Distributed and Dynamic Mobility Management in Mobile Internet: Current Approaches and Issues

H Anthony Chan
University of Cape Town, Rondebosch, South Africa
Huawei Technologies, Plano, USA
Email: h.a.chan@ieee.org

Hidetoshi Yokota
KDDI R&D Laboratories, Inc., Fujimino, Japan
Email: yokota@kddilabs.jp

Jiang Xie
University of Northern Carolina, Charlotte, USA
Email: linda.xie@uncc.edu

Pierrick Seite
France Telecom – Orange, Cesson-Sevigne, France
Email: Pierrick.seite@orange-ftgroup.com

Dapeng Liu
China Mobile, Beijing, China
Email: liudapeng@chinamobile.com

Abstract—**Cellular networks have been hierarchical so that mobility management have primarily been deployed in a centralized architecture. More flattened network architecture for the mobile Internet is anticipated to meet the needs of rapidly increasing traffic from the mobile users and to reduce cost in the core network. Distributing the mobility management functions as opposed to centralizing them at the root of the network hierarchy is more compatible with a flat network architecture. Mobility management may be distributed at different levels: core level, access router level, access level, and host level. It may also be partially distributed or fully distributed. A distributed mobility management architecture avoids unnecessarily long routes, is more scalable with the increasing number of mobile users, and is a convenient platform for dynamic mobility management which means providing mobility support to mobile users only when they need the support. Dynamic mobility management can avoid waste of resources and also reduce signaling overhead and network cost. The desired distributed and dynamic mobility management needs to solve existing problems, meet the needs of changes in traffic and network architecture, and be simple and inexpensive to deploy. This paper surveys existing mobility management solutions in mobile Internet, explains the limitations of a centralized mobility management approach, and discusses potential approaches of distributing mobility management functions. The issues and challenges in the design of distributed and dynamic mobility management are also described.**

Index Terms—**distributed mobility management, dynamic mobility management, mobility anchor, Mobile IP, proxy Mobile IP.**

I. Introduction

The Internet, as it converges with mobile wireless networks, has been experiencing numerous new challenges requiring various extensions to the base mobility protocols. While extensions are needed to optimize handover performance, more extensions are needed with the proliferation of multiple-interface devices utilizing heterogeneous wireless access networks. Deployments missing the appropriate extensions can result in sub-optimal performance.

Further challenges to the deployment of mobility support in mobile Internet is in the content delivery network (CDN) environment in which the content servers are moving closer to access networks. In addition, a new trend for mobile networks is to become more flat in network architecture, i.e., to have fewer levels of hierarchy. The impact of these changes to existing deployment of mobility protocols should be understood in order to ensure better performance optimization. Distributed mobility management [1] with local content servers and the flattened network may integrate the needed extensions to ease deployment and optimize the performance.

Manuscript received Dec 8 2010; accepted Jan 13 2011.

doi:10.4304/jcm.6.1.4-15

Meanwhile, the mobile user traffic volume is increasing much more rapidly than the revenue, and service/network providers are already experiencing need to selectively offload traffic through alternative access networks. In addition, dynamic mobility management is needed with the co-existence of mobile nodes running applications that are actively using mobility support from the network and those that are not [2] [3]. A further selective capability arises when, after route optimization, it is desirable for the end nodes to communicate in peer-to-peer mode without the need for signaling message exchanges to establish and to periodically refresh a bi-directional security association between them. These selective mechanisms are called dynamic mobility management.

Most existing IP mobility solutions are derived from Mobile IP (MIP) [4] [5] principles where a given mobility anchor, e.g., the home agent (HA) in Mobile IP or the local mobility agent (LMA) in Proxy Mobile IPv6 (PMIPv6) [6], maintains mobile nodes (MNs) bindings. Data traffic is then encapsulated between an MN or its access router (AR), e.g., the mobile access gateway (MAG) in PMIPv6, and its mobility anchor. These approaches have been implemented in a centralized architecture where both the mobility context and traffic encapsulation are maintained at a central network entity, the mobility anchor.

Such centralized implementation of mobility management provides the ability to route packets to an MN wherever the MN is located and to maintain IP session continuity during handovers, i.e., when the MN changes its IP point of attachment. However, compared with a distributed approach, a centralized approach has several issues or limitations affecting its performance and scalability, which requires costly network dimensioning and engineering to fix them.

The rest of this paper is organized as follows. After explaining the background on mobility management in Section II, this paper discusses the issues with centralized IP mobility management, as compared with distributed and dynamic mobility management in Section III. It then discusses the networks for which distributed mobility management is relevant in Section IV and categorizes different approaches in distributed and dynamic mobility management in Section V. Finally, some challenges to the work of distributed and dynamic mobility management are stated in Section VI.

II. Background on Mobility Management

This background section explains the basic concepts of mobility management and summarizes the options for mobility management. network-layer mobility management including mobile IP and proxy mobile IP, centralized and distributed mobility management.

A. Session Continuity and Mobility Management

Mobility management provides the mechanisms for maintaining active session continuity while a user switches between different communication channels, locations, protocols, networks, and physical environments. It offers seamless access and connectivity to users across personal, local, and wide area networks without interruption. The main concerned issues involved in mobility management include handovers, routing/re-routing, location management, address management, session identification, session migration, etc.

B. Mobility Management Options

The mobility management functions in data networks may reside in different protocol layers according to the design [7]. At the link layer, mobility concerns a change in the access point. In this case, a handover is triggered to carry out the detach/attach operations to different access points. Information on the physical characteristics of the network, such as received signal strength, channel condition, and bit error rate, is usually needed during the handover process. IEEE 802.11 and 802.16 standards have amendments to introduce the link-layer handover procedures, including IEEE 802.11r [8] and IEEE 802.16-2009 [9]. Numerous papers have proposed different link-layer fast handover schemes in IEEE 802.11-based wireless networks [10] [11]. In addition, the IEEE Std. 802.21-2008 [12] provides a framework for handovers between heterogeneous wireless networks such that the same framework is applicable to different network types which differ at the link-layer and below.

At the network layer, mobility concerns a change in the subnet, i.e., the location change in the Internet. Network-layer mobility solutions adopt two basic approaches: routing-based approach and mapping-based approach [13]. Under the routing-based approach, a mobile keeps its IP address unchanged regardless of its location changes. Thus, the IP address is used to both identify the mobile and to deliver packets to it. In this case, the routing system must keep tracking the most up-to-date location of the mobile and update the routing tables to deliver packets with the unchanged IP address to the new location, which is not scalable to increasing number of MNs. Routing-based network-layer mobility solutions include Cellular IP [14], HAWAII [15], and TIMIP [16]. Under the mapping-based approach, the IP address of a mobile dynamically changes to reflect the current location of the mobile. In this case, an explicit mapping function in the system is needed to map the stable identifier of the mobile to its changing IP address for packet delivery. Mapping-based network-layer mobility solutions include Mobile IP (MIP) [4] [5], Proxy Mobile IP (PMIP) [6], and Hierarchical Mobile IP (HMIP) [17].

Mobility management at the transport layer focuses on end-to-end mobility for TCP connections. Mobility solutions handle how to migrate TCP connections when the IP address of a TCP end-node changes. Transport-layer mobility solutions, such as the end-to-end approach proposed in [18] and Mobile Stream Control Transmission Protocol (M-SCTP) [19], use dynamic domain name system (DNS) to track the changing IP address of a mobile but at the same time keep the ongoing TCP connection unaffected. Other solutions, such as the Indirect TCP (I-TCP) [20] and MSOCKS [21], split a TCP connection at intermediate agents so that the connection between the mobile and the intermediate

agent is updated during mobility, while the correspondent node (CN) is unaware of the movement.

Mobility solutions at the application layer let specific applications manage mobility using end-to-end signaling. One approach is to implement a specific mobility scheme for every application, such as the Session Initiation Protocol (SIP) [22][23]. Another approach is to implement middleware between applications of two end-nodes to deal with mobility, such as WiSwitch [24]. In this case, the middleware should be able to know which applications require mobility management.

Irrespective to the protocol layers, mobility management design options include host based and network based approaches. A host-based mobility management protocol provides mobility support at the mobile node. A network-based mobility solution resides in the network only. It therefore enables mobility for existing hosts with existing network applications, which are already in deployment but may lack such mobility support in them.

In what follows, this paper focuses on network-layer mobility management. Also, only "mapping based" schemes are further studied.

C. Network-layer Mobility Solutions

Since network layer is present in all Internet nodes, network-layer mobility solutions are studied the most in mobile Internet. Network-layer mobility solutions can offer transparent mobility support to all applications. Among all the solutions, Mobile IP (MIP) [4] [1][5] is the most well-known network-layer mobility solution. In addition, variants of Mobile IP are also proposed, including Mobile IP Regional Registration [25], Hierarchical Mobile IP (HMIP) [17], Fast Handover in Mobile IP (FMIP) [26] [27], Dual-Stack Mobile IP (DSMIP) [28] [29], Proxy Mobile IP (PMIP) [6] and Fast Handovers for PMIP (PFMIP) [30].

1) Mobile IP

Mobile IP (MIP) [4] [1][5] as well as its many variants decouples the session identity, in a home address (HoA), and the routing address, in a care-of-address (CoA). A mobile node (MN) acquires a HoA from its home network and a CoA when it is attached to a visited network. With Mobile IP, the HoA takes the role of session identifier whereas the CoA takes the role of routing address. The binding between them is maintained at the home agent (HA), which is the MN's mobility anchor.

As an MN attaches to a different IP network, its routing address, i.e., the CoA, changes. MIP enables an MN to keep its session identifier by routing via a mobility anchor at its home network so that ongoing sessions may survive the routing address changes.

Figure 1(a) shows the protocol stack as a packet traverses from a CN to an MN. The network layer in the figure shows the destination IP address of the packet. The packet from the CN is destined to the HoA and is intercepted by a HA. The HA tunnels the packet to the MN or a foreign agent (FA) (not shown) by encapsulating the packet using the CoA as the destination address. The MN or its FA will receive this IP-in-IP tunneled packet and de-capsulates the packet to retrieve the original packet.

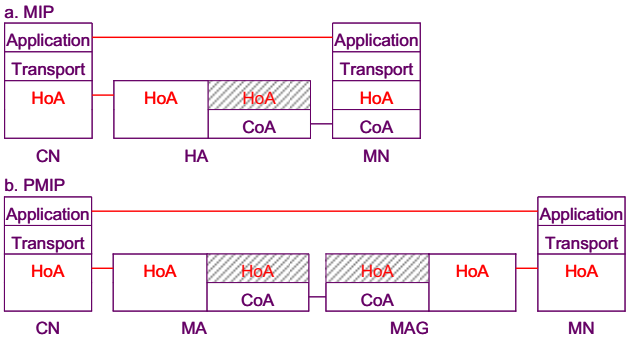

Figure 1. Architectural view and protocol stacks of (a) MIP and (b) PMIP.

The above basic Mobile IP is host-based, requiring MNs to possess the capability of supporting this protocol.

2) Proxy Mobile IP

Network-based mobility management protocol, such as Proxy Mobile IPv6 (PMIPv6) [6], employs network elements to perform the mobility management functions on behalf of the MN and therefore removes the need to add the MIP function to every MN. In PMIP (Figure 1(b)), a CN and an MN communicates with each other using the MN's HoA with today's unmodified IP protocol stack. The MIP function that is needed in the MN for the host-based MIP is moved to a network element called mobile access gateway (MAG). The MN connects to a visited network through the MAG, which provides a proxy CoA, so that the MN can continue to use its own HoA to attach to the visited network and communicate with the CN. The CN only knows the MN's HoA, and its packets destined to the MN are first intercepted by the Mobility Anchor (MA). (MA is originally called local mobility anchor in PMIPv6. The word "local" is dropped in this paper.) MAG and MA manage the binding between the HoA and the CoA, perform encapsulation and decapsulation, and are the tunneling endpoints for the traffic between the MN and the CN. Between the MA and the MAG, packets are tunneled with the proxy CoA as the destination IP address in the outer header of the IP tunnel. The inner header uses the HoA as the destination IP address which is shaded in the figure and is not visible while a packet is being tunneled.

D. Centralized Mobility Management

Mobility management functions in a network may be centralized or distributed. With centralized mobility management, the mapping information for the stable session identifier and the changing IP address of an MN is kept at a centralized mobility anchor. Packets destined to an MN are routed via this anchor. In other words, such mobility management systems are centralized in both the control plane and the data plane.

Many existing mobility management deployments leverage on centralized mobility anchoring in a hierarchical network architecture, as shown in Figure 2. Examples of such centralized mobility anchors are the

home agent (HA) and local mobility anchor (LMA) in Mobile IP [5] and Proxy Mobile IP [6], respectively. Current mobile networks such as the Third Generation Partnership Project (3GPP) [31] UMTS networks, CDMA networks, and 3GPP Evolve Packet System (EPS) networks also employs centralized mobility management, with Gateway GPRS Support Node (GGSN) and Serving GPRS Support Node (SGSN) in the 3GPP UMTS hierarchical network and with Packet data network Gateway (P-GW) and Serving Gateway (S-GW) in the 3GPP EPS network.

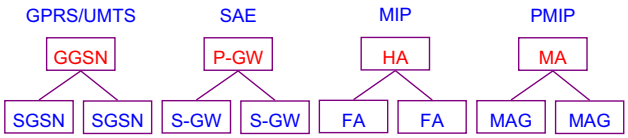

Figure 2. Centralized mobility management deployments.

E. Distributed Mobility Management

Mobility management functions may also be distributed to multiple locations in different networks as shown in Figure 3, so that an MN in any of these networks may be served by a closeby mobility function (MF). Distributed mobility management may be partially distributed, i.e., only the data plane is distributed, or fully distributed where both the data plane and control plane are distributed. These different approaches are described in detail in Section V.

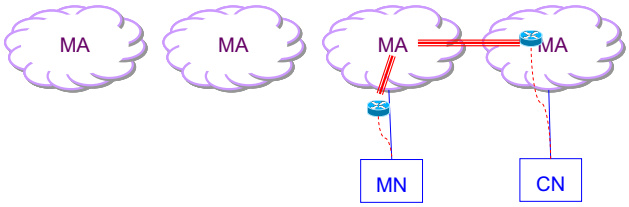

Figure 3. Distributed mobility management deployment. Mobile node in any network is served by a closeby mobility anchor function.

A distributed mobility management scheme is proposed in [32] for future flat IP architecture consisting of access nodes. The benefits of this design over centralized mobility management are also verified through simulations in [33].

While it is possible to design new mobility management protocols for the future flat IP architecture, one may first ask whether the existing mobility management protocols that have already been deployed for the hierarchical mobile networks can be extended to serve the flat IP architecture. Indeed, MIPv4 has already been deployed in 3GPP2 networks, and PMIPv6 has already been adopted in WiMAX Forum [34] and in 3GPP standards. Using MIP or PMIP for both centralized and distributed architectures will then ease the migration of the current mobile networks towards the future flat architecture. It has therefore been proposed to adapt MIP or PMIPv6 to achieve distributed mobility management by using a distributed mobility anchor architecture [35] [36].

In [35], the HA functionality is copied to many locations. The HoA of all MNs are anycast addresses, so that a packet destined to a HoA from any CN from any network can be routed via the nearest copy of the HA. In addition, distributing the function of HA using a distributed hash table structure is proposed in [37]. A lookup query to the hash table will find out where the location information of an MN is stored.

In [36], only the mobility routing (MR) function is duplicated and distributed in many locations. The location information for any MN that has moved to a visited network is still centralized and kept at a location management (LM) function in the home network of the MN. The LM function at different networks constitutes a distributed database system of all the MNs that belong to any of these networks and have moved to a visited network. The location information is maintained in the form of a hierarchy: the LM at the home network, the CoA of the MR of the visited network, and then the CoA to reach the MN in the visited network. The LM in the home network keeps a binding of the HoA of the MN to the CoA of the MR of the visited network. The MR keeps the binding of the HoA of the MN to the CoA of the MN in the case of MIP, or the proxy-CoA of the Mobile Access Gateway (MAG) serving the MN in the case of PMIP.

III. Limitations of Centralized Approach

This section describes the problems or limitations in a centralized mobility approach and compares it against the distributed approach.

A. Non-optimal Routes

Routing via a centralized anchor often results in a longer route. Figure 4 shows two cases of non-optimized routes.

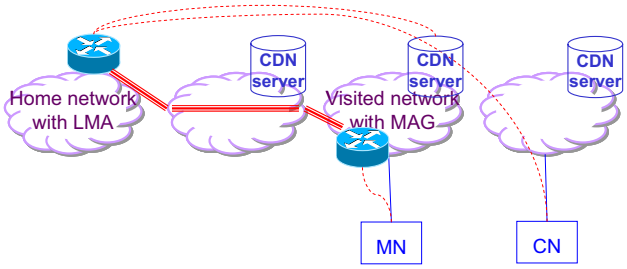

Figure 4. Non-optimized route when communicating with CN and when accessing local content.

In the first case, the MN and the CN are close to each other but are both far from the mobility anchor [38]. Packets destined to the MN need to be routed via the mobility anchor, which is not in the shortest path. The second case involves a content delivery network (CDN) [39]. A user may obtain content from a server, such as when watching a video. As such usage becomes more popular, resulting in an increase in the core network traffic, service providers may relieve the core network traffic by placing these contents closer to the users in the access network in the form of cache or local CDN servers. Yet as the MN is getting content from a local or

cache server of a CDN, even though the server is close to the MN, packets still need to go through the core network to route via the mobility anchor in the home network of the MN, if the MN uses the HoA as the session identifier.

In a distributed mobility management design, mobility anchors are distributed in different access networks so that packets may be routed via a nearby mobility anchor function, as shown in Figure 3.

Due to the above limitation, with the centralized mobility anchor design, route optimization extensions [40] to mobility protocols are therefore needed. Whereas the location privacy of each MN may be compromised when the CoA of an MN is given to the CN, those mobility protocol deployments that lack such optimization extensions will encounter non-optimal routes, which affect the performance. In contrast, route optimization may be naturally an integral part of a distributed mobility management design.

B. *Non-optimality in Evolved Network Architecture*

Centralized mobility management is currently deployed to support the existing hierarchical mobile data networks. It leverages on the hierarchical architecture. However, the volume of wireless data traffic continues to increase exponentially. The data traffic increase would require costly capacity upgrade of centralized architectures. It is thus predictable that the data traffic increase will soon overload the centralized data anchor point, e.g., the P-GW in 3GPP EPS. In order to address this issue, a trend in the evolution of mobile networks is to distribute network functions close to access networks. These network functions can be the content servers in a CDN, and also the data anchor point.

Mobile networks have been evolving from a hierarchical architecture to a more flattened architecture. In the 3GPP standards [31], the GPRS network has the hierarchy GGSN – SGSN – RNC – NB (Node B). In 3GPP EPS networks, the hierarchy is reduced to P-GW – S-GW – eNB (Evolved NB). In some deployments, the P-GW and the S-GW are collocated to further reduce the hierarchy. Reducing the hierarchy this way reduces the number of different physical network elements in the network, contributing to easier system maintenance and lower cost. As mobile networks become more flattened, the centralized mobility management can become non-optimal. Mobility management deployment with distributed architecture is then needed to support the more flattened network and the CDN networks.

C. *Low Scalability of Centralized Route and Mobility Context Maintenance*

Special routes are set up to enable session continuity when a handover occurs. Packets sent from the CN need to be tunneled between the HA and FA in MIP and between the LMA and MAG in PMIP. However, these network elements at the ends of the tunnel are also routers performing the regular routing tasks for ordinary packets not involving a mobile node. These ordinary packets need to be directly routed according to the routing table in the routers without tunneling. Therefore, the network must be able to distinguish those packets requiring tunneling from the regular packets. For each packet that requires tunneling owing to mobility, the network will encapsulate it with a proper outer IP header with the proper source and destination IP addresses. The network therefore needs to maintain and manage the mobility context of each MN, which is the relevant information needed to characterize the mobility situation of that MN to allow the network to distinguish their packets from other packets and to perform the required tunneling.

Setting up such special routes and maintaining the mobility context for each MN is more difficult to scale in a centralized design with a large number of MNs. Distributing the route maintenance function and the mobility context maintenance function among different networks can be more scalable.

D. *Wasting Resources to Support Mobile Nodes Not Needing Mobility Support*

The problem of centralized route and mobility context maintenance is aggravated when the via routes are set up for many more MNs that are not requiring IP mobility support. On the one hand, the network needs to provide mobility support for the increasing number of mobile devices because the existing mobility management has been designed to always provide such support as long as a mobile device is attached to the network. On the other hand, many nomadic users connected to a network in an office or meeting room are not even going to move for the entire network session. It has been studied that over two-thirds of a user mobility is local [41]. In addition, it is possible to have the intelligence for applications to manage mobility without needing help from the network, such as those proposed in [42] [43] [44] [45]. Network resources are therefore wasted to provide mobility support for the devices that do not really need it at the moment.

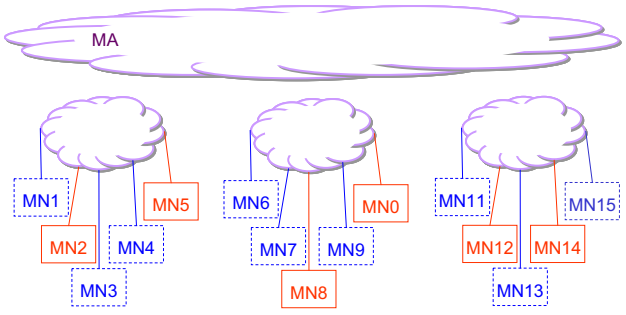

Figure 5. Coexistence of nodes requiring mobility support (in solid line) and those not (in dashed line).

It is necessary to dynamically set up the via routes only for MNs that actually undergo handovers and lack higher-layer mobility support. With distributed mobility anchors, such dynamic mobility management mechanism may then also be distributed. Therefore, dynamic mobility and distributed mobility may complement each other and may be integrated.

E. *Complicated Deployment with Too Many Variants and Extensions of MIP*

Mobile IP (MIP), which has primarily been deployed in a centralized manner for the hierarchical mobile networks, already has numerous variants and extensions including PMIP, Fast MIP (FMIP) [26] [27], Proxy-based FMIP (PFMIP)[30], hierarchical MIP (HMIP) [17], Dual-Stack Mobile IP (DSMIP) [28] [29] and there may be more to come. These different modifications or extensions of MIP have been developed over the years owing to the different needs that are found afterwards. Deployment can then become complicated, especially if interoperability with different deployments is an issue.

While adaptations for MIP are being proposed for the deployment in a distributed manner for more flattened networks, it is desirable to also take a holistic view of different networks and scenarios and integrate the different MIP extensions. The result will then be a more comprehensive mobility solution with options that can be turned on or off depending on different scenarios.

A desirable feature of mobility management is to be able to work with network architectures of both hierarchical networks and flattened networks, so that the mobility management protocol possesses enough flexibility to support different networks. In addition, one goal of dynamic mobility management is the capability to selectively turn on and off mobility support and certain different mobility signaling. Such flexibility in the design is compatible with the goal to integrate different mobility variants as options. Some additional extensions to the base protocols may then be needed to improve the integration.

F. *Mobility Signaling Overhead with Peer-to-Peer Communications*

In peer-to-peer communications, end users communicate by sending packets directly addressed to each other's IP address. However, they need to find each other's IP address first through signaling in the network. While different schemes for this purpose may be used, MIP already has a mechanism to locate an MN and may be used in this way. In particular, MIPv6 Route Optimization (RO) mode enables a more efficient data packets exchange than the bidirectional tunneling (BT) mode, as shown in Figure 6.

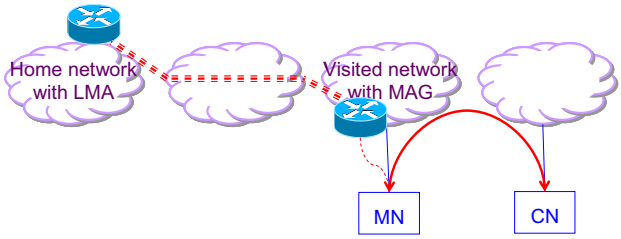

Figure 6. Using MIP to locate hosts with peer-to-peer communications.

This RO mode is expected to be used whenever possible unless the MN is not interested in disclosing its topological location, i.e., the CoA, to the CN (e.g., for privacy reasons) or some other network constraints are put in place. However, MIPv6 RO mode requires exchanging a significant amount of signaling messages in order to establish and periodically refresh a bidirectional security association (BSA) between an MN and its CN. While the mobility signaling exchange impacts the overall handover latency, the BSA is needed to authenticate the binding update and acknowledgment messages (note that the latter is not mandatory). In addition, the amount of mobility signaling messages increases further when both endpoints are mobile.

A dynamic mobility management capability to turn off these signaling when they are not needed will enable the RO mode between two mobile endpoints at minimum or no cost. It will also reduce the handover latency owing to the removal of the extra signaling. These benefits for peer-to-peer communications will encourage the adoption and large-scale deployment of dynamic mobility management.

G. *Single Point of Failure and Attack*

A centralized mobility anchoring architecture is generally more vulnerable to a single point of failure or attack, requiring duplication and backups of the support functions. On the other hand, a distributed mobility management architecture has intrinsically mitigated the problem to a local network which is then of a smaller scope. In addition, the availability of such functions in neighboring networks has already provided the needed architecture to support protection.

IV. Applicable Networks for DMM

Distributed Mobility Management (DMM) can be applied at different parts of a mobile network (see Figure 7). This section introduces possible scenarios for introducing DMM.

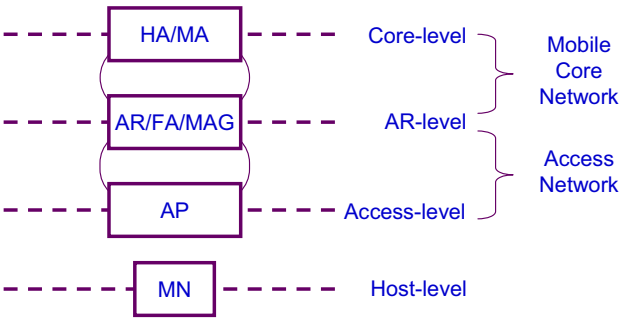

Figure 7. Multiple levels of mobility management distribution.

A. *Distribution in Mobile Core Network*

Conventional mobility management assumes a single mobility anchor per MN, such as the HA, which has been regarded as a negative aspect due to the cause of concentration of mobile data traffic and a single point of failure. By topologically distributing mobility anchors, MNs can be managed in a decentralized way and mobile data traffic can also be distributed (i.e., the "Core-level" distribution in Figure 8). If each mobility anchor covers specific geographical area and an MN crosses this boundary, change of the mobility anchor occurs, and this

handover must be handled properly by, for example, transferring the binding information of the MN from the old to the new mobility anchor. When different mobility anchors manage different blocks of IP addresses, packet delivery to/from the MN must also be assured after handover.

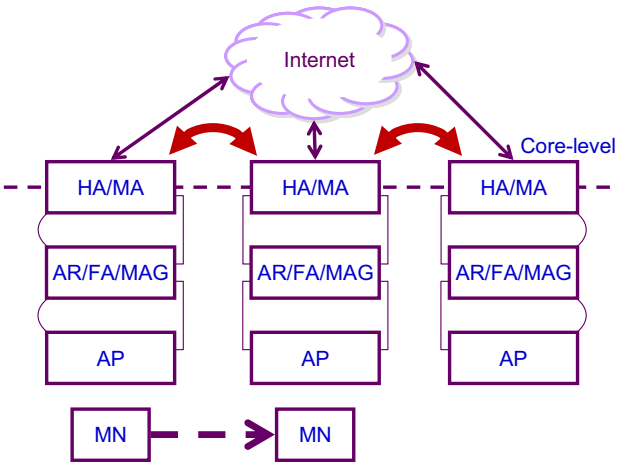

Figure 8. Core-level distribution.

If the mobile network adopts a hierarchical architecture, such as home agent (HA) and foreign agent (FA) in Mobile IPv4 or local mobility anchor (LMA) and mobile access gateway (MAG) in PMIPv6, more flat architecture can be considered by confining the mobility management within a specific region and directly exchanging mobile data at a specific level of hierarchy (i.e., the "AR-level" distribution in Figure 9).

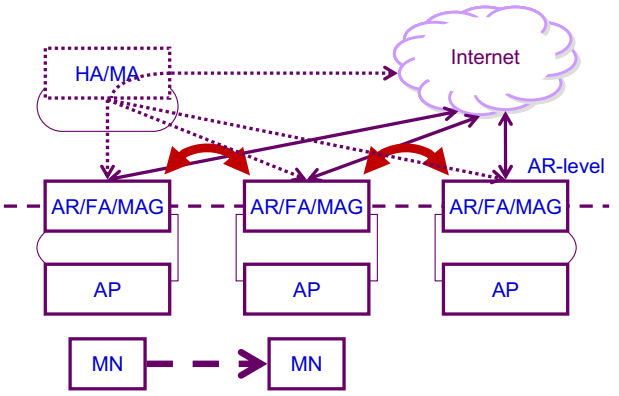

Figure 9. Access router - level distribution: The role of HA/MA is regressed: traffic and mobility bindings are distributed over the ARs, and signaling and routing are needed between the ARs.

The former approach is regarded as localized mobility management and the latter as route localization. Several methods and protocols have been proposed, but no universal and self-contained protocol exists. Moreover, there are different possibilities to distribute mobility functions at the AR-level. The mobility anchor (MA) may be confined down to each of the routers closest to the MNs, i.e., the first AR, resulting in a flat mobility architecture of having only one level of ARs in the mobile access networks interconnecting with each other and to the Internet. It is also possible to have one MA for several ARs. The result is then a less flat mobility architecture with the MA as the next level of hierarchy above the ARs.

B. Distribution in Access Network

The location of information content is getting distributed and closer to users. Consumer Generated Media (CGM) contributed by end users can be innately located in a distributed manner. Content Delivery Network (CDN), which has been constructed near the backbone network, is getting more distributed along with the cache technology and closer to the access network for further efficient use of network resources. As a wireless access method, WiFi is rapidly prevailing and its access points (APs) are being more installed in residential and public areas. As for the cellular system as well, picocells or femtocells are gaining higher attention for more efficient spectrum usage and data traffic offload, examples of which are in 3GPP Local IP Access (LIPA) and Selected IP Traffic Offload (SIPTO) [46]. These access nodes (ANs) basically have layer-2 capability, but by adding layer 3 capability, they can handle IP-level mobility management working as, for example, an FA or MAG, which is shown as the access-level distribution in Figure 10. This approach is employed in [47] for femtocell-based route optimization. The same protocol can be applied as in Section IV.A, but the number of ANs (i.e., WiFi APs or pico/femto-cells) is much larger than the number of the HAs, thus more frequent handover is likely to happen. Therefore, scalability and signaling overhead are the main design issues which should be more carefully considered.

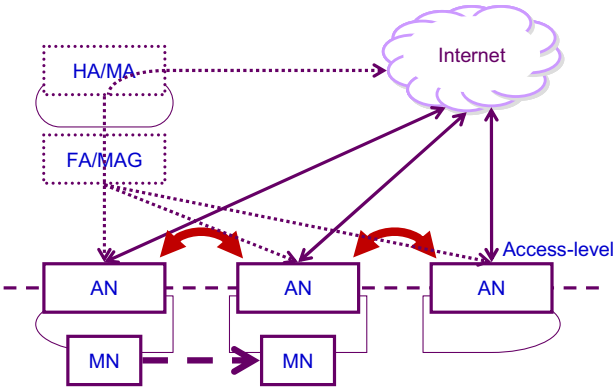

Figure 10. Access - level distribution.

C. Distribution in Host

This is a more peer-to-peer approach, whereby once the corresponding host is found, both hosts directly communicate, which is shown as the "host-level" distribution in Figure 11. In order to discover the peer host, information server such as DNS is required in the network, which can be centralized or distributed. While MIPv6 [5] is not a peer-to-peer communication protocol, its route optimization mechanism can provide a host-to-host communication leveraging the HA.

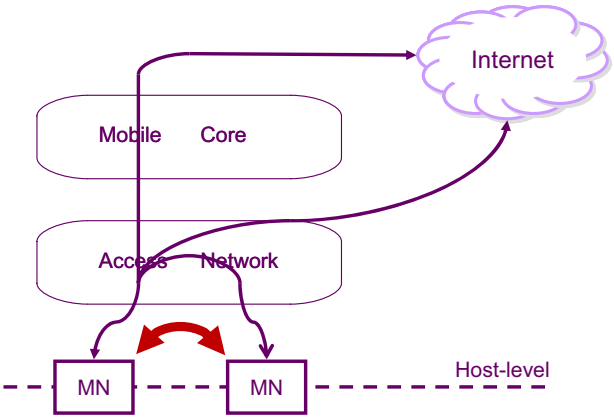

Figure 11. Host - level distribution.

V. Approaches for DMM

Distributed mobility management may be partially distributed or fully distributed.

A. Partially Distributed Approach

Distributed mobility management can be applied partially (1) by considering the separation of control and data planes while taking advantage of the differences in traffic volume or host behavior, and/or (2) by providing mobility support only to the hosts who really need it, thereby saving resources for mobility management.

1) Control/Data Plane Separation

Conventional mobility management protocols such as Mobile IP (MIP) or Proxy Mobile IP (PMIP) combine the control and data planes, which means that all signaling packets and data packets go through the HA or local mobility anchor (MIPv6 route optimization [5] is not included). The volume of data traffic is much higher than that of control traffic, so by separating the control and data planes and applying a distributed architecture to the data plane, effective traffic distribution can be achieved without reallocating mobility anchors during the session, as described in Section IV.A. This simplifies the interaction between distributed mobility anchors (MAs), but new signaling between the control and data plane functional entities is required.

A partially distributed mobility management scenario is depicted in Figure 12 with a centralized control plane and a distributed data plane. There are multiple MAs in the network to route the data traffic. In this example, the routing function of the MA is confined with the access router, but less flat deployment is also possible (see Section IV.A). If an MN attaches to MA1 and initiates an IP communication with a CN, the traffic will be anchored to MA1. When performing a handover to MA2, the control function updates the routing state of MA1 in order to forward packets to the new location. Registration update to the control function may be initiated by the MN or controlled by the network. An example of separating control plane and data plane using PMIP is proposed in [36].

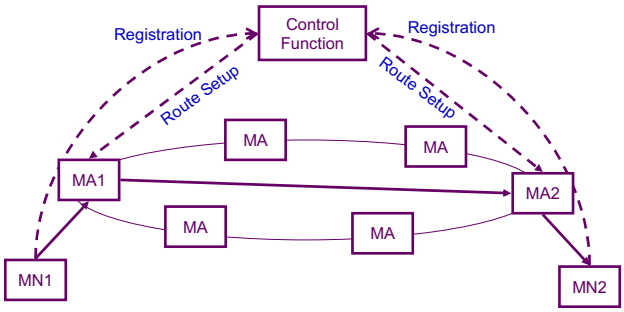

Figure 12. Control/data plane separation scenario with signaling (dashed line) in a centralized control plane and data traffic (solid line) in a distributed data plane.

2) Dynamic Mobility Management

If an MN is nomadic meaning that once attached, it rarely moves, or is idle most of time, it should be enough to provide handover capability only when it is really needed. This can save signaling traffic and network resources for maintaining mobility bindings.

The purpose of dynamic mobility management is to provide mobility support only to those applications and to those MNs that really need it. An important case of such need is to ensure session continuity during handovers. Unless an MN needs to use static IP address, many applications that are initiated after a handover do not really need mobility support. One scenario to avoid providing mobility support to such applications not needing mobility support is depicted in Figure 13. Here, an MN acquires an IP address (IP1) from the local access router (AR1). When this MN moves to another network, this local access router plays a role of the HA to this MN and interacts with the access router (AR2) in the new network for continuous packet delivery. The MN has also acquired another IP address (IP2) in the new network. Communications newly initiated with IP2 while the MN is attached to AR2 will be routed in a standard way as that used to route any other IP packets not involving mobility. In other words, the MN plays with an IP flow to AR1 (the IP flow initiated while attached to AR1) and an IP flow via AR2. If the MN moves away from AR2, while maintaining communications, two mobility anchors will come into play: the data traffic will be anchored in AR1 for communications initiated via AR1 and in AR2 for communications initiated via AR2. An example of dynamic mobility management is the Dynamic Mobility Anchoring (DMA), as proposed in [32] and depicted in Figure 13.

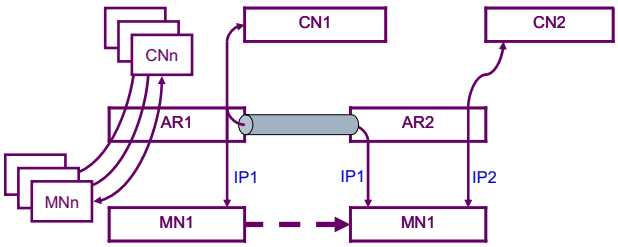

Figure 13. A dynamic mobility management scenario: network sessions initiated after MN1 has moved to a new network uses the new IP address (IP2) which it acquires from the new network.

Dynamic mobility management can be combined with the approaches for control/data plane distribution considered in this paper (i.e., separation of control and data planes in Section V.A.(1) and fully distributed approach in Section V.B). An example of combining dynamic mobility management with fully distributed PMIP is explained in [3].

B. Fully Distributed Approach

The distribution scheme is applied to both control and data planes in a fully distributed approach. One of the most significant issues of the distributed control plane (e.g., distributed HAs), is that a special mechanism is needed to identify the exact mobility anchor that maintains the mobility binding of each MN. A possible solution is to replicate the HA in many networks and use anycast to direct packets destined to the HoA of an MN from any network to the nearest HA [35].

In a fully distributed mobility management, the routing and control functions of the mobility anchor (MA) may be confined within an access router, but less flat deployment is also possible (see Section IV.A). If an MN attaches to an MA and initiates an IP communication with a CN, the traffic is anchored to this MA. When performing a handover to another MA, the control information to support this movement will be shared by these two MAs or all MAs or may be handled independently. Thus, the control function is distributed without relying on a centralized entity. The previous MA can forward packets to the new location of the MN with this control information, which means that the data plane is also distributed. Registration updates to the control function can be initiated by the host or controlled by the network.

The following subsections provide clues to implementing fully distributed mobility management schemes.

1) P2P Type of Approach (Search and Delivery)

This approach searches for the correct MA for an MN before delivering packets.

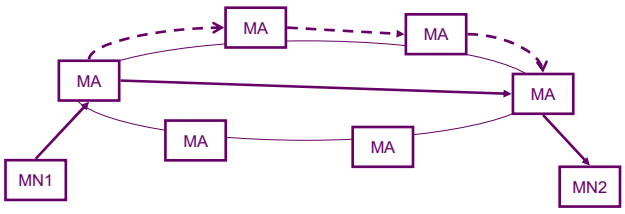

Figure 14. P2P type of fully distributed mobility management with signaling traffic (dashed line) in a distributed control plane and data traffic (solid line) in a distributed data plane.

Distributed Hash Table (DHT) is a popular search mechanism for its efficiency and can be used for searching MAs. However, as the number of MAs increases, the number of hops increases and the MA search time cannot be ignored. In addition, when an MN moves to another network, the location information of the MN needs to be updated. According to the search scenario, this location information may need to be disseminated among multiple distributed MAs, which generates additional signaling traffic among MAs. The user data can be continuously delivered to the MN in the new location by, for example, the new MA's searching for the old MA and getting data forwarded from it.

2) Broadcast/Multicast Type of Approach (Multiple Delivery)

In this approach, data packets are delivered to all or multiple MAs and only the corresponding MA delivers the packets to the MN. This approach does not require an MA search mechanism and the signaling between MAs is not mandatory when the MN moves to a new location. However, the use of the network resources is not efficient since the packets are multiply delivered. This approach is only effective and feasible in relatively limited areas; from local to metropolitan areas, but not suitable for the global area network.

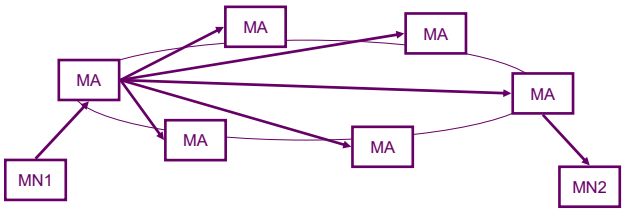

Figure 15. Broadcast/Multicast type of fully distributed mobility management.

VI. Challenges and Issues

Distributed and dynamic mobility management is a promising research direction. Currently, there are several issues and challenges in the design of distributed and dynamic mobility management, which the solutions need to address.

A. Traffic/Mobility Model Has Changed

The very fast growth in mobile Internet traffic is posing new challenges to mobile operator's network. Mobile network operators provide many services to their users using the hierarchical core network, such as that in the 3GPP EPS and UMTS. The traffic from MNs is therefore all going through the mobile core network in order to receive these services when needed. However, the majority of the traffic needs only simple Internet connectivity and does not really need the mobile core network services. In addition, the model of user mobility has also changed, with many low mobility users and smart applications that can handle location changes without relying on the mobility support from the network.

All these changes require conventional mobility management to be re-visited and re-designed. As explained in Section III.D, the existing mobility management design to always provide mobility support to all applications sent to or from any MN wastes network resources especially for users who do not move often. New mobility solutions should consider the change in traffic and mobility models in the current mobile Internet, which also makes dynamic mobility management an attractive option.

B. Network Architecture is Evolving

Mobile operators are offloading traffic, which leads to the evolution of the network towards a flat architecture. However, the current mobility management solutions are not optimized for this flat architecture, as explained in Section III.B. Distributed mobility management can be a good option for this evolved flat network architecture. Based on the network architecture, the level of the distribution (as explained in Section IV) and the approach of distribution (as explained in Section V) should be carefully considered in order to optimize the mobility management performance for the flat architecture.

C. Operator Needs to Simplify Network and Reduce Cost

The hierarchical core network has been deployed with centralized mobility anchor and has provided the high level of telecom grade services. However, it is much more expensive and more complicated than the Internet.

The Internet and the mobile wireless networks have been very different. Their convergence brings together numerous challenges. The Internet is less expensive than the conventional cellular operator networks, and it can be desirable to support mobility in the future Internet [48].

The rapid growth in mobile Internet traffic is already overloading the mobile core network, and it is expensive to expand the capacity of the hierarchical mobile core network. A more flattened network architecture for mobile Internet with distributed mobility management can address these issues, offering simplicity and cost reduction.

D. Performance Optimization is Needed

The CDN servers and content caching are already moving towards the network edge to reduce the delay for users to receive contents from the Internet. Distributing the mobility anchors to the network edge may seem a similar strategy to improve the performance of mobility support. It can, for example, support offloading traffic to the Internet. However, providing such a much larger number of mobility anchors must also be accomplished with cost reduction and optimized performance in mind. In addition, as explained in Section III.D and III.E, a comprehensive mobility solution with options that can be turned on or off depending on different scenarios is desirable. A distributed and dynamic mobility solution must then possess enough flexibility and be able to provide optimized mobility performance for each MN.

VII. Conclusion

In this paper, distributed and dynamic mobility management for mobile Internet was introduced. The current approaches and issues of implementing distributed and dynamic mobility management were also discussed. Distributed mobility management is a promising research direction. It has many features that are compatible with the evolution trend of mobile networks and mobile data traffic. It also has the potential to overcome many limitations of centralized mobility management, if carefully designed. This paper aims to shed light on this promising research direction and more research works on distributed mobility management are expected to emerge in the next few years.

Acknowledgment

The authors wish to thank the following for valuable discussions and comments to this work: Charles Perkins from Tellabs Inc., Melia Telemaco from Alcatel-Lucent Bell Labs, Wassim Michel Haddad from Ericsson, Elena Demaria from Telecom Italia, Philippe Bertin from France Telecom – Orange, and Zhen Cao from China Mobile.

References

[1] D. Liu, Z. Cao, P. Seite, and H. Chan, "Distributed mobility management," IETF draft-liu-distributed-mobility-02, July 2010, work in progress.

[2] M. Kassi-Lahlou, C. Jacquenet, L. Beloeil, and X. Brouckaert, "Dynamic Mobile IP (DMI)," IETF draft-kassi-mobileip-dmi-01.txt, January 2003, work in progress.

[3] P. Seite and P. Bertin, "Dynamic Mobility Anchoring," IETF draft-seite-netext-dma-00.txt, May 2010, work in progress.

[4] C. Perkins, editor, "IP Mobility Support for IPv4," IETF RFC 5944, November 2010.

[5] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support for IPv6," IETF RFC 3775, June 2004.

[6] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, "Proxy Mobile IPv6," IETF RFC 5213, August 2008.

[7] I. F. Akyildiz, J. Xie, and S. Mohanty, "A Survey of Mobility Management in Next-Generation All-IP-Based Wireless Systems," *IEEE Wireless Communications*, vol. 11, no. 4, pp. 16-28, August 2004.

[8] IEEE Std 802.11r-2008: IEEE standard for local and metropolitan area networks – Part 11 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 2: Fast Basic Service Set (BSS) Transition, IEEE Standards Association, July 2008.

[9] IEEE Std 802.16-2009: IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, IEEE Standards Association, May 2009.

[10] S. Pack, J. Choi, T. Kwon, and Y. Choi, "Fast-Handoff Support in IEEE 802.11 Wireless Networks," *IEEE Communications Surveys & Tutorials*, vol. 9, no. 1, pp. 2-12, 1st Quarter, 2007.

[11] H. Yokota, A. Idoue, T. Hasegawa, and T. Kato, "Link Layer Assisted Mobile IP Fast Handoff Method over Wireless LAN Networks," Proceedings or ACM MOBICOM 2002, pp. 131-139, September 2002.

[12] IEEE Std 802.21-2008: IEEE standard for local and metropolitan area networks – Part 21: Media Independent Handover Services, IEEE Standards Association, January 2009.

[13] Z. Zhu, R. Wakikawa, and L. Zhang, "A Survey of Mobility Support in the Internet," IETF draft-zhu-mobility-survey-03.txt, June 2010, work in progress.

[14] A. G. Valko, "Cellular IP: A New Approach to Internet Host Mobility," *ACM SIGCOMM Computer Communication Review*, 1999.

[15] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S.-Y. Wang, and T. La Porta, "HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless

Networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396-410, June 2002.
[16] A. Grilo, P. Estrela, and M. Nunes, "Terminal Independent Mobility for IP (TIMIP)," *IEEE Communications Magazine*, vol. 39, no. 12, pp. 34-41, December 2001.
[17] H. Soliman, C. Castelluccia, K. ElMalki, and L. Bellier, "Hierarchical Mobile Ipv6 (HMIPv6) Mobility Management, IETF RFC 5380, October 2008.
[18] A. C. Snoeren and H. Balakrishnan, "An End-to-end Approach to Host Mobility," *Proceedings of ACM International Conference on Mobile Computing and Networking (MOBICOM 2000)*, August 2000, pp. 155-166.
[19] M. Riegel and M. Tuexen, "Mobile SCTP," IETF draft-riegel-tuexen-mobile-sctp-09.txt, November 2007, expired.
[20] A. Bakre and B. Badrinath, "I-TCP: Indirect TCP for mobile Hosts," *Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS'95)*, pp. 136-143, 1995.
[21] D. A. Maltz and P. Bhagwat, "MSOCKS: An Architecture for Transport Layer Mobility," *Proceedings of IEEE INFOCOM'98*, vol. 3, pp. 1037-1045, 1998.
[22] J. Rosemberg, H. Schulzrinne, G. Camarillo, A. Johnston, R. Sparks, A. Handley, and E. Schooler, "SIP: Session Initiation Protocol," IETF RFC 3261, June 2002.
[23] H. Schulzrinne and E. Wedlund, "Application-layer Mobility Using SIP," *Mobile Computing and Communication Review*, vol. 4, no. 3, pp. 47-57, July 2000.
[24] S. Giordano, D. Lenzarini, A. Puiatti, and S. Vanini, "WiSwitch: Seamless Handover between Multi-provider Networks," *Proceedings of Second Annual Conference on Wireless On-demand Network Systems and Services (WONS 2005)*, pp. 224-235, January, 2005.
[25] E. Fogelstroem, A. Jonsson, and C. Perkins, "Mobile IPv4 Regional Registration," IETF RFC 4857, June 2007.
[26] R. Koodli, "Fast Handover for Mobile IPv6," IETF RFC 4068, July 2005.
[27] R. Koodli and C. Perkins, "Mobile IPv4 Fast Handovers," IETF RFC 4988, October 2007.
[28] G. Tsirtsis, V. Park, and H. Soliman, "Dual-Stack Mobile IPv4," IETF RFC 5454, March 2009.
[29] H. Soliman, Ed., "Mobile IPv6 Support for Dual Stack Hosts and Routers," IETF RFC 5555, June 2009.
[30] H. Yokota, K. Chowdhury, R. Koodli, B. Patil, and F. Xia, "Fast Handovers for Proxy Mobile IPv6," IETF RFC 5949, September 2010.
[31] Third Generation Partnership Project (3GPP), http://www.3gpp.org .
[32] P. Bertin, S. Bonjour, and J-M Bonnin, "A Distributed Dynamic Mobility Management Scheme Designed for Flat IP Architectures," *Proceedings of 3rd International Conference on New Technologies, Mobility and Security, (NTMS 2008)*.
[33] P. Bertin, S. Bonjour, and J-M Bonnin, "Distributed or Centralized Mobility?" *Proceedings of Global Communications Conference (GlobeCom 2009)*, Honolulu, Hawaii, 30 Nov – 4 Dec 2009.
[34] WiMAX Forum, http://www.wimaxforum.org
[35] R. Wakikawa, G. Valadon, and J. Murai, "Migrating Home Agents Towards Internet-scale Mobility Deployments," *Proceedings of the ACM 2nd CoNEXT Conference on Future Networking Technologies*, Lisboa, Portugal. 4-7 December 2006.
[36] H. Chan, "Proxy Mobile IP with Distributed Mobility Anchors," GlobeCom 2010 Workshop on Seamless Wireless Mobility, Miami, USA, 6-10 December 2010.
[37] M. Fisher, F.U. Anderson, A. Kopsel, G. Schafer, and M. Schlager, "A Distributed IP Mobility Approach for 3G SAE," 19th International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC 2008).
[38] C. Perkins and D. Johnson, "Route Optimization for Mobile IP," *Cluster Computing*, vol. 1, no. 2, pp. 161-176, (1998).
[39] A. Barbir, B. Cain, R. Nair, and O. Spatscheck, "Known Content Network (CN) Request-Routing Mechanisms," IETF RFC 3568, July 2003.
[40] J. Arkko, C. Vogt, and W. Haddad, "Enhanced Route Optimization for Mobile IPv6," IETF RFC 4866, May 2007.
[41] G. Kirby, "Locating the User," *Communication International*, 1995.
[42] M. Chang, H. Lee, M. Lee: "A Per-application Mobility Management Platform for Application-specific Handover Decision in Overlay Networks," *Computer Networks*, vol 53-11, July 2009, pp. 1846-1858.
[43] M. Chang, M. Lee, H. Lee: "Per-Application Mobility Management with Cross-Layer Based Performance Enhancement," *IEEE WCNC 2008*., March 31 2008-April 3 2008, pp 2822–2827.
[44] S. Salsano, C. Mingardi, S. Niccolini, A. Polidoro, L. Veltri "SIP-based Mobility Management in Next Generation Networks", *IEEE Wireless Communication*, vol. 15-2, April 2008.
[45] M. Bonola, S. Salsano, A. Polidoro, "UPMT: Universal Per-application Mobility management using Tunnels," *Proceedings of Global Communications Conference (GlobeCom 2009)*, Honolulu, Hawaii, 30 Nov – 4 Dec 2009.
[46] 3GPP TS 23.829: GPP Local IP Access and Selected IP Traffic Offload, September 2010, http://www.3gpp.org .
[47] T. Chiba and H. Yokota, "Efficient Route Optimization Methods for Femtocell-based All IP Networks," *Proceedings of the 5th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2009)*, Marrakech, Morocco, 12 - 14 October, 2009.
[48] L. Zhang, R. Wakikawa, and Z. Zhu, "Support Mobility in the Global Internet," *Proceedings of ACM Workshop on MICNET*, MobiCom 2009, Beijing, China, 21 September 2009.

H. Anthony Chan received his PhD in physics at Univ. of Maryland, College Park in 1982 and then continued post-doctorate research there in basic science.

After joining the former AT&T Bell Labs in 1986, his work moved to industry and manufacturing oriented research in areas of electronic packaging and reliability, and then moved again to network management, network architecture and standards. He was visiting Endowed Pinson Chair Professor in Networking at San Jose State University during 2001-2003 and was professor at University of Cape Town during 2004-2007. His current research in Huawei Technologies is in emerging broadband wireless network technologies, and he is contributing to IEEE 802.21 and IETF standards activities.

Dr. Chan is a Fellow of IEEE and is a distinguished speaker of IEEE CPMT Society and of IEEE Reliability Society.

Hidetoshi Yokota received his B.E., M.E. and Ph.D. degrees from Waseda University, Tokyo, in 1990, 1992, and 2003, respectively.

He started working for KDDI R&D Laboratories, Japan, in 1992. From 1995 to 1996 he was with SRI International, in Menlo Park, CA as an International Fellow. He is currently senior manager of Mobile Network Laboratory in KDDI R&D Labs. He is actively involved with IETF and has published several RFCs on fast handovers for Mobile IP. His current research interests include mobile communications and network virtualization.

Dr. Yokota is a member of IEEE as well as a member of IEICE and IPSJ Japan.

Jiang Xie received her B.E. degree from Tsinghua University, Beijing, China, in 1997, M.Phil. degree from Hong Kong University of Science and Technology in 1999, and M.S. and Ph.D. degrees from Georgia Institute of Technology, Atlanta, Georgia, in 2002 and 2004, respectively, all in electrical and computer engineering.

She joined the Department of Electrical and Computer Engineering at the University of North Carolina at Charlotte as an Assistant Professor in August 2004. Currently, she is an Associate Professor. Her current research interests include resource and mobility management in wireless networks, QoS provisioning, and the next-generation Internet.

Dr. Xie is on the Editorial Boards of *IEEE Communications Surveys& Tutorial, Computer Networks (Elsevier), Journal of Network and Computer Applications (Elsevier), and Journal of Communications (Academy Publisher)*. She has served as a Symposium Co-Chair for the Wireless Networking Symposium of IEEE GLOBECOM 2009 and 2010 conferences. Dr. Xie received an NSF Faculty Early Career Development (CAREER) Award in 2010, a Best Paper Award from IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2010), an Outstanding Leadership Award from IEEE GLOBECOM 2010, a Best Symposium Co-Chair Award from IEEE GLOBECOM 2009, and a Lee College of Engineering Graduate Teaching Excellence Award from UNC-Charlotte in 2007. She is a senior member of IEEE and a member of ACM.

Pierrick Seïté, received the Ph.D in electronic and digital communications from the university of Metz (France) in 1995.

He joined France Telecom in 2001 being in charge of architecture studies dealing with mobility management over heterogeneous access systems. He is currently focusing on network-controlled mobility management and convergent fixed/mobile network architectures. He was involved in European collaborative projects (IST/CISMUNDUS and IST/OVERDRIVE) and is now contributing to CELTIC/MEVICO. He is also involved in standardization bodies; among them is the IETF where he is contributing to mobility or multiple interfaces terminals related working groups (NetExt, MEXT, MultiMob and MIF).

Dapeng Liu received his master degree in communication system at Bejing Jiaotong University, Beijing, China, in 2006.

He joined Institute of Computing Technology, Chinese Academic of Sciences in 2006 and involved in projects related to Internet routing and IPv4/IPv6 transition technology. He joined Research Institute of China Mobile as project manager in 2008. His current research interests include mobile Internet architecture, mobility management, IPv4/IPv6 transition etc. He has been contributing to IETF and IEEE 802 standards activities since 2008.