively, the least expensive technology available should always be employed for data transmission in order to minimize the communication cost. This naive approach, however, does not guarantee QoS requirement, since the low-cost networks may not provide sufficient bandwidth to achieve the required QoS. At the same time, the low-cost networks are not available anytime anywhere either. For example, when MH X in Fig. 1 is covered by high-cost cells only, continuing aggressive data transmission will noticeably increase the total communication cost. Thus, MH X may defer its data transmission or decrease its data rate until it enters the coverage of high-speed low-cost cells, as long as the delay budget can be met. Although a simple greedy algorithm can be derived based on intuition, it is nontrivial to develop effective schemes in minimizing the communication cost while simultaneously meeting QoS constraints. In this paper, we propose a linear-programming (LP) [16] algorithm that takes into consideration such constraints as channel bandwidth, link costs, delay budget, and user routes. In case of insufficient bandwidth for communication with the core network, prefetch is employed to fully utilize the wireless-network capacity. When multiple routes are available, a probability-based approach is taken for cost minimization (CM). Extensive simulations are carried out to evaluate the performance of the proposed approach. Our results show that the LP algorithm can effectively reduce the overall cost of data transmission.

The rest of this paper is organized as follows. Section II discusses background and related work. Section III introduces the system architecture and signaling protocols. Section IV defines the CM problem and proves it to be NP-Hard. Section V presents our proposed LP algorithms. Simulation results are illustrated in Section VI. Finally, Section VII concludes this paper.

II. Related Work

A series of efforts have been made, so far, to enable and enhance the integration of heterogeneous wireless networks, focusing on the system framework that supports interoperable signaling, seamless roaming, Internet accessing, security, authentication, authorization, and billing. For example, led by telecommunication industry, some initial steps have been taken for integrating the emerging wireless LAN's and the well-established cellular systems [17], [18]. On the other hand, there are several proposals from the Internet society [19]–[21], aiming to efficiently implement Internet protocols in wireless networks for connecting mobile ad hoc networks to the Internet. In the meantime, new techniques, such as iCAR [22], MACA [23], PARCELS [24], and Multi-hop Cellular [25]–[27], have been proposed for employing *ad hoc* technology to improve the performance of cellular systems, or vice versa.

CM has been long recognized as a critical design issue in heterogeneous wireless networks. Stemm and Katz [28] consider an overlaid structure of room-size (e.g., infrared LAN), building-size (e.g., IEEE 802.11b wireless LAN), and wide-area data networks (e.g., 2.5G or 3G cellular systems), where the lower the level of overlay, the smaller area it covers and the higher data-rate per-unit-coverage area it has. "Vertical handoff" is employed to support MH roaming from one network to another. The MH always switches to the lowest reachable overlay to achieve the highest data-rate per-unit-coverage area. In order to adapt to the system dynamics (e.g., the varying traffic load), a policy-enabled handoff scheme is proposed in [29] to take multiple factors into considerations. Specifically, a cost function is defined as the weighted sum of network bandwidth, the MH's power consumption, and the network-access cost. The available network with the least cost is chosen for communication.

The pioneering work of Katz *et al.* has motivated our research on CM in heterogeneous wireless networks. The approaches considered in [28] and [29], however, only take into consideration the local networks being accessed, which may not result in optimized overall performance (see the example discussed in Section I). Thus, a "longer term planning" strategy for the entire data-transmission period (rather than only at a given snapshot) is needed to achieve global optimization. In addition, the QoS requirement of the MH is not considered in [28] and [29]. In this paper, we establish a more general CM model based on not only system information (e.g., available bandwidth, access costs, network congestion) but also on a mobile users' status (e.g., traffic type/amounts and moving routes) in order to minimize the overall communication cost while simultaneously meeting the QoS requirement.

III. System Architecture

In this section, we give an overview of the system architecture considered in this paper. Five components are involved in CM: 1) the MH, 2) the communication peer, 3) the access point

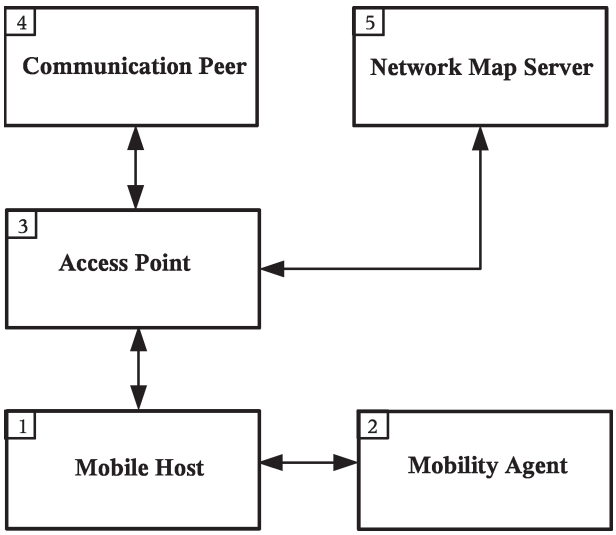

Fig. 2. Signaling protocols for CM problem.

(AP), 4) the network-map server (NMS), and 5) the mobility agent, as shown in Fig. 2. The MH receives data from, or sends data to, its communication peer (e.g., a node in a P2P network or a data server in the Internet). An AP is a unit that provides wireless access to the MH. It could be a wireless LAN access point, a cellular base station, an IEEE 802.16 base station, an IEEE 802.20 base station, or a satellite. The NMS maintains a database of every AP in the system. The mobility agent helps the MH obtain its mobility information.

Before initiating data transmission, the MH first negotiates with its communication peer to acquire data-traffic information (e.g., the total amount of data to be transmitted and its delay budget). If the amount of transmitted data is less than a predefined threshold, no attempt in minimizing the communication cost is made, which is due to overhead involved. Otherwise, CM is initiated. More specifically, the MH collects needed information, including MH mobility and network topology, from the mobility agent and the NMS, respectively, and runs the CM algorithm (to be discussed in Section V).

The network topology and user mobility are important inputs in running our proposed CM algorithm. In the rest of this section, we briefly discuss the signaling protocols to maintain that information, although their design and implementation are not the focus of this paper.

1) Signaling Protocol for Network-Map Maintenance: When an AP is established, it sends to its associated NMS a network-update message, which includes its location (e.g., GPS coordinates), estimated coverage area, transmission technology, communication costs, and data rates. The coverage area could be estimated as an ideal circle or in any arbitrary shape depending on the environment and the estimation model. Upon receiving a network-update message, the NMS creates an entry in its database, with a timer associated. The AP periodically sends update messages to refresh the timers. If no new update message is received when the timer expires, the entry will be removed from the database. Note that once a cell has been established, no significant changes on the AP's location, coverage, data rates, and/or communication costs are expected in the near future. Thus, the timer can be set to a large value, and the AP can send update messages to NMS infrequently, resulting in ignorable overhead. Based on information maintained in the NMS database, a 2-D network map can be established. Upon receiving a request from the MH, the NMS replies with a (partial) network map and information associated with the AP. Such information can also be buffered by the MH for future use, where the MH only needs to contact the NMS to check if the buffered network map is out of date.

2) Signaling Protocol for User Mobility: The mobile user's moving pattern (i.e., the traveling route) could be predicted precisely, roughly, or may be unpredictable at all. Three typical types of user mobility are discussed as follows:

1) Predetermined routes and speeds. The mobile user may follow a predetermined route and move at predetermined speeds. A typical example is a railway system (which is a main carrier in many metropolitan areas) transporting multitudinous people every day. For instance, 500 000 passengers take trains daily in New York's Pennsylvania Station; tens of millions of people travel by train daily in other major cities around the world; moreover, all of these passengers are current or potential mobile-network users. The train follows predetermined routes and precise schedules and may travel through many cells (in systems such as cellular, IEEE 802.20, or satellite) and stop at a number of stations, where, for instance, wireless LAN's are available.
2) Predetermined routes but variable speeds. For example, buses follow regular routes, but depending on the traffic condition, they may move faster or slower than the predetermined schedule. Another example is self-guided touring. A tourist may carry an electronic cicerone provided by the tourist agent and follow the recommended route to visit the scenic landmarks. The MH moving schedule can be estimated according to the average pacer speed. The actual moving speeds of the tourists, however, may vary, resulting in deviation from the estimated schedules.
3) Uncertain routes. The moving route of a mobile user is sometimes undetermined. For example, there may be two (or more) alternative routes available from home to office, and the mobile user may choose either one of them with a certain probability. A self-learning model (to be discussed next) may be adopted to learn the possible routes of the mobile users and the probability that a route is taken.

The MH needs to acquire route information as the input of the CM algorithm. Two approaches for mobility management are considered in this paper: 1) system-mobility management, where mobility information is maintained by the system and provided to the MH via mobility agents that are installed on the carriers (e.g., trains, buses, or cars) to broadcast mobility information periodically and 2) autonomous-mobility management, where mobility information is maintained by every MH via a self-learning model. More specifically, the moving routes are mapped to a graph, where the edges represent routes traversed by the MH, and the vertices represent the intersection points. A visiting counter and a stale timer are associated with each edge. The visiting counter is increased by one whenever the

corresponding route is taken by the MH and decreased by one whenever the stale timer expires. At any vertex, the probability of taking a particular edge for next movement is calculated by dividing the visiting counter associated with this edge by the sum of the counters of all edges connecting to this vertex. As a result, a route is associated with a higher probability if it was traversed more frequently in the recent past.

IV. Problem Formulations and NP-Hard Proof

In this section, we formulate the CM problem and prove its NP-hardness.

A. Problem Formulation

To facilitate our discussion, we consider an abstract model of the heterogeneous wireless system. Specifically, we assume that the system is divided into N cells, where each cell (e.g., cell $i, 1 \leq i \leq N$) is served by and only by a set of APs that are denoted by $\{A_{ij} \,|1 \leq j \leq |A_i|\}$, where $|A_i|$ is the number of APs serving cell i. Clearly, the cells discussed here are different from the nature cells formed by the coverage of APs. The former can be obtained by dividing the whole area according to the boundaries of the natural cells. An AP A_{ij} connects to the core network through a dedicated communication link that may be either wired or wireless, with a cost rate of p_{ij}^c and an average data rate of r_{ij}^c b/s. At the same time, A_{ij} provides wireless access for the MH within its coverage, with a cost rate of p_{ij}^m and an average data rate of r_{ij}^m b/s. The cost rate is the weighted sum of several factors, such as service charge, signaling overhead, and power consumption, measured by the cost per-unit amount of data.

We consider an MH with β programmable wireless-network interface cards, each of which can tune to different wireless technology. Note that one such wireless interface card can be used to access one AP at a given time only, although it can switch from one to another at different times. Assume that the MH sends or receives a sequence of K data blocks with M bits of data in total, while traveling in the coverage area of the heterogeneous wireless system. Let M_k denotes the size (in terms of bits) of data block $k(1 \leq k \leq K)$. Clearly, $M = \sum_{k=1}^{K} M_k$. A data block k is assumed to have a delay bound d_k, indicating the time by which data block k should be received.

We denote τ_i as the dwelling time of the MH in each cell i. τ_i is obtained from the calculation based on the network map and the user mobility. When an MH is in cell i, it may or may not communicate with A_{ij} during the entire period of τ_i in order to minimize the communication cost. We denote t_{ij} to be the time that the MH connects to A_{ij}. Clearly, $t_{ij} \leq \tau_i$. The total amount of data downloaded during t_{ij} is denoted by L_{ij}. We also introduce a prefetch methodology where a certain amount of data may be prefetched by APs in order to enhance the CM process. We denote f_{ij} as the amount of prefetched data at A_{ij} with a corresponding cost g_{ij}. The cost of data transmission/reception through A_{ij} is denoted as C_{ij} and calculated based on t_{ij}, r_{ij}^c, p_{ij}^c, r_{ij}^m, and p_{ij}^m. The goal of the CM algorithm is to determine the values of t_{ij}, L_{ij}, f_{ij}, and g_{ij} such that the total cost $C = \sum_{i=1}^{N} \sum_{j=1}^{|A_i|} C_{ij}$ is minimum, with all delay constraints satisfied. The inputs and outputs of the CM algorithms are summarized in Fig. 3.

INPUT:
1. N: number of cells;
2. K: number of data blocks to be transmitted;
3. β: number of wireless interface cards per MH;
4. M: total amount of data (in terms of bits) to be transmitted;
5. M_k $(1 \leq k \leq K)$: size (in terms of bits) of data block k;
6. d_k $(1 \leq k \leq K)$: time delay bound of data block k;
7. r_{ij}^c : data rate between the core network and A_{ij};
8. r_{ij}^m : data rate between the MH and A_{ij};
9. p_{ij}^c : communication cost rate between the core network and A_{ij};
10. p_{ij}^m : communication cost rate between the MH and A_{ij};
11. τ_i : dwelling time during which the MH travels in cell i;

OUTPUT:
1. t_{ij} : communication time between the MH and A_{ij};
2. L_{ij} : amount of data downloaded between the MH and A_{ij};
3. f_{ij} : amount of prefetched data at A_{ij};
4. g_{ij} : cost associated with f_{ij};
5. C: minimized total communication cost.

Fig. 3. Summary of inputs and outputs for the CM problem.

B. NP-Hard Proof

In this section, we prove that the CM problem is NP-hard by providing a polynomial reduction from a known NP-hard problem, which is known as the continuous multiple-choice knapsack (CMCK) problem, to the CM problem.

1) CMCK Problem: The CMCK problem was first formulated by Ibaraki *et al.* in 1978 [30] and proven to be NP-hard. The CMCK problem considers N groups, where each group i has K_i items. The item j in group i has a value of ν_{ij} and a size of s_{ij}. The CMCK problem is to select a fraction $x_{ij}(0 \leq x_{ij} \leq 1)$ of, at most, one item from each of the N groups, in order to satisfy both the value and the size constraints. More specifically, the sum of the size is less than a size threshold S, and the sum of the value is greater than a value threshold V. The CMCK problem is defined as follows.

Given $V > 0$ and $S > 0$, does a set of $0 \leq x_{ij} \leq 1$ exist, such that

$$\sum_{i=1}^{N} \sum_{j=1}^{K_i} \nu_{ij} x_{ij} > V \tag{1}$$

and

$$\sum_{i=1}^{N} \sum_{j=1}^{K_i} s_{ij} x_{ij} < S \tag{2}$$

hold, where, at most, one $\{x_{ij} \mid 1 \le j \le K_i\}$ is nonzero, for $\forall\, 1 \le i \le N$?

2) Polynomial Reduction From CMCK to CM Problem: Now, we consider a simplified scenario of the CM problem, where only one delay bound needs to be satisfied for completing the transmission of all data (M). The simplified CM problem with an objective minimum-cost value (denoted by C) can be defined as follows.

Given $M > 0$ and $C > 0$, does a set of $0 \le x_{ij} = t_{ij}/\tau_i \le 1$ exist, such that

$$\sum_{i=1}^{N} \sum_{j=1}^{|A_i|} r_{ij}^m \tau_i x_{ij} > M \tag{3}$$

and

$$\sum_{i=1}^{N} \sum_{j=1}^{|A_i|} \left(p_{ij}^c + p_{ij}^m\right) r_{ij}^m \tau_i x_{ij} < C \tag{4}$$

hold, where $\sum_{j=1}^{|A_i|} x_{ij} \le 1$, for $\forall\, 1 \le i \le N$?

In the above CM problem, the definition of r_{ij}^m, τ_i, p_{ij}^c, and p_{ij}^m can be found in Fig. 3. We assume that data is downloaded to the MH at r_{ij}^m b/s. t_{ij} is represented as $\tau_i x_{ij}$, where x_{ij} is between zero and one. To facilitate the discussion, we reiterate (3) and (4) by letting $V' = M$; $S' = C$; $x'_{ij} = x_{ij}$; $K'_i = |A_i|$; $\nu'_{ij} = r_{ij}^m \tau_i$; and $s'_{ij} = (p_{ij}^c + p_{ij}^m) r_{ij}^m \tau_i$; then, we can reformulate the minimum-cost problem as follows.

Given $V' > 0$ and $S' > 0$, does a set of $0 \le x'_{ij} \le 1$ exist, such that

$$\sum_{i=1}^{N} \sum_{j=1}^{K'_i} \nu'_{ij} x'_{ij} > V' \tag{5}$$

and

$$\sum_{i=1}^{N} \sum_{j=1}^{K'_i} s'_{ij} x'_{ij} < S' \tag{6}$$

hold, where $\sum_{j=1}^{K'_i} x_{ij} \le 1$, for $\forall\, 1 \le i \le N$?

Clearly, the parameters of (5) and (6) can be mapped to that of (1) and (2), and we can see that the difference between the CMCK and minimum-cost problems is the chosen of x_{ij} or x'_{ij}. In CMCK, at most, one item in group i can be chosen, i.e., $x_{ij} \ne 0$ is true for, at most, one j in $1 \le j \le K_i$; in minimal-cost problem, multiple items can be chosen in each group i, i.e., there can be more than one $x'_{ij} \ne 0$ for $1 \le j \le K'_i$. One additional constraint of the CMCK problem is that $\sum_{j=1}^{K'_i} x'_{ij} \le 1$.

Recall that a polynomial reduction from problem A to problem B requires mapping the YES instance of A to the YES instance of B, i.e., YES $\rightarrow$ YES, and mapping the YES instance of B to the YES instance of A, i.e., YES $\leftarrow$ YES.

We first prove YES $\rightarrow$ YES: given a set of x_{ij} that satisfies inequations (1) and (2), can we construct an instance of the minimum-cost problem that satisfies inequations (5) and (6)? This can be easily achieved by matching the variables of inequalities (1) and (2) to inequalities (5) and (6), respectively. That is to say, we let $V' = V$; $S' = S$; $x'_{ij} = x_{ij}$; $K'_i = K_i$; $\nu'_{ij} = \nu_{ij}$; and $s'_{ij} = s_{ij}$; then, we can verify that $\sum_{j=1}^{K'_i} x'_{ij} \le 1$ holds since, at most, one $x'_{ij} \ne 0$ for $1 \le j \le K'_i$, and all x'_{ij} is between zero and one.

We then prove YES $\leftarrow$ YES: given a set of x'_{ij} that satisfies inequations (5) and (6), can we construct an instance of the CMCK problem that satisfies inequalities (1) and (2)? Starting from inequality (5), we first find the largest value of $\nu'_{ij}x'_{ij}$ in each group i, and we denote this index j as p_i. Then, we can rewrite the left-hand side of (5) as $\text{LHS}_{\text{eq5}} = \sum_{i=1}^{N} \nu'_{ip_i} x'_{ip_i} + \sum_{i=1}^{N} \sum_{j=1, j\ne p_i}^{K'_i} \nu'_{ij} x'_{ij}$. Notice that $v'_{ip_i} x'_{ip_i}$ is the largest value of $\nu'_{ij}x'_{ij}$ in group i for $1 \le j \le K'_i$, we have: $\text{LHS}_{\text{eq5}} = \sum_{i=1}^{N} \nu'_{ip_i} x'_{ip_i} + \sum_{i=1}^{N} \sum_{j=1, j\ne p_i}^{K'_i} \times \nu'_{ij} x'_{ij} \le \sum_{i=1}^{N} \nu'_{ip_i} x'_{ip_i} + \sum_{i=1}^{N} (K'_i - 1) \nu'_{ip_i} x'_{ip_i} = \sum_{i=1}^{N} \times \nu'_{ip_i} x'_{ip_i} K'_i \le K'_{i\max} \sum_{i=1}^{N} \nu'_{ip_i} x'_{ip_i}$, where $K'_{i\max} = \max_{i=1}^{N} K'_i$. Thus, we have $K'_{i\max} \sum_{i=1}^{N} \nu'_{ip_i} x'_{ip_i} \ge \text{LHS}_{\text{eq5}} > V' \rightarrow \sum_{i=1}^{N} \nu'_{ip_i} x'_{ip_i} > (V'/K_{i\max}) = V'' > 0$.

In addition, it is easy to verify that $\sum_{i=1}^{N} s'_{ip_i} x'_{ip_i} < S'$, since $\sum_{i=1}^{N} s'_{ip_i} x'_{ip_i} \le \sum_{i=1}^{N} \sum_{j=1}^{K'_i} s'_{ij} x'_{ij} < S'$. Now, we summarize the YES $\leftarrow$ YES proof as follows: given a set of x'_{ij} that satisfies inequalities (5) and (6), we first find out the p_i for each group i in polynomial time, such that $\nu'_{ip_i} x'_{ip_i}$ is the largest among $\nu'_{ij}x'_{ij}$, for $1 \le j \le K'_i$. Then, we construct the following inequations in polynomial time, which match inequations (1) and (2) of CMCK problem: 1) $\sum_{i=1}^{N} \nu'_{ip_i} x'_{ip_i} > V'' > 0$, where $V'' = (V'/K_{i\max})$, $K'_{i\max} = \max_{i=1}^{N} K'_i$; and 2) $\sum_{i=1}^{N} s'_{ip_i} x'_{ip_i} < S'$. We have proven that the above two inequalities hold. Thus, we finish the YES $\leftarrow$ YES proof.

V. Proposed LP Algorithms for CM

In this section, we examine several cases of the CM problem with increasing levels of complexity and discuss the proposed LP algorithms for each of them.

A. Single Route

We first consider the scenario where an MH traverses a predetermined route and follows predetermined speeds. Note that the MH's moving speed is not necessarily constant. As long as the speeds (which may vary with time) are predetermined, one can precisely calculate the dwelling time of the MH in each cell i, i.e., τ_i. To facilitate our discussion, we perform a preprocessing on the network map. More specifically, we divide the N cells into N' small "cells" so that all delay bounds are due at the boundary of the small cells.

1) Base Case: We start with a base case, where the following assumption is made: the wireless link between AP and the MH is the performance bottleneck, i.e., $r_{ij}^m \le r_{ij}^c (1 \le j \le |A_i|, 1 \le i \le N)$. Thus, AP_{ij} can always communicate with the MH at r_{ij}^m b/s.

We have developed an LP model, as shown in Fig. 4, for the base case. The primary objective of this LP model is to minimize the total communication cost C by choosing a

OBJECTIVE: minimize C;
VARIABLES: t_{ij}, $1 \le i \le N'$ and $1 \le j \le |A_i|$;
CONSTRAINTS:
1: $C = \sum_{i=1}^{N'} \sum_{j=1}^{|A_i|} C_{ij}$;
2: $C_{ij} = (p_{ij}^m + p_{ij}^c) \times L_{ij}$;
3: $L_{ij} = r_{ij}^m \times t_{ij}$;
4: $M = \sum_{i=1}^{N'} \sum_{j=1}^{|A_i|} L_{ij}$;
5: $\sum_{u \in \theta(d_k)} \sum_{j=1}^{|A_u|} L_{uj} \ge \sum_{v=1}^{k} M_v$, for $1 \le k \le K$;
6: $t_{ij} \le \tau_i$;
7: $\sum_{j=1}^{|A_i|} t_{ij} \le \tau_i \times min(\beta, |A_i|)$;
– $\theta(x)$ is a function that returns the set of all cells that the MH has visited by time x.

Fig. 4. LP model for predetermined route (base case).

proper value for each t_{ij}, such that all constraints 1–7 (given in Fig. 4) are satisfied. The first three constraints define the communication cost, which is calculated based on the unit communication cost between A_{ij} and the core network (i.e., p_{ij}^c), the unit communication cost between A_{ij} and the MH (i.e., p_{ij}^m), the data rate r_{ij}^m, and the communication time t_{ij}. The fourth constraint ensures the total transmitted data to be M. The delay budget is specified in the fifth constraint, which ensures the transmission of the kth data block to be finished no later than d_k. The sixth constraint limits t_{ij}, the data-transmission time between the MH and A_{ij}, to be no longer than the dwelling time of the MH in cell i (i.e., τ_i).

An AP is available for an MH, if and only if the MH is within the coverage of the AP, and the MH has the suitable wireless interface to access the AP. Note that, even when multiple APs are available, hardware may limit the number of simultaneous wireless connections that can be established by an MH. More specifically, with β-programmable wireless interface cards, the MH may communicate with, at most, $\min(\beta, |A_i|)$ APs in cell i. With the seventh constraint, the total data-transmission time between the MH and the access points in cell i should not exceed τ_i multiplied by $\min(\beta, |A_i|)$. Finally, $\theta(x)$ is a function that returns the set of all cells that the MH has visited by time x. The formulated LP model in Fig. 4 is solved by using LP-SOLVE [31], yielding optimal values of $\{t_{ij} | 1 \le j \le |A_i| \ and \ 1 \le i \le N'\}$ that minimize the total communication cost.

In frequency-division duplex system where uplink channel and downlink channel are separated, the optimization of uplink cost is just a dual problem of the downlink case. In a time-division duplex (TDD) system, where the uplink and downlink share the same channel, if we try to optimize both the uplink and downlink cost, then we need to let the uplink and downlink channel share the MH's cell dwelling time in each AP, which can be easily implemented in the LP model by adding the term for uplink dwelling time in constraint 7 of Fig. 4. Other modifications of constrains in Fig. 4 are also straightforward, such as adding terms for data and delay requirements of uplink and the associated cost. Since the LP model in considering both uplink and downlink is similar to that of downlink only, we do not bother to list the solution for considering both uplink and downlink. Notice that in a TDD system, the individual optimization goals of uplink and downlink may conflict with each other, and the yielded optimized solution of considering both uplink and downlink is a compromise of the two individual goals.

2) Advanced Case: In the base case, we have assumed $r_{ij}^m \le r_{ij}^c$. Now, we nullify this assumption in order to establish a more realistic model. If $r_{ij}^m > r_{ij}^c$, the data-rate limitation between A_{ij} and the core network may lead to inefficiency, because A_{ij} cannot draw down data blocks from the core network at a rate as high as r_{ij}^m and, thus, limits the data transmission to the MH. Two approaches have been considered to address this problem, as in the following discussions.

1) Reduce data rate: When $r_{ij}^m > r_{ij}^c$, A_{ij} may lower the data rate at its wireless interface by setting $r_{ij}^m = r_{ij}^c$. This approach is simple, but it sacrifices the channel efficiency and results in reduced throughput. Moreover, it may also increase the communication cost, because the data rate in the low-cost cell is limited by r_{ij}^c, and thus, the low-cost resource cannot be utilized thoroughly.
2) "Prefetch": When $r_{ij}^m > r_{ij}^c$, A_{ij} may draw down a certain number of data blocks beforehand and store them in buffer so that A_{ij} can transmit data to the MH at a high data rate when the MH enters cell i.

To enable efficient prefetch, we divide t_{ij} into two parts t_{ij1} and t_{ij2} with $t_{ij} = t_{ij1} + t_{ij2}$. t_{ij1} represents the time period when A_{ij} transmits at data rate r_{ij}^m, with the help of the prefetched data. t_{ij2} represents the time period when A_{ij} transmits at data rate r_{ij}^c. The optimal values of t_{ij1} and t_{ij2} will be determined by solving the LP model in Fig. 5. The amount of prefetched data f_{ij} equals $(r_{ij}^m - r_{ij}^c) \times t_{ij1}$. If $t_{ij1} = 0$; no data is prefetched. Prefetch should be completed before the MH enters the coverage area of A_{ij}, as specified by constraint 3.2 in Fig. 5. Given the assumption that the route and the speed of the MH are predetermined, the exact time to start prefetching is not a crucial issue, as long as prefetch can be finished before data-transmission begins.

B. Multiple Possible Routes

We now consider the case where the mobile user's moving route is undetermined when the data transmission is initiated. Assume that the mobile user may take one of W possible routes, with a probability of P_w associated with Route $w (1 \le w \le W)$ according to the mobility-management schemes discussed in Section III. The multiple possible routes may have common or uncommon paths. Common paths refer to those that are shared by more than one route, whereas uncommon paths are those completely disjoint from each other.

There are two issues that need to be handled properly in multiple possible routes scenario. First, the delay bounds should be satisfied in the worst case. In other words, the data blocks are to be transmitted within their delay budgets, even the route with the lowest capacity is eventually taken by the mobile user. Second, a prefetch has two impacts on the overall communication cost. On the one hand, prefetch may reduce the communication cost by enabling the thorough use of low-cost resources, but on the other hand, the prefetched data in the untaken routes

OBJECTIVE: minimize C;
VARIABLES: t_{ij} and f_{ij}, $1 \le i \le N'$ and $1 \le j \le |A_i|$;
CONSTRAINTS:
1: $C = \sum_{i=1}^{N'} \sum_{j=1}^{|A_i|} C_{ij}$;
2: $C_{ij} = (p_{ij}^m + p_{ij}^c) \times L_{ij}$;
3.1: for non-prefetch A_{ij} : $L_{ij} = r_{ij}^m \times t_{ij}$;
3.2: for prefetch A_{ij} :
3.2.1: $L_{ij} = r_{ij}^m \times t_{ij1} + r_{ij}^c \times t_{ij2}$;
3.2.2: $f_{ij} = (r_{ij}^m - r_{ij}^c) \times t_{ij1}$;
3.2.3: $f_{ij} \le r_{ij}^c \times \sum_{l=1}^{i-1} \tau_l$;
3.2.4: $t_{ij} = t_{ij1} + t_{ij2}$;
4: $M = \sum_{i=1}^{N'} \sum_{j=1}^{|A_i|} L_{ij}$;
5: $\sum_{u \in \theta(d_k)} \sum_{j=1}^{|A_u|} L_{uj} \ge \sum_{v=1}^{k} M_v$, for $1 \le k \le K$;
6: $t_{ij} \le \tau_i$;
7: $\sum_{j=1}^{|A_i|} t_{ij} \le \tau_i \times min(\beta, |A_i|)$;
– $\theta(x)$ is a function that returns the set of all cells that the MH has visited by time x;

Fig. 5. LP model for predetermined route with prefetch (advanced case).

become wasteful and, thus, may increase the communication cost. It is nontrivial to decide whether to employ prefetch or not and how to use it.

If prefetch is disabled for simplicity, handling multiple possible routes is not much different from the single route case, as one can always optimize each route separately without considering other routes.

At the presence of prefetch, however, it is inefficient to take the simple approach that considers each route separately by following the LP model presented in Fig. 5, since the cost of wasted prefetch data is not taken into account. Our proposed approach is summarized in Fig. 6, where all routes and their probabilities are considered in order to minimize the overall communication cost. We denote τ_{iw} as the dwelling time of the MH in cell i if Route w is taken. The total communication cost is the weighted sum of the costs of all possible routes. The communicating cost of Route w (denoted by C_w) is the sum of its communication cost if it is taken with a probability of P_w and the cost of wasted prefetch data if the route is not taken with a probability of $(1 - P_w)$, as shown in Fig. 6. The cost of the wasted prefetch data is $p_{ij}^c \times f_{ij}$, since only the communication link to the core network is used for prefetching data. Based on the above cost calculation, a penalty is added to those routes with high prefetch costs. As a result, the route with lower probability will prefetch a less amount of data to contain the cost. For the common path of multiple routes, there is only one set of variables to represent its network parameters and prefetch data, and the yielding result of this common path will reflect the compromises of multiple routes. The LP model aims to minimize the overall cost, based on the probability of each route when there are multiple possible routes, and it does not intend to decide for the MH which route to take. Once the MH chooses any one of the possible routes, the LP model can be rerun for further optimization of the cost. However, there is no technical hurdle to let AP decide for MH which route to take. For example, AP can suggest to MH the route with the lowest cost.

Given W possible routes,
OBJECTIVE: minimize $C = \sum_{w=1}^{W} C_w$;
VARIABLES: t_{ij} and f_{ij}, $1 \le i \le N'$ $1 \le j \le |A_i|$;
For each route w, where $1 \le w \le W$,
LOOP: Compute C_w , with the following
CONSTRAINTS:
1: $C_w = \sum_{i=1}^{N'} \sum_{j=1}^{|A_i|} [C_{ij} \times P_w + g_{ij} \times (1 - P_w)]$;
2: $C_{ij} = [(p_{ij}^m + p_{ij}^c) \times L_{ij}$;
3.1: for non-prefetch A_{ij} :
3.1.1: $L_{ij} = r_{ij}^m \times t_{ij}$;
3.1.2: $g_{ij} = 0$, since there is no prefetch data;
3.2: for prefetch A_{ij} :
3.2.1: $L_{ij} = r_{ij}^m \times t_{ij1} + r_{ij}^c \times t_{ij2}$;
3.2.2: $f_{ij} = (r_{ij}^m - r_{ij}^c) \times t_{ij1}$;
3.2.3: $f_{ij} \le r_{ij}^c \times \sum_{l=1}^{i-1} \tau_{lw}$;
3.2.4: $t_{ij} = t_{ij1} + t_{ij2}$;
3.2.5: $g_{ij} = p_{ij}^c \times f_{ij}$;
4: $M = \sum_{i=1}^{N'} \sum_{j=1}^{|A_i|} L_{ij}$;
5: $\sum_{u \in \theta(d_k)} \sum_{j=1}^{|A_u|} L_{uj} \ge \sum_{v=1}^{k} M_v$, for $1 \le k \le K$;
6: $t_{ij} \le \tau_{iw}$;
7: $\sum_{j=1}^{|A_i|} t_{ij} \le \tau_{iw} \times min(\beta, |A_i|)$;
END LOOP;
– $\theta(x)$ is a function that returns the set of all cells that the MH has visited by time x;

Fig. 6. LP model for multiple possible routes.

In the above discussion, all possible routes have been taken into consideration. To reduce computing complexity, one may consider the routes with high probabilities only. Specifically, the possible routes are sorted via a decreasing order of their probabilities. The routes are chosen for consideration from the top until $\sum_w P_w \ge \Gamma$, where Γ is a predefined constant. The LP model yields results based on currently available route information. While the MH moves, the number of possible routes may change, or one predetermined route may not be taken. The LP model is then rerun based on an updated route information. The probability-based approach for multiple-route scenario does not aim to deal with very large (nonpolynomial) number of routes. First, nonpolynomial number of routes is not reasonable in real scenarios. The number of routes for an MH is usually limited. Second, the probability of each route will drop with the increase of the number of possible routes, and the prefetch data in each route will decrease to near zero if the probability is low. We assume that the MH's dwelling time in each cell is known, and this information in needed as the input of the LP model. If the MH's actual dwelling time in any cell is different from the predicted value, rerun of the LP model is necessary.

C. Variable Speed

In the above discussion, we have assumed that the MH's moving speeds are predetermined. This assumption does not always hold in practical applications. In this section, we study possible speed variations and their impacts on CM.

Note that not all speed variations affect CM. For example, the MH may vary its speed in a cell i. As long as the average speed equals the expected value, τ_i does not change. As a result,

TABLE I
NETWORK PARAMETERS FOR SIMULATION SETUP

Wireless networks	802.11 WiFi	3G Cellular	802.16 WiMax
Cell radius	200m	1000m	3000m
Wireless access data rate (r_{ij}^m)	5Mbps	1Mbps	6Mbps
Core network data rate (r_{ij}^c)	3Mbps	3Mbps	9Mbps
Core network data cost (p_{ij}^c)	0.4/MBytes	0.7/MBytes	1.9/MBytes
Total data cost ($p_{ij}^m + p_{ij}^c$)	1/MBytes	1.5/MBytes	4/MBytes

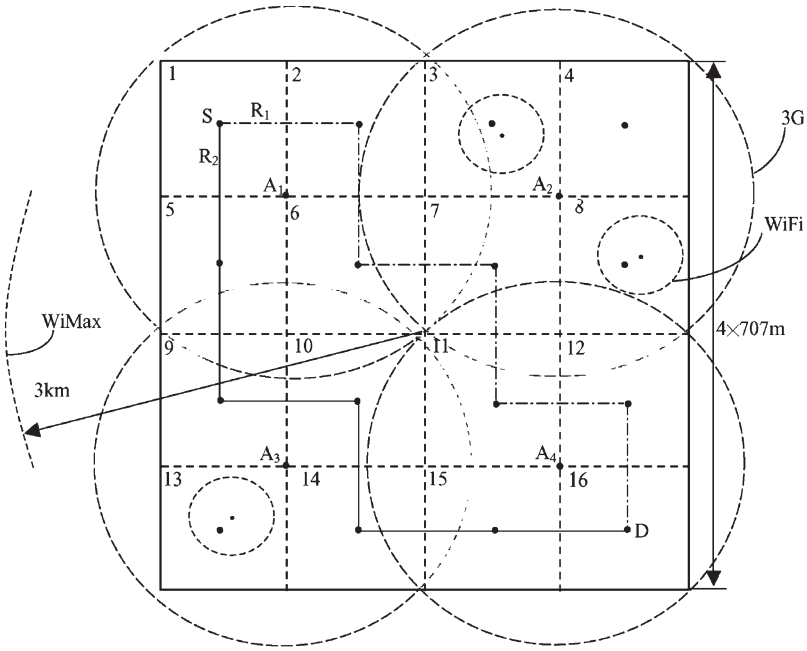

Fig. 7. Network topology with one WiMax cell, four 3G cells, and 16 WiFi cells.

the same CM schemes, as discussed in Sections V-A and B, can be employed to yield the same results as the case with a predetermined speed. If τ_i changes, however, the CM scheme will be different.

The above observation motivates us to develop an approach for the scenario with variable MH moving speed, centering at the variation of τ_i. More specifically, the MH maintains an expected moving speed (V). Based on the moving speed, the moving route (either single route or multiple possible routes) and the network map $\{\tau_i | 1 \leq i \leq N'\}$ can be calculated. Upon entering a cell (e.g., cell i), the MH keeps tracking its actual average speed in this cell (denoted by $\overline{V}$). The MH compares $\overline{V}$ with V at either of the following two occasions, whichever arrives earlier: 1) when the MH has spent τ_i in cell i or 2) when the MH has finished traversing cell i. If $\overline{V} \neq V$, τ_i changes. As a result, the MH updates V with $\overline{V}$ and performs CM based on the updated τ_i $(1 \leq i \leq N')$ and the remaining data to be delivered. Note that, since the MH already has the moving route and the network map, no extra signaling overhead is incurred. Although additional delay is expected to run the LP algorithm, such a computation delay is short, as will be discussed next in Section VI.

VI. SIMULATIONS AND DISCUSSION

We have carried out extensive simulations to evaluate the effectiveness and efficiency of our proposed minimum-cost data-delivery algorithm. Three typical wireless technologies with different coverage areas, data-transmission rates, and costs are considered in our simulation. Specifically, we simulate a number of IEEE 802.11b WiFi, 3G Cellular, and IEEE 802.16 WiMax cells. Their network parameters, including cell radii, data rates, and costs are summarized in Table I.

WiMax targets at providing wide-area broadband access, with the longest cell radius and the highest wireless data rate among the three, as indicated in Table I. We set the data rate of a WiMax user to be 6 Mb/s, although IEEE 802.16 can reach a peak rate of 75 Mb/s. WiFi cells locate in the hot-spot areas, with the smallest cell size. The average user data rate of WiFi is assumed to be 5 Mb/s, given its maximum raw data rate of 11 Mb/s. The peak wireless data rate of 3G is 2 Mb/s. We assume its average data rate to be 1 Mb/s for mobile users. In addition, we notice that r_{ij}^c is usually higher than r_{ij}^m in WiMax and 3G systems, and as a result, prefetch is not necessary. The communication cost has no unit because it is a weighted sum of service charge and power consumption. It is a separate issue (outside the scope of this paper) to establish efficient models for determining cost rates.

The network topology and parameters in our simulations are shown in Fig. 7. It is a square area with diagonal length of 4 km. We divide this square area into 16 small square blocks, and each block has a diagonal length of 1 km. We number these 16 blocks from 1 to 16, as indicated in the upper left corner of each block. A WiMax AP is placed at the center of the square

and covers the entire square area. Four 3G APs are placed at locations A_1, A_2, A_3, and A_4, respectively. Since the radius of a 3G AP is 1 km, each 3G AP covers its surrounding four blocks. For example, 3G AP at location A_1 covers blocks 1, 2, 5, and 6. There are 16 WiFi APs in total. Each of them is randomly and uniformly distributed inside a block. In Fig. 7, only three WiFi APs are drawn to avoid too many items included in the figure. A handoff procedure happens between two APs of different types or if they are of the same type.

Without loss of generality, we set the start point of all the routes at the center of block 1 and the destination at the center of block 16, which are denoted by S and D in Fig. 7, respectively. We let the MH travel along the straight line connecting the centers of two adjacent blocks. When arriving at the center of a block, the MH can choose any of its neighboring blocks to be the next block to traverse, except for the block from which this MH enters the current block. Two random routes are shown as examples in Fig. 7 and are marked as R_1 in a dash dot line and R_2 in a solid line. The speed of the MH can be constant (at 10 m/s) or variable. The MH's dwelling time in each cell (i.e., τ_i, $1 \leq i \leq N'$, where N' is the number of cells in the route) can be calculated, given the route and the speed information. The amount of data (M) to be delivered and its associated delay bounds may vary in different simulation scenarios. We define two thresholds M_l and M_h. M_l is the total amount of data that can be downloaded if the least expensive AP in each cell is always selected (i.e., by following a greedy algorithm that selects the AP with the lowest cost). M_h is the total amount of data that can be downloaded if the AP with the highest data rate in each cell is always chosen (i.e., by following a greedy algorithm for minimizing delay). Obviously, $M_l \leq M_h$. If $M_l = M_h$, it implies that the least expensive AP is the AP with the highest data rate in every cell, and thus, the optimization problem becomes trivial. If the total amount of data M falls below M_l, then a greedy algorithm choosing the least expensive AP in each cell yields the optimal solution. For M between M_l and M_h, the greedy algorithm selects the access technology with the highest data rate in each cell. If M is greater than M_h, then there is no feasible solution because even data are downloaded at the highest data rate in each cell; the delay budget still cannot be met. Thus, for a nontrivial CM problem, we let $M_l < M < M_h$.

We study the performance of the proposed LP model under four different scenarios: random single route, random map, random multiple routes, and random distribution of cell dwelling time. For each scenario, we generate 20 different samples and average their results. In these dynamic scenarios, the routes, the maps, or the MH's cell dwelling time change randomly. Random single route is created by randomly generating a path from S to D in Fig. 7. Random map is created by randomly distributing the location of each WiFi AP inside its block. Random multiple-route scenario is created by combining multiple single routes together and then feeding them to the multiple routes LP model in Fig. 6. We have several patterns for the random distribution of the MH's cell dwelling time. One is to let the dwelling time conform to a certain distribution, e.g., a uniform distribution. Other patterns include increasing or decreasing the cell dwelling time by a certain amount. Due to the change in cell dwelling time, the MH will rerun the LP model for a new download scheme when it is necessary. Next, we discuss the simulations in detail.

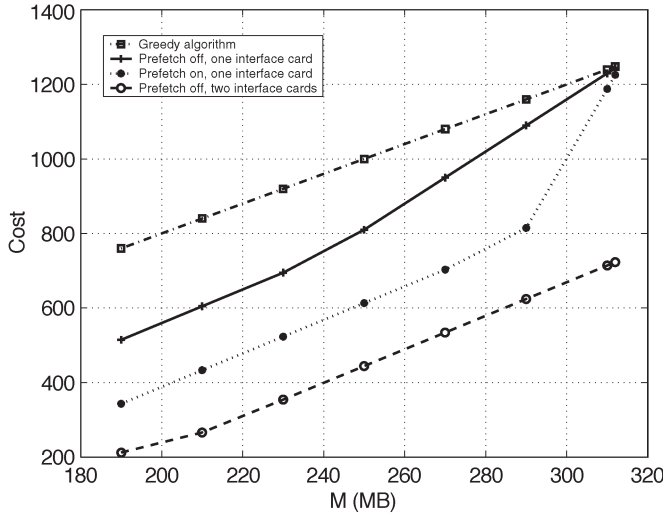

Fig. 8. Performance averaged over random routes.

In the random single-route scenario, we keep the same map but generate different route randomly for each scenario. Only one delay bound is specified at the end of route for delivering a total of M megabytes data, where M varies from 190 to 312 MB. Fig. 8 compares the performances of the LP model and the greedy algorithm. As can be seen, the LP model without prefetch achieves a cost reduction of up to 32% (or 20% in average), as compared to the greedy algorithm. Prefetch can also help to further reduce the communication cost in any cell i with $r_{ij}^m > r_{ij}^c$. By employing prefetch, A_{ij} may draw down a certain number of data blocks beforehand and store them in buffer so that A_{ij} can transmit to the MH at a high data rate when the MH enters cell i. In our simulation, prefetch is employed at the WiFi APs only. As shown in Fig. 8, prefetch reduces the overall communication cost by 27% in average, compared to the nonprefetch case. Although results are not shown here, the benefit of prefetch increases with the increase of the gap between r_{ij}^m and r_{ij}^c. Prefetch incurs an extra buffering cost at the access point, which is assumed negligible. We also consider the scenario where the MH is equipped with more than one programmable wireless interface card. Fig. 8 shows the result of employing two interface cards, which lead to a drop of 50% in average in the communication cost, compared to the one-interface-card scenario. In the random-maps simulation, we fix one route and generate random maps. The result shown in Fig. 9 indicates that the LP model achieves similar cost reduction, as we discussed in Fig. 8.

In the above discussion, we have ignored the possible overhead that may result in additional communication cost. The proposed CM approach introduces three types of overhead, i.e., signaling, computing, and handoff overheads. They are elaborated below. First, the signaling overhead is there since the MH needs to obtain such information as network map and MH mobility before initiating CM. The impact of the signaling overhead on the communication cost includes the extra transmission time and the transmission cost associated with this

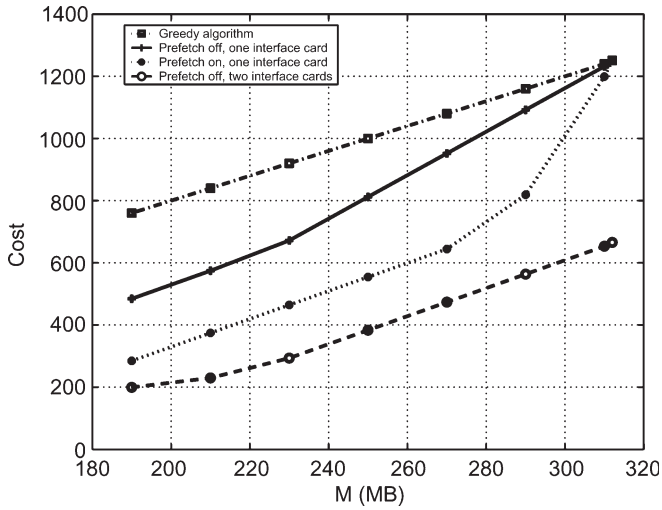

Fig. 9. Performance averaged over random maps.

signaling data. With proper design of data structure, the total amount of signaling data is far less than 1 kB, which takes less than 3 ms of transmission time, given the core network data rate in Table I. The actual transmission cost of the signaling data can be calculated according to the core-network data cost in Table I. The second type of overhead is introduced by the computing time needed to solve the LP model. Our simulation shows that the computation delay is usually between 10 to 20 ms on a Celeron 2-GHz CPU by using the LP-SOLVE [31] tool. The third type of overhead is caused by handoff, which can be either introsystem or intersystem handoff. The exact handoff delay depends on the handoff strategy and the network-traffic condition. Typical handoff delays among the three different types of APs are listed below: The handoff delay between WiFi and 3G is around 400 ms [32]; the handoff delay between 3G cells is usually less than 100 ms [33]; and the handoff delay between WiFi cells is less than 100 ms [34]. Although without known experimental results, we assume that the handoff delay of WiMax is similar to that of WiFi, i.e., 100 ms between WiMax cells, 100 ms between WiMax cell and WiFi cell, and 300 ms between WiMax cell and 3G cell. Now, we want to estimate the delay by taking the above three different types of overhead into consideration. When the MH is inside a cell, the longest handoff delay this MH can experience is to hand off from WiFi to 3G and then to WiMax, which takes 700 ms. Considering the handoff delay of 100 ms to the next cell, we have 800 ms in total. Since the sum of the computing delay and the signaling delay is less than 100 ms, the total delay is less than 900 ms in a cell. For the worst-case estimation, we assume a delay of 1 s in each cell introduced by the three types of system overhead. Given this delay being subtracted from τ_i in each cell i and the transmission cost of signaling data counted into the total cost, we can compute the optimal data-delivery scheme by running the LP model. We fix one route, generate random maps, and then average their results. As shown in Table II, the increase in total communication cost due to system overhead is no more than 3%.

Now, we study the scenario with multiple delay bounds. We fix the route and the map and introduce three delay bounds,

TABLE II
IMPACT OF SYSTEM OVERHEAD ON COMMUNICATION COST

M(MB)	190	210	230	250	270	290
cost w/o overhead	285	374	464	554	644	819
cost with overhead	288	381	474	567	659	837
cost increased	1.1%	1.9%	2.2%	2.3%	2.3%	2.2%

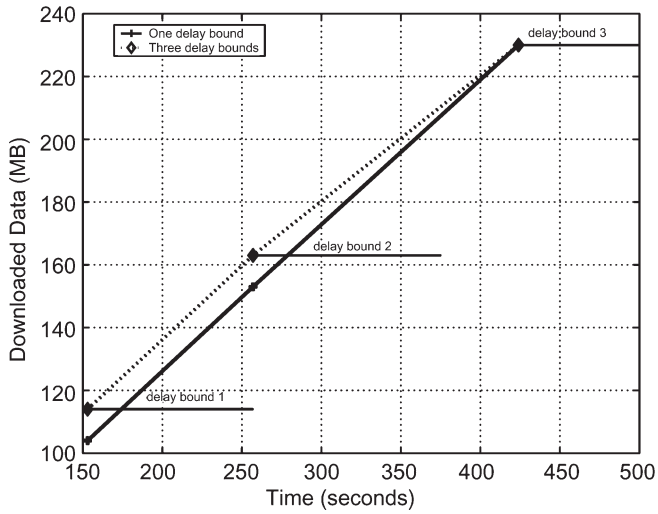

Fig. 10. Impact of delay bounds on downloaded data.

at $t_1 = 153$ s for 114 MB, at $t_2 = 257$ s for 163 MB, and at $t_3 = 424$ s for 230 MB. In Fig. 10, the downloaded data is compared with the scenario where only one delay bound is specified at time $T = 424$ s for $M = 230$ MB data. As shown in Fig. 10, an extra amount of data must be downloaded by time t_1 and t_2 in order to satisfy the corresponding delay bounds; thus, the MH is not able to postpone data transmission to future low-cost cells. Consequently, the overall communication cost rises in the multiple-delay-bounds scenario.

We have also studied the scenario where multiple possible routes exist. In particular, we choose three candidate routes and keep them fixed in the simulation. Associated with each route is its probability, which is denoted by P_1, P_2, and P_3, respectively. In our simulation, P_1 increases from 0.1 to 0.8, while $P_2 = P_3 = (1 - P_1)/2$. One delay bound at $T = 424$ s is set for $M = 230$ MB data, and one wireless interface card is used. It is indicated in Fig. 11 that, with the increase of P_1, the cost of route 1 decreases, since the optimization algorithm is more in favor of the route with higher probability. At the same time, the costs of routes 2 and 3 increase with P_1, and the overall cost decreases slightly. Two curves are shown in Fig. 12 for the cost of prefetched data of route 1 and the average cost of prefetched data of routes 2 and 3. As shown in Fig. 12, the increase of P_1 results in more prefetch in route 1 and less prefetch in routes 2 and 3. In general, a higher taken probability of a route leads to a more aggressive use of this route and a lower total communication cost in this route.

We change the MH's cell dwelling time to study the impact of variable moving speed of the MH. Let τ_i denotes the cell dwelling time of cell i, and δ denotes the percentage of variation of τ_i. We create three different patterns for the variation of the cell dwelling time. The first pattern is to increase τ_i by δ, ranging from 10% to 50%. The second pattern is to decrease

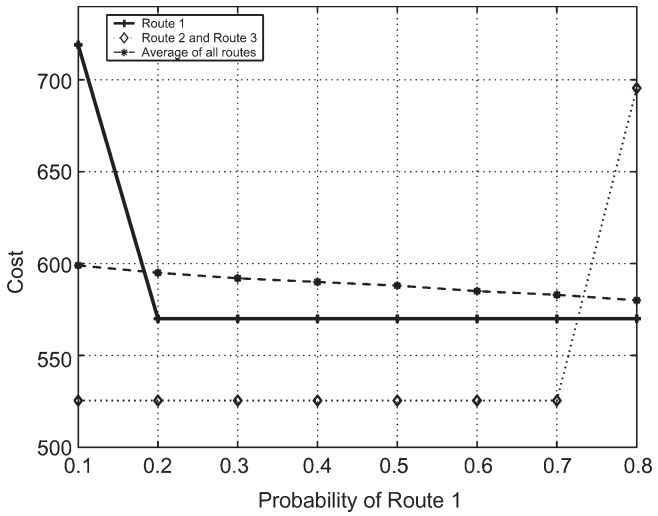

Fig. 11. Impact of route probability on communication cost.

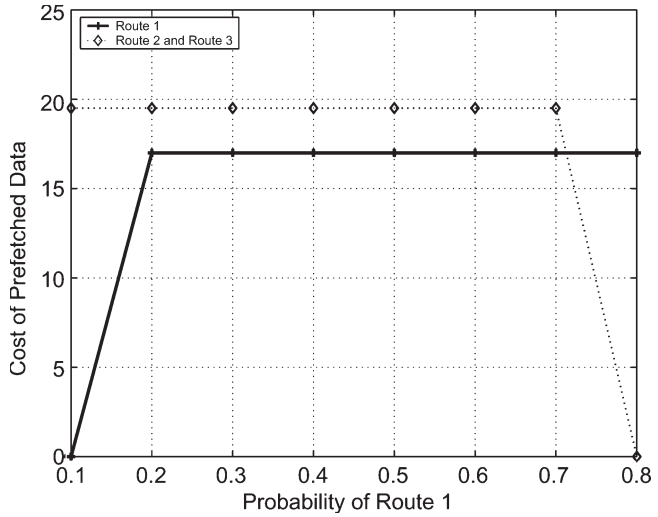

Fig. 12. Impact of route probability on prefetched data.

τ_i by δ, ranging from 10% to 50%. The third pattern is to let the cell dwelling time be uniformly distributed between $(1-\delta)\times\tau_i$ and $(1+\delta)\times\tau_i$. In our simulation, we fix a route, and let $M = 230$ MB. When the MH detects a change in τ_i, the LP model will be regenerated and rerun.

Simulation results of patterns 1 and 2 are shown in Figs. 13 and 14, respectively. We show the communication cost versus the number of reruns in these two figures. As shown in Fig. 13, when τ_i increases by δ from 10% to 50%, the corresponding communication cost drops accordingly. The reason is that the MH has more time to finish the download task due to the increase of τ_i in the cell, and thus, the MH can resort to cheaper AP to download data instead of more expensive ones. It is also true that the communication cost reduces with the increase of rerun times for any δ value. On the contrary, as shown in Fig. 14, if τ_i decreases by δ from 10% to 50%, the communication cost rises, or the MH cannot even finish the downloading task because it becomes infeasible for the data left to be downloaded in the remaining time.

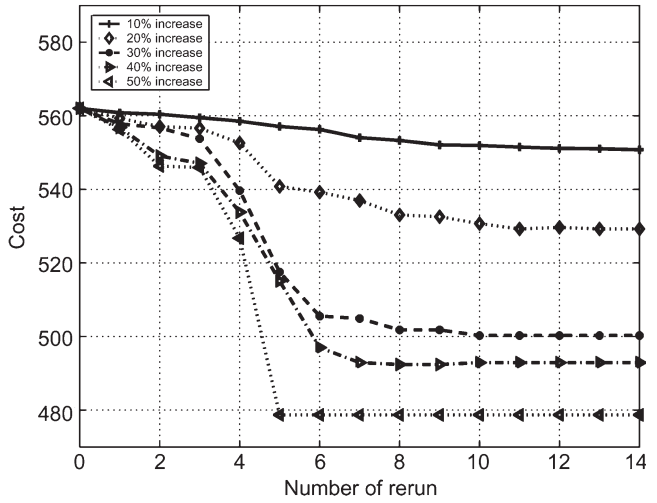

Fig. 13. Impact of increasing cell dwelling time.

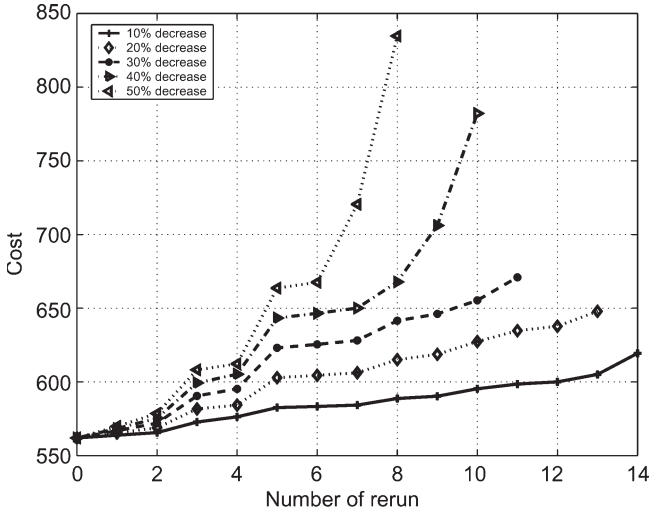

Fig. 14. Impact of decreasing cell dwelling time.

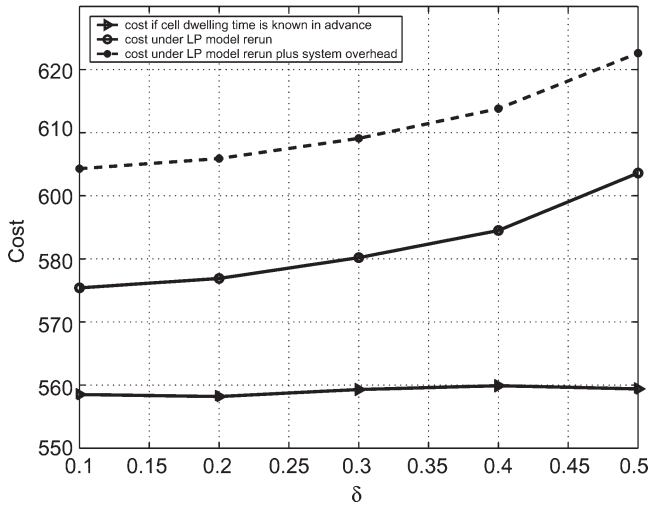

Fig. 15. Cost comparison when cell dwelling time varies randomly.

Fig. 15 shows the results under pattern 3, which compares the actual communication cost with the ideal communication cost if we knew the exact cell dwelling time in advance. The

motivation for such comparison is to learn the impact of variable speed on the proposed LP model. As depicted in Fig. 15, the difference in communication cost between these two scenarios increases from 3% to 6% when δ goes from 10% to 50%. The result shows that the proposed LP model works effectively by rerunning the LP model, even when the MH significantly varies its speed. Fig. 15 also shows that the system overhead due to handoff, rerunning the LP model, and signaling overhead is around 3%. This observation conforms to the simulation results of system overhead in Table II.

VII. Conclusion

This paper aims to minimize the communication cost in an overlay heterogeneous wireless network. We have investigated and formulated the CM problem, which has been proven to be NP-hard. We have proposed an efficient minimum-cost data-delivery algorithm based on LP, with various constraints, such as channel bandwidth, link costs, delay budget, and user mobility, taken into consideration. In case of insufficient bandwidth to the core network, we have proposed a prefetch scheme in order to fully utilize the wireless-network capacity. When multiple routes are available, a probability-based approach is adopted for CM. Extensive simulation has been carried out to evaluate the proposed CM schemes. Our results show that the proposed LP approach can effectively reduce the overall communication cost in various application scenarios, with small overhead ($< 3\%$) for signaling, computing, and handoff. The proposed algorithm can be integrated into future heterogeneous wireless networks and emerging 4G wireless systems for minimum-cost data delivery.

References

[1] V. MacDonald, "The cellular concept," *Bell Syst. Tech. J.*, vol. 58, no. 1, pp. 15–43, 1978.

[2] T. Rappaport, *Wireless Communications: Principles and Practice*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[3] R. Prasad, W. Konhauser, and W. Mohr, *Third Generation Mobile Radio Systems*. Norwood, MA: Artech House, 2000.

[4] *IEEE Standard for Local and Metropolitan Area Networks—Standard Air Interface for Mobile Broadband Wireless Access Systems Supporting Vehicular Mobility—Physical and Media Access Control Layer Specification*, 2002, IEEE 802.20. [Online]. Available: http://www.ieee802.org/20/

[5] IEEE 802.16e*IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Fixed Broadband Wireless Access Systems—Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, 2005, IEEE 802.16e.

[6] *IEEE Standard for Local and Metropolitan Area Networks—Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, 2004, IEEE 802.16.

[7] *IEEE Standard for Local and Metropolitan Area Networks—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Higher-Speed Physical Layer (PHY) Extension in the 2.4 GHz Band*, 1999, IEEE 802.11b.

[8] *IEEE Standard for Local and Metropolitan Area Networks—Part 15.1: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs)*, 2005, IEEE 802.15.1.

[9] *Cellular WiFi*. [Online]. Available: http://www.troposnetworks.com/

[10] J. Gibson, *The Mobile Communications Handbook*. Boca Raton, FL: CRC, 1996.

[11] J. Mitola, III, "The software radio architecture," *IEEE Commun. Mag.*, vol. 33, no. 5, pp. 26–38, May 1995.

[12] A. Munro, "Mobile middleware for the reconfigurable software radio," *IEEE Commun. Mag.*, vol. 38, no. 8, pp. 152–161, Aug. 2000.

[13] P. Rodriguez, R. Chakravorty, J. Chesterfield, I. Pratt, and S. Banerjee, "MAR: A commuter router infrastructure for the mobile Internet," in *Proc. 2nd Int. Conf. MobiSys*, 2004, pp. 217–230.

[14] R. Schollmeier, "A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications," in *Proc. 1st Int. Conf. Peer-to-Peer Comput.*, 2001, pp. 101–102.

[15] G. Fox, "Peer-to-peer networks," *IEEE Comput. Sci. Eng.*, vol. 3, no. 3, pp. 75–77, May/Jun. 2001.

[16] W. Winston and M. Venkataramanan, *Introduction to Mathematical Programming*. Pacific Grove, CA: Brooks/Cole, 2003.

[17] J. Blau, "Wi-Fi hotspot networks sprout like mushrooms," *IEEE Spectr.*, vol. 39, no. 9, pp. 18–20, Sep. 2002.

[18] X. Gao, G. Wu, and T. Miki, "QoS framework for mobile heterogeneous networks," in *Proc. IEEE Int. Conf. Commun.*, 2003, pp. 933–937.

[19] J. J. Garcia-Luna-Aceves, C. Fullmer, E. Madruga, D. Beyer, and T. Frivold, "Wireless Internet gateway (WINGS)," in *Proc. Mil. Commun. Conf.*, 1997, pp. 1271–1276.

[20] A. Campbell, J. Gomez, and A. Valko, "An overview of cellular IP," in *Proc. IEEE WCNC*, 1999, pp. 606–610.

[21] S. Bae, S. Lee, and M. Gerla, "Unicast performance analysis of extended ODMRP in a wired-to-wireless hybrid ad-hoc network," in *Proc. Mil. Commun. Conf.*, 2002, pp. 1228–1232.

[22] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad-hoc relaying systems: iCAR," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2105–2115, Oct. 2001.

[23] X. Wu, B. Mukerjee, and S. G. Chan, "MACA—An efficient channel allocation scheme in cellular networks," in *Proc. IEEE Global Telecommun. Conf.*, 2000, vol. 3, pp. 1385–1389.

[24] J. Zhou and Y. Yang, "Parcels: Pervasive *ad-hoc* relaying for cellular systems," in *Proc. 1st Annu. Mediterranean Ad Hoc Netw. Workshop*, 2002.

[25] Y. Lin and Y. Hsu, "Multihop cellular: A new architecture for wireless communication," in *Proc. IEEE INFOCOM*, 2000, pp. 1273–1282.

[26] H. Hsieh and R. Sivakumar, "On using the ad-hoc network model in cellular packet data networks," in *Proc. ACM MobiHOC*, 2002, pp. 36–47.

[27] B. Bhargava, X. Wu, Y. Lu, and W. Wang, "Integrating heterogeneous wireless technologies: A cellular aided mobile ad hoc network (CAMA)," *ACM Mobile Netw. Appl.*, vol. 9, no. 4, pp. 393–408, Aug. 2004.

[28] M. Stemm and R. H. Katz, "Vertical handoffs in wireless overlay networks," *ACM Mobile Netw. Appl.—Special Issue Mobile Networking Internet*, vol. 3, no. 4, pp. 335–350, 1998.

[29] H. J. Wang, R. H. Katz, and J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks," in *Proc. 2nd IEEE WMCSA*, 1999, pp. 51–60.

[30] T. Ibaraki, T. Hasegawa, K. Teranaka, and J. Iwase, "The multiple choice knapsack problem," *J. Oper. Res. Soc. Jpn.*, vol. 21, pp. 59–93, 1978.

[31] *LP-SOLVE*. [Online]. Available: http://sourceforge.net/projects/lpsolve

[32] R. Rajavelsamy, V. Jeedigunta, B. Holur, M. Choudhary, and O. Song, "Performance evaluation of VoIP over 3G-WLAN interworking system," in *Proc. IEEE WCNC*, 2005, pp. 2312–2317.

[33] J. Wang, J. C. L. Liu, and Y. Cen, "Handoff algorithms in dynamic spreading WCDMA system supporting multimedia traffic," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 10, pp. 1652–1662, Dec. 2003.

[34] I. Samprakou, C. Bouras, and T. Karoubalis, "Fast IP handoff support for VoIP and multimedia applications in 802.11 WLANs," in *Proc. 6th IEEE Int. Symp. WoWMoM*, 2005, pp. 332–337.

Haining Chen (S'04) received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 1996 and the M.S. degree in electrical engineering from the Chinese Academy of Sciences, Beijing, in 1999. He has been working toward the Ph.D. degree in computer engineering with the Center for Advanced Computer Studies, University of Louisiana, Lafayette, since 2003.

From 1999 to 2002, he was a Wireless-Networking Engineer in China. His current research interests include quality of service in broadband wireless-access networks, integration of heterogeneous wireless networks, and optimization for the next-generation wireless networks.

Hongyi Wu (M'02) received the B.S. degree in scientific instruments from Zhejiang University, Hangzhou, China, in 1996 and the M.S. degree in electrical engineering and the Ph.D. degree in computer science from the State University of New York, Buffalo, in 2000 and 2002, respectively.

He is currently an Assistant Professor with the Center for Advanced Computer Studies, University of Louisiana, Lafayette. His research interests include wireless mobile *ad hoc* networks, wireless-sensor networks, next-generation cellular systems, and integrated heterogeneous wireless systems. He has published more than 50 technical papers in leading journals and conference proceedings.

Dr. Wu was a Chair and technical committee member of several IEEE conferences and the Guest Editor of two Special Issues of *ACM MONET*. He was the recipient of the National Science Foundation CAREER Award in 2004.

Sundara Kumar (S'98) received the B.S. and M.S. degrees in electronic engineering and computer science from Bharathiyar University, Coimbatore, India, in 1999 and 2001, respectively. He has been working toward the Ph.D. degree in computer science with the Center for Advanced Computer Studies, University of Louisiana, Lafayette, since 2002.

Since 2005, he contributed to his startup business in the technology sector. His current research interests include integration and optimized use of the heterogeneous wireless-network environments and on developing a unifying and efficient IP protocol to access the heterogeneous wireless network.

Nian-Feng Tzeng (S'85–M'86–SM'92) received the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign.

Since 1987, he has been with the Center for Advanced Computer Studies, University of Louisiana, Lafayette, where he is currently a Professor. His current research interests are in the areas of computer communications and networks, high-performance computer systems, parallel and distributed processing, and fault-tolerant computing.

Dr. Tzeng was on the Editorial Board of the IEEE TRANSACTIONS ON COMPUTERS from 1994 to 1998 and the Editorial Board of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS from 1998 to 2001. He was a Distinguished Visitor of the IEEE Computer Society from 1994 to 1997 and was the Chair of Technical Committee on Distributed Processing of the IEEE Computer Society from 1999 to 2002. He was on the technical program committees of various conferences and served as the Technical Program Chair of the 10th International Conference on Parallel and Distributed Systems in July 2004. He was the recipient of the Outstanding Paper Award of the 10th International Conference on Distributed Computing Systems in May 1990. He was also the recipient of the University Foundation Distinguished Professor Award in 1997.