Flat Architectures: Towards Scalable Future Internet Mobility

László Bokor, Zoltán Faigl, and Sándor Imre

Budapest University of Technology and Economics, Department of Telecommunications
Mobile Communication and Computing Laboratory – Mobile Innovation Centre
Magyar Tudosok krt. 2, H-1117, Budapest Hungary
{goodzi, szlaj, imre}@mcl.hu

Abstract. This chapter is committed to give a comprehensive overview of the scalability problems of mobile Internet nowadays and to show how the concept of flat and ultra flat architectures emerges due to its suitability and applicability for the future Internet. It also aims to introduce the basic ideas and the main paradigms behind the different flat networking approaches trying to cope with the continuously growing traffic demands. The discussion of the above areas will guide the readers from the basics of flat mobile Internet architectures to the paradigm's complex feature set and power creating a novel Internet architecture for future mobile communications.

Keywords: mobile traffic evolution, network scalability, flat architectures, mobile Internet, IP mobility, distributed and dynamic mobility management

1 Introduction

Mobile Internet has recently started to become a reality for both users and operators thanks to the success of novel, extremely practical smartphones, portable computers with easy-to-use 3G USB modems and attractive business models. Based on the current trends in telecommunications, vendors prognosticate that mobile networks will suffer an immense traffic explosion in the packet switched domain up to year 2020 [1–4]. In order to accommodate the future Internet to the anticipated traffic demands, technologies applied in the radio access and core networks must become scalable to advanced future use cases.

There are many existing solutions aiming to handle the capacity problems of current mobile Internet architectures caused by the mobile traffic data evolution. Reserving additional spectrum resources is the most straightforward approach for increasing the throughput of the radio access, and also spectrum efficiency can be enhanced thanks to new wireless techniques (e.g., High Speed Packet Access, and Long Term Evolution). Heterogeneous systems providing densification and offload of the macrocellular network throughout pico, femtocells and relays or WiFi/WiMAX interfaces also extend the radio range. However, the deployment of novel technologies providing higher radio throughput (i.e., higher possible traffic rates) easily generates new

J. Domingue et al. (Eds.): Future Internet Assembly, LNCS 6656, pp. 35–50, 2011.

usages and the traffic increase may still accelerate. Since today's mobile Internet architectures have been originally designed for voice services and later extended to support packet switched services only in a very centralized manner, the management of this ever growing traffic demand is quite hard task to deal with. The challenge is even harder if we consider fixed/mobile convergent architectures managing mobile customers by balancing user traffic between a large variety of access networks. Scalability of traffic, network and mobility management functions has become one of the most important questions of the future Internet.

The growing number of mobile users, the increasing traffic volume, the complexity of mobility scenarios, and the development of new and innovative IP-based applications require network architectures able to deliver all kind of traffic demands seamlessly assuring high end-to-end quality of service. However, the strongly centralized nature of current and planned mobile Internet standards (e.g., the ones maintained by the IETF or by the collaboration of 3GPP) prevents cost effective system scaling for the novel traffic demands. Aiming to solve the burning problems of scalability from an architectural point of view, flat and fully distributed mobile architectures are gaining more and more attention today.

The goal of this chapter is to provide a detailed introduction to the nowadays emerging scalability problems of the mobile Internet and also to present a state of the art overview of the evolution of flat and ultra flat mobile communication systems. In order to achieve this we first introduce the issues relating to the continuously growing traffic load inside the networks of mobile Internet providers in Section 2. Then, in Section 3 we present the main evolutionary steps of flat architectures by bringing forward the most important schemes, methods, techniques and developments available in the literature. This is followed, in Section 4, by an introduction of distributed mobility management schemes which can be considered as the most essential building block of flat mobile communications. As a conclusion we summarize the benefits and challenges concerning flat and distributed architectures in Section 5.

2 Traffic Evolution Characteristics and Scalability Problems of the Mobile Internet

2.1 Traffic Evolution Characteristics of the Mobile Internet

One of the most important reasons of the traffic volume increase in mobile telecommunications is demographical. According to the current courses, world's population is growing at a rate of 1.2 % annually, and the total population is expected to be 7.6 billion in year 2020. This trend also implies a net addition of 77 million new inhabitants per year [5]. Today, over 25% of the global population – this means about two billion people – are using the Internet. Over 60% of the global population – now we are talking about five billion people – are subscribers of some mobile communication service [1][6]. Additionally, the number of wireless broadband subscriptions is about to exceed the total amount of fixed broadband subscriptions and this development

becomes even more significant considering that the volume of fixed broadband subscriptions is gathering much slower.

The expansion of wireless broadband subscribers not only inflates the volume of mobile traffic directly, but also facilitates the growth in broadband wireless enabled terminals. However, more and more devices enable mobile access to the Internet, only a limited part of users is attracted or open to pay for the wireless Internet services meaning that voice communication will remain the dominant mobile application also in the future. Despite this and the assumption of [5] implying that the increase in the number of people potentially using mobile Internet services will likely saturate after 2015 in industrialized countries, the mobile Internet subscription growth potential will be kept high globally by two main factors. On one hand the growth of subscribers continues unbrokenly in the developing markets: mobile broadband access through basic handhelds will be the only access to the Internet for many people in Asia/Pacific. On the other hand access device, application and service evolution is also expected to sustain the capability of subscriber growth.

The most prominent effect of services and application evolution is the increase of video traffic: it is foreseen that due to the development of data-hungry entertainment services like television/radio broadcasting and VoD, 66% of mobile traffic will be video by 2014 [2]. A significant amount of this data volume will be produced by mobile Web-browsing which is expected to become the biggest source of mobile video traffic (e.g., YouTube). Cisco also forecasts that the total volume of video (including IPTV, VoD, P2P streaming, interactive video, etc.) will reach almost 90 percent of all consumer traffic (fixed and mobile) by the year 2012, producing a substantial increase of the overall mobile traffic of more than 200% each year [7]. Video traffic is also anticipated to grow so drastically in the forthcoming years that it could overstep Peer-to-Peer (P2P) traffic [4]. Emerging web technologies (such as HTML5), the increasing video quality requirements (HDTV, 3D, SHV) and special application areas (virtual reality experience sharing and gaming) will further boost this process and set new challenges to mobile networks. Since video and related entertainment services seems to become dominant in terms of bandwidth usage, special optimization mechanisms focusing on content delivery will also appear in the near future. The supposed evolution of Content Delivery Networking (CDN) and smart data caching technologies might have further impact on the traffic characteristics and obviously on mobile architectures.

Another important segment of mobile application and service evolution is social networking. As devices, networks and modes of communications evolve, users will choose from a growing scale of services to communicate (e.g., e-mail, Instant Messaging, blogging, micro-blogging, VoIP and video transmissions, etc.). In the future, social networking might evolve even further, like to cover broader areas of personal communication in a more integrated way, or to put online gaming on the next level deeply impregnated with social networking and virtual reality.

Even though video seems to be a major force behind the current traffic growth of the mobile Internet, there is another emerging form of communications called M2M (Machine-to-Machine) which has the potential to become the leading traffic contributor in the future. M2M sessions accommodate end-to-end communicating devices

without human intervention for remote controlling, monitoring and measuring, road safety, security/identity checking, video surveillance, etc. Predictions state that there will be 225 million cellular M2M devices by 2014 with little traffic per node but resulting significant growth in total, mostly in uplink direction [3]. The huge number of sessions with tiny packets creates a big challenge for the operators. Central network functions may not be as scalable as needed by the increasing number of sessions in the packet-switched domain.

As a summary we can state that the inevitable mobile traffic evolution is foreseen thanks to the following main factors: growth of the mobile subscriptions, evolution of mobile networks, devices, applications and services, and significant device increase potential resulted by the tremendous number of novel subscriptions for Machine-to-Machine communications.

2.2 Scalability Problems of the Mobile Internet

Existing wireless telecommunication infrastructures are not prepared to handle this traffic increase, current mobile Internet was not designed with such requirements in mind: mobile architectures under standardization (e.g., 3GPP, 3GPP2, WiMAX Forum) follow a centralized approach which cannot scale well to the changing traffic conditions.

On one hand user plane scalability issues are foreseen for anchor-based mobile Internet architectures, where mechanisms of IP address allocation and tunnel establishment for end devices are managed by high level network elements, called anchor points (GGSN in 3GPP UMTS, PDN GW in SAE, and CSN for WiMAX networks). Each anchor point maintains special units of information called contexts, containing binding identity, tunnel identifier, required QoS, etc. on a per mobile node basis. These contexts are continuously updated and used to filter and route user traffic by the anchor point(s) towards the end terminals and vice versa. However, network elements (hence anchor points too) are limited in terms of simultaneous active contexts. Therefore, in case of traffic increase new equipments should be installed or existing ones should be upgraded with more capacity.

On the other hand, scalability issues are also foreseen on the control plane. The well established approach of separating service layer and access layer provides easy service convergence in current mobile Internet architectures but introduces additional complexity regarding session establishment procedures. Since service and access network levels are decomposed, special schemes have been introduced (e.g., Policy and Charging Control architecture by 3GPP) to achieve interaction between the two levels during session establishment, modification and release routines. PCC and similar schemes ensure that the bearer established on the access network uses the resources corresponding to the session negotiated at the service level and allowed by the operator policy and user subscription. Due to the number of standardized interfaces (e.g., towards IP Multimedia Subsystem for delivering IP multimedia services), the interoperability between the service and the access layer can easily cause scalability and QoS issues even in the control plane.

As a consequence, architectural changes are required for dealing with the ongoing traffic evolution: future mobile networks must specify architecture optimized to maximize the end-user experience, minimize CAPEX/OPEX, energy efficiency, network performance, and to ensure mobile networks sustainability.

3 Evolution of Flat Architectures

3.1 Evolution of the Architecture of 3GPP Mobile Networks

Fixed networks were firstly subject to similar scalability problems. The evolution of DSL access architecture has shown in the past that pushing IP routing and other functions from the core to the edge of the network results in sustainable network infrastructure. The same evolution was started to happen within the wireless telecommunication and mobile Internet era.

The 3GPP network architecture specifications having the numbers 03.02 [8] and 23.002 [9] show the evolution of the 3GPP network from GSM Phase 1 published in 1995 until the Evolved Packet System (EPS) specified in Release 8 in 2010. The core part of EPS called Evolved Packet Core (EPC) is continuously extended with new features in Release 10 and 11. The main steps of the architecture evolution are summarized in the followings. Fig. 1 illustrates the evolution steps of the packet-switched domain, including the main user plane anchors in the RAN and the CN.

In Phase 1 (1995) the basic elements of the GSM architecture have been defined. The reasons behind the hierarchization and centralization of the GSM architecture were both technical and economical. Primarily it offloaded the switching equipments (cross-bar switch or MSC). In parallel, existing ISDN switches could be re-used as MSCs only if special voice encoding entities were introduced below the MSCs, hence further strengthening the hierarchical structure of the network. However, with the introduction of the packet-switched domain (PS) and the expansion of the PS traffic the drawbacks of this paradigm started to appear very early.

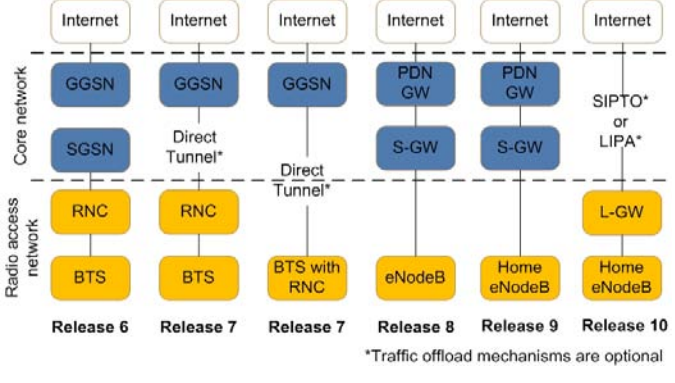

Fig. 1. The evolution of the packet-switched domain of the 3GPP architecture, including the main user plane anchors in the RAN and the CN.

The main driver to introduce packet-switching was that it allowed multiplexing hence resources could be utilized in a greater extent. In Phase 2+ (1997) the PS domain is described, hence centralized General Packet Radio Service (GPRS) support nodes are added to the network. Release 1999 (2002) describes the well known UMTS architecture clearly separating the CS and PS domains. Seeing that UMTS was designed to be the successor of GSM, it is not strange that the central anchors remained in place in 3G and beyond.

Progress of mobile and wireless communication systems introduced some fundamental changes. The most drastic among them is that IP has become the unique access protocol for data networks and the continuously increasing future wireless traffic is also based on packet data (i.e., Internet communication). Due to the collateral effects of this change a convergence procedure started to introduce IP-based transport technology in the core and backhaul network: Release 4 (2003) specified the Media gateway function, Release 5 (2003) introduced the IP Multimedia Subsystem (IMS) core network functions for provision of IP services over the PS domain, while Release 6 standardized WLAN interworking and Multimedia Broadcast Multicast Service (MBMS).

With the increasing IP-based data traffic flattening hierarchical and centralized functions became the main driving force in the evolution of 3GPP network architectures. Release 7 (also called Internet HSPA, 2008) supports the integration of the RNC with the NodeB providing a one node based radio access network. Another architectural enhancement of this release is the elaboration of Direct Tunnel service [10][11]. Direct Tunnel allows to offload user traffic from SGSN by bypassing it. The Direct Tunnel enabled SGSNs can initiate the reactivation of the PDP context to tunnel user traffic directly from the RNC to the GGSN or to the Serving GW introduced in Release 8. This mechanism tries to reduce the number of user-plane traffic anchors. However it also adds complexity in charging inter-PS traffic because SGSNs can not account the traffic passing in direct tunnels. When Direct Tunnel is enabled, SGSNs still handle signaling traffic, i.e., keep track of the location of mobile devices and participate in GTP signaling between the GGSN and RNC.

Release 8 (2010) introduces a new PS domain, i.e., the Evolved Packet Core (EPC). Compared to four main GPRS PS domain entities of Release 6, i.e. the base station (called NodeB), RNC, SGSN and GGSN, this architecture has one integrated radio access node, containing the precious base station and the radio network control functions, and three main functional entities in the core, i.e. the Mobility Management Entity (MME), the Serving GW (S-GW) and the Packet data Network GW (PDN GW).

Release 9 (2010) introduces the definition of Home (e)NodeB Subsystem. These systems allow unmanaged deployment of femtocells at indoor sites, providing almost perfect broadband radio coverage in residential and working areas, and offloading the managed, pre-panned macro-cell network [14].

In Release 10 (2010) Selective IP Traffic Offload (SIPTO) and Local IP Access (LIPA) services have been published [15]. These enable local breakout of certain IP traffic from the macro-cellular network or the H(e)NodeB subsystems, in order to offload the network elements in the PS and EPC PS domain. The LIPA function enables an IP capable UE connected via Home(e)NodeB to access other IP capable

entities in the same residential/enterprise IP network without the user plane traversing the core network entities. SIPTO enables per APN and/or per IP flow class based traffic offload towards a defined IP network close to the UE's point of attachment to the access network. In order to avoid SGSN/S-GW from the path, Direct Tunnel mode should be used.

The above evolutionary steps resulted in that radio access networks of 3GPP became flattened to one single serving node (i.e., the eNodeB), and helped the distribution of previous centralized RNC functions. However, the flat nature of LTE and LTE-A architectures concerns only the control plane but not the user plane: LTE is linked to the Evolved Packet Core (EPC) in the 3GPP system evolution, and in EPC, the main packet switched core network functional entities are still remaining centralized, keeping user IP traffic anchored. There are several schemes to eliminate the residual centralization and further extend 3GPP.

3.2 Ultra Flat Architecture

One of the most important schemes aiming to further extend 3GPP standards is the Ultra Flat Architecture (UFA) [16–20]. Authors present and evaluate an almost green field approach which is a flat and distributed convergent architecture, with the exception of certain control functions still provided by the core. UFA represents the ultimate step toward flattening IP-based core networks, e.g., the EPC in 3GPP. The objective of UFA design is to distribute core functions into single nodes at the edge of the network, e.g., the base stations. The intelligent nodes at the edge of the network are called UFA gateways. Fig. 2 illustrates the UFA with HIP and PMIP-based mobility control.

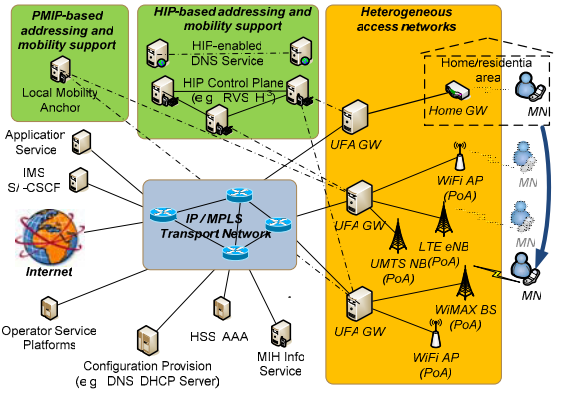

Fig. 2. The Ultra Flat Architecture with HIP and PMIP-based mobility control

Since mobility introduces frequent IP-level handovers a Session Initialization Protocol (SIP) based handover procedure has been described in [16]. It has been shown by a numerical analysis, and in a later publication with measurements on a testbed [17] that seamless handovers can be guaranteed for SIP-based applications. SIP Back-to-

Back User Agents (B2BUAs) in UFA GWs can prepare for fast handovers by communicating the necessary contexts, e.g., the new IP address before physical handover. This scheme supports both mobile node (MN) and network decided handovers.

In the PS domain, IP multimedia services require a two-level session establishment procedure. First, the MN and the correspondent node (CN) negotiate the session parameters using SIP on the service level, then the Policy and Charging Control (PCRF), ensures that the bearer established in the access layer uses the resources corresponding to the negotiated session. The problem is that service level is not directly notified about access layer resource problems, and, e.g., it is difficult to adapt different application session components of the same service to the available resources in the access layer. In order to solve this problem, a novel SIP-based session establishment and session update procedure is introduced in [16] for the UFA.

Interworking with Internet applications based on non SIP control protocol is a technical challenge for mobile operators. One of their aims is to provide seamless handovers for any application. IP-mobility control can be provided by protocols below the application layer. A Mobile IPv6 and a Host Identity Protocol (HIP) based signaling scheme alternative has been introduced for UFA by Z. Faigl et al. [18]. L. Bokor et al. describe a new HIP extension service which enables signaling delegation [19]. This service is applied in HIP-based handover and session establishment procedures of UFA, to reduce the number of HIP Base Exchanges in the access and core network, and to enable delegation of HIP-level signaling of the MN by the UFA GWs. Moreover, a new cross-layer access authorization mechanism for L2 and HIP has been introduced, to replace certificate-based access authorization with a more lightweight access authorization. In [20] authors clearly define the terminal attachment, session establishment and handover procedures, further enhance the original idea by providing two integrated UFA schemes (i.e., SIP–IEEE 802.21–HIP and SIP–IEEE 802.21–PMIP) and analyze the suitability of the two solutions using the Multiplicative Analytic Hierarchy Process.

4 Distributed Mobility Management in Flat Architectures

4.1 Motivations for Distributing Mobility Functions

Flat mobile networks not only require novel architectural design paradigms, special network nodes and proprietary elements with peculiar functions, but also demand certain, distinctive mobility management schemes sufficiently adapted to the distributed architecture. In fact the distributed mobility management mechanisms and the relating decision methods, information, command and event services form the key routines of the future mobile Internet designs. The importance of this research area is also emphasized by the creation of a new IETF non-working group called Distributed Mobility Management (DMM) in August 2010, aiming to extend current IP mobility solutions for flat network architectures.

Current mobility management solutions rely on hierarchical and centralized architectures which employ anchor nodes for mobility signaling and user traffic forwarding. In 3G UMTS architectures centralized and hierarchical mobility anchors are

implemented by the RNC, SGSN and GGSN nodes that handle traffic forwarding tasks using the apparatus of GPRS Tunneling Protocol (GTP). The similar centralization is noticeable in Mobile IP (MIP) [21] where the Home Agent –an anchor node for both signaling and user plane traffic– administers mobile terminals' location information, and tunnels user traffic towards the mobile's current locations and vice versa. Several enhancements and extensions such as Fast Handoffs for Mobile IPv6 (FMIP) [22], Hierarchical Mobile IPv6 (HMIP) [23], Multiple Care-of Addresses Registration [24], Network Mobility (NEMO) Basic Support [25], Dual-Stack Mobile IPv6 [26], and Proxy Mobile IPv6 (PMIP) [27], were proposed to optimize the performance and broaden the capabilities of Mobile IP, but all of them preserve the centralized and anchoring nature of the original scheme.

There are also alternate schemes in the literature aiming to integrate IP-based mobility protocols into cellular architectures and to effectively manage heterogeneous networks with special mobility scenarios. Cellular IP [28] introduces a gateway router dealing with local mobility management while also supporting a number of handoff techniques and paging. A similar approach is the handoff-aware wireless access Internet infrastructure (HAWAII) [29], which is a separate routing protocol to handle micromobility. Terminal Independent Mobility for IP [30] combines some advantages from Cellular IP and HAWAII, where terminals with legacy IP stacks have the same degree of mobility as terminals with mobility-aware IP stacks. Authors of [31] present a framework that integrates 802.21 Media Independent Handover [32] and Mobile IP for network driven mobility. However, these proposals are also based on centralized functions and generally rely on MIP or similar anchoring schemes.

Some of the above solutions are already standardized [12][13][33] for 3G and beyond 3G architectures where the introduced architectural evolution is in progress: E-UTRAN (Evolved Universal Terrestrial Radio Access Network) or LTE (Long Term Evolution) base stations (eNodeBs) became distributed in a flatter scheme allowing almost complete distribution of radio and handover control mechanisms together with direct logical interfaces for inter-eNodeB communications. Here, traffic forwarding between neighboring eNodeBs is temporarily allowed during handover events providing intra-domain mobility. However, traffic forwarding and inter-gateway mobility operations remain centralized thanks to S-GW, PDN-GW, Local Mobility Anchor and Home Agent, responsible for maintaining and switching centralized, hierarchical and overlapping system of tunnels towards mobile nodes. Also, offloading with LIPTO and SIPA extensions cannot completely solve this issue: mobility management mechanisms in current wireless and mobile networks anchor the user traffic relatively far from users' location. This results in centralized, unscalable data plane and control plane with non-optimal routes, overhead and high end-to-end packet delay even in case of motionless users, centralized context maintenance and single point of failures. Anchor-based traffic forwarding and mobility management solutions also cause deployment issues for caching contents near the user..

To solve all these problems and questions novel, distributed and dynamic mobility management approaches must be envisaged, applicable to intra- and inter-technology mobility cases as well.

4.2 Application Scenarios for DMM Schemes

The basic idea is that anchor nodes and mobility management functions of wireless and mobile systems could be distributed to multiple locations in different network segments, hence mobile nodes located in any of these locations could be served by a close entity.

A first alternative for achieving DMM is core-level distribution. In this case mobility anchors are topologically distributed and cover specific geographical area but still remain in the core network. A good example is the Global HA to HA protocol [34], which extends MIP and NEMO in order to remove their link layer dependencies on the Home Link and distribute the Home Agents in Layer 3, at the scale of the Internet. DIMA (Distributed IP Mobility Approach) [35] can also be considered as a core-level scheme by allowing the distribution of MIP Home Agent (the normally isolated central server) to many and less powerful interworking servers called Mobility Agents (MA). These new nodes have the combined functionality of a MIP Home Agent and HMIP/PMIP Mobility Anchor Points. The administration of the system of distributed MAs is done via a distributed Home Agent overlay table structure based on a Distributed Hash Table (DHT) [36]. It creates a virtual Home Agent cluster with distributed binding cache that maps a mobile node's permanent identifier to its temporary identifier.

A second alternative for DMM is when mobility functions and anchors are distributed in the access part of the network. For example in case of pico- and femto cellular access schemes it could be very effective to introduce Layer 3 capability in access nodes to handle IP mobility management and to provide higher level intervention and even cross-layer optimization mechanisms. The concept of UMTS Base Station Router (BSR) [37] realizes such an access-level mobility management distribution scheme where a special network element called BSR is used to build flat cellular systems. BSR merges the GGSN, SGSN, RNC and NodeB entities into a single element: while a common UMTS network is built from a plethora of network nodes and is maintained in a hierarchical and centralized fashion, the BSR integrates all radio access and core functions. Furthermore, the BSR can be considered a special wireless edge router that bridges between mobile/wireless and IP communication. In order to achieve this, mobility support in the BSR is handled at three layers: RF channel mobility, Layer 2 anchor mobility, and Layer 3 IP mobility. The idea of Liu Yu et al. [38] is quite similar to the BSR concept. Here a node called Access Gateway (AGW) is introduced to implement distributed mobility management functionalities at the access level. The whole flat architecture consists of two kinds of elements, AGW on the access network side and terminals on the user side. Core network nodes are mainly simple IP routers. The scheme applies DHT and Loc/ID separation: each mobile node has a unique identifier (ID) keeping persistent, and an IP address based locator (Loc) changed by every single mobility event. The (Loc,ID) pair of each mobile is stored inside AGW nodes and organized/managed using DHTs.

A third type of DMM application scenarios is the so-called host-level or peer-to-peer distributed mobility management where once the correspondent node is found, communicating peers can directly exchange IP packets. In order to find the correspondent node, a special information server is required in the network, which can also

be centralized or distributed. A good example for host-level schemes in the IP layer is MIPv6 which is able to bypass the user plane anchor (i.e., Home Agent) due to its route optimization mechanism, therefore providing a host-to-host communication method. End-to-end mobility management protocols working in higher layers of the TCP/IP stack such as Host Identity Protocol (HIP) [39], TCP-Migrate [40], MSOCKS [41], Stream Control Transmission Protocol (SCTP) [42], or Session Initiation Protocol (SIP) [43] can also be efficiently employed in such schemes.

4.3 Distribution Methods of Mobility Functions

Mobility management functions can be distributed in two main ways: partially and fully.

Partially distributed schemes can be implemented either by distinguishing signaling and user planes based on their differences in traffic volume or end-host behavior (i.e., only the user plane is distributed), or by granting mobility support only to nodes that actually need it (i.e., actually eventuate mobility event), hence achieving more advanced resource management. Note that these two approaches may also be combined.

Today's mobility management protocols (e.g., Mobile IP, NEMO BS and Proxy Mobile IP without route optimization) do not separate signaling and user planes which means that all control and data packets traverse the centralized or hierarchized mobility anchor. Since the volume of user plane traffic is much higher compared to the signaling traffic, the separation of signaling and user planes together with the distribution of the user plane but without eliminating signaling anchors can still result in effective and scalable mobility management. This is exploited by the HIP based UFA scheme [18–20] where a relatively simple inter-UFA GW protocol can be used thanks to the centralized HIP signaling plane, but the user plane is still fully distributed. Mobile IP based DMM solutions also rely on the advantages of this partial distribution concept when they implement route optimization, hence separate control packets from data messages after a short period of route optimization procedure.

The second type of partially distributed mobility management is based on the capability to turn off mobility signaling when such mechanisms are not needed. This so-called dynamic mobility management dynamically executes mobility functions only for mobile nodes that are actually subjected to handover event, and lack transport or application-layer mobility support. In such cases, thanks to the removal of unwanted mobility signaling, handover latency and control overhead can be significantly reduced. Integrating this concept with distributed anchors, the algorithms supporting dynamic mobility could also be distributed. Such integration is accomplished in [44][45] where authors introduce and evaluate a scheme to dynamically anchor mobile nodes' traffic in distributed Access Nodes (AN), depending on mobiles' actual location when sessions are getting set up. The solution's dynamic nature lies in the fact that sessions of mobile nodes are dynamically anchored on different ANs depending on the IP address used. Based on this behavior, the system is able to avoid execution of mobility management functions (e.g., traffic encapsulation) as long as a particular mobile node is not moving. The method is simultaneously dynamic and dis-

tributed, and because mobility functions are fully managed at the access level (by the ANs), it is appropriate for flat architectures. Similar considerations are applied in [46] for MIP, in [47] for HMIP and in [48] for PMIP. The MIP-based scheme introduces a special mode for the mobility usage in IP networks: for all the IP sessions opened and closed in the same IP sub-network no MIP functions will be executed even if the mobile node is away from its home network; standard MIP mechanisms will be used only for the ongoing communications while the mobile node is in motion between different IP sub-networks. The HMIP-based method proposes a strategy to evenly distribute the signaling burden and to dynamically adjust the micromobility domain (i.e., regional network) boundary according to real-time measurements of handover rates or traffic load in the networks. The PMIP-based solution discusses a possible deployment scheme of Proxy Mobile IP for flat architecture. This extension allows to dynamically distributing mobility functions among access routers: the mobility support is restricted to the access level, and adapted dynamically to the needs of mobile nodes by applying traffic redirection only to MNs' flows when an IP handover event occurs.

Fully distributed schemes bring complete distribution of mobility functions into effect (i.e., both data plane and control plane are distributed). This implies the introduction of special mechanisms in order to identify the anchor that manages mobility signaling and data forwarding of a particular mobile node, and in most cases this also requires the absolute distribution of mobility context database (e.g., for binding information) between every element of the distributed anchor system. Distributed Hash Table or anycast/broadcast/multicast communication can be used for the above purposes. In such schemes, usually all routing and signaling functions of mobility anchor nodes are integrated on the access level (like in [49]), but less flat architectures (e.g., by using Hi3 [50] for core-level distribution of HIP signaling plane) are also feasible.

5 Conclusion

Flat architectures infer high scalability because centralized anchors – the main performance bottlenecks – are removed, and traffic is forwarded in a distributed fashion. The flat nature also provides flexibility regarding the evolution of broadband access, e.g., the range extension of RANs with unmanaged micro-, pico- and femtocells, without concerns of capacity in centralized entities covering the actual area in a hierarchical structure.

In flat architectures, integrated and IP-enabled radio base station (BS) entities are directly connected to the IP core infrastructure. Therefore, they provide convenient and implicit interoperability between heterogeneous wireless technologies, and facilitate a convenient way of sharing the infrastructure for the operators. Flattening also infers the elimination of centralized components that are access technology specific. Thanks to the integrated, "single box" nature of these advanced base stations, the additional delay that user and signaling plane messages perceive over a hierarchical and multi-element access and core network (i.e., transmission and queuing delays to a central control node) are also reduced or even eliminated. This integrated design of

BS nodes also minimizes the feedback time of intermodule communication, i.e., signaling is handled as soon as it is received locally, on the edge of the operator's network, and enables to incorporate sophisticated cross-layer optimization schemes for performance improvements.

The application of general-purpose IP equipments produced in large quantities has economic advantages as well. In flat architectures the radio access network components could be much cheaper compared to HSPA and LTE devices today because of the economy of scale. Also operational costs can be reduced as a flat network has fewer integrated components, and lacks of hierarchical functions simultaneously influenced by management processes. The higher competition of network management tools due to the apparition of tools developed formerly for the Internet era may reduce the operational expenditures as well.

Failure tolerance/resistance, reliability and redundancy of networks also can be refined and strengthen by flat design schemes. Anchor and control nodes in hierarchical and centralized architectures are often single point of failures and their shortfall can easily cause serious breakdowns in large service areas. Within flat architectures no such single points of failure exist, and the impact of possible shortfalls of the distributed network elements (i.e., BSs) can smoothly narrowed to a limited, local area without complex failure recovery operations.

Another important benefit of flat architectures is the potential to prevent suboptimal routing situations and realize advanced resource efficiency. In a common hierarchical architecture, all traffic passes through the centralized anchor nodes, which likely increases the routing path and results in suboptimal traffic routing compared to the flat use-cases.

However, in order to exploit all the above benefits and advantages, some challenges that flat architectures face must be concerned.

In flat architectures, network management and configuration together with resource control must be done in a fully distributed and decentralized way. It means that self-configuration and self-optimization capabilities are to be introduced in the system. Closely related to self-optimization and self-configuration, self-diagnosis and self-healing is essential for continuous and reliable service provision in flat networking architectures. This is reasoned by the fact that IP equipments are more sensible to failures: due to lack of core controller entities base stations are no more managed centrally; hence failure diagnostics and recovery must be handled in a fully distributed and automated way. This is a great challenge but it comes with the benefits of scalability, fault tolerance and flexibility.

Optimization of handover performance is another key challenge for flat networks. Unlike in hierarchical and centralized architectures which usually provide efficient fast handover mechanisms using Layer 2 methods, in flat architectures IP-based mobility management protocol – with advanced micromobility extension – must be used. Since all the BSs are connected directly to the IP core network, hiding mobility events from the IP layer is much harder.

Last but not least Quality of Service provision is also an important challenge of flat architectures. This problem emerges because current QoS assurance mechanisms in the IP world require improvements to replace the Layer 2 QoS schemes of the tradi-

tional hierarchical and centralized mobile telecommunication architectures. The IP network that deals with the interconnection of base stations in flat networks must be able to assure different QoS levels (e.g., in means of bandwidth and delay) and manage resources for adequate application performance.

Based on the collected benefits and the actual challenges of flat architectures we can say that applying flat networking schemes together with distributed and dynamic mobility management is one of the most promising alternatives to change the current mobile Internet architecture for better adaptation to future needs.

Acknowledgments. This work was made in the frame of Mobile Innovation Centre's 'MEVICO.HU' project, supported by the National Office for Research and Technology (EUREKA_Hu_08-1-2009-0043) under the co-operation of the Celtic Call7 Project MEVICO.

References

1. UMTS Forum White Paper: Recognising the Promise of Mobile Broadband (June 2010)
2. Cisco VNI: Global Mobile Data Traffic Forecast Update, 2009-2014 (Feb. 2010)
3. Dohler, M., Watteyne, T., Alonso-Zárate, J.: Machine-to-Machine: An Emerging Communication Paradigm, Tutorial. In: GlobeCom'10 (Dec. 2010)
4. Schulze, H., Mochalski, K.: Ipoque, Internet Study 2008/2009, Ipoque (Jan. 2011)
5. UMTS Forum, REPORT NO 37, Magic Mobile Future 2010-2020 (April 2005)
6. International Telecommunication Union, Press Release: ITU sees 5 billion mobile subscriptions globally in 2010 (February 2010)
7. Cisco VNI: Hyperconnectivity and the Approaching Zettabyte Era (June 2010)
8. ETSI GTS GSM 03.02-v5.1.0: Digital cellular telecommunications system (Phase 2+) - Network architecture (GSM 03.02) (1996)
9. 3GPP TS 23.002: Network architecture, V10.1.1, Release 10 (Jan. 2011)
10. 3GPP TR 23.919: Direct Tunnel Deployment Guideline, Release 7, V1.0.0 (May 2007)
11. 3GPP TS 23.401: General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access, Rel.8, V8.12 (Dec. 2010)
12. 3GPP TS 29.275, Proxy Mobile IPv6 (PMIPv6) based Mobility and Tunneling protocols, Stage 3, Release 10, V10.0.0 (Dec. 2010)
13. 3GPP TS 24.303, Mobility management based on Dual-Stack Mobile IPv6, Stage 3, Release 10, V10.1.0 Dec (2010)
14. FemtoForum: Femtocells – Natural Solution for Offload – a Femto Forum brief (June 2010)
15. 3GPP TR 23.829: Local IP Access and Selected IP Traffic Offload, Release 10, V1.3 (2010)
16. Daoud, K., Herbelin, P., Crespi, N.: UFA: Ultra Flat Architecture for high bitrate services in mobile networks. In: Proc. of PIMRC'08, Cannes, France, pp. 1–6 (2008)
17. Daoud, K., Herbelin, P., Guillouard, K., Crespi, N.: Performance and Implementation of UFA: a SIP-based Ultra Flat Mobile Network Architecture. In: Proc. of PIMRC (Sep. 2009)
18. Faigl, Z., Bokor, L., Neves, P., Pereira, R., Daoud, K., Herbelin, P.: Evaluation and comparison of signaling protocol alternatives for the Ultra Flat Architecture, ICSNC, pp. 1–9 (2010)

19. Bokor, L., Faigl, Z., Imre, S.: A Delegation-based HIP Signaling Scheme for the Ultra Flat Architecture. In: Proc. of the 2nd IWSCN, Karlstad, Sweden, pp. 9–16 (2010)
20. Faigl, Z., Bokor, L., Neves, P., Daoud, K., Herbelin, P.: Evaluation of two integrated signalling schemes for the ultra flat architecture using SIP, IEEE 802.21, and HIP/PMIP protocols. In: Journal of Computer Networks (2011), doi:10.1016/j.comnet.2011.02.005
21. Johnson, D., Perkins, C., Arkko, J.: IP Mobility Support in IPv6, IETF RFC 3775 (2004)
22. Koodli, R. (ed.): Fast Handoffs for Mobile IPv6, IETF RFC 4068 (July 2005)
23. Soliman, H., Castelluccia, C., El Malki, K., Bellier, L.: Hierarchical Mobile IPv6 Mobility Management (HMIPv6), IETF RFC 4140 (Aug. 2005)
24. Wakikawa, R. (ed.): V. Devarapalli, G. Tsirtsis, T. Ernst, K. Nagami: Multiple Care-of Addresses Registration, IETF RFC 5648 (October 2009)
25. Devarapalli, V., Wakikawa, R., Petrescu, A., Thubert, P.: Network Mobility (NEMO) Basic Support Protocol, IETF RFC 3963 (Jan. 2005)
26. Soliman, H. (ed.): Mobile IPv6 Support for Dual Stack Hosts and Routers, IETF RFC 5555 (June 2009)
27. Gundavelli, S. (ed.): K. Leung, V. Devarapalli, K. Chowdhury, B. Patil: Proxy Mobile IPv6, IETF RFC 5213 (Aug. 2008)
28. Valko: Cellular IP: A New Approach to Internet Host Mobility, ACM SIGCOMM Comp. Commun. Rev., 29 (1), 50-65 (1999)
29. Ramjee, R., Porta, T.L., Thuel, S., Varadhan, K., Wang, S.: HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-area Wireless Networks. In: IEEE Int. Conf. Network Protocols (1999)
30. Grilo, A., Estrela, P., Nunes, M.: Terminal Independent Mobility for IP (TIMIP). IEEE Communications Magazine 39(12), 34–41 (2001)
31. Melia, T., de la Oliva, A., Vidal, A., Soto, I., Corujo, D., Aguiar, R.L.: Toward IP converged heterogeneous mobility: A network controlled approach. Com. Networks 51 (2007)
32. IEEE, IEEE Standard for Local and metropolitan area networks- Part 21: Media Independent Handover, IEEE Std 802.21-2008 (Jan. 2009)
33. 3GPP TS 23.402, Architecture enhancements for non-3GPP accesses, Rel.10,V10.2 (2011)
34. Thubert, P., Wakikawa, R., Devarapalli, V.: Global HA to HA protocol, IETF Internet-Draft, draft-thubert-nemo-global-haha-02.txt (Sept. 2006)
35. Fischer, M., Andersen, F.-U., Kopsel, A., Schafer, G., Schlager, M.: A Distributed IP Mobility Approach for 3G SAE. In: Proc. of 19th PIMRC, ISBN: 978-1-4244-2643-0 (Sept. 2008)
36. Farha, R., Khavari, K., Abji, N., Leon-Garcia, A.: Peer-to-peer mobility management for all-ip networks. In: Proc. of ICC ’06, V. 5, pp. 1946–1952 (June 2006)
37. Bauer, M., Bosch, P., Khrais, N., Samuel, L.G., Schefczik, P.: The UMTS base station router. Bell Labs Tech. Journal, I. 11(4), 93–111 (2007)
38. Liu Yu, Zhao Zhijun, Lin Tao, Tang Hui: Distributed mobility management based on flat network architecture. In: Proc. of 5th WICON, pp. 1-5, Singapore (2010)
39. Moskowitz, R., Nikander, P., Jokela, P. (eds.): T. Henderson: Host Identity Protocol, IETF RFC 5201 (April 2008)
40. Snoeren, A.C., Balakrishnan, H.: An End-to-End Approach to Host Mobility. In: Proc. of MobiCom (Aug. 2000)
41. Maltz, D., Bhagwat, P.: MSOCKS: An Architecture for Transport Layer Mobility. In: Proc. INFOCOM, pp. 1037-1045 (Mar 1998)
42. Stewart, R. (ed.): Stream Control Transmission Protocol, IETF RFC 4960 (Sept. 2007)
43. Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., Schooler, E.: SIP: Session Initiation Protocol, IETF RFC 3261 (June 2002)

44. Bertin, P., Bonjour, S., Bonnin, J.-M.: A Distributed Dynamic Mobility Management Scheme Designed for Flat IP Architectures. In: Proc. of NTMS '08, pp.1-5 (2008)
45. Bertin, P., Bonjour, S., Bonnin, J.: Distributed or centralized mobility? In: Proc. of the 28th IEEE conference on Global telecommunications (GLOBECOM'09), Honolulu, HI (2009)
46. Kassi-Lahlou, M., Jacquenet, C., Beloeil, L., Brouckaert, X.: Dynamic Mobile IP (DMI), IETF Internet-Draft, draft-kassi-mobileip-dmi-01.txt (Jan. 2003)
47. Song, M., Huang, J., Feng, R., Song, J.: A Distributed Dynamic Mobility Management Strategy for Mobile IP Networks. In: Proc. of 6th ITST, pp. 1045-1050 (June 2006)
48. Seite, P., Bertin, P.: Dynamic Mobility Anchoring, IETF Internet-Draft (May 2010)
49. Yan, Z., Lei, L., Chen, M.: WIISE - A Completely Flat and Distributed Architecture for Future Wireless Communication Systems, Wireless World Research Forum (Oct. 2008)
50. Gurtov, A., et al.: Hi3: An efficient and secure networking architecture for mobile hosts. Journal of Computer Communications 31(10), 2457–2467 (2008)