An Application Layer Architecture for Seamless Roaming

Adrian Popescu[†], Dragos Ilie[†], David Erman[†], Markus Fiedler[†], Alexandru Popescu[†‡] and Karel de Vogeleer[†]

[†] *Dept. of Telecommunication Systems*
School of Engineering
Blekinge Institute of Technology
371 79 Karlskrona, Sweden

[‡] *Dept. of Computing*
School of Informatics
University of Bradford
Bradford, West Yorkshire BD7 1DP, United Kingdom

***Abstract*—**

The paper advances a new architecture for seamless roaming, which is implemented at the application layer. This architecture is subject for the research projects PERIMETER and MOBICOME, recently granted by the EU STREP FP7 and EUREKA, respectively. The research challenges are on mobility management, security, QoE management, overlay routing, node positioning, mobility modeling and prediction, middleware and handover.

The foundation of seamless handover is provided by several components, the most important ones being the handover, mobility management, connectivity management and Internet mobility. The paper provides an analysis of these components as well.

I. Introduction

Future mobile networks are expected to be all-IP-based heterogeneous networks that allow users to use any system anytime and anywhere. They consist of a layered combination of different access technologies, e.g., UMTS, WLAN, WiMAX, WPAN, which are connected via a common IP-based core network to provide interworking. These networks are expected to provide high usability (anytime, anywhere, any technology), support for multimedia services, and personalization. Key features are user friendliness and personalization as well as terminal and network heterogeneity. The most important technologies are multicarrier modulation, smart antenna techniques, OFDM-MIMO techniques, adaptive modulation and coding with time-slot scheduler, cooperative communication services and local/triangular retransmissions, software-defined radio and cognitive radio [1].

The main requirements for handover are in terms of service continuity, provision of horizontal and vertical handover, provision of security, policy-based handover, flexibility, transparency to user and design of the system architecture such as it is independent of the (wireless) access technology. Particular focus must be given to mobility management aspects (e.g., access network location, paging and registration) as well as provision of QoS, user and network security [2].

There are many types of handover systems existing today, which can be partitioned based on several dimensions like, e.g., domain, system, overlay, technology. The IETF document RFC 3753 on "Mobility Related Terminology" is perhaps one of the best documents that defines terms for mobility related terminology [3]. The document covers specific terminology used in handover in a heterogeneous environment as well as in mobile ad-hoc networking.

There are three possibilities to handle movement: at the link layer (L2), network layer (L3) and application layer (L5) in the TCP/IP protocol stack. The complexity of handover is large and demands for solving problems of different nature. Accordingly, a number of standard bodies have been working on handover, e.g., IEEE, 3GPP, 3GPP2, WiMAX, IETF. L2 mobility across different access technologies is covered by 3GPP, 3GPP2 and WiMAX in a number of documents. L3 mobility is addressed by IETF. Therefore, the IP Multimedia Subsystem (IMS), which is acting as a service layer, does not need to cover mobility issues related to access but other mobility issues.

The paper advances a new architecture for seamless roaming, which is implemented at the application layer. This architecture is subject for the research projects PERIMETER and MOBICOME, recently granted by the EU STREP FP7 and EUREKA, respectively. The research challenges are on mobility management, security, QoE management, overlay routing, node positioning, mobility modeling and prediction, middleware and handover.

The rest of the paper is as follows. Section II is about seamless handover and the solutions existent today with a particular focus on their limitations. Section III describes the main elements involved in mobility management. Section IV describes the algorithms that can be used for connectivity management in connection with mobility. Section V is about Internet mobility and the most important solutions used to solve this. Section VI advances a new architecture for seamless mobility, which is implemented at L5. Section VII presents the main research challenges of this architecture. Finally, section VIII concludes the paper.

II. Seamless Handover - Situation Today

Most of the existent solutions attempt to solve the handover problem at L2 (access and switching) and L3 (IP) with particular consideration given to L4 (transport). Some of the most important requirements are on seamless handover,

efficient network selection, security, flexibility, transparence with reference to access technologies and provision of QoS.

Typically, the handover process involves the following phases: handover initiation; network and resource discovery; network selection; network attachment; configuration (identifier configuration; registration; authentication and authorization; security association; encryption); and media redirection (binding update; media rerouting).

The basic idea of L2/L3 handover is using Link Event Triggers (LET) fired at Media Access Control (MAC) layer, and sent to a handover management functional module such as L3 Mobile IP (MIP), L3 Fast MIP (FMIP) or IEEE 802.21 Information Server (IS). LET is used to report on changes with regard to L2 or L1 conditions, and to provide indications regarding the status of the radio channel. The purpose of these triggers is to assist IP in handover preparation and execution.

The type of handover (horizontal or vertical) as well as the time needed to perform it can be determined with the help of neighbor information provided by the Base Station (BS) or Access Point (AP) or the IEEE 802.21 Media Independent Handover Function (MIHF) Information Server (IS).

Given the extreme diversity of the access networks, the initial model was focused on developing common standards across IEEE 802 media and defining L2 triggers to make Fast Mobile IP (FMIP) work well. Connected with this, media independent information needs to be defined to enable mobile nodes to effectively detect and select networks. Furthermore, appropriate ways need to be defined to transport the media independent information and the triggers over all 802 media.

In reality, however, the situation is much more challenging. This is because of the extreme diversity existent today with reference to access networks, standard bodies and standards as well as architectural solutions. Other problems are because of the lack of standards for handover interfaces, lack of interoperability between different types of vendor equipment, lack of techniques to measure and assess the performance (including security), incorrect network selection, increasing number of interfaces on devices and the presence of different fast handover mechanisms in IETF, e.g., MIPv4, Fast MIPv6 (FMIPv6), Hierarchical MIPv6 (HMIPv6), Fast Hierarchical MIPv6 (FHMIPv6).

IETF anticipated L2 solutions in standardized form (in the form of triggers, events, etc), but today the situation is that we have no standards and no media independent form [4]. Other problems are related to the use of L2 predictive trigger mechanisms, which are dependent on L1 and L2 parameters. Altogether, the consequence is in form of complexity and dependence on the limitations of L1, L2 and L3. The existent solutions are generally not yet working properly, which may result in service disruptions. Because of this, it is important to develop cross-layer architectural solutions where cooperation is established between L2 and L3 to assist the IP handover process and to improve the performance. Even better would be to develop architectural solutions where IP has control over specific L2 handover-related actions.

Today, user mobility across different wireless networks is mainly user centric, thus not allowing operators a reasonable control and management of inherently dynamic users. This is the reason for why IEEE 802.21 Working Group is doing an effort to ratify the Media Independent Handover (MIH) standard, to enhance the user centric mobility handovers and enable network controlled handovers across heterogeneous networks [5]. In parallel to this, IETF addresses the IP level support for mobile heterogeneous access like, e.g., the Working Group on "The Mobility for IP: Performance, Signaling and Handoff Optimization (MISHOP)". This WG regards the delivery of information for MIH services at L3 or above. The L3 discovery component is also defined. The target is to enable MIH services even in the absence of the corresponding L2 support. The security issue is addressed as well.

IEEE 802.21 defines a framework to support information exchange regarding mobility decisions, irrespective of media. The goal is to facilitate handovers among heterogeneous access networks. Handover decisions are taken based on information collected from both mobile nodes and network, e.g., link type, link identifier, link availability, link quality.

IEEE 802.21 MIH is targeted at optimizing L3 handovers and above. It acts across 802 networks and extends to cellular networks like 802.3, 802.11, 802.16. The 802.21 Media Independent Handover Function (MIHF) Information Server (IS) has information about location of PoA, list of available networks, cost, L2 information (neighbor maps), higher layer services and others. Key benefits are optimum network selection, seamless roaming and low power operation for multi-radio devices.

It is also important to point out that the traditional TCP/IP protocol stack was not designed for mobility but for fixed computer networks. This is particularly shown by the fact that the responsibility of individual layers is ill-defined with reference to mobility. The main consequence is that problems in lower layers related to mobility may create bigger problems in higher layers. Higher layer mobility schemes are therefore expected to better suit Internet mobility.

Better prediction mechanisms and especially some form of movement prediction would definitely improve the handover performance in the sense that this may compensate for errors connected with delay in the handover process and the associated service disruptions. This kind of solutions opens up for research and development of new architectural solutions for handover based on movement, possibly implemented at L5 in the TCP/IP protocol stack.

III. Mobility Management

Mobility management refers to the problem of managing the mobility of users in the context of diverse computing and networking environments. Considerations must be given in this case to elements like location-aware services, system capacity and application demands.

There are two major elements involved in mobility management, i.e., handover management and location management [2]. Handover management refers to the way the network acts to keep mobile users connected when they move and change

their position and access points in the network. For instance, in the case of UMTS, there are two types of handover: intra-cell handover and inter-cell handover. Intra-cell handover refers to the situation when the mobile user changes the communication channel to one with a better signal strength at the same Base Station (BS). Inter-cell handover occurs when a user moves from one cell to another. In this case, another BS takes over the control of the user connection.

Location management refers to the process used by a network to find out the current attachment point of a mobile user and provide call delivery. There are two phases involved in location management: location registration or update and paging. Location registration means that the mobile user periodically notifies the network about the new access point and the network uses this information to authenticate users and to update the location profile. Paging means that the network is queried for the user location profile so that the current position is found.

The standard solution existent today for Location Area (LA) based location update does not allow adaptation to the mobility characteristics of the mobile node. Many research efforts have been done over the last years to improve the performance by designing dynamic location update mechanisms and paging algorithms. The basic idea is that these mechanisms consider user mobility and accordingly optimize the signaling cost associated with location update and paging. The goal is to reduce the costs associated with these mechanisms to a minimum. Examples of such algorithms are distance-, time- and movement-based location update, movement threshold and information theoretic [2].

A very important research issue is therefore regarding location modeling and mobility modeling and prediction. Location modeling refers to how to describe the position of a mobile user, whether it is a one-, two- or three-dimension system. Different methods can be used for location modeling, which depend upon the specific network infrastructure. Usually, the position of a mobile user can be specified at three levels: location area, cell ID and the position inside the cell. Furthermore, one should also mention that a more precise location modeling (i.e., within a cell or a WLAN rather than finding the residing cell) may demand for solving a so-called geo-location problem.

Mobility modeling and prediction strongly influences the performance of other resource management elements like call admission control, routing and handover. Diverse criteria can be used for mobility modeling like, e.g., dimension, scale, randomness and geographical constraints. The most popular models are fluid-flow, random-walk, random-waypoint, Gaussian-Markov, geographic-based, group-mobility and kinematic mobility models [2]. These models have specific advantages and drawbacks, and each of them is usually used in specific cases only.

IV. Connectivity Management

The extreme heterogeneity existing today with reference to access networks and network technologies has had as a consequence that the problem of mobility management has now become more complex. Today, mobility refers not only to the user geographic position but also to the change of a logical location with respect to network access points.

There are two aspects that must be considered in vertical handover. These are regarding handover at device level and handover at flow level [6]. Device level handover refers to the situation when data transfers are switched over from one network interface to another within the same mobile node. On the other hand, flow level handover refers to the situation when the network interface is selected based on the specific traffic flow and every individual traffic flow takes own handover decisions. Multi-homing handover is possible in this case when multiple network connections are simultaneously used.

There are two general classes of algorithms used in the vertical handover, which are based on traditional algorithms and context based algorithms.

Traditional algorithms are typically used in horizontal handover and focus mainly on L1 and L2 parameters like link quality conditions, e.g., Received Signal Strength Indicator (RSSI), Signal to Noise Ratio (SNR), frame error rate and base station workload. These parameters can be used in vertical handover as well. The target in this case is to minimize the number of unnecessary handovers while maintaining throughput and latency constraints.

Context based algorithms target at always providing best possible QoS and user-perceived Quality of Experience (QoE). High level information like user preferences, cost, application features, device capacity, bandwidth, security are considered in this case. The target is to provide the so-called "Always Best Connected (ABC)" paradigm in the handover procedure.

There are three categories of context based algorithms. These are traffic flow based, Simple Additive Weighting (SAW) and Advanced Multiple Criteria Decision Making (MCDM) algorithms [6].

Traffic flow algorithms classify the packets based on their traffic class field, IP address, port number and protocol. Different network interfaces are assigned to different traffic flows based on the characteristics of applications like, e.g., real-time and non-real-time.

SAW-based algorithms use weights assigned to parameters considered relevant for a specific handover mechanism. Weighted sums are computed based on all normalized factor values for the specific parameters. Based on this, individual scores are computed and the network interfaces are ranked based on the scores resulted from the evaluation [7].

MCDM-based algorithms are quite sophisticated. The handover decision is treated in this case as a MCDM problem, which is solved using classical MCDM methods and including techniques like Analytic Hierarchy Process (ARP), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) and Grey Relation Analysis (GRA) [6].

V. Internet Mobility

Internet mobility refers to providing support for communication continuity when an IP-based mobile node moves to

different networks and it changes the point of attachment. There are in this case several basic requirements on the TCP/IP protocol stack and networks, which refer to handover and location management, support for multihoming, support for current services and applications, support for security, avoidance of using third-party for routing and security purposes as well as easy integration in the existent infrastructure.

The traditional TCP/IP protocol stack and networks have been designed and developed for fixed computer networks. This means that several limitations must be addressed when further developing the system to provide support for mobility. These limitations are because of physical and link layer, IP layer, lack of cross-layer awareness and cooperation, transport layer and applications.

Today, wireless access techniques are typically providing mobility of homogeneous networks at link layer only. On the other hand, Internet mobility across heterogeneous networks demands for mobility support provided in higher layers as well. Furthermore, radio channels typically show limitations when compared to fixed networks. They are characterized by lower bandwidth, higher bit error rates, faded and interfered signal. These limitations degrade the performance of transport protocols.

The main limitation related to the IP layer is that IP addresses play the roles of both locator and identifier. In a mobile environment the IP address of a mobile node must be changed when moving to another network to reflect the change of the point of attachment. This feature is in conflict with the situation at fixed networks, where the IP addresses never change.

Other important limitations are the lack of cross-layer awareness and cooperation. For instance, the congestion control mechanism of TCP is not able to distinguish packet losses due to link properties from those due to handover. Because of this, TCP does not perform well for seamless roaming. In a similar way, the lack of L2/L3 cross-layer interaction further deteriorates the performance. Another fundamental limitation of transport protocols is because they can not deal with mobility on their own.

Limitations due to improper design of applications for mobile environments are important as well. For instance, applications like Domain Name System (DNS) and Session Initiation Protocol (SIP) have characteristics that are not favorable for mobility. The best example is given by DNS, where the Fully Qualified Domain Name (FQDN) is usually statically bound to an IP address of a node. This is not favorable in the case of mobility, where mobile nodes change IP addresses. Further, the main limitation of SIP is because of the relatively large delays associated with SIP transactions.

A number of solutions have been suggested and developed to solve the problem of Internet mobility. They can be partitioned into four classes:

- Mobility support at L3, e.g., MIPv4, MIPv6, Location Independent Network Architecture for IPv6 (LIN6)
- Mobility support at L4, e.g., improving TCP performance for mobility (e.g., Mobile TCP - MTCP) or mobility extension to TCP (e.g., MSOCK, Mobile UDP - MUDP, Mobile SCTP - MSCTP)
- New layer between L3 and L4, where the Internet mobility is deployed, e.g., Host Identity Protocol (HIP), Multiple Address Service for Transport (MAST)
- Mobility support at L5, e.g., Dynamic Updates to DNS (DDNS), Session Initiation Protocol (SIP), MOBIKE

Detailed description of these protocols, together with their limitations, is provided in [8]. As a general comment, it is observed that none of the available solutions fulfills all requirements for mobility. For instance, the network layer solutions do not support multihoming, the transport layer solutions do not support location management, application layer solutions are only appropriate for specific applications and so on.

VI. BTH Architecture

A new architectural solution is suggested by Blekinge Institute of Technology (BTH) for seamless handover, which is implemented at L5. Compared to the existent L2/L3 handover solutions, this solution offers the advantage of less dependence on physical parameters and more flexibility in the design of architectural solutions. By this, the convergence of different technologies is simplified. Furthermore, by using an architecture based on middleware and overlays, we have the possibility to combine the services offered by different (present and future) overlays. This offers the advantage of flexibility in the development of new services and applications. The suggested architecture resembles the Android mobile development platform developed by Google [9], opening thus up for similar architectural solutions developed in the terminal and in the network. By this, new applications and services can be easily designed and developed, which can, e.g., be written once and deployed in many phones. This facility is today prevented because of the current mobile technical fragmentation.

The suggested architectural solution is shown in Fig. 1. It is based on using a middleware (with a common set of APIs), a number of overlays and a number of underlays. By middleware, we refer to software that bridges and abstracts underlying components of similar functionality and exposes the functionality through a common API. On the other hand, by overlay we refer to any network that implements its own routing or other control mechanisms over another already existing substrate, e.g., TCP/IP, Gnutella. Finally, by underlays we refer to substrates, which are abstracted.

The underlays can be either structured or unstructured. Structured overlays are networks with specific type of routing geometry decided by the Distributed Hash Table (DHT) algorithm they use. Structured underlays use keys for addressing like, e.g., Chord [12]. In unstructured overlays the topology can be viewed as emergent instead of being decided before hand. Unstructured overlays can use IP addresses or other forms of addressing, e.g., Gnutella, which uses Universal Unique IDs (UUIDs) for addressing.

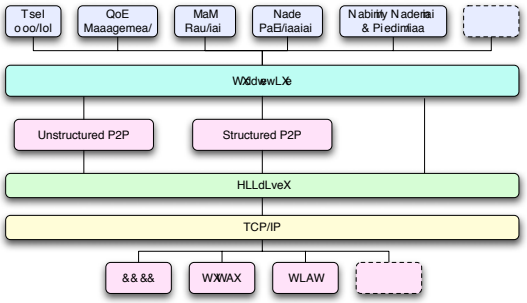


Fig. 1. BTH architecture

An important goal of the middleware is to abstract structured and unstructured underlays as well as overlays. The BTH research group uses this API architecture in different projects like, e.g., QoS routing [10], [11].

There are a number of research challenges that must be solved. These are regarding SIP and delay measurements, security, Quality of Experience (QoE) management, overlay routing, node positioning, mobility modeling and prediction, middleware and handover.

VII. Research Challenges

A. Mobility Management

An application layer mobility system together with IEEE 802.21 and Media-independent Pre-Authentication (MPA) is suggested. Application layer mobility refers to using the application protocol Session Initiation Protocol (SIP) [13]. This solution offers the advantage of eliminating the need for a mobility stack in mobile nodes and also does not demand for any other mobility elements in the network. Simple IP is used in this case together with a SIP protocol stack. No additional elements are needed to support application layer mobility. This solution is very suitable for applications like VoIP.

SIP-based handover also has drawbacks. These are mainly because SIP is an application protocol and therefore involves large delays in handover, due to application layer processing. There are several solutions to reduce the handover delays, and one of the most efficient is to develop a tight-coupled interworking architecture like, e.g., in the case where the WLAN Access Points are integrated into the UMTS network architecture [1].

Another drawback is because the existing client frameworks do not accommodate IETF SIP [13] and 3GPP SIP [14] within the same framework. The consequence is that one needs two different sets of client frameworks on the mobile, one for the mobile domain (e.g., UMTS) and the other one for the fixed domain (e.g., fixed broadband access in combination with WLAN). Furthermore, it is also important to do delay measurements and to analyze the SIP transactions and potential weaknesses (e.g., large delays in handover, security issues) as well as ways to compensate for these limitations.

B. Security

An important problem is given by the compatibility problems related to the authentication used in WLAN and mobile networks. Today, the authentication schemes used in the WLAN hotspots vary widely. Even worse, they are different from the authentication schemes used in mobile networks.

The security schemes are different for network access and for intra- and inter-technology handovers [15]. Typically, network access involves the following security steps: network access authentication; secure association; and access control and encryption.

Network access security is basically about how to bind these steps together to provide appropriate security properties for network access with the use of security associations. An important challenge is to reduce the security signaling latency, which originates (up to 90 %, with values of hundreds of ms) from the Extensible Authentication Protocol (EAP) signaling.

There are several solutions existent today for handover security in intra-technology handovers like, e.g., Access Point (AP) to AP, Base Station (BS) to BS and typically within the same Authorization, Authentication and Accounting (AAA) domain. The challenge in this case is to reduce the security-related signaling delay, particularly for the case of single-radio handover. This is because handover techniques that assume concurrent radio usage can not be used in this case, with the consequence of service disruptions. On the other hand, the situation is relaxed in the case of dual-radio devices, although the signaling delay needs to be reduced as well. Service disruptions can be avoided in this case.

Other important research challenges are regarding intra- and inter-AAA-domain handover transitions. Typically, such problems demand for pre-authentication based solutions [15].

C. User Control

This overlay handles the actions related to user control. These actions refer basically to informing the particular overlays (e.g., QoE management, QoS routing) about user preferences and other relevant information. In other words, this overlay helps in offering the end user the possibility to make Always Best Connected (ABC) decisions with the help of generic QoE models and distributed QoE measurements and data exchange. The user is expected to take handover decisions with the help of specific Graphical User Interfaces (GUIs) in multiple-access, multiple-operator environments. These decisions refer to diverse parameters, e.g., QoS, cost, service availability, security and privacy levels. At the same time, the user is expected to do own measurements and to contribute so to the database of the ”QoE Management” overlay. These measurements are tagged with access provider identification as well as location information.

When the end user is roaming among different networks, it requests the ”QoE Management” overlay to provide information about the average QoE aggregated for different operators in the neighborhood. Decision making algorithms are then applied based on the user preferences, to find out the next network. Finally, this information is transferred to the ”Overlay

Routing" overlay, to control the routing and to avoid so service disruption.

D. QoE Management

There are two fundamental functions for QoE management: data collection and adaptation, and data processing. Two distinct systems are suggested for these units. A distributed system is used for the data collection and adaptation unit. Diverse collection and adaptation modules can be placed in different places in a network/underlay, which are suitable for the particular measurement task. On the other hand, a centralized system is used for the data processing unit, and the QoE data should be finally available in BS or AP. QoE may in this case reflect mean values of different QoS parameters, taken over a particular network/underlay and a particular time interval.

E. Overlay Routing

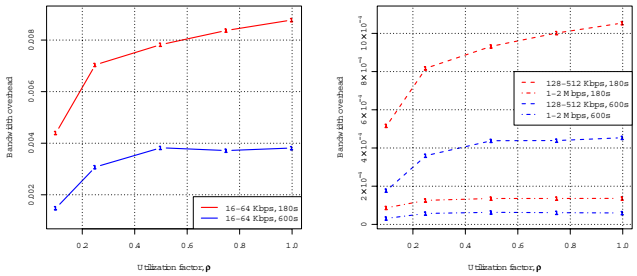
Unicast QoS routing is one of the most important overlays in the BTH architecture. A protocol called Overlay Routing Protocol (ORP) has been developed to implement the unicast QoS routing [11]. The main purpose of ORP is to provide soft QoS to end-users. ORP nodes establish paths with each other on demand, subject to constraints on bandwidth, delay, loss, jitter, etc. The main areas of use for ORP are VoIP, videoconferencing and large data transfers.

ORP itself is concerned with the problem of finding and maintaining QoS-constrained paths only. It relies on other service to build and maintain the overlay. For example, in a Gnutella environment, ORP uses the Gnutella overlay and piggybacks its messages on the Gnutella messages. As long as the service provides ways to address and transport messages in the overlay, ORP can be adapted to use it.

In ORP, each node manages its own traffic flows as well as traffic flows from other nodes. The ORP framework consists of two protocols: Route Discovery Protocol (RDP) and Route Maintenance Protocol (RMP) [11]. RDP is used to find a QoS-constrained path in the overlay. It does this by forwarding a path request on all links that can satisfy the request. When path requests reach the destination, acknowledgments are sent back to the source over the feasible path. An interesting solution is, e.g., to modify RDP to use Gnutella's dynamic query method. This method has the advantage that it reduces the total traffic volume required to satisfy a path query.

RMP is used to handle churn. Each ORP node is responsible for a number of QoS paths, i.e., its own and those belonging to other nodes that use the current node as transit node. The ORP node exchanges link-state information with each node on each of the paths it is responsible for. This is done by using a link-vector algorithm. The ORP node computes the K shortest paths for each destination node it knows about from the link-vector algorithm. These are backup paths. If the next hop on a QoS path exits the overlay, the ORP node spreads the traffic on the backup paths and sends a message to the source asking it to recompute a QoS path. If such a path is found then the traffic can be rerouted on it.

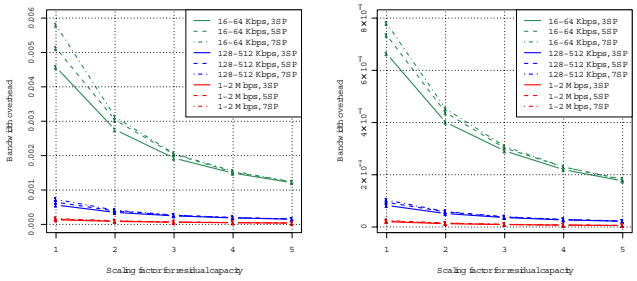
The performance of ORP has been evaluated through a comprehensive simulation study and a large set of results are reported in [11]. The study has showed for instance that QoS paths can be established and maintained as long as one is willing to accept a protocol overhead of maximum 1.5 % of the network capacity. It has been also observed that RDP can find bandwidth-constrained paths with no more than 0.03 % overhead in a network with 1000 nodes when the Time To Live (TTL) is 8 (Fig. 2). The call blocking ratio depends on the amount of available bandwidth in the network and on the TTL value.



(a) 16–64 Kbps. (b) 128–512 Kbps and 1–2 Mbps.

Fig. 2. RDP bandwidth overhead.

Furthermore, RMP is used to restore RDP paths when the original paths are broken, which may include the case when the path QoS constraints can no longer be satisfied. It has for instance been observed that, in conditions of aggressive churn, RMP is able to restore up to 40 % of broken paths used for transporting 1-2 Mbps flows, with approximatively 0.02 % bandwidth overhead (Fig. 3).



(a) Exp(30 s) churn (b) Exp(300 s) churn

Fig. 3. RMP bandwidth overhead.

Another important problem is related to the path selection algorithm. In general, the path selection problem is posed in the form of an optimization problem. A network can be represented in the form of a directed graph G = (V, E), where V is a set of V nodes/vertices and E is a set of E directed links/edges. Each link has a number of additive QoS metrics (e.g., delay) as well as non-additive QoS metrics (e.g., bandwidth, error rate). Problems involving constraints on non-additive metrics can be resolved by, e.g., pruning the links of

the graph that do not satisfy the constraints. On the other hand, additive metrics are more difficult to handle and demand for the so-called Multi-Constrained Path Optimization (MCPO) algorithms. Moreover, there may also be an objective function that needs to be optimized, e.g., a global cost function. This problem is even more complicated given the particular conditions existent in such cases, i.e., multiple constraints, dynamic environments, "real-time" performance demand [10].

Several popular optimization algorithms considered for path selection are Self-Adaptive Multiple Constraints Routing Algorithm (SAMCRA), the Simplex Method (popular method of mathematical programming for linear optimization problems with linear constraints), Gradient Projection Method (for unconstrained optimization problems) and Conjugate Gradient Method (for unconstrained optimization problems) [10].

A number of algorithms have been implemented so far, such as Breadth-First-Search (BFS), Depth-First Search (DFS), topology closure algorithm, among others. Today, the optimization algorithms use only bandwidth information. Changing them to use delay and packet loss is straightforward. It is also important to mention that optimal routing may involve multiple constraints, e.g., minimum delay across the network in addition to bandwidth constraints. The consequence is that nonlinear optimization algorithms need to be developed.

Other important research activities are regarding the extension of the above-mentioned routing protocols to handle node mobility, the development of online methods to measure QoS metrics as well as to build up a separate overlay for network embedding These algorithms must finally be tested in real network environments like, e.g., PlanetLab [16].

Finally, it is also important to develop real-world simulation models (particularly for WLAN) by including obstacles and developing realistic, generic and comprehensive mobility and signal propagation models that emulate properties of fading in the presence of obstacles. Such models should allow the placement of obstacles that restrict movement and obstruct the signal propagation.

F. Node Positioning

The issues of modeling and management of location and mobility represent today some of the most challenging research issues. Location modeling is dealing with how to model the location of mobile nodes and their relationships in space. On the other hand, mobility modeling and prediction specify the dynamic characteristics of node movement, which is very useful, e.g., in seamless roaming, design and performance modeling of wireless networks, routing and network planning.

Today, there are different positioning systems, which depend upon the particular wireless system, e.g., UMTS, WLAN and the standards used for location management. Accordingly, the location of a Mobile Node (MN) can be modeled and described in different ways depending upon the network infrastructure. For instance, a base station in cellular networks serves as an access point in delivering radio services. This means that the location of a MN is limited to one cell in cellular networks. Furthermore, the exact position of a MN in a cell can be determined by solving a so-called geolocation problem [2]. On the other hand, the location of MNs can not be determined with reference to cells in the case of WLANs and ad-hoc networks. In such cases, geolocation algorithms can still be used, the difference however is that the MNs are used for routing as well. This means that positioning systems are more sophisticated in this case [17].

We suggest a solution where the "Node Positioning" overlay collects positioning information from different underlays (using different positioning systems) and transforms these positions into positions placed in a geographic positioning system created by the "Node Positioning" overlay. This information is used by other overlays like "Mobility Modeling and Prediction". Information regarding diverse geographic circumstances, e.g., street number, distance to a city sign, can today be easily collected from different sources. A geographic positioning system offers important advantages like, e.g., implementing geographic routing, geographic-based roaming, better facilities for QoS, provision of location-aware services.

G. Mobility Modeling and Prediction

The basic function of this overlay is to avoid the "Break Before Make (BBM)" phenomena in the handover procedure. This means that, with reference to the current geographic position and the prediction of user mobility, the time is computed for doing handover such as to avoid service interruptions. This is particularly important in BBM networks like WLAN and WiMAX, where local conditions may create abrupt service disruptions [18].

There are several dimensions that are relevant for modeling and predicting mobility [19]. For instance, a mathematical analysis can be done at different levels of detail, i.e., microscopic (individual behavior), mesoscopic (reflecting the homogenized movement behavior of several nodes), macroscopic (global parameters). Another dimension refers to the tools used for analysis, like Markovian models, transportation theory models, flow traffic models. Time dependency is also important in analysis, with processes that can behave stationary or non-stationary. Geographical areas considered for analysis is an important dimension as well. Typical geographical areas are hot spots, urban, highway/main road, open/rural.

Several criteria can be used in the modeling and prediction of a MN movement, like dimension (one-, two- or three-dimension), scale of mobility (micro- or macro-mobility), degree of randomness used in modeling, geographical constraints (indoor, outdoor or vehicular), destination oriented parameters and expected change of parameters (e.g., kinetic models, speed decrease) [2].

The size of cells used in modeling can be different, from macro-cells (40 km–1 km, used in rural areas, umbrellas in urban areas) to micro-cells (1 km–100 m, used for streets, main roads, avenues) and further on to pico-cells (100 m–10 m, used in airports, railway stations, business areas, indoors). To make things even more complicated, the cell shape is irregular in practice (due to features like propagation, shadowing), although regular shape is typically used in modeling.

Furthermore, there are many other parameters that need to be considered in developing a model for mobility, e.g., cell residence time, call holding time, channel holding time, handover time, handover area residence time, traffic density, and speed [19]. The metrics used for modeling mobility are required to capture diverse parameters, e.g., spatial dependence, temporal dependence, geographical restrictions, relative velocity. The complexity is therefore very high.

H. Middleware

The main goal of the project is to develop a testbed to facilitate the development, testing, evaluation and performance analysis of different solutions for user-centric mobility, while requiring minimal changes to the applications using the platform. In other words, we implement a software system with two sets of APIs, one for application writers and another one for interfacing various overlay and underlay systems.

Current overlay implementations are built with incompatible language specific frameworks on top of the low level networking abstractions, e.g., YOID, i3, JXTA [12], [20]. This complicates the design of overlays and their comparison as well as the integration of different overlays. We therefore suggest a middleware based on the Key-Based Routing (KBR) layer of the common API framework suggested in [20]. By doing so, independent development of overlay protocols, services and applications is facilitated.

The middleware is designed to work on top of both structured and unstructured underlays. Structured underlays can be used to construct services such as Distributed Hash Tables (DHT), scalable group multicast/anycast and decentralized object location. The advantage is that they support highly scalable, resilient, distributed applications like cooperative content distribution and messaging. Unstructured overlays do not have such facilities, but they tend to have less overhead in handling churn and keyword searches [21].

By using a common API, we can develop applications by using combinations of arbitrary overlays and underlays. This facility allows us to design a testbed where we can investigate interoperability issues and performance of different combinations of protocols. This also allows us to have overlays that export APIs that other overlays can use. For instance, we can have the "QoE Management" export an API that can be used by the "QoS Routing" and "Handover" overlays.

I. Handover

BTH has developed an interesting solution for vertical handover, which is called Network Selection Box (NSB) [7]. Tunneling is used to send the packets over the interfaces encapsulated in UDP. The NSB can today be used for the transport over WLAN, UMTS and GPRS. The solution automatically switches to the best network available, detects when a network connection is lost, performs handover during ongoing communication without breaking the session is transparent to user, and, allows applications to determine the quality of connections.

VIII. Conclusions

The paper has two parts. The first part is on developments and challenges related to seamless handover, namely L2/L3 handover, mobility management, connectivity management and Internet mobility. The second part is dedicated to an architectural solution suggested for L5 handover. The research challenges are on mobility management, security, QoE management, QoS routing, node positioning, mobility modeling and prediction, handover and middleware. The paper has developed on these challenges as well.

References

[1] Garg V.K., *Wireless Communications and Networking*, Morgan Kaufmann, 2007

[2] Katsaros, D., Nanopoulos A. and Manolopoulos Y., *Wireless Information Highway*, IRM Press, 2005

[3] Manner J. and Kojo M., *Mobility Related Terminology*, IETF RFC 3753, http://www.ietf.org

[4] Gupta V., Williams M.G., Johnston D.G., McCann S., Barber P. and Ohba Y., *802.21 - Overview of Standard for Media Independent Handover Services*, IEEE 802 tutorial, http://ieee802.org/802_tutorials/index.html

[5] IEEE, *Draft IEEE Standard for Local and Metropolitan Area Networks: Media Independent Handover Services*, IEEE P802.21/D04.00, IEEE, February 2007

[6] Sun J-Z., *A Review of Vertical Handoff Algorithms for Cross-Domain Mobility*, International Conference on Wireless Communications, Networking and Mobile Computing, Shangai, China, September 2007

[7] Isaksson L., *Seamless Communications Seamless Handover Between Wireless and Cellular Networks with Focus on Always Best Connected*, PhD thesis, BTH, Karlskrona, Sweden, March 2006

[8] Le D., Fu X. and Hogrefe D., *A Review of Mobility Support Paradigms for the Internet*, IEEE Communications Surveys and Tutorials, Volume 8, No. 1, 1st Quarter 2006

[9] Android, An Open Headset Alliance Project, http://code.google.com/android/

[10] Ilie D. and Popescu A., *A Framework for Overlay QoS Routing*, 4th Euro-FGI Workshop on "New Trends in Modelling, Quantitative Methods and Measurements" (WP IA.7.1), Ghent, Belgium, May/June 2007

[11] Ilie D., *On Unicast QoS Routing in Overlay Networks*, PhD thesis, BTH, Karlskrona, Sweden, October 2008

[12] Stoica I., Morris R., Karger D., Kaashoek F. and Balakrishnan H., *Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications*, ACM SIGCOMM 2001, San Diego, USA, August 2001

[13] Rosenberg G., Schulzrinne H., Camarillo G., Johnston A., Peterson J., Sparks R., Handley M. and Schooler E., *SIP: Session Initiation Protocol*, IETF RFC 3261, http://www.ietf.org

[14] ETSI/3GPP, *Universal Mobile Telecommunication System (UMTS): Signaling Interworking Between the 3GPP Profile of the Session Initiation Protocol (SIP) and non-3GPP SIP Usage*, 3GPP TR 29.962 version 6.1.0 Release 6, http://www.3gpp.org/ftp/specs/html-info/29962.htm

[15] Ohba Y., Meylemans M. and Das S., *Media Independent Handover Security*, tutorial, IEEE 802.21, March 2008, http://ieee802.org/802_tutorials/index.html

[16] PlanetLab, http://www.planet-lab.org/

[17] Siva Ram Murthy C. and Manoj B.S., *Ad Hoc Wireless Networks Architectures and Protocols*, Prentice Hall, 2004

[18] Yoo S-J., Cypher D. and Golmie N., *Predictive Link Trigger Mechanism for Seamless Handovers in Heterogeneous Wireless Networks*, Wireless Communications and Mobile Computing Journal, John Wiley & Sons, Vol. 8, Issue 7, September 2008

[19] Gavrilovska L. and Prasad R., *Ad Hoc Networking Towards Seamless Communications*, Springer, 2006

[20] Dabek F., Zhao B., Druschel P., Kubiatowicz J. and Stoica I., *Towards a Common API for Structured Peer-to-Peer Overlays*, Proceedings of IPTPS, Berkeley, CA, USA, February 2003

[21] Chawathe Y., Ratnasamy S., Breslau L., Lanham N. and Shenker S., *Making Gnutella-Like P2P Systems Scalable*, ACM Conference on Applications, Technologies, Architectures and Protocols for Computer Communications, Karlsruhe, Germany, 2003