Assignment Strategies for Mobile Data Users in Hierarchical Overlay Networks: Performance of Optimal and Adaptive Strategies

Thierry E. Klein, *Member, IEEE,* and Seung-Jae Han, *Member, IEEE*

***Abstract*—Hierarchical wireless overlay networks have been proposed as an attractive alternative and extension of cellular network architectures to provide the necessary cell capacities to effectively support next-generation wireless data applications. In addition, they allow for flexible mobility management strategies and quality-of-service differentiation. One of the crucial problems in hierarchical overlay networks is the assignment of wireless data users to the different layers of the overlay architecture. In this paper, we present a framework and several analytical results pertaining to the performance of two assignment strategies based on the user's velocity and the amount of data to be transmitted. The main contribution is to prove that the minimum average number of users in the system, as well as the minimum expected system load for an incoming user, are the same under both assignment strategies. We provide explicit analytical expressions as well as unique characterizations of the optimal thresholds on the velocity and amount of data to be transmitted. These results are very general and hold for any distribution of user profiles and any call arrival rates. We also show that intelligent assignment strategies yield significant gains over strategies that are oblivious to the user profiles. Adaptive and on-line strategies are derived that do not require any *a priori* knowledge of the user population and the network parameters. Extensive simulations are conducted to support the theoretical results presented and conclude that the on-line strategies achieve near-optimal performance when compared with off-line strategies.**

***Index Terms*—Adaptive control, assignment strategy, decision threshold, hierarchical wireless networks, macrocell, microcell.**

I. Introduction

THE GROWTH in next-generation wireless networks is driven by the ever increasing popularity of wireless data applications. In addition to supporting voice services, future mobile networks are envisioned to be truly multimedia networks, which integrate a variety of different applications. Their success depends on the ability of future network architectures to provide the necessary capacities to support high data rate services, while at the same time dealing with the different mobility patterns of the users. To this effect, *hierarchical network architectures* with different cell layers have been proposed [3], [4]. Different layers are distinguished by their respective cell sizes, their maximum throughput, and the number of supportable users. For simplicity, we only consider two layers of cells, generically called the *macro*cell and the *micro*cell layers. In such schemes, macrocells are typically designed for universal coverage of the geographical region, while microcells provide high throughputs in local hot spots. In order to maximize the system performance, the assignment of users to the different layers of the architecture (i.e., to the different base stations), as well as the potential switching between layers during the operation of the network, is of crucial importance. The problem is of course only relevant to users with multiple connection choices.

In traditional cellular voice networks, the cell selection problem in the area with multiple cell coverage has been studied mostly around the concept of "directed retry" [5]–[8]. In this approach, a mobile station compares the channel condition of the cells that it can reach, and initiates a call setup request to a cell with the best channel quality. If the selected cell has a free channel, the request is accepted, but if it does not, the cell provides that user with a list of neighboring cells. The mobile station then sends a retry message to the cell, which has the best channel condition among the list. If the channel quality of all cells in the list is below the acceptable level or no cell has an available channel, the call setup request is rejected. As an enhancement, load balancing can be applied to the basic scheme. That is, when the number of users in a cell exceeds a certain threshold, the new users and/or the already existing users are advised to switch to an adequate neighboring cell. This approach is not suitable to data applications because (non-real-time) data applications do not require a fixed amount of resources and, therefore, a cell cannot make a simple decision whether to accept or reject a data connection request.

Another approach is the "velocity-sensitive" cell selection method. In this approach, the cell assignment decision is made based on the mobile's estimated velocity, so that slow mobiles can be assigned to microcells, while fast mobiles are assigned to macrocells. In [9], all newly arrived users are assigned to a microcell by default. When a user moves out of the coverage of his/her current cell, the cell dwell time is compared with a predetermined threshold. If a user is slowly moving (i.e., if the cell dwell time is larger than the threshold), the user is handed over to a neighboring microcell. Otherwise, the user is handed up to the macrocell. In [10], the channel condition information is utilized in conjunction with mobile velocity information. Essentially, the cell selection is done by comparing the signal strength of the

Manuscript received May 30, 2003; revised November 11, 2003. This paper was presented in part at the 2003 Conference on Information Sciences and Systems (CISS 2003), The Johns Hopkins University, Baltimore, MD, March 2003 [1] and in part at the IEEE 2003 Global Communications Conference (GLOBECOM 2003), San Francisco, CA, December 2003 [2].

The authors are with the Wireless Research Laboratory, Bell Laboratories-Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: tek@lucent.com; sjhan@lucent.com).

Digital Object Identifier 10.1109/JSAC.2004.826922

microcell and the macrocell, so that the cell with the stronger signal is selected. A certain negative offset is applied to the signal strength of the microcell, which becomes effective when the mobile enters the coverage of microcells and is reduced over time. Due to this negative offset, mobiles are more likely to select the macrocell at first, if the signal strengths of macrocells and microcells are comparable. However, as the offset decreases, more mobiles select the microcell if they remain in the coverage of microcells. As a result, slow mobiles would choose microcells, while the fast mobiles choose the macrocell. In [11], a systematic method to determine the threshold of mobile velocity for assignment decision is proposed. It probably is closest to our approach in that it uses the threshold-based method for optimal user assignment. That work, however, is aimed at cellular voice networks, focusing on minimizing the call blocking probability, which may not be the most relevant metric for data users. Other literature on velocity-based handoff techniques includes [12]–[14]. A comparative study is given in [15].

There exists a broad range of literature on the channel management problem in cellular networks. The reader is referred to [4] for an extensive survey of the literature. This issue, however, is not directly related to the problem that we tackle in this paper. The channel planning is more of a network planning problem, rather than a run-time user assignment problem. Nevertheless, both problems share the common goal of maximizing system efficiency in hierarchical wireless networks.

In a broad sense, our solution approach is somewhat similar to the optimal load balancing algorithms for distributed computer systems [16]. The contribution of our paper is the derivation of load balancing metrics which optimize certain system goals for hierarchical wireless data networks. In a narrower sense, our scheme can be classified as a form of optimal queue control [17]. [18] proposes a greedy run-time user assignment algorithm for wireless networks. In this myopic approach, users are assigned to a cell with minimum load without conducting repacking. It is different from our approach in that it requires continuous tracking of current actual load conditions of each cell, whereas our approach derives an optimal threshold from the user profile distribution and applies the same threshold unless the user population changes substantially.

The main contribution of this paper is to develop an analytical framework to evaluate the performance of assignment strategies for nonreal-time data users in hierarchical and other overlay networks. By nonreal-time users, we mean users who have a fixed amount of data to transmit and remain connected to the system until all of the data is transmitted to the intended receiver. In other words, the user connection time depends on the feasible transmission rate and the capacity awarded to each user, as well as the assignment strategy employed. Such a model is for example applicable to e-mail applications or file transfers, but would not apply to voice communications or video streaming for example. We analytically show that two schemes based on velocity and amount of data to be transmitted achieve the same system performance (in terms of the average number of users in the system and the expected system load seen by an incoming new user), and also yield the same stability region of call arrival rates. The optimal threshold values on velocity and data amount are calculated explicitly. Our results are very general and do not depend on the statistical description of the users' profiles or the underlying link-layer technologies. Unique characterizations of the optimal decision thresholds are derived and used to devise adaptive and on-line assignment strategies that do not require any *a priori* knowledge of the user profiles, the cell capacities or the call arrival rates.

The remainder of this paper is organized as follows. In Section II, we present a general description of our model and the set of assumptions made throughout the paper. The two assignment strategies based on velocity and amount of data are presented in Section III. In Section IV, the main theoretical results of this paper are derived. Numerical results obtained through simulations are shown in Section V. Conclusions follow in Section VI.

II. System Model and Assumptions

In this section, we describe our hierarchical network model of macrocell and microcell layers and state the general assumptions made throughout the paper. This setup might correspond to a macro-micro or a micro-pico cellular network context. However, the proposed strategy and the corresponding results are more general and may very well be extended to multitiered architectures. The proposed framework may equally well model the integration of different network technologies, such as, for example, a 3G-802.11 integrated network [19]. We assume that the network topology and the layout of the macrocells and microcells is fixed and known to the decision process. The microcells geographically underlie a macrocell but do not necessarily cover the entire macrocell. We also assume that an underlying control structure, instructing users to connect to the macrocells or the microcells, is available. Mobile devices may transmit (and receive) to (from) either the macrocell or the microcell (but not simultaneously).

The users' behavior is modeled by several stochastic processes that facilitate the analysis of the proposed assignment strategies. First of all, the *aggregate data traffic* is modeled as two Poisson call arrival[1] processes in the coverage region of the microcells and the macrocells, of rates λ_1 and λ_2, respectively. In general, λ_1 and λ_2 may be different and allow us to model hot spots with high call volumes in the geographical region of the microcell. At this point, we note that a call may originate in the micro coverage region, but still be assigned to the macrocell (if this is deemed to increase the system-wide performance). Thus, λ_1 and λ_2 are not the actual arrival rates (after assignment) to the microcells and macrocells. The mobility of user k is measured by its *average velocity* V_k. It is recognized that the instantaneous speed of mobiles changes continuously, but only the average speed (calculated over an appropriate time horizon) should influence the assignment decision. A choice of a small time horizon induces frequent reassignments and switching between layers, leading to excessive signaling overhead, potential handoff failures, and instability. The amount of data (in bits) that user k wants to transmit is exponentially distributed with mean D_k. Both V_k and D_k are random variables and each user's *velocity-data profile* is chosen according to the joint probability

[1]Even though we consider the context of data users, we refer to the beginning of a data session as a call arrival.

density function (pdf) $f_{V,D}(v,d)$, defined on the two-dimensional (2-D) velocity and data amount plane and assumed to be continuous and differentiable. The maximum velocity and data amount are denoted by $V_{\max}$ and $D_{\max}$. Without loss of generality, the respective minimum values are set to $V_{\min} = 0$ and $D_{\min} = 0$. R_f (in bits per second) is the *maximum feasible transmission rate* for each user. R_f is independent of the users' profiles, but depends on the number of users at each layer in the system and the underlying technologies and influences the connection time of the users. Indeed the time required to transmit the amount of data of user k is exponentially distributed with mean $\overline{T}_k = D_k/R_f$ (and, thus, with rate $\eta_k = R_f/D_k$).

We identify the *state of the system* (i, j) as the number of users in the microcell and macrocell.[2] The system remains in a certain state until either a call in the microcell or the macrocell is completed (meaning all the data for that user has been transmitted), a new call comes in or there is a mobility-induced handoff between macrocells and microcells. The call termination process now is the superposition of multiple Poisson processes and is, therefore, again a Poisson process. For example, the rate of the call termination process out of the macrocell is the sum of the rates of the corresponding j users in state (i, j), namely $\eta_c^{(m)}(j) = \sum_{k=1}^{j} \eta_k$. It is immediately seen that this rate depends on the set of users in the state (and not just on the number of these users). We do not model this fine granularity of the system and, therefore, are not able to distinguish between different combinations of users assigned to the cell. As an approximation for our theoretical calculations, we identify each user with the average behavior among all the users and further approximate the call duration as an exponentially distributed random variable with mean $\overline{D}/R_f$, where $\overline{D}$ is the average data size, averaged over the user profiles. This approximation is acceptable for the numerical values considered for the typical range of D_k and R_f, and becomes more accurate as the average time required to transmit the data becomes smaller, i.e., as R_f increases or as $D_{\max}$ decreases. We continue to make this approximation in order to derive our theoretical results. The validity of this approximation is assessed in the simulations which show a close match between the theoretical results and those of the simulations (conducted without this approximation). Thus, from now on, we assume that the call termination process (when j users are in the macrocell) is modeled as a Poisson process of rate

$$\eta_c^{(m)}(j) = \frac{jR_f^{(m)}(j)}{\overline{D}_m} \tag{1}$$

where $\overline{D}_m$ is the average amount of data to be transmitted by users in the macrocell (averaged over the corresponding profiles). Similarly, the call termination rate out of the microcell with i users is

$$\eta_c^{(\mu)}(i) = \frac{iR_f^{(\mu)}(i)}{\overline{D}_\mu}. \tag{2}$$

[2]The state (i, j) actually denotes the number of ongoing sessions in the microcell and the macrocell. We allow users to simultaneously open multiple sessions and treat multiple sessions of the same user as different users. With a slight abuse of notation, we continue to refer to (i, j) as the number of users in the given state.

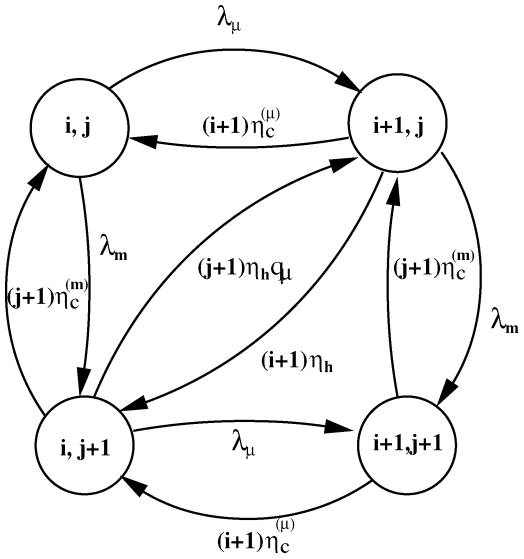

Fig. 1. Markov chain model for dynamic behavior of macrocell-microcell assignment schemes.

$R_f^{(m)}(j)$ and $R_f^{(\mu)}(i)$ are the maximum per user transmission rates of the macrocell and microcell with j and i users, respectively, and depend on the physical- and link-layer technologies (e.g., available bandwidth, multiaccess protocol, scheduling algorithms, power constraints). We are now in the position to conclude that the dynamic behavior of the system can be analyzed by a 2-D Markov chain with an infinite number of states in each dimension, a generic step of which is shown in Fig. 1. The transition probabilities are related to the call arrival, call holding and mobility parameters. They also critically depend on the assignment schemes. Let λ_m and λ_μ denote the call arrival rates to the macrocell and microcell. η_h is the mean outgoing handoff rate per calling mobile from a microcell. Similar to [11], we assume that the handoff process is Poisson and approximate the handoff rate by the mean boundary crossing rate η_b

$$\eta_h \approx \eta_b = \frac{2\overline{V}_\mu}{\pi r_\mu}$$

where $\overline{V}_\mu$ is the average velocity of users with microcell profiles and r_μ is the radius of the microcell (assumed to be circular for convenience). We point out that λ_m, λ_μ, $\eta_c^{(m)}$, $\eta_c^{(\mu)}$, and η_h all depend on the assignment strategy under consideration.

We emphasize the crucial distinction between the user's *location*, the *profile*, and the *assignment*. The user's location refers to the geographical position of the user and is independent of the assignment strategy. On the other hand, a user has a macrocell/microcell profile if its (V, D) characteristic is consistent with the requirements for assignment to the macrocell or the microcell with respect to the assignment scheme under consideration. The profile refers to the *preferred* assignment of the user. Finally, the user's assignment refers to the outcome of the decision process (i.e., the *actual* assignment). The subtle distinction is especially important when the macrocell is not fully covered by microcells. In this case, users with micro profiles who are not in the coverage of the microcells are assigned to the macrocell. In other words, a user is only assigned to the microcell if it has a micro profile and is in the coverage region of the microcell.

Otherwise, the user is always assigned to the macrocell. Denote by q_m and q_μ the probabilities that a given user has a macrocell, respectively, a microcell profile. Of course, q_m and q_μ also depend on the assignment strategy. Then, the probabilities that a user is assigned to the macrocell, respectively, the microcell are given as

$$\Pr(m) = q_m + \frac{\lambda_2}{\lambda_1 + \lambda_2} q_\mu \tag{3}$$

$$\Pr(\mu) = \frac{\lambda_1}{\lambda_1 + \lambda_2} q_\mu. \tag{4}$$

Therefore, the call arrival rates to the macrocell and the microcell can be computed as follows:

$$\lambda_m = \lambda_2 + \lambda_1 q_m \tag{5}$$

$$\lambda_\mu = \lambda_1 q_\mu. \tag{6}$$

In the proof of Theorem 1, we show how $\overline{D}_m$ and $\overline{D}_\mu$ are computed as functions of the assignment strategies. The calculation of $\overline{V}_\mu$ involves the same logical steps.

III. Assignment Strategies

We now introduce the two *assignment strategies* considered in this paper. An assignment strategy is a rule to divide the entire user population into two subsets (in the case of a two-tiered architecture). In general, the user population has to be divided in as many subsets as there are layers in the architecture. For ease of exposition, we restrict our attention to two-tiered architectures, but we emphasize that all the concepts and results naturally extend to more general architectures as well. In this paper, we only consider assignment strategies that depend on one of the parameters of the user's profile, either the velocity or the data size, but not on the joint profile information. Intuitively it makes sense to consider threshold-based assignment strategies and in fact, in Section IV, we show the optimality of such assignment strategies.

The first strategy is a *velocity-based assignment strategy* (VAS), characterized by a velocity threshold V_0: users with average velocity $V > V_0$ ($V \le V_0$) have a macro (micro) profile. The rationale behind this scheme is that assigning fast-moving users to macrocells and slower users to microcells reduces the number of required handoffs. Rather than basing the decision on full instantaneous knowledge (position, instantaneous velocity, direction of movement) of the wireless network, the decision depends solely on the average velocity of the users. The velocity of the users can be estimated from the cell sojourn times [11]. Other methods include measuring the propagation delay periodically, estimating the maximum Doppler spread [20], estimating the velocity based on higher order crossings [21], level crossing and zero crossing rates [22], or based on a covariance approximation [22], [23]. The profile probabilities q_m and q_μ can now be computed as

$$q_m = \int_{v=V_0}^{V_{\max}} \int_{d=0}^{D_{\max}} f_{V,D}(v,d)dv\,dd = \int_{v=V_0}^{V_{\max}} f_V(v)dv$$

$$q_\mu = \int_{v=0}^{V_0} \int_{d=0}^{D_{\max}} f_{V,D}(v,d)dv\,dd = \int_{v=0}^{V_0} f_V(v)dv.$$

The second assignment strategy, called the *data-based assignment strategy* (DAS), is based on the average amount of data to be transmitted by the user. One would envision a strategy that assigns users with small data amounts to the macrocells, while users with large data amounts are assigned to the microcells. The rationale behind this strategy comes from the fact that the inherent capacity of microcells is typically larger than that of macrocells. Instead of a velocity threshold, we now consider a threshold D_0 on the average amount of data. Users, whose amount of data is larger (smaller) than the threshold, have a micro (macro) profile. The macro/micro profile probabilities are now computed as

$$q_m = \int_{d=0}^{D_0} \int_{v=0}^{V_{\max}} f_{V,D}(v,d)dv\,dd = \int_{d=0}^{D_0} f_D(d)dd$$

$$q_\mu = \int_{d=D_0}^{D_{\max}} \int_{v=0}^{V_{\max}} f_{V,D}(v,d)dv\,dd = \int_{d=D_0}^{D_{\max}} f_D(d)dd.$$

Note that the profile probabilities only depend on the marginal profile distributions and not on the joint distribution (even when the velocity and the data size are not assumed to be independent of each other). The main objective of the paper is to determined the optimal decision thresholds V_0^* and D_0^* for two different system-wide objectives.

IV. Analytical Results

We now derive some analytical results pertaining to the performances of the VAS and DAS strategies. In particular, our main theorem shows that the minimum average number of users in the system is the same under the optimal VAS and DAS schemes, and the optimal thresholds V_0^* and D_0^* are explicitly computed. Similarly, we also show that the same qualitative conclusion holds when the objective is to minimize the expected system load seen by an incoming new user. We then derive unique characterizations of the optimal decision thresholds which are used to devise adaptive and on-line assignment strategies. As opposed to the optimal off-line strategies, these on-line strategies do not require any *a priori* knowledge of the statistical description of the user population, the call arrival rates or the cell capacities. In order to derive the theoretical results in this section, we need to make the following additional assumptions. Our simulation results in Section V, however, show that the qualitative results presented here hold true even without these assumptions.

Assumption 1: There are no handoffs between macrocells and microcells.

Assumption 2: The cell capacities at the macro and micro layers are constant and independent of the number of users. In other words, $i\ R_f^{(\mu)}(i) \doteq C_\mu$ and $j\ R_f^{(m)}(j) \doteq C_m, \forall\, i, j$.

Assumption 1 is reasonable when the user mobility is limited with respect to the data transmission time. Assumption 2 is reasonable for systems in which a fixed air-link capacity is shared evenly between the users. This assumption may not hold for systems that exploit multiuser diversity.

A. Minimum Average Number of Users in Systems

In this first section, the objective is to assign the users to the different layers in the hierarchical architecture so as to minimize the total average number of users in the system. More efficient assignment strategies allow admitted users to depart quicker, thereby freeing available resources for the remaining users, which then depart from the system faster than under nonoptimal assignment strategies. Minimizing the average number of users in the system is, therefore, a desirable objective as it allows for maximum utilization of the available resources and, therefore, maximizes the network capacity (in terms of the maximum number of subscribers that can stably be supported by the network). Let $E[N_m]$ and $E[N_\mu]$ be the average number of users in the macro, respectively, the microcell under a given assignment strategy, parameterized by V_0 or D_0. $E[N_{sys}] = E[N_m] + E[N_\mu]$ denotes the total average number of users in the hierarchical network. The following theorem is the main result of this section.

Theorem 1: Under assumptions 1 and 2, the minimum average number of users in the system is the same under the optimal DAS and VAS strategies. We distinguish three cases.

Case 1) When $(\lambda_1 + \lambda_2)\ E[D] < \sqrt{C_m}\ [\sqrt{C_m} - \sqrt{C_\mu}]$, the objective is monotonically decreasing in D_0 and monotonically increasing in V_0. Thus, the optimal thresholds are $D_0^* = D_{\max}$ and $V_0^* = 0$ and the corresponding minimum average number of users in the system is

$$E\left[N_{\text{sys}}^*\right] = \frac{(\lambda_1 + \lambda_2)E[D]}{C_m - (\lambda_1 + \lambda_2)E[D]}.$$

Case 2) When $(\lambda_1\ E[D] - C_\mu)\ \sqrt{C_m} < (\lambda_2\ E[D] - C_m)\ \sqrt{C_\mu}$, the objective is monotonically increasing in D_0 and monotonically decreasing in V_0. Thus, the optimal thresholds are $D_0^* = 0$ and $V_0^* = V_{\max}$, leading to a minimum average number of users in the system of

$$E\left[N_{\text{sys}}^*\right] = \frac{\lambda_1 E[D]}{C_\mu - \lambda_1 E[D]} + \frac{\lambda_2 E[D]}{C_m - \lambda_2 E[D]}.$$

Case 3) Otherwise, there exist optimal thresholds given as the solution of the following integral equations

$$\int_0^{D_0^*} x f_D(x)dx = \frac{C_m\sqrt{C_\mu} - \sqrt{C_m}C_\mu - (\lambda_2\sqrt{C_\mu} - \lambda_1\sqrt{C_m})E[D]}{(\sqrt{C_m} + \sqrt{C_\mu})\lambda_1} \quad (7)$$

$$\int_0^{V_0^*} f_V(x)dx = \frac{(\lambda_1+\lambda_2)E[D]\sqrt{C_\mu} - C_m\sqrt{C_\mu} + C_\mu\sqrt{C_m}}{(\sqrt{C_m} + \sqrt{C_\mu})\lambda_1 E[D]}. \quad (8)$$

The minimum average number of users in the system is then determined as

$$E\left[N_{\text{sys}}^*\right] = \frac{2(\lambda_1 + \lambda_2)E[D] + 2\sqrt{C_m C_\mu} - C_m - C_\mu}{C_m + C_\mu - (\lambda_1 + \lambda_2)E[D]}.$$

Proof: The details of the proof are found in the Appendix. However, we note that the minimum average number of users only depends on the cell capacities, the call arrival rates, and the average data size, but does not depend explicitly on the user profile distribution, although the optimal thresholds do depend on the profile distribution.[3] ■

Proposition 1: Threshold-based assignment strategies are optimal in the sense that they achieve the minimum total average number of users in the system for the DAS and the VAS assignment strategies.

Proof: The proof is obtained along the same lines as that of Theorem 1 by considering multiple division points in the intervals $[0, D_{\max}]$ and $[0, V_{\max}]$, and showing that the optimal system performance coincides with that of threshold-based assignment strategies. ■

Proposition 2: Under the optimal decision thresholds V_0^* and D_0^*, load balancing is achieved between the macrocells and microcells such that

$$\text{VAS}: \quad \frac{C_m - g_m(V_0^*)}{\sqrt{C_m}} = \frac{C_\mu - g_\mu(V_0^*)}{\sqrt{C_\mu}} \quad (9)$$

$$\text{DAS}: \quad \frac{C_m - g_m(D_0^*)}{\sqrt{C_m}} = \frac{C_\mu - g_\mu(D_0^*)}{\sqrt{C_\mu}} \quad (10)$$

where $g_m = \lambda_m \overline{D}_m$ and $g_\mu = \lambda_\mu \overline{D}_\mu$. This balancing metric uniquely characterizes the optimal decision thresholds.

Proof: The proof of this result is obtained as an intermediate step of the proof of Theorem 1. Note that this result only applies to the third case as described in Theorem 1, i.e., when the optimal thresholds are not equal to the minimum or maximum permissible values. ■

The above proposition implicitly defines the balancing metric when the objective is to minimize the average number of users in the system. In other words, the balancing metric at the macrocells and microcells are calculated as $X_m = ((C_m - g_m)/\sqrt{C_m})$ and $X_\mu = ((C_\mu - g_\mu)/\sqrt{C_\mu})$. It is noteworthy that, even though the objective is to minimize the average number of users in the system, the balancing metric is in general not equal to the number of users at the macro and microlayers. Indeed, one could have expected that the optimal decision thresholds would aim at equalizing the number of users in each layer of the architecture. This simpler rule is easily implemented and would simply compare the number of

[3]We also point out that the results of Theorem 1 can be used for dimensioning and provisioning of the network. In some of today's systems [19], users would always be assigned to the microcell whenever they are in the coverage region of the microcell. This corresponds to choosing $D_0^* = 0$, or equivalently, $V_0^* = V_{\max}$. The result of Theorem 1, especially the characterization of Case 2) can be used to determine the region of cell capacities and/or arrival rates for which this strategy is optimal. Such information is important for service providers when deploying and dimensioning their networks for a given traffic load and call arrival rates (and equivalently for call admission control for given network provisioning and cell capacities).

users in each cell and require that the threshold be adjusted in order to ensure that $E[N_m] = E[N_\mu]$. Proposition 2 shows that such a rule is in general suboptimal, unless the cell capacities at the macro and microlayers are equal. However, we will see in Section IV-B that this is the optimal balancing metric and the characterization of the optimal decision thresholds when the objective is to minimize the expected system load.

Proposition 3: The set of arrival rates (λ_1, λ_2) for which the system (i.e., the Markov chains) is stable can be explicitly determined and only depends on the cell capacities and the average data size to be transmitted. We distinguish two cases.

Case 1) If $C_m \geq C_\mu$

$$\begin{cases} \lambda_2 < \frac{C_m}{E[D]} \\ \lambda_1 + \lambda_2 < \frac{C_m + C_\mu}{E[D]}. \end{cases}$$

Case 2) If $C_m < C_\mu$

$$\begin{cases} \lambda_1 < \frac{C_\mu}{E[D]} \\ \lambda_1 + \lambda_2 < \frac{C_m + C_\mu}{E[D]}. \end{cases}$$

Proof: The result of this proposition is relatively straightforward to derive by considering the three different cases given in Theorem 1, and ensuring that the denominators in each of the expressions for the average number of users in the system remains positive. These latter conditions correspond of course to the stability condition in $M/M/1$ chains. ■

Note that the stability region is convex as it is determined by two linear constraints on λ_1 and λ_2. The two constraints, however, are not symmetric and, therefore, the stability region is not symmetric in λ_1 and λ_2. This could be expected, as the arrival rates to the macrocells and microcells (i.e., the arrival rates in the two Markov chains, λ_m and λ_μ) are not symmetric in λ_1 and λ_2. This follows from the fact that a user is only assigned to the microcell if it is in the micro coverage region *and* if it has a micro profile. Some of the arrivals from the micro coverage region may be "overflown" to the macrocell through the intelligent choice of the thresholds (i.e., through the characterization of macro and micro profiles), whereas the reverse is not possible.

B. Minimum Expected System Load

We now consider a different objective and show very similar results to the ones derived in Section IV-A. Indeed in this section, the objective of the assignment strategies is to minimize the expected system load seen by an incoming new user. While minimizing the average number of users in the system can be viewed as a network-wide objective, the expected system load seen by an incoming user is a user-centric objective. The expected system load is defined as the average number of bits in the system awaiting transmission when a new call arrives, and is mathematically calculated as

$$E[L_{\text{sys}}] = \Pr(m)\overline{D}_m E[N_m] + Pr(\mu)\overline{D}_\mu E[N_\mu]$$

where $\Pr(m)$ and $\Pr(\mu)$ are the probabilities that a user is assigned to the macrocell, respectively, the microcell. $\overline{D}_m$ and $\overline{D}_\mu$ are the average amounts of data to be transmitted by users assigned to the macrocell, respectively, the microcell. $E[N_m]$ and $E[N_\mu]$ are the average numbers of users in the macrocell and the microcell. The following theorem is the main result of this section and the equivalent of Theorem 1. It is important to point out that, while the optimal performance of the VAS and DAS strategies are the same, the optimal thresholds are computed according to different integral equations, depending on the objective function under consideration.

Theorem 2: Under assumptions 1 and 2, the minimum expected system load seen by an incoming new user is the same under the optimal DAS and VAS strategies. We distinguish two cases.

Case 1) When $\lambda_1 C_m \leq \lambda_2 C_\mu$, the optimal thresholds are $D_0^* = 0$ and $V_0^* = V_{\max}$, leading to an expected system load of

$$E\left[L_{\text{sys}}^*\right] = \frac{1}{\lambda_1 + \lambda_2}\left\{\frac{[\lambda_1 E[D]]^2}{C_\mu - \lambda_1 E[D]} + \frac{[\lambda_2 E[D]]^2}{C_m - \lambda_2 E[D]}\right\}.$$

Case 2) When $\lambda_1 C_m > \lambda_2 C_\mu$, we have that

$$E\left[L_{\text{sys}}^*\right] = \frac{(\lambda_1 + \lambda_2)\left(E[D]\right)^2}{C_m + C_\mu - (\lambda_1 + \lambda_2)E[D]}.$$

The optimal velocity and data thresholds are computed as solutions to the following integral equations

$$\int_0^{D_0^*} x f_D(x)dx = \frac{(\lambda_1 C_m - \lambda_2 C_\mu)E[D]}{(C_m + C_\mu)\lambda_1}$$

$$\int_0^{V_0^*} f_V(x)dx = \frac{(\lambda_1 + \lambda_2)C_\mu}{(C_m + C_\mu)\lambda_1}.$$

Proof: The proof follows the same logical steps as for Theorem 1 and is omitted here. ■

The following propositions are easily obtained as corollaries of the proof of the above theorem.

Proposition 4: Under the optimal decision thresholds V_0^* and D_0^*, load balancing is achieved between the macrocells and microcells such that $E[N_m^*] = E[N_\mu^*]$.

Proof: The result of the corollary follows directly from an intermediate step of the proof of Theorem 2 and the fact that the average number of users in the macrocells and microcells are given, respectively, by $E[N_m] = (g_m/(C_m - g_m))$ and $E[N_\mu] = (g_\mu/(C_\mu - g_\mu))$. Accordingly, when the system objective is to minimize the expected system load, the balancing metric is chosen as $X_m = N_m$ and $X_\mu = N_\mu$, where N_m and N_μ are the number of users in the macro, respectively, the microcell, and are averaged over an appropriate time horizon. ■

Proposition 5: The set of arrival rates (λ_1, λ_2) for which the system is stable can be explicitly determined and only depends on the cell capacities and the average data size to be transmitted

$$\begin{cases} \lambda_2 < \frac{C_m}{E[D]} \\ \lambda_1 + \lambda_2 < \frac{C_m + C_\mu}{E[D]}. \end{cases}$$

C. Adaptive Assignment Strategies

The main results of the previous subsections show that the optimal performances (both in terms of minimizing the average number of users in the system and of minimizing the expected system load) are the same for both VAS and DAS. The optimal thresholds are explicitly computed as solutions of integral equations. This derivation, however, requires knowledge of the macrocell and microcell capacities, the profile distribution of velocity and data amount across the user population, and the call arrival rates. Unfortunately in practical systems, this knowledge may not be available and, thus, would have to be estimated and is typically time varying. Therefore, the need arises for on-line strategies that do not require such *a priori* knowledge, but still achieve the same performance as if this knowledge were available. Equally important is the development of adaptive strategies that adjust the decision thresholds to the network conditions and the user behavior. In Propositions 2 and 4, we have identified unique characterizations of the system that are only achieved when the optimal thresholds are selected. These characterizations are now used to adapt the decision thresholds and compute them in an on-line fashion.

Consider an *update interval* of length T_{update}. The thresholds are held constant during an update interval and updated only at the beginning of each update interval. Let $X_m[k]$ and $X_\mu[k]$ be the measured or estimated values of the balancing metric at the macrocells and microcells at the beginning of the kth update interval, as defined in Propositions 2 and 4. The decision thresholds are then updated proportionally to the imbalance between $X_m[k]$ and $X_\mu[k]$. Specifically, the velocity threshold is updated according to the following rule:

$$V_0[k+1] = \min\left\{ \max\left\{0, V_0[k] + \frac{\beta_v}{k^{\gamma_v}}\left[X_m[k] - X_\mu[k]\right]\right\}, V_{\max}\right\} \quad (11)$$

where the *update magnitude parameter* β_v is a parameter to be tuned in order to regulate the speed of convergence of the algorithm. γ_v is the *time discounting factor*, which results in the algorithm making smaller adjustments as the number of updates performed is increased (and, hence, as the velocity threshold becomes closer to the intended value). The updated value of the velocity threshold is of course constrained to be in the interval $[0, V_{\max}]$. The rationale for this rule is as follows: If $X_m[k] > X_\mu[k]$, then there are too many users assigned to the macrocell resulting in an imbalance in favor of the microcell. Hence, the threshold should be adjusted in such a way as to increase the number of users in the microcell, i.e., V_0 should be increased.[4] Similarly, if $X_m[k] < X_\mu[k]$, V_0 should be decreased. A drawback of the above method is that it is very reactive to the measurements in the last update interval. A somewhat smoother update rule considers the exponentially weighted moving average of the difference in the balancing metric to update the threshold. Let

$$\Delta[k+1] = (1-\alpha_v)\left[X_m[k] - X_\mu[k]\right] + \alpha_v \Delta[k] \quad (12)$$

[4]We implicitly assume here that the balancing metric at the macro, respectively, the microcell, is an increasing function of the number of users assigned to the macro, respectively, the microcell. This assumption is indeed verified for the balancing metrics derived in this paper.

with $\Delta[0] = 0$ and $\alpha_v \in [0,1)$ a *smoothing factor* that can be tuned to give more or less weight to the past measurements. The update is then performed according to the rule

$$V_0[k+1] = \min\left\{\max\left\{0, V_0[k] + \frac{\beta_v}{k^{\gamma_v}}\Delta[k+1]\right\}, V_{\max}\right\}. \quad (13)$$

Note that the two update rules coincide when $\alpha_v = 0$. It is obviously seen that, when the algorithm converges, $\Delta[k] = 0$, leading to $X_m[k] = X_\mu[k]$ as desired. The term β_v/k^{γ_v} is needed in order to guarantee convergence of the update algorithm and to avoid limit cycles, by making the incremental change of the velocity threshold smaller. Another objective of our algorithm is to react to changes in the arrival rate or the user profile distribution. Such changes are indirectly detected by an imbalance of the balancing metric, which leads to larger values of $\Delta[k]$, which in return trigger an update of the velocity threshold. However, in order to allow the algorithm to react relatively quickly to such changes, we need to avoid small incremental updates. This is accomplished by periodically resetting the value of k to 1. Let T_{reset} be the number of update steps of the algorithm before k is reset to 1. Ideally, we would like to reset k right before a change in the optimal/target velocity threshold is warranted. In a completely on-line and adaptive solution, these changes are unpredictable and one solution is to periodically reset k. Another solution would be to reset k to 1 whenever the difference between $X_m[k]$ and $X_\mu[k]$ becomes very small. This latter method is preferred in our simulations.

We now provide a similar update rule for the data threshold in the context of DAS assignment strategies, which mirrors the update rule for VAS

$$D_0[k+1] = \min\left\{\max\left\{0, D_0[k] - \frac{\beta_d}{k^{\gamma_d}}\Delta[k+1]\right\}, D_{\max}\right\}. \quad (14)$$

$\Delta[k+1] = (1-\alpha_d)[X_m[k] - X_\mu[k]] + \alpha_d \Delta[k]$ with $\Delta[0] = 0$ and $\alpha_d \in [0,1)$. α_d, β_d, and γ_d are parameters to be tuned (and which *a priori* should be different from α_v, β_v and γ_v). Note that the only difference with the update rule for V_0 is the minus sign. This difference comes from the fact that, in order to increase the number of users in the microcell, the data threshold D_0 should be decreased (whereas the velocity threshold V_0 should be increased). All of the comments, especially related to the resetting of k and the role of the different parameters, equally well apply here.

V. Numerical Results

In this section, we present some simulation results to support the statements of the main theorems of this paper, even when assumptions 1 and 2 are not made. For the sake of brevity of the paper and to avoid unnecessary duplication, we choose to concentrate on the objective of minimizing the average number of users in the system. Very similar results can be obtained for the second objective, but they do not add any further insight into the performance of the assignment strategies.

A. Fixed Cell Capacities

We first present some numerical results to demonstrate that the achieved performance and the optimal thresholds are ac-

TABLE I
PERFORMANCE OF VAS AND DAS FOR FIXED CELL CAPACITIES. THE FIRST VALUE IN EACH TABLE ENTRY IS THE VALUE OF THE DECISION THRESHOLD. THE SECOND VALUE IN PARENTHESES IS THE ACHIEVED AVERAGE NUMBER OF USERS, I.E., THE VALUE OF THE OBJECTIVE FUNCTION

	λ_1/λ_2	*Theory*	*Simulations*	*Adaptive*
VAS	5 / 0	16.6 (0.91)	16.5 (0.91)	16.6 (0.91)
	5 / 1	18.9 (1.24)	19.0 (1.23)	18.9 (1.21)
	5 / 2	20.0 (1.67)	20.0 (1.65)	20.0 (1.62)
DAS	5 / 0	0.414 (0.91)	0.400 (0.92)	0.414 (0.91)
	5 / 1	0.233 (1.24)	0.200 (1.24)	0.233 (1.21)
	5 / 2	0 (1.67)	0 (1.65)	0 (1.62)

curately predicted by our theorem and propositions and match those obtained by our simulations. They also serve to verify that indeed the system performances achieved by the VAS and DAS schemes are the same and that the adaptive strategies converge to the optimal off-line strategies. In this first part, we continue to assume that the macrocell and microcell capacities are fixed and independent of the number of users and their locations. We also assume that the amounts of data to be transmitted (or equivalently the data transmission times) are exponentially distributed. The analytical results derived in this paper are valid for any joint probability density function of the users' profiles of velocity and average data size. However, in order to illustrate the results without unnecessarily burdening the computational complexity, we assume that the user profiles are uniformly distributed between 0 and $V_{\max} = 20$ m/s and between 0 and $D_{\max} = 1$ Mb and that the cell capacities are fixed at $C_m = 2.5$ Mb/s and $C_\mu = 5$ Mb/s.

The reader is referred to [1] for additional results with different cell capacities and for several plots showing the average number of users in the system as functions of the thresholds V_0 and D_0. Table I summarizes the VAS and DAS performance results for different arrival rates (expressed in calls/s) and compares the theoretical performance with that obtained by simulations. The optimal threshold value for the simulations is obtained by exhaustive search over 40 different threshold values in the intervals $[0, V_{\max}]$, respectively, $[0, D_{\max}]$. This quantization to 40 possible thresholds explains part of the difference between the theoretical and the simulation results. The table shows the velocity threshold (in m/s), respectively, the data threshold (in Mb) and, in parentheses, the achieved average number of users in the system. We also include the performance results of the adaptive assignment strategies.

In Fig. 2, we show the performance of the adaptive VAS rule over time, for different values of the arrival rates. It is observed that the velocity threshold converges very quickly to its final value (which coincides with the optimal threshold, obtained both by exhaustive search over the set of possible thresholds and by our theoretical calculations). Fig. 3 shows a similar plot for the adaptive DAS rule. The initial values of the adaptive rules are chosen as $V_0[0] = (V_{\max}/2)$ m/s and $D_0[0] = (D_{\max}/2)$ Mb and the tunable parameters are chosen as $T_{\text{update}} = 20$ s, $\alpha_v = \alpha_d = 0.3$, $\beta_v = (V_{\max}/2) = 10$ m/s, $\beta_d = (D_{\max}/2) = 0.5$ Mb, and $\gamma_v = \gamma_d = 0.7$.

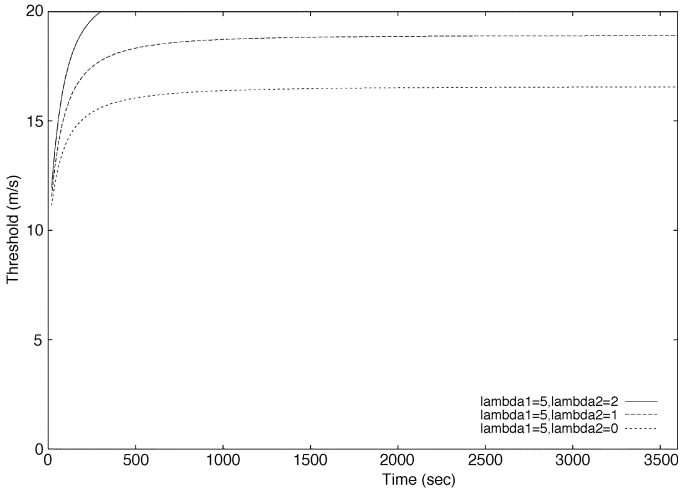

Fig. 2. Evolution of the adaptive velocity threshold under the adaptive VAS strategy for different arrival rates.

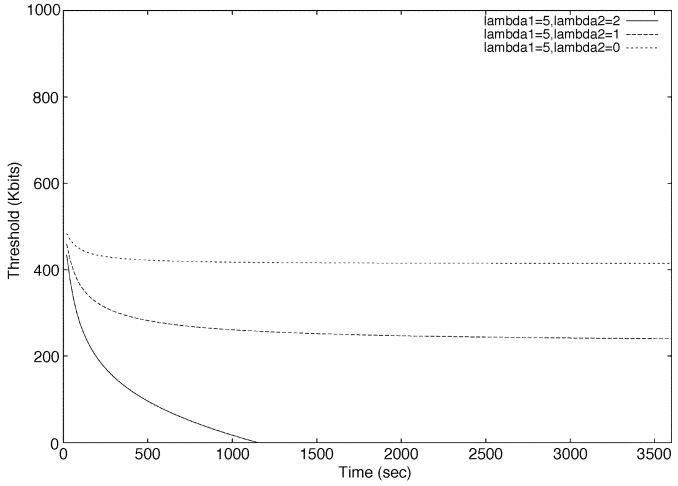

Fig. 3. Evolution of the adaptive data threshold under the adaptive DAS strategy for different arrival rates.

So far, we have only demonstrated the on-line behavior of our schemes. We now focus on a truly adaptive scenario and assume that the arrival rates change over time. We also fix the arrival rate in the macro-only region to $\lambda_1 = 5$ calls/s and consider three different values for λ_2. The arrival rate is assumed to be constant at $\lambda_2 = 2$ calls/s, then switch to $\lambda_2 = 1$ calls/s, further decrease to $\lambda_2 = 0$ calls/s, before increasing again to $\lambda_2 = 1$ call/s, and finally returning to its original value $\lambda_2 = 2$ calls/s. The arrival rate is assumed to remain constant at each value for 5000 s. In Fig. 4, we show the performance of the adaptive VAS rule under this dynamic scenario. We observe that the adaptive rule reacts quickly to changes in the arrival rate and converges to the new desired value of the decision threshold, even in this scenario when the changes in the arrival rate are instantaneous and quite significant. A similar graph is obtained for the adaptive DAS strategy and the conclusions mirror those of the VAS strategy. We observe that similar results to those presented here can be obtained if the arrival rate λ_1 or if both λ_1 and λ_2 are varied. Although not reported here, the rules equally well adjust the decision thresholds if the profile distribution of

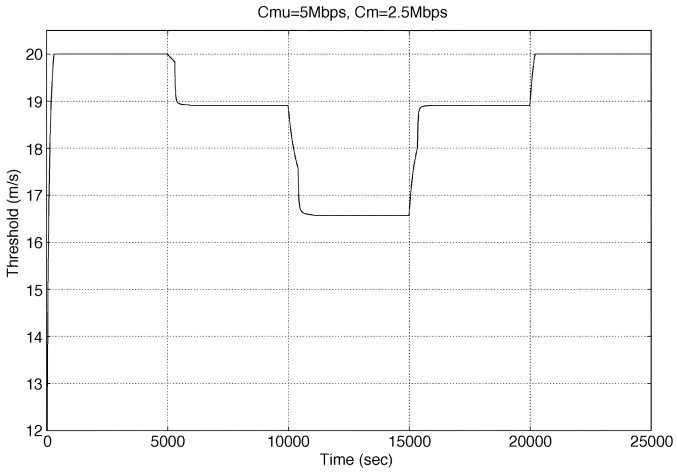

Fig. 4. Dynamic behavior of the adaptive VAS strategy under time-varying arrival rate λ_2.

velocity and data amount is changed over time, or when the cell capacities are changed over time.

B. Variable Cell Capacities and Multiuser Diversity

Next, we study the performance of threshold based assignment strategies under more realistic user behavior and system assumptions. We consider that the amount of data to be transmitted is distributed according to a log-normal distribution [24] with average data size $E[D] = 302$ Kb and truncated at $D_{\max} = 2$ Mb. Further, we implement a data queue for each user and decrease the remaining amount of data to be transmitted at every time slot by the actual transmission rate (time slots are normalized to have unit length). Similarly, we consider that the velocity distribution is piecewise uniform with 40% of the users having uniform velocity less than 3.6 m/s and 60% having uniform velocity between 3.6 and 20 m/s. This situation might correspond to an urban environment with a large stationary or slow-moving population of users.

Further, we suppose that the cell capacities (and, thus, the feasible transmission rates) depend on the number of users in each cell and that both the macrocells and the microcells exploit multiuser diversity through opportunistic scheduling techniques (such as the well-known proportional fair scheduling algorithm implemented in the HDR system or 1X-EV-DO [25]). We take this scheduler as an example to illustrate the system-wide performance of the assignment strategies and emphasize that our methods are applicable to other scheduling disciplines as well and the qualitative (if not the quantitative) results remain valid. In any given time slot, only one user is allowed to transmit for each cell. The transmission rate of the selected user depends on the user's location (and, thus, the user's channel condition) and is taken from the discrete set of allowable rates for HDR [26]. The initial user location is chosen uniformly across the geographical region with a uniformly selected direction of movement at the speed corresponding to the user's profile. The radii of the macrocell and microcell are, respectively, $r_m = 2000$ m and $r_\mu = 500$ m. All rates are possible when a user is transmitting in the macrocell. However, we assume that only rates larger than 614.4 Kb/s are used in the microcell. This is consistent with the fact that the inherent microcell capacity is larger than that of the macrocell and that the microcell has smaller coverage region.

Our objective is to show that the general qualitative conclusions remain valid in these very realistic scenarios. It is recognized that the derived Markov chain results are no longer valid and, therefore, the theoretical calculations in Section IV can no longer be applied. The optimal thresholds have to be determined through exhaustive search over 40 different threshold values in the intervals $[0, V_{\max}]$, respectively, $[0, D_{\max}]$. Further, even in this more general situation, we continue to adapt the decision thresholds according to the rules in (13) and (14). In Table II, we report the performance results under the optimal and the adaptive VAS and DAS strategies for different arrival rates. We consider two adaptive strategies based on two different balancing metrics. "Adaptive" refers to the "ideal" balancing metric in (9) and (10), whereas "Adaptive*" refers to a simpler and intuitive metric of balancing the number of users in each cell. Finally, we also present the results of the simple strategy that would always assign a user to the microcell, whenever the user is in the coverage region of the microcell.

The main conclusions to be drawn from these experiments are the following.

1) The performances achieved by the VAS and DAS strategies are equal or indeed very close to each other, even in these more general and realistic scenarios when assumptions 1 and 2 are no longer applicable.
2) The adaptive and on-line strategies also achieve close-to optimal performance without any *a priori* knowledge of the system parameters and quickly adapt to changes in these parameters and the resulting target (i.e., the optimal) decision thresholds.
3) The adaptive assignment strategies based on the statistical description of the user's profiles and behavior lead to increased system performance when compared with a simple strategy that would always assign users to the microcell when they are in the coverage region of the microcell. This result clearly demonstrates the effectiveness of intelligent assignment strategies that explicitly take into account the user's profile and behavior. While the "always-micro" policy seems intuitively appealing and may be simpler to implement, our results show a performance degradation of up to 230% compared with the optimal intelligent assignment schemes. The performance improvement is especially significant when the arrival rate in the macro-only region is small compared with the micro region. If the average number of users in the system is smaller, each user can be allocated more bandwidth and achieve greater throughput. Therefore, under an efficient assignment strategy, the service completion time required until all the user's data is transmitted is reduced.
4) The adaptive strategy based on the simpler metric of balancing the number of users in each cell achieves slightly better performance than the one based on the optimal balancing metric in (9) (at least in the case when an HDR scheduler is implemented). This is somewhat surprising as the metric in (9) corresponds to the characterization of the optimal thresholds in Theorem 1 (and in fact achieves the optimal performance when the cell capacities are

TABLE II
PERFORMANCE COMPARISON FOR VAS AND DAS WITH VARIABLE CELL CAPACITIES AND MULTIUSER DIVERSITY. THE FIRST VALUE IN EACH TABLE ENTRY IS THE VALUE OF THE DECISION THRESHOLD. THE SECOND VALUE IN PARENTHESES IS THE ACHIEVED AVERAGE NUMBER OF USERS, I.E., THE VALUE OF THE OBJECTIVE FUNCTION

	Strategy \ λ_1 / λ_2	3 / 0.3	3 / 0.2	3 / 0.1	3 / 0.0
VAS	Optimal	16.0 (4.01)	6.5 (2.35)	4.0 (1.48)	2.5 (0.87)
	Adaptive	19.8 (4.40)	13.6 (2.56)	8.6 (1.61)	4.3 (1.00)
	Adaptive*	15.1 (4.03)	8.2 (2.38)	4.0 (1.48)	2.9 (0.93)
	Always micro	20.0 (4.58)	20.0 (3.61)	20.0 (3.15)	20.0 (2.84)
DAS	Optimal	0.200 (4.01)	0.450 (2.37)	0.600 (1.46)	0.950 (0.86)
	Adaptive	0.021 (4.50)	0.255 (2.56)	0.409 (1.60)	0.550 (1.00)
	Adaptive*	0.180 (4.03)	0.422 (2.38)	0.585 (1.46)	0.779 (0.88)
	Always micro	0 (4.58)	0 (3.61)	0 (3.15)	0 (2.84)

fixed). We, therefore, attribute the performance degradation to the fact that the cell capacities are no longer fixed quantities, but depend on the number of users in the system and their locations. Thus, the performance of the adaptive rule depends on how the cell capacities are estimated in the calculation of the balancing metric. The assignment strategy "Adaptive*" does not suffer from this caveat as it only needs to calculate the average number of users in each cell for the calculation of its simplified balancing metric.

C. Influence of System Parameters on Performance of Adaptive Rules

In this final section, we investigate the impact of some of the tunable system parameters on the performance of the adaptive assignment strategies. These parameters influence the speed of convergence of the algorithms and determine how quickly the algorithms react to changes in the user behavior, if the algorithms are rather sluggish or very reactive, whether there are any oscillations around the target thresholds before the algorithms settle on the optimal values, and how large such oscillations might be. For the sake of brevity of the paper, we limit ourselves to exclusively study the influence of the parameters for the adaptive VAS assignment strategy defined in (9). Similar results and conclusions hold true for the adaptive DAS strategy. Furthermore, we consider the scenario when the cell capacities are fixed to $C_m = 5$ Mb/s and $C_\mu = 2.5$ Mb/s, as the HDR scheduler does not add any further insight into the role of the algorithm parameters, and fix the arrival rates at $\lambda_1 = 5\,\text{calls/s}$ and $\lambda_2 = 1\,\text{call/s}$. Note that the parameters have to be set independently of the arrival rates and the cell capacities, as these quantities may not be known *a priori* or may be time-varying in a real system.

In Fig. 5, we show the evolution of the velocity threshold as a function of time for different values of T_{update}. In particular, it is observed that, as T_{update} becomes larger, the algorithm takes longer to converge to the optimal threshold. This is expected, as in that case, updates are performed less frequently. Making T_{update} large leads to sluggish behavior of the adaptive rule. However, making T_{update} small is not desirable either, as the adaptive rule now becomes too reactive to instantaneous changes that may not reflect a corresponding change in system

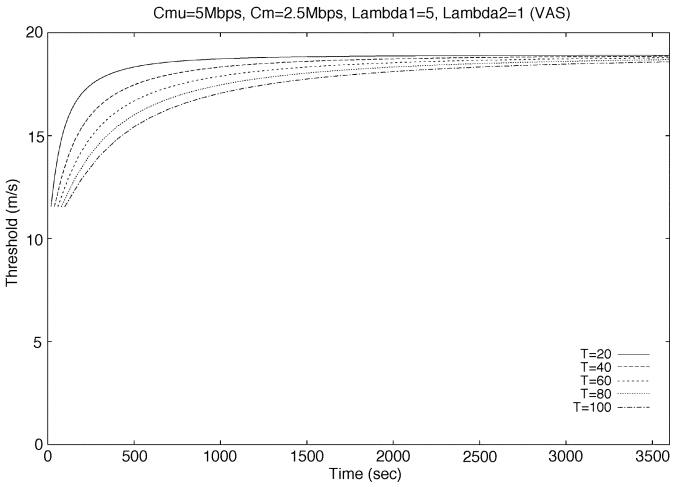

Fig. 5. Evolution of the velocity decision threshold for different values of the update interval T_{update}.

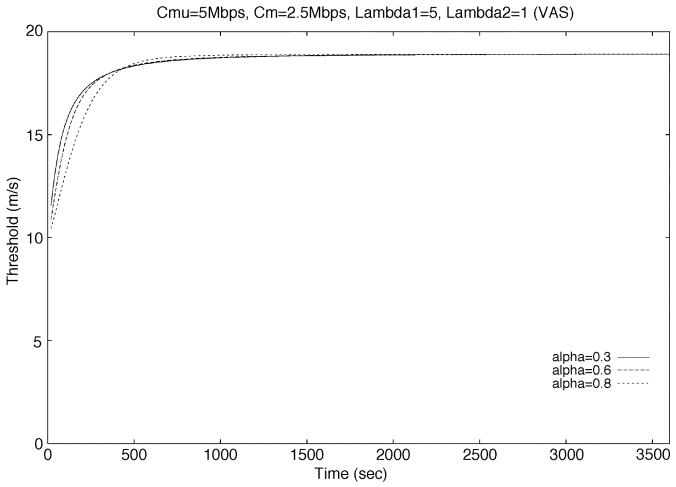

Fig. 6. Evolution of the velocity decision threshold for different values of the weighting factor α_v.

parameters or profile distribution. In Fig. 6, we show the convergence of the velocity threshold for different values of α_v. It is observed that if α_v is larger, more weight is put on the past measured values in the exponential smoothing of the balancing metric. Or equivalently, less weight is placed on the most recent value of the balancing metric. This implies that the adaptive

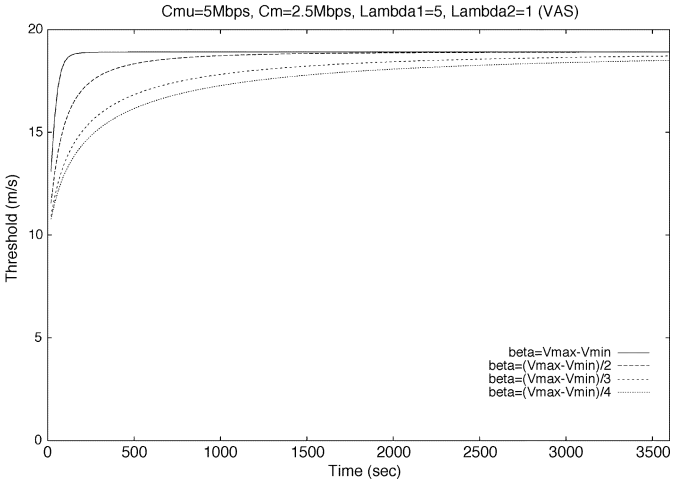

Fig. 7. Evolution of the velocity decision threshold for different values of the update magnitude parameter β_v.

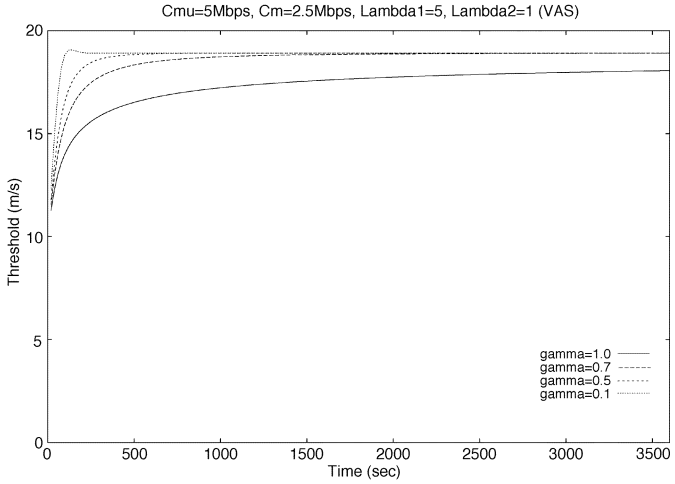

Fig. 8. Evolution of the velocity decision threshold for different values of the time discounting factor γ_v.

rule is less reactive to changes in the balancing metric and, thus, the updating of the decision threshold and the convergence of the algorithm is slower. This is clearly observed in Fig. 6. The above argument would suggest making α_v very small. If α_v is too small, however, the exponential smoothing is ineffective and the history of the measurements of the balancing metric is not taken into account. The update mechanism becomes very (too) aggressive and very (too) reactive to instantaneous changes in the balancing metric that may not be representative of the general call arrival patterns and user profiles. Fig. 7 shows the evolution of the velocity threshold for different values of the update magnitude parameter β_v. The graphs confirm the following interpretation of the role of β_v. If β_v is chosen to be very large, the magnitude of the update step is large, leading to large changes in the decision thresholds if an imbalance is detected in the balancing metric. Therefore, the update mechanism and its convergence to the optimal thresholds is faster if β_v is larger. On the other hand, if β_v is too large, the change in the value of the decision threshold could be too large, leading to potential overshoot of the optimal threshold. The resulting effect is that oscillations around the optimal value may occur, which may hamper the speed of convergence. Finally, in Fig. 8, we show the evolution of the velocity threshold for different values of the time discounting factor γ_v. The interpretation of the role of γ_v into the convergence of the adaptive rule is very similar (but opposite) to that of β_v.

VI. Conclusion

In this paper, we have considered assignment strategies for mobile data users based on their velocity and the amount of data to be transmitted by each user. The main contributions of the paper are to show analytically that the minimum average number of users in the system and the expected system load are the same under both strategies. The optimal thresholds are explicitly calculated and unique characterizations of the thresholds are provided. These characterizations are then used to devise adaptive and on-line assignment strategies that achieve the same performance as the optimal off-line strategies. Extensive simulation results are presented to support these statements. In this paper, we have exclusively investigated nonreal-time data users with a fixed amount of data to transmit. A second study within the same framework considers real-time users with a fixed connection time. Throughput maximizing assignment strategies for real-time data users have been examined in [28]. The objective of future research would be to study a system in which both real-time and nonreal-time users are competing for the available resources.

Appendix
Proof of Theorem 1

In this appendix, we provide the proof of Theorem 1. We first compute the performance of the DAS strategy for a given threshold D_0 and then we minimize the average number of users to obtain the optimal threshold. The first step involves the calculation of the conditional average data amount, given that a user is assigned to the macro, respectively, the microcell. Using Bayes' rule to calculate the conditional probabilities $\Pr(D \le d|m)$ and $\Pr(D \le d|\mu)$, we compute the conditional mean data sizes as

$$\overline{D}_m = E[D|m] = \frac{\lambda_1 \int_0^{D_0} x f_D(x)dx + \lambda_2 E[D]}{\lambda_1 \int_0^{D_0} f_D(x)dx + \lambda_2}$$

$$\overline{D}_\mu = E[D|\mu] = \frac{\int_{D_0}^{D_{\max}} x f_D(x)dx}{\int_{D_0}^{D_{\max}} f_D(x)dx}.$$

Following these preliminaries, we now turn to the heart of the proof of the theorem. Assumptions 1 and 2 are equivalent to saying that the 2-D Markov chain can be decoupled into two independent $M/M/1$ chains. It is well known [27] that the average number of users in an $M/M/1$ system is given by $E[N_m] = (\lambda_m/(\nu_m - \lambda_m))$, where ν_m is the "service rate" of the M/M/1 chain. In our case, $\nu_m = (C_m/\overline{D}_m)$. Similarly, we have that $E[N_\mu] = (\lambda_\mu/(\nu_\mu - \lambda_\mu))$ with $\nu_\mu = (C_\mu/\overline{D}_\mu)$. As a reminder, the call arrival rates to the macrocells and microcells are computed as $\lambda_m = \lambda_2 + \lambda_1 q_m$ and $\lambda_\mu = \lambda_1 q_\mu$. In order to simplify the derivations, we introduce the following notation:

$$g_m(D_0) \doteq \lambda_m \overline{D}_m = \lambda_1 \int_0^{D_0} x f_D(x)dx + \lambda_2 E[D] \quad (15)$$

$$g_\mu(D_0) \doteq \lambda_\mu \overline{D}_\mu = \lambda_1 \int_{D_0}^{D_{\max}} x f_D(x) dx. \quad (16)$$

The average number of users in the system is, therefore, determined by the expression

$$E[N_{\text{sys}}] = \frac{g_m(D_0)}{C_m - g_m(D_0)} + \frac{g_\mu(D_0)}{C_\mu - g_\mu(D_0)}.$$

To find stationary points of the objective function, we take the first order derivative and set it to zero. It is straightforward to obtain that

$$\begin{aligned}\frac{\partial E[N_{\text{sys}}]}{\partial D_0} &= \frac{C_m g_m'(D_0)}{[C_m - g_m(D_0)]^2} + \frac{C_\mu g_\mu'(D_0)}{[C_\mu - g_\mu(D_0)]^2} \\ &= \lambda_1 D_0 f_D(D_0) \\ &\quad \times \left[\frac{C_m}{[C_m - g_m(D_0)]^2} - \frac{C_\mu}{[C_\mu - g_\mu(D_0)]^2} \right] \quad (17)\end{aligned}$$

where we have used the fact that, for any D_0, $g_m'(D_0) = -g_\mu'(D_0) = \lambda_1 D_0 f_D(D_0)$. This is directly obtained from the definition of $g_m(D_0)$ and $g_\mu(D_0)$ in (15) and (16). Furthermore $f_D()$ is a probability density function and, therefore, nonnegative. We may now determine stationary points of the objective function by setting the first order derivative to 0, using the definitions in (15) and (16) and solving for the corresponding data threshold. Assuming that the profile distribution is in fact positive, we obtain the following integral equation for the stationary data thresholds:

$$\begin{aligned}&\int_0^{D_0^*} x f_D(x) dx \\ &= \frac{C_m\sqrt{C_\mu} - \sqrt{C_m}C_\mu - (\lambda_2\sqrt{C_\mu} - \lambda_1\sqrt{C_m})E[D]}{(\sqrt{C_m} + \sqrt{C_\mu})\lambda_1}. \quad (18)\end{aligned}$$

Note that the threshold D_0 is restricted to be contained in the interval $[0, D_{\max}]$. Hence, the above integral equation only yields an acceptable threshold value if the right-hand side of the equation lies in the interval $[0, E[D]]$. We distinguish two special cases when this is not verified. Specifically, it is easily shown that when the right-hand side of (18) is greater than $E[D]$, then the first order derivative of the objective function in (17) is negative, implying that the objective function is monotonically decreasing as a function of D_0. Thus, the minimum is achieved when $D_0^* = D_{\max}$. The required condition can be translated into the following requirement that:

$$(\lambda_1 + \lambda_2)E[D] < \sqrt{C_m}[\sqrt{C_m} - \sqrt{C_\mu}].$$

In this case, we immediately obtain that: $g_m(D_0^*) = (\lambda_1 + \lambda_2)E[D]$ and $g_\mu(D_0^*) = 0$. The corresponding minimum average number of users is obtained upon substitution of these quantities. Note that, in general, we expect that $C_\mu > C_m$ and that this special case does not apply. Similarly, when the right-hand side of (18) is less than 0, it is straightforward to show that the first order derivative in (17) is positive, and the objective is an increasing function of D_0. Thus, the optimal threshold is $D_0^* = 0$ and correspondingly we have that: $g_m(D_0^*) = \lambda_2 E[D]$ and $g_\mu(D_0^*) = \lambda_1 E[D]$. The related condition on the system parameters is obtained by enforcing that the right-hand side of (18) be less than 0

$$(\lambda_1 E[D] - C_\mu)\sqrt{C_m} \le (\lambda_2 E[D] - C_m)\sqrt{C_\mu}.$$

We now return to the general (and more interesting) case when the right-hand side of (18) is in the interval $[0, E[D]]$, leading to a nontrivial value of the data threshold. We now show that (18) has a unique solution which is a global minimum. Taking the derivative with respect to D_0 in (17) and evaluating it at a stationary point, i.e., at a solution of (18), yields

$$\begin{aligned}\left.\frac{\partial^2 E[N_{\text{sys}}]}{\partial D_0^2}\right|_{D_0^*} &= 2\left[\lambda_1 D_0^* f_D(D_0^*)\right]^2 \\ &\quad \times \left[\frac{C_m}{[C_m - g_m(D_0^*)]^3} + \frac{C_\mu}{[C_\mu - g_\mu(D_0^*)]^3} \right]. \quad (19)\end{aligned}$$

Assuming that $D_0^* > 0$ and that $f_D(D_0^*) > 0$, we conclude that we have a local minimum. This second condition can in general be guaranteed if $f_D(x) > 0$ for any $x \in [0, \ldots, D_{\max}]$. Therefore, we have shown that any solution to the integral equation is a local minimum. We now show that under the same conditions, the integral equation has only a single solution. However, this last statement follows immediately from the fact that the profile distribution is assumed to be positive and, therefore, $h(D_0^*) = \int_0^{D_0^*} x f_D(x) dx$ is a positive, monotonically increasing function of D_0^* with $\lim_{D_0^* \to D_{\max}} h(D_0^*) = E[D]$. Since the right-hand side of (18) is in the interval $[0, E[D]]$, we conclude that the integral equation as a unique solution D_0^*. Hence, the corresponding local minimum is in fact a global minimum of the objective function. Finally, the optimal value of the objective function can be computed upon substitution of the optimal values of $g_m(D_0^*)$ and $g_\mu(D_0^*)$. This concludes the proof to compute the data threshold that minimizes the average number of users in the system. When a velocity-based assignment strategy is used, the proof follows the same steps and is omitted here for the sake of brevity of the paper. However, we note that we again need to distinguish three cases, depending on the relative values of the arrival rates, the average data size, and the cell capacities. The minimum average number of users is given by the same expressions as for DAS, proving that the optimal VAS and DAS strategies achieve the same system performance.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the anonymous reviewers for their valuable comments and to M. Haner for his continued support and his encouragements of this research.

REFERENCES

[1] T. E. Klein and S. Han, "Assignment strategies in wireless overlay networks: Stability and performance analysis," in *Proc. 2003 Conf. Information Sciences Systems (CISS 2003)*. Baltimore, MD, Mar. 2003.

[2] S. Han and T. E. Klein, "Performance analysis of adaptive and online assignment strategies for wireless data users in hierarchical overlay networks," in *Proc. IEEE GLOBECOM*, San Francisco, CA, Dec. 2003.

[3] B. Ahn, H. Yoon, and J. W. Cho, "A design of macro-microcells CDMA cellular overlays in the existing big urban areas," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 2094–2104, Oct. 2001.

[4] L. Ortigoza-Guerrero and A. H. Aghvami, *Resource Allocation in Hierarchical Cellular Systems,* Mobile Communications Series. Norwood, MA: Artech House, 1999.
[5] B. Eklundh, "Channel utilization and blocking probability in a cellular mobile telephone system with directed retry," *IEEE Trans. Commun.*, vol. 34, pp. 329–337, Apr. 1986.
[6] J. Karlsson and B. Eklundh, "A cellular mobile telephone system with load sharing—an enhancement of directed retry," *IEEE Trans. Commun.*, vol. 37, pp. 530–535, May 1989.
[7] X. Lagrange and B. Jabbari, "Fairness in wireless microcellular networks," *IEEE Trans. Veh. Technol.*, vol. 47, pp. 472–479, May 1998.
[8] T. P. Yum and K. L. Yeung, "Blocking and handoff performance analysis of directed retry in cellular mobile systems," *IEEE Trans. Veh. Technol.*, vol. 44, pp. 645–650, Aug. 1995.
[9] W. Jolley and R. Warfield, "Modeling and analysis of layered cellular mobile networks," *Teletraffic Datatraffic in a Period Change*, vol. ITC-13, pp. 161–166, 1991.
[10] M. Benveniste, "Cell selection in two-tier microcellular/macrocellular systems," in *Proc. IEEE GLOBECOM*, 1995, pp. 1532–1536.
[11] K. L. Yeung and S. Nanda, "Channel management in micro/macrocellular radio systems," *IEEE Trans. Veh. Technol.*, vol. 45, pp. 601–612, Nov. 1996.
[12] R. Beraldi, S. Marano, and C. Mastroianni, "A reversible hierarchical scheme for microcellular systems with overlaying macrocells," in *Proc. IEEE INFOCOM*, 1996, pp. 51–58.
[13] B. Jabbari and W. F. Fuhrmann, "Teletraffic modeling and analysis of flexible hierarchical cellular networks with speed-sensitive handoff strategy," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1539–1548, Oct. 1997.
[14] S. Rappaport and L. R. Hu, "Microcellular communication systems with hierarchical macrocell overlays: Traffic performance models and analysis," *Proc. IEEE*, vol. 82, pp. 1383–1397, Sept. 1994.
[15] C. W. Sung and W. S. Wong, "User speed estimation and dynamic channel allocation in hierarchical cellular system," in *Proc. IEEE Vehicular Technology Conf.*, 1994, pp. 91–95.
[16] H. Kameda, J. Li, C. Kim, and Y. Zhang, *Optimal Load Balancing in Distributed Computer Systems*. New York: Springer-Verlag, 1997.
[17] T. Crabill, D. Gross, and M. J. Magazine, "A classified bibliography of research on optimal design and control of queues," *Oper. Res.*, vol. 25, no. 2, pp. 219–232, Mar. 1977.
[18] M. Alanyali and B. Hajek, "On simple algorithms for dynamic load balancing," in *Proc. IEEE INFOCOM*, 1995, pp. 230–238.
[19] M. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, S. Miller, and L. Salgarelli, "Integration of 802.11 and third-generation wireless data networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, 2003.
[20] C. Tepedelenlioglu and G. B. Giannakis, "On velocity estimation and correlation properties of narrowband mobile communication channels," *IEEE Trans. Veh. Technol.*, vol. 50, pp. 1039–1052, July 2001.
[21] A. Abdi and M. Kaveh, "A new velocity estimator for cellular systems based on higher order crossings," in *Proc. 32nd Asilomar Conf. Signals, Systems, Computers*, 1998, pp. 1423–1427.
[22] M. D. Austin and G. L. Stuber, "Velocity adaptive handoff algorithms for microcellular systems," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 549–561, Aug. 1994.
[23] K. D. Anim-Appiah, "On generalized covariance-based velocity estimation," *IEEE Tran. Veh. Technol.*, vol. 48, pp. 1546–1557, Sept. 1999.
[24] M. Molina, P. Castelli, and G. Foddis, "Web traffic modeling exploiting TCP connections' temporal clustering through HTML-REDUCE," *IEEE Network*, vol. 14, pp. 46–55, May 2000.
[25] *CDMA2000 1X-DO Standard*, www.3gpp2.org, Oct. 2001.
[26] P. Bender *et al.*, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, vol. 38, pp. 70–77, July 2000.
[27] D. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[28] T. E. Klein and S. Han, "Assignment strategies for throughput maximization of real-time data users in wireless overlay networks," Tech. Memo., Lucent Technologies-Bell Labs., Murray Hill, NJ, Mar. 2002.

Thierry E. Klein (S'97–M'00) was born in Ettelbruck, Luxembourg, on October 7, 1971. He received the B.S. and the M.S. degrees in mechanical engineering from the Universite de Nantes, Nantes, France, in 1993 and 1994, respectively, and the E.E. degree in automatics from Ecole Centrale de Nantes, Nantes, France, in 1995 (ranked first in class), and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, in October 2000.

From 1990 to 1992, he attended the "Classes Preparatoires" at Lycee Louis-Le-Grand, Paris, France, in preparation for the National Admission Contest for Ecole Centrale. Since January 2001, he has been a Member of Technical Staff in the Networking Infrastructure Research Department, Wireless Research Laboratory, Bell Laboratories, Lucent Technologies, Murray Hill, NJ. His research interests include information and communication theory, mobility management, and resource allocation in wireless networks, as well as end-to-end data performance analysis and cross-layer optimizations.

Seung-Jae Han (M'02) received the B.S. and M.S. degrees in computer engineering from Seoul National University, Seoul, Korea, and the Ph.D. degree in Computer Science and Engineering from the University of Michigan, Ann Arbor.

He is a Member of Technical Staff of the Wireless Research Laboratory, Bell Laboratories, Lucent Technologies, Murray Hill, NJ. His research interests include QoS networks, wireless networks, and fault-tolerant systems.