Handover in Mobile Communication Networks: Who is in Control Anyway?

Frank A. Zdarsky and Jens B. Schmitt
Distributed Computer Systems Lab
University of Kaiserslautern, Germany
{zdarsky, jschmitt}@informatik.uni-kl.de

Abstract

The migration from 3rd generation mobile communication networks to 4th generation, purely IP-based networks leads to new chances, but also great risks for the traditional mobile network operators, as competition increases dramatically. While this competition potentially improves the cost-value ratio for mobile users, they might well find their freedom of seamless session mobility abruptly end at the administrative borders of Wireless Internet Service Providers. In this paper, we will describe a possible and in our opinion likely scenario for future mobile wireless access, point out the inherent problem of current IP mobility management schemes when it comes to inter-provider handovers, have a look at why control over this process is currently shared between the network and the mobile end-devices, and describe approaches to shift the major part of control back to the user.

1. Introduction

Current mobile communication networks still come in two flavors: mobile cellular (telecommunication) networks such as GSM and UMTS and wireless data networks such as IEEE 802.11 wireless LANs.

Mobile cellular networks are operated by classical Telecommunication Service Providers (TSPs). Their wireless access networks are connected via an operator's own core network and together form a closed and controlled system. Until UMTS Release 4, the core network is still divided into a circuit-switched domain for telephony services and packet-switched domain for data services. The wireless access networks use relatively homogeneous access technologies (GSM upgraded to UMTS in Europe and IS-95 upgraded to cdma2000 in the U.S.); the "intelligence" is located in the network, rather than on the end-devices (usually mobile phones), which means the network is responsible for control functions such as handovers, session setup, etc. The wireless access technologies operate in licensed frequency bands and their use is strictly regulated by public authorities such as the FCC in the United States. The wireless access points are densely deployed on a large scale, usually spanning an area that covers a high percentage of a country's population. Due to the high cost of deploying such a large-scale, closed system and the regulation of radio frequencies in the used frequency bands, the barrier for entering into the mobile cellular network market is very high. As a result, competition in this market is comparatively low (oligopolistic structure).

Wireless data networks, on the other hand, can already be deployed by small Internet Service Providers (ISPs), who may not have a backbone of their own, but rely on the Internet for wide-area connectivity, and therefore on an open system over which the single ISP has little control. Access and core networks are purely packet-switched and the "intelligence" is assumed to be on the end-devices, rather than in the network. The choice of wireless access technologies is potentially very wide, as long as they allow for IP on the network layer. Still the most widely deployed access points follow the IEEE 802.11{a,b,g} standards. These operate in the Industrial, Science, and Medicine (ISM) frequency bands and do not require a license for operation. Also, access points are usually sparsely deployed and cover only small hot-spot areas, such as a city block. The sparse deployment, the possibility to use the Internet as backbone network and the lack of licensing of the frequency bands means a low barrier of market-entry and therefore potentially a very high competition.

In this paper, we will discuss a possible and in our opinion likely and sensible scenario for future mobile wireless access. This scenario is characterized by high competition between all types of wireless Internet service providers and users who benefit significantly from being able to dynamically select the best provider for their needs at any time. We also discuss, which existing technologies are required and how they would have to be extended to allow *free provider selection*.

The paper is organized as follows: We first introduce the terminology used in the remainder of the paper (Section 2)

Proceedings of the 30th EUROMICRO Conference (EUROMICRO'04)
1089-6503/04 $ 20.00 IEEE

	Mobile Cellular Networks	Wireless Data Networks
Examples	GSM/GPRS, UMTS	IEEE 802.11, HIPERLAN2
Typical Operator	Large TSPs with own backbone infrastructure	Small ISPs, usually without own backbone infrastructure
Typical Network	closed network; circuit- and packet-switched domains	open network (Internet); purely packet-switched
Frequency Allocation	licensed bands	unlicensed ISM band
Cell Size and Density	500m-35km diameter; densely deployed	about 50m diameter; sparsely deployed
Market-Entry Cost Factors	license, nation-wide access network, backbone	no license, scalable system cost

Table 1. Comparison of mobile cellular networks and Wireless Data Networks.

and describe our scenario for future mobile wireless access, the developments leading to it, and the requirements and success factors for letting the user benefit from this scenario (Section 3). A short overview of handover mechanisms in wireless networks is given in Section 4. We then discuss different options for handover control and which of these options fits best to the scenario's requirements (Section 5). We present solutions for issues raised throughout the paper in Section 6 and finally conclude in Section 7.

2. Terminology

As the terminology used in connection with mobile cellular neworks and wireless data networks varies strongly, we first introduce some common terminology used in the remainder of the paper:

- Telecommunication Service Provider (TSP): A provider that primarily offers classical telephony services via a public switched telephone network and/or a mobile cellular network. A TSP typically owns or leases a backbone infrastructure that is large enough to offer nation-wide connectivity.
- Internet Service Provider (ISP): A provider of access to the Internet. Some ISPs own wide-area backbone networks, others have to lease them. Usually, ISPs rely on a TSP's network for the "last mile" to the user.
- Wireless Internet Service Provider (WISP): A provider of Internet access via a wireless access technology such as IEEE 802.11. It may have an own backbone network just like the ISP, but does not require a TSP for the last mile, as users are connected wirelessly.
- Mobile Host (MH): An IP-enabled wireless, mobile end-point (mobile phone, notebook, etc.). It is equivalent to a "mobile station" in GSM or IEEE 802.11 terminology or "user equipment" in UMTS terminology.
- Access Point (AP): A layer 2 device that offers a wireless link connection to a mobile host, roughly equivalent to an access point transceiver in IEEE 802.11, a base station in GSM or a NodeB in UMTS.
- Access Router (AR): A router which provides layer 3 connectivity between the mobile access network and the Internet and to which one or more APs are connected. From the viewpoint of a mobile host the AR is the first-hop router.

3. Scenario: future mobile wireless access

3.1. Observable trends

Mobile cellular networks and wireless data networks are continually converging: Starting from Release 5 of the UMTS standard, the core network is no longer divided into a circuit-switched and a packet-switched domain, but will be based on a packet-switched IP network alone. This approach is commonly called "All-IP", although IP is still not used as a network protocol in the radio access network, but is instead stacked on top of the UMTS network protocol. Wireless data networks, on the other hand, are developing from local-area to wide-area networks. It is a common vision that 4th generation mobile communication networks will be purely based on IP technology (i.e. also in the wireless access networks), as it has several advantages for service providers, such as higher resource efficiency, the integration of circuit-switched and packet-switched networks as well as the possibility to leverage the cost-effect of "off-the-shelf" hardware.

However, the migration to IP technology in mobile communication networks may be a threat to TSPs as well. This becomes evident when taking a short excursion into the known evolution in fixed telecommunication networks, which can roughly be divided into three phases:

1. When the first analogue telephone networks were deployed, huge investments were needed to build up an infrastructure that provided basic communication services to the whole population. For this it was beneficial to have a single provider, usually state-owned, so that the investments paid-off quickly (monopolistic phase).

2. In the second phase the public TSPs were privatized and the national telecommunication markets were deregulated, allowing for more competition and lower prices for consumers. Still the barrier of entry for new TSPs was comparatively high, as a new TSP either had to provide its own telephone network including interconnectivity with other networks or to act as a re-seller of the former monopolist's telephony products (deregulation phase).

3. The third phase is finally marked by the increasing availability of high-speed Internet connections for private households and therefore the viability to use voice over IP services. This allows small, regional Internet Service Providers to offer telephony services to their customers as well, since they do not necessarily need a public switched telephone network (PSTN) infrastructure of their own, but instead use their existing IP backbone and possibly have extra cost for interconnection gateways to the PSTN if receivers of their customers' calls do not yet use VoIP. Overall, it is fair to say that the large-scale proliferation of high-speed Internet connections in fixed network leads to a significant increase in competition for TSPs (high-competition phase), which in turn results in a reduction in service prices and an increase in service quality.

3.2. Extrapolation or looking into the crystal ball

Coming back to mobile cellular networks we expect an analogous development. Currently only a small number of TSPs is building up and operating 3G networks, due to the very high investments necessary not only for obtaining licenses for radio frequencies, but also for the radio access network infrastructure. A UMTS license grants the license holder a nation-wide exclusive right to use a pair of frequency bands (one for the uplink, one for the downlink). At the same time the license holder commits itself to build up an infrastructure that allows a sufficiently high percentage of the whole population to benefit from UMTS services within a sufficiently short period of time. This way the applicants for a license could not pick out only the highly populated and most lucrative large cities for installing UMTS access points, but had to provide service on a large geographical scale, while on the other hand their investments were protected.

We envision that in a couple of years the mobile communication market will undergo deregulation just like it happened in the market for fixed telecommunications, as there are already more (2G) mobile phones than fixed phones in Europe and world-wide[1]. Several options exist for deregulation: There might be additional licenses for UMTS which are restricted to smaller geographic regions or the TSPs might be forced to open their access networks to other providers who can then act as re-sellers. It is also possible that new competition arises in the form of wireless wide area data networks such as IEEE 802.20, also known as Mobile Broadband Wireless Access (MBWA)[4], which will most likely also operate in licensed frequency bands. It might, however, just as well happen that the TSPs themselves create the foundation for new competitors:

Already TSPs are starting to deploy WLAN access points to complement their 2.5G and 3G mobile cellular networks. The result is a "wireless overlay network"[2], in which their customers can obtain low-bandwidth, high-latency access to data services almost everywhere through GPRS, medium bandwidths in urban and sub-urban areas through UMTS and high bandwidths via WLAN in hot-spot areas (see Figure 1). As the coverage areas overlap, the user is supposed to be able to dynamically switch between different access technologies of the same provider, depending on the available portfolio of access technologies at the current location and a user's personal preferences. This "vertical handover"[1] requires mobile devices capable of connecting to GPRS/UMTS networks as well as WLANs. The latter, however, is a wireless access technology that does not require any licenses to be obtained prior to deployment, and there is already an increasing number of commercial wireless Internet service providers (WISPs). As the number of WLAN hot-spots increases, more and more locations will exist where a mobile user will not only have the possiblity to choose between different wireless access technologies of the *same* provider, but also between *different* service providers and therefore different quality of service (QoS) and pricing models. Finally, when VoIP services become more popular, users will be able to make normal phone calls from any hot-spot, especially near their house or workplace with a private WLAN access point installed, and all this finally over one single "phone".

Mobile users would benefit significantly from this development, as they could choose the provider and access technology that best fits their needs at any given moment. While *free provider selection* is an opportunity for new WISPs, it is also a threat for established WISPs who might try to erect new barriers to lock-in their customers.

If users shall fully benefit from increased competition between WISPs, they must be enabled to

- discover all available access points at a location independent of their administrative domain,
- query the capabilities of each access point, and

1 A "vertical handovers" denotes a handover between APs of different access technologies, as opposed to a "horizontal handover" which occurs when moving between APs of the same access technology.

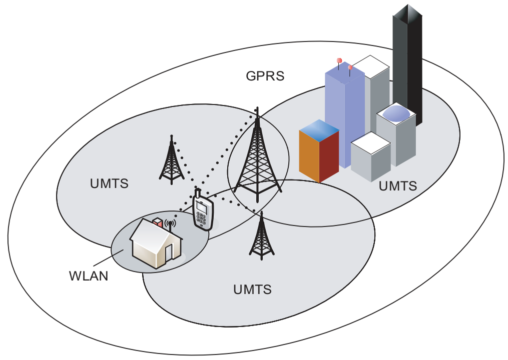

Figure 1. A wireless overlay network

- perform a handover to the best access point available (if possible seamlessly).

Assuming that 3G networks continue to evolve towards pure IP, MobileIP will be the mechanism that allows switching between networks of different providers dynamically, even during a session. How noticeable this inter-provider handover will be to the user and whether it is possible and acceptable to users at all depends on the performance of handover mechanisms. We will give a short overview of handover mechanisms next.

4. Handovers in wireless access networks

4.1. Layer 2 handovers

A layer 2 handover occurs when a mobile host's connection changes from one access point to another. This type of handover involves releasing the connection to the old access point, establishing a connection to the new access point (though not necessarily in this order), and updating the binding between a mobile host's IP address and its temporary layer 2 identifier (e.g. a MAC address) on the subnetwork at the access router. If the old connection is released before the new connection is established (break-before-make handover), data reception is interrupted for a short period of time. The service disruption can be avoided if the mobile host is capable of establishing connections to multiple access points simultaneously. It can then connect to the new access point before breaking its connection to the old access point (make-before-break handover). If both access points involved in a layer 2 handover are part of the same subnetwork, the mobile host's IP address can remain the same. In this case the handover has no effect outside the subnetwork and is usually fast enough to go unnoticed by the user. An exception to this is if the new access point is not able to provide sufficient resources for the new mobile host, so that it suffers from serious service degradation.

4.2. Layer 3 handovers

Switching between access points on different subnetworks requires a handover on the network layer (layer 3 handover). This can be the result of the mobile node moving out of the geographical area that is covered by access points of the current subnetwork. It might also happen when switching between access points of different wireless access technologies (vertical handover), as they are usually—but not necessarily—on different subnetworks. Finally, a handover to an access point in the domain of a different WISP always involves a change in subnetworks and therefore a layer 3 handover.

Before initiating a layer 3 handover, the mobile host has to choose a suitable target access point. It then establishes a new layer 2 connection to that access point and acquires a valid temporary IP address in the new subnetwork, either from the foreign agent of that subnet (MobileIPv4) or by stateful address autoconfiguration such as DHCPv6 (MobileIPv6). Finally, it can initiate the layer 3 handover by sending a binding update message with the new care-of address to its home agent and possibly the correspondent node as well.

The choice of a suitable target access point—and with it that of the access router which connects the access point to the Internet—will depend on various factors, such as

- the price structure for accessing the (Internet) service over a particular access router,
- the security level,
- the available QoS levels,
- the provider itself, etc.

The mobile host decides whether or not the access router of the next subnetwork meets its requirements profile.

4.3. Discovery of access routers for layer 3 handovers

Therefore, the problem is how to find candidate access routers for a layer 3 handover and how to query their capabilities. The IETF SEAMOBY working group is currently working on a Candidate Access Router Discovery (CARD) protocol[5] that shall allow just that. A mobile host periodically activates its wireless network interfaces to scan the medium for available access points. These send out beacon signals from which the mobile host recognizes their presence and also their unique layer 2 identification (such as the MAC address in IEEE 802.11). Using CARD, the mobile host can then query its current access router for the IP address of the access router that controls the access point with that layer 2 identification. The result is a set of candidate access routers for handover. The mobile host can then either

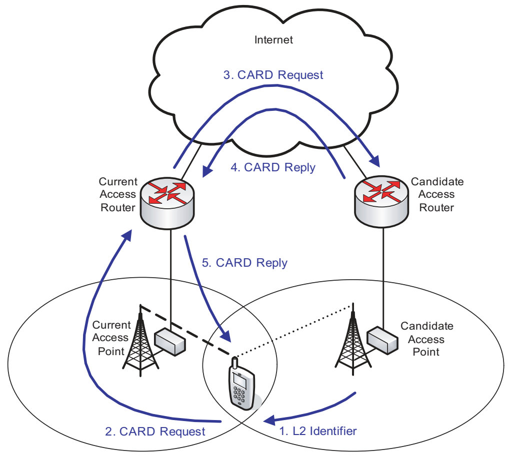

Figure 2. CARD protocol operation

query the capabilities of each candidate itself or use the information that its current access point has already collected about the candidates. This information may not be up-to-date, though. From the candidate set the mobile host then chooses the access router that fits its requirements best.

This approach has some weaknesses, though. First of all, the mobile host needs to be connected to an access router to resolve the layer 2 identifier of an access point into the IP address of the access router controlling it. If the mobile host is not connected—at session start or when it looses its current connection unexpectedly—it has to connect to each available access point to query the capabilities of the access routers manually. It might have cached the required information long before the connection was lost, but the cached information might be out-of-date by the time it is needed. It would furthermore be useful to have timely information about the current load status of the access points, so that the chances of a successful handover attempt can be estimated. Two even more serious problems of the CARD protocol arise due to the competition between WISPs. It is not evident that a WISP would be willing to provide its competitors with a list of L2 identifiers of its access points and the mapping to an access router. More importantly, a WISP might not be willing to aid its customers in performing a handover to the network of a competitor by mapping L2 identifiers to IP addresses.

The customer of mobile Internet services can only benefit from the increased competition between WISPs if he has the freedom of choice of selecting the service that matches his requirements best. On the other hand, WISPs are likely to try to use lock-in strategies to prevent their customers from changing to other providers. It is therefore worth having a look at who has how much control over the handover process anyway.

5. Handover-control

5.1. Discussion of handover-control approaches

The handover process consists of two phases, each of which can be controlled either by the mobile host or an entity in the network, e.g. a single access router or a group of access points cooperatively[7]:

1. In the *handover decision* phase, decisions are made on whether to initiate a handover or not, which access point should be the target of the handover, and at what time the handover should be initiated. If it is the mobile host that makes this decision, we denote this a Mobile-Initiated HandOver (MIHO), otherwise a Network-Initiated HandOver (NIHO). Usually the handover decision is based on information from both the network and the mobile host. If the mobile host passes information about its current status to the network, which the network then includes in its handover decision, the handover is called a Mobile-Assisted HandOver (MAHO), and it is called a Network-Assisted HandOver (NAHO) if the information is passed the other way around.
2. In the *handover execution* phase, either the network or the mobile host executes and controls the actual handover process (Network-Controlled HandOver (NCHO) versus Mobile-Controlled HandOver (MCHO)).

The handover decision can be based on a variety of criteria. These criteria can be the same for all mobile hosts or be chosen and weighted differently depending on user preferences (“policy-enabled handover” [9]).

The most influential criteria in mobile cellular networks are the downlink quality (from the access point) and the uplink quality (to the access point), respectively. The link quality can be measured in terms of received signal strength, signal-to-interference ratio, bit error rate, etc. When the quality of the radio link between mobile host and access point falls below a minimum acceptable value, often due to the mobile host moving away from the access point, a handover to a different access point with better link quality is initiated. For NIHO, a mobile host periodically measures the link quality to each access point which it currently receives and sends the results to the network, for MIHO the network sends its measurements to the mobile host. In terms of signaling and processing effort, both approaches are roughly the same, but MIHO allows the processing effort to be distributed between mobile hosts, rather than having to do it centrally in the access network. Including link quality as criterion in the handover decision is beneficial both for the provider and the user of the improved service.

Another decision criterion is the current load at each access point, that is how many mobile hosts currently use how many of an access point's wireless link resources. From a provider perspective this information is important, as it allows access routers to transfer mobile hosts that are reachable from several access points from a very busy access point to a less busy one (load balancing). This way the access point whose load is reduced can better service newly arriving hosts, which increases the utilization of the whole access network and therefore the provider's utility. For the mobile host it can be beneficial to transfer to a less loaded access point as well, as it improves the QoS of its connection. Again, both the provider and the user have an interest in handovers depending on current AP load.

In addition, a mobile host has a strong interest in conserving battery power. Therefore, it may wish to handover to a closer AP to be able to reduce transmission power, or to switch to a wireless access technology that may be more energy-efficient, but offers less bandwidth. For similar reasons a mobile host might wish to remain attached to the same AP, even though other criteria would suggest a transfer. In particular, including this criterion and optimizing for low energy consumption counteracts the effects of load balancing. While this can be acceptable to the user, the provider's utility might decrease.

Another criterion for a handover decision is the mobile host's current position and velocity, which together with measurements of link quality allows a more accurate prediction of the next handover time, which improves the timing of handover execution.

Furthermore and maybe most importantly, only the mobile host can know the context and the requirements of its applications with respect to link quality and tolerance towards handover delays at a specific moment. It therefore has a strong interest in being able to delay or expedite the handover process. This can be viewed as yet another incarnation of the end-to-end argument which is often considered to lie at the heart of the Internet's success[8]. Both this and the previous criterion only affect the handover timing, not the target. Therefore, the effects on the provider utility are negligible, but the user's utility might improve significantly through better service.

Especially in a scenario with heterogeneous access technologies and multiple service providers, there are two more aspects influencing a mobile host's desire to handover to an access point controlled by a different access router. Another access router might have capabilities that better match the mobile host's requirements, like better security, more reliable QoS, a wider coverage via its access points, etc. Also, the pricing model for the access service might be more favorable. These criteria enable the user to choose a target access point that provides the best cost-benefit ratio.

Most of these criteria mainly improve the user's utility by increasing his received QoS, reducing energy consumption and/or improving the cost-value ratio. The providers benefit primarily from the higher resource utilization of load balancing. Link quality and improved handover timings are mainly beneficial in improving the quality of the offered service. Considering the battery status as handover decision criterion partially counteracts load balancing and therefore might reduce the access network's resource utilization. Finally, whether more information about the capabilities of a network's access router and the service pricing model benefits a particular provider or not depends on how this provider's service compares to that of others.

5.2. Choosing the preferable handover-control approach

To be able to consider all these criteria in the handover decision process, the necessary information has to be cumulated at the mobile host in the case of MIHO and in the network in the case of NIHO, respectively. The answer to the question of who should be in control of handovers partly depends on how much effort is involved in cumulating this information in one location or the other. Yet there are more aspects to consider, so we will now have a look at the different aspects in more detail:

- Radio link resources are scarce, signaling over the radio link should therefore be kept at a minimum necessary level. Looking at the decision criteria above, link quality information could be cumulated either way without much difference in signaling. Information about the battery status, the position and movement of the mobile host, and the application context is generated on each mobile host individually. Transfering it to the network for a handover decision would involve frequent updates from every mobile host, with substantial amounts of data from each. On the other hand, transferring information about load status, AR capabilities and service price model to mobile hosts could be accomplished by broadcast, as it is the same for all mobile hosts[2]. Also, the AR capabilities and the service price model is likely to be updated on a low frequency. Therefore, we prefer MIHO as solution with respect to transfer of handover decision criteria.
- Some of the criteria described above can only be reasonably incorporated into the decision process if the user's utility function is known. For example, it is not clear whether a user prefers low service cost at the price of an increased energy consumption or whether he can tolerate to stay connected to a low-speed network, as he will soon move away from the area where

2 Sending individualized information to mobile hosts, e.g. additional price models for premium customers, is not possible at this stage, as the listening mobile hosts are not yet associated with the advertising access point and can therefore not be discriminated.

high-speed access is available anyway. For reasons of privacy the utility functions should not be transfered to the network, so again MIHO is preferable over NIHO here.

- From a provider perspective it is little reassuring to transfer control to the mobile hosts, especially with respect to load balancing. NIHO would be a preferred solution here.
- The choice between MIHO and NIHO does not affect the provider's ability to do charging and accounting, as both network and mobile host cooperate in the handover and can measure exactly when a connection to an AP is established and torn down. For the same reason, both approaches allow the mobile host to verify the charging and accounting by the network.
- Mobile hosts which have multiple wireless network interfaces can initiate seamless handovers to a different access network using IP diversity, i.e. sending packets over multiple paths, which is an advantage of MIHO.
- A strong argument in favor of MIHO is also the possibility to perform handovers to access points in different administrative domains to benefit from cheaper access rates and higher QoS of other providers. It is unlikely that providers would support this type of handover when NIHO is used.
- In terms of security implications and possiblities for manipulation we consider NIHO and MIHO to be the same.
- With respect to handover reliability we see a small advantage of MIHO, as it would allow a mobile host to resume sessions that were interrupted due to a handover failure upon reconnection.

As the list shows, MIHO is the preferred solution for the majority of the considered aspects, except for the aspect of load balancing, with which the provider can increase the resource utilization of his access networks and therefore its profits. In the next section, we will present a possible solution approach that has the same effect even with handover control by mobile hosts. Another result is that MIHO is a must, if users shall be able to handover between access networks of different providers to benefit from increased competition. Yet, it is of no use if the mobile hosts can decide on a handover to a different provider if they are not also in control of the handover execution process. Overall, we believe that freedom of provider selection mandates that handovers be mobile-initiated, network-assisted, and mobile-controlled. We admit, though, that it can be advantageous to let the network initiate and control soft handovers (i.e. those using macro diversity) between access points where possible, as these handovers are masked to the mobile host and let a group of cells appear as one large virtual cell.

6. Open issues & solution approaches

The previous sections have described the mechansisms, such as MobileIP in connection with the CARD protocol, that will allow L3 handovers between access routers of different subnetworks, in principle even if these are in different administrative domains. However, if providers are not cooperative, which is likely in a highly competitive scenario, these mechanisms will not be able to guarantee mobile users a free choice of the wireless access provider that best fits their needs anymore. In this section we describe solution approaches to some remaining open issues.

6.1. Discovery of AR capabilities under provider competition

In the CARD protocol a mobile host has to query its current provider's access router for the IP addresses and capabilities of nearby access routers even if these belong to one of the provider's competitors. To avoid potential manipulation of the query results by the current provider, who has a strong interest not to advertise the services of competitors, the mobile host should be enabled to query nearby ARs directly. For efficiency reasons this should not require establishing layer 2 connections and sending router solicitations first. Instead we propose to broadcast the AR's IP address together with a capability vector periodically in the beacon signal of the APs. There might be cases in which a provider cannot or does not want to advertise its services via layer 2 beacons. It could then be useful to have a mechanism that allows nearby mobile hosts to exchange information and recommendations about access points on the application level.

6.2. Load balancing in the MIHO paradigm

As we have shown in Section 5, MIHO is the preferable solution from the viewpoint of mobile users. Providers, on the other hand, will be reluctant to give away control over the handover process, as they will not be able to perform load balancing themselves anymore. If MIHO shall be acceptable to providers, they need to have a way to influence the handover decision processes of mobile hosts in their access networks in such a way that the load is still balanced between access points and resource utilization remains high. We propose to let APs signal to mobile hosts that they currently experience high loads, even though neighboring APs are only slightly loaded. A mobile host that ignores this signal and stays in the highly loaded cell could then be penalized by an additional service charge or a reduction in service quality. As mobile hosts consider this negative incentive in their decision process, some of them will decide to handover to less loaded APs. Using this mechanism we transform the optimization problem that was solved in a centralized way in NIHO into a decentralized optimization problem. If the "price" for staying in a highly loaded network cell is chosen appropriately, the results of this decentralized load balancing approach can come very close to those of load balancing by the provider itself, similar to [6, 3].

6.3. Charging under provider competition

Apart from the technical requirements for free provider selection, sustaining competition between wireless access providers also requires common charging solutions. Currently, for example, a user who wants to connect to a mobile cellular network needs to have a SIM card issued by that network's operator. This SIM card uniquely identifies the user and enables the provider to charge the correct account. In order to use the network of a different provider, the user can either obtain a SIM card of that provider as well or hope that both providers have made roaming agreements. In the latter case one provider does the charging for both networks and afterwards settles its accounts with the other provider. Obviously, roaming makes using a different network more expensive and therefore as unattractive as obtaining a second SIM card. This is why under these circumstances users will mostly choose a provider that is large enough to offer access everywhere the user requires it, again impeding free provider selection. Ideally, users would have a universal SIM card that is independent of a particular service provider and can be used in various mobile communication networks. As electronic personal IDs issued by public authorities become more widely available, these personal IDs could be the basis for provider-independent charging. Alternatives are widely accepted clearing house or micropayment solutions.

7. Conclusions

In this paper, we have described a possible and in our opinion likely scenario for future mobile access: As current 3G mobile cellular networks evolve towards pure IP networks, they also build the foundation for massively increasing competition. The result is that in many locations user will not only have the choice between various wireless access technologies, but also between a number of different service providers. Ideally, users could then dynamically switch to the provider and access technology that best fits their needs at any given moment.

In order to benefit from this increased competition, mobile users need the ability to discover all available access points at a given location independent of their owner, to query the capabilities of each access point, and to perform a handover to the best access point available.

We have shown, however, that current mobility management in the form of MobileIP in combination with the proposed "candidate access router discovery" (CARD) protocol is not suitable when it comes to inter-provider handovers. The reason is that CARD assumes cooperative access routers that voluntarily provide information about other access routers—which they are unlikely to do if the access routers are owned by competing providers.

Furthermore, we have discussed the advantages of including criteria such as mobile hosts' battery state, current application context, access point capability, pricing, etc. in the handover decision process. By comparing handover control by the network to that of the mobile host, we have advocated that for reasons of efficiency, privacy, and free provider selection it is preferable to let the mobile host control the handover process.

Finally, we have described approaches to solve the shortcomings of current handover mechanisms in the context of high competition between providers, which are based on the idea to let access points provide more information already on the link layer, instead of over the network layer and via other access routers.

References

[1] G. Jenkins. GSM White Paper - Brilliant Past, Bright Future. Deutsche Bank White Paper, http://www.gsmworld.com/GSM White Paper.pdf, Feb. 2004.

[2] R. H. Katz and E. A. Brewer. The Case for Wireless Overlay Networks. In *Proceedings of the SPIE Multimedia and Networking Conference (MMNC'96)*, San Jose, CA, USA, Jan. 1996.

[3] F. Kelly. Models for a self-managed Internet. *Philosophical Transactions of the Royal Society*, A358:2335–2348, 2000.

[4] M. Klerer. Introduction to IEEE 802.20 – Technical and Procedural Orientation. IEEE 802.20 WG Document, http://grouper.ieee.org/groups/802/20/P_Docs/IEEE 802.20 PD-04.pdf, Mar. 2003.

[5] M. Liebsch, A. Singh, H. Chaskar, et al. Candidate Access Router Discovery Draft 6. IETF Internet-Draft, draft-ietf-seamoby-card-protocol-06.txt, Dec. 2003.

[6] S. H. Low and D. E. Lapsley. Optimization Flow Control, I: Basic Algorithm and Convergence. *IEEE/ACM Transactions on Networking*, 7(6):861–875, Dec. 1999.

[7] J. Manner et al. Mobility Related Terminology. IETF Internet-Draft, draft-ietf-seamoby-mobility-terminology-06.txt, Feb. 2004.

[8] J. H. Saltzer, D. P. Reed, and D. D. Clark. End-To-End Arguments in System Design. *ACM Transactions on Computer Systems*, 2(4):277–288, Nov. 1984.

[9] H. J. Wang, J. Giese, and R. H. Katz. Policy-Enabled Handoffs Across Heterogeneous Wireless Networks. In *2nd IEEE Workshops on Mobile Computing and Applications (WMCSA 1999)*, New Orleans, LA, USA, Feb. 1999.