ITU Centres of Excellence for Europe

Next Generation Mobile and Wireless Networks

Module 1:

IMT-Advanced: the ITU concept for the Next Generation Mobile and Wireless Networks

Table of contents

1.1. ITU IMT-Advanced requirements for 4G 2
1.2. 3GPP LTE and LTE-Advanced technologies 9
1.3. IEEE wireless technologies (802.11n, 802.16m)...15
    1.3.1. IEEE 802.11n standard 16
    1.3.2. IEEE 802.16m standard 23
1.4. ITU NGN model and mobile networks...36
1.5. All-IP concept for 4G...48
1.6. QoS and mobility management concepts for 4G...54
1.7. Control and signaling protocols for 4G..66
    1.7.1. SIP 66
    1.7.2. DIAMETER 72
    1.7.3. Mobile IP...77
References 85

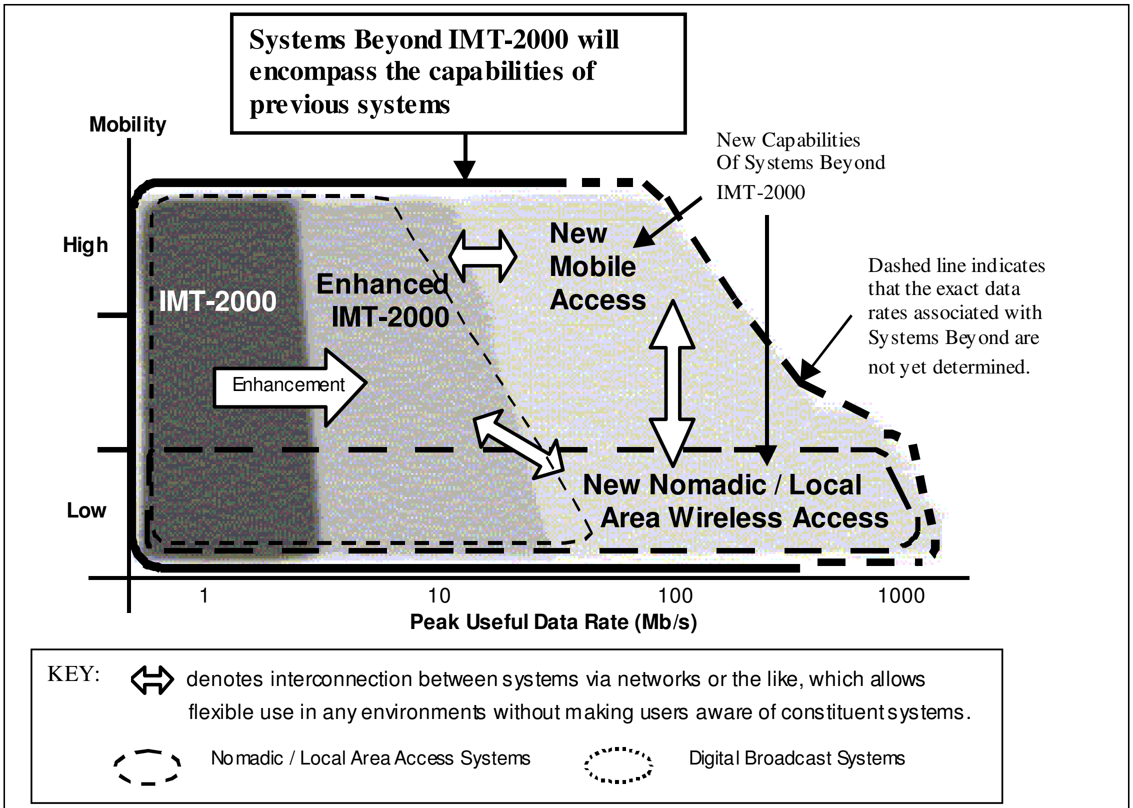
1.1. ITU IMT-Advanced requirements for 4G

The International Mobile Telecommunications-Advanced (IMT-Advanced) systems are mobile systems that include the new capabilities of IMT that go beyond those of IMT-2000. Such systems provide access to a wide range of telecommunication services including advanced mobile services, supported by mobile and fixed networks, which are increasingly packet-based.

IMT-Advanced systems support low to high mobility applications and a wide range of data rates in accordance with user and service demands in multiple user environments. IMT Advanced also has capabilities for high-quality multimedia applications within a wide range of services and platforms, providing a significant improvement in performance and Quality of Service (QoS). Moreover, the consumer demands will shape the future development of IMT-2000 and IMT Advanced. Recommendation ITU-R M.1645 describes these trends in detail, some of which include the growing demand for mobile services, increasing user expectations, and the evolving nature of the services and applications that may become available. Also, Report ITU-R M.2072 details the market analysis and forecast of the evolution of the mobile market and services for the future development of IMT-2000, IMT-Advanced and other systems. This Report provides forecasts for the year 2010, 2015, and 2020 timeframes.

IMT-2000 systems provide access to a wide range of telecommunication services, supported by the fixed telecommunication networks (e.g. PSTN/ISDN/IP), and to other services which are specific to mobile users. To meet the ever increasing demand for wireless communication (e.g. increased no. of users, higher data rates, video or gaming services which require increased quality of service, etc.), IMT-2000 has been, and continues to be, enhanced.

The Figure 1.1 is taken directly from Recommendation ITU-R M.1645 and reflects the terminology in use at the time of its adoption. Resolution ITU-R 56 defines the relationship between “IMT-2000”, the future development of IMT-2000 and “systems beyond IMT-2000” for which it also provides a new name: IMT-Advanced. Resolution ITU-R 56 resolves that the term IMT 2000 encompasses also its enhancements and future developments. The term “IMT Advanced” should be applied to those systems, system components, and related aspects that include new radio interface(s) that support the new capabilities of systems beyond IMT-2000. The term “IMT” is the root name that encompasses both IMT-2000 and IMT-Advanced collectively. In October 2010, only two technologies are accepted within the IMT-Advanced umbrella: **LTE-Advanced (**LTE Release 10 & beyond**) and Mobile WiMAX 2.0 (802.16m**, also known as WirelessMAN-Advanced**)**. Moreover, ITU IMT-Advanced defines the **4G** mobile networks.



Dark gray color indicates existing capabilities, medium gray color indicates enhancements to IMT-2000, and the lighter gray color indicates new capabilities of Systems Beyond IMT-2000.

The degree of mobility as used in this figure is described as follows: Low mobility covers pedestrian speed, and high mobility covers high speed on highways or fast trains (60 km/h to ~250 km/h, or more).

Figure 1.1. Relationship between IMT-2000 (3G LTE) and IMT-Advanced (4G).

For the last 20 years, ITU has been coordinating efforts of government and industry and private sector in the development of a global broadband multimedia international mobile telecommunication system, known as IMT. Since 2000, the world has seen the introduction of the first family of standards derived from the IMT concept. ITU estimates that worldwide mobile cellular subscribers are likely to reach the 4 billion mark before the end of this year of which IMT systems technology will constitute a substantial part considering that already in 2007 (during the ITU World Radiocommunication Conference (WRC-07) in Geneva), there were more than 1 billion IMT-2000 subscribers in the world. It is realised that by the year 2010 there are 1 700 million terrestrial mobile subscribers worldwide. And moreover, it is envisaged that, by the year 2020, potentially the whole population of the world could have access to advanced mobile communications devices, subject to, amongst other considerations, favourable cost structures being achieved. There are already more portable handsets than either fixed line telephones or fixed line equipment such as PCs that can access the Internet, and the number of mobile devices is expected to continue to grow more rapidly than fixed line devices. Mobile terminals will be the most commonly used devices for accessing and exchanging information. User expectations are continually increasing with regard to the variety of services and applications. In particular, users will expect a dynamic, continuing stream of new applications, capabilities and services that are ubiquitous and available across a range of devices using a single subscription and a single identity (number or address). Versatile communication systems offering customized and ubiquitous services based on diverse individual needs will require flexibility in the technology in order to satisfy multiple demands simultaneously.

Moreover, the multimedia traffic is increasing far more rapidly than speech, and will increasingly dominate traffic flows. There will be a corresponding change from predominantly circuit-switched to packet-based delivery. This change will provide the user with the ability to more efficiently receive multimedia services, including e mail, file transfers, messaging and distribution services. These services can be either symmetrical or asymmetrical, and real-time or non real-time. They can consume high bandwidths, resulting in higher data rate requirements in the future.

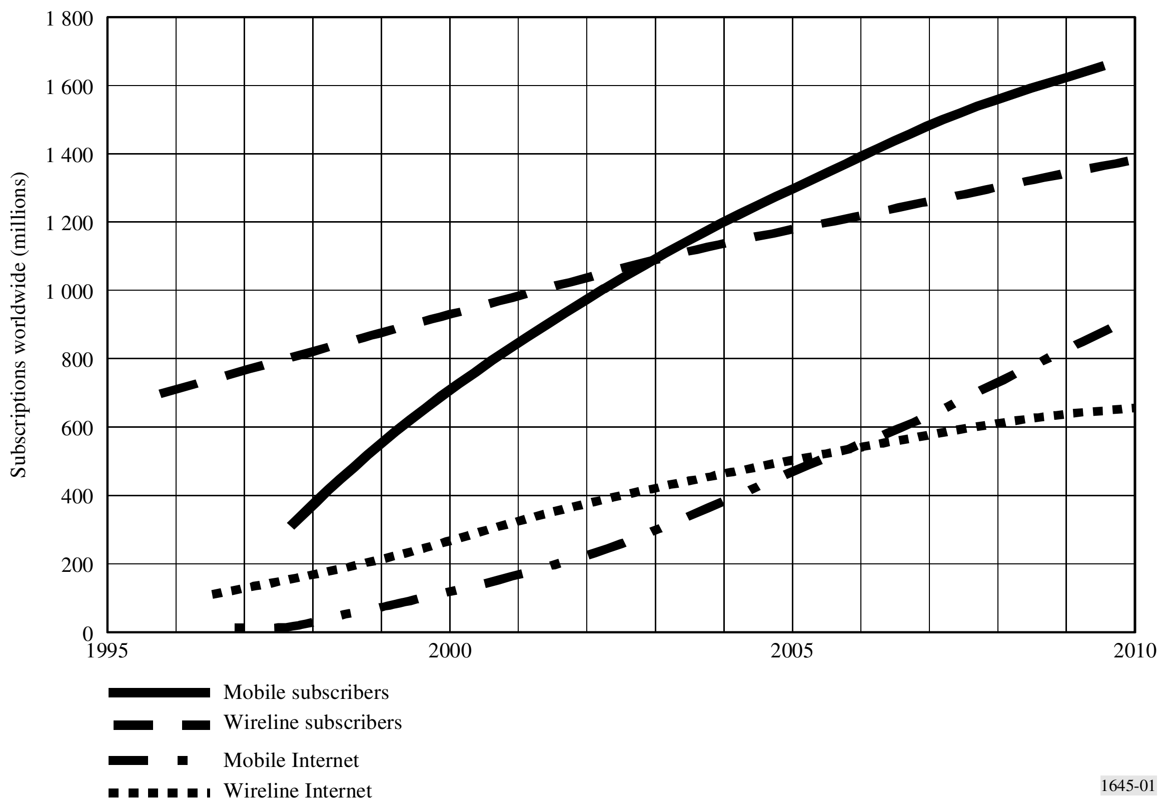


Figure 1.2. Illustration of the global growth of mobile and wireline subscribers.

In Figure 1.2. the global growth of mobile and wireline subscribers up to 2010 is illustrated. However, in planning process for the future development of IMT-2000 and IMT-Advanced, it is important to consider the timelines associated with their realization, which depend on a number of factors:

– user trends, requirements and user demand;

– technical capabilities and technology developments;
– standards development;
– spectrum availability, including allowing sufficient time to re-locate systems that may be using proposed bands;
– regulatory considerations;
– system (mobile and infrastructure) development and deployment.

All of these factors are interrelated. The first five have been and will continue to be addressed within ITU. System development and deployment relates to the practical aspects of deploying new networks, taking into account the need to minimize additional infrastructure investment and to allow time for customer adoption of the services of a major new system, such as IMT-2000.

The timelines associated with these different factors are depicted in Figure 1.3. When discussing the time phases for systems beyond IMT-2000, it is important to specify the time at which the standards are completed, when spectrum must be available, and when deployment may start.

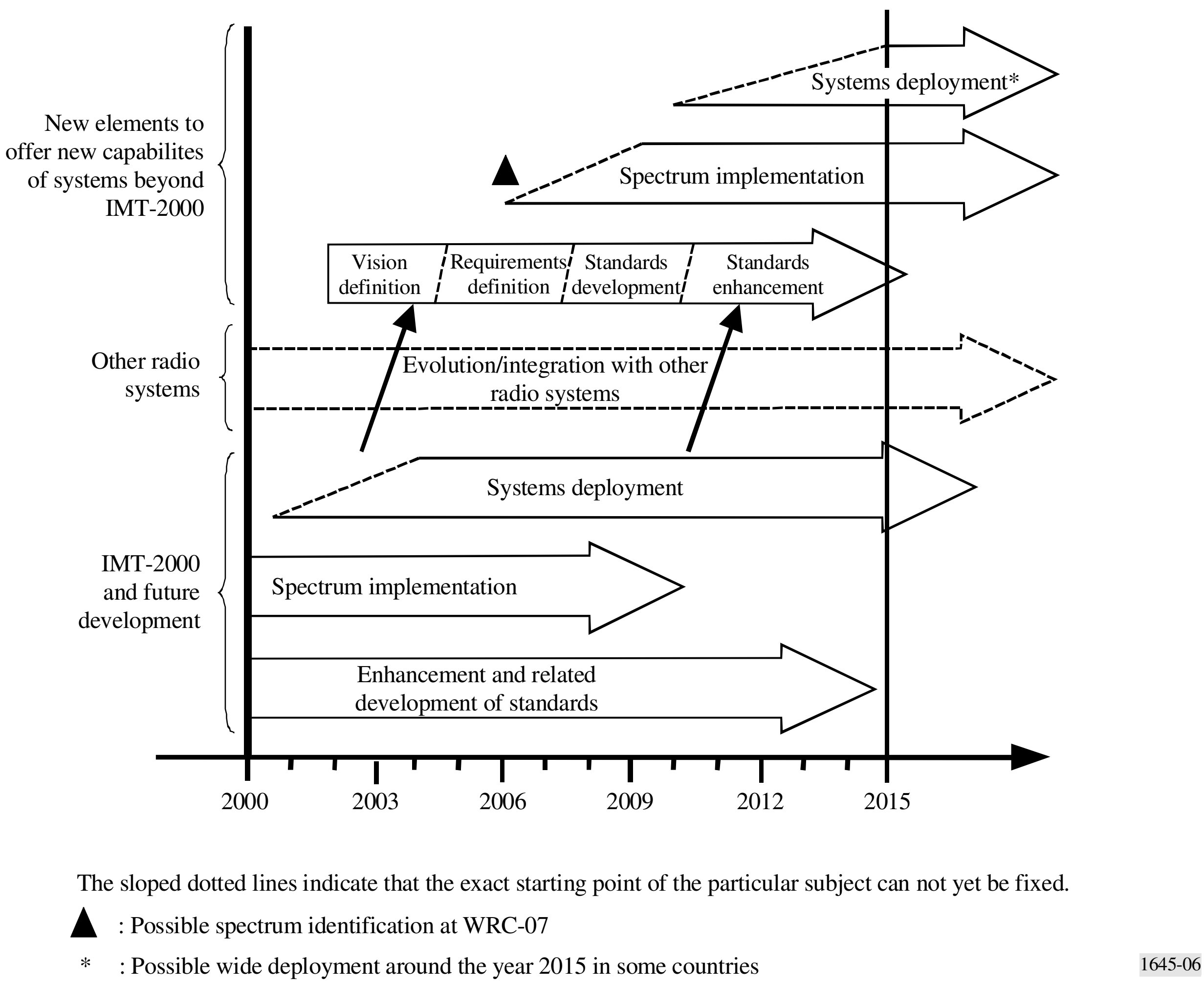


Figure 1.3. Phases and (expected) timelines for future development of IMT-2000 and IMT-Advanced.

Furthermore, the IMT-Advanced can be considered from multiple perspectives, including the users, manufacturers, application developers, network operators, and service and content providers as it is summarized in Table 1.1. Therefore, it is recognized that the technologies for IMT-Advanced can be applied in a variety of deployment scenarios and can support a range of environments, different service capabilities, and technology options. Consideration of every variation to encompass all situations is therefore not possible; nonetheless the work of the ITU-R has been to determine a representative view of IMT-Advanced consistent with the process defined in Resolution ITU-R 57 – Principles for the process of development of IMT-Advanced.

Table 1.1. Objectives from multiple perspectives in IMT-Advanced

Perspective	Objectives
END USER	Ubiquitous mobile access Easy access to applications and services Appropriate quality at reasonable cost Easily understandable user interface Long equipment and battery life Large choice of terminals Enhanced service capabilities User-friendly billing capabilities
CONTENT PROVIDER	Flexible billing capabilities Ability to adapt content to user requirements depending on terminal, location and user preferences Access to a very large marketplace through a high similarity of application programming interfaces
SERVICE PROVIDER	Fast, open service creation, validation and provisioning Quality of service (QoS) and security management Automatic service adaptation as a function of available data rate and type of terminal Flexible billing capabilities
NETWORK OPERATOR	Optimization of resources (spectrum and equipment) QoS and security management Ability to provide differentiated services Flexible network configuration Reduced cost of terminals and network equipment based on global economies of scale Smooth transition from IMT-2000 to systems beyond IMT-2000 (IMT-Advanced) Maximization of sharing capabilities between IMT-2000 and 4G IMT-Advanced systems (sharing of mobile, UMTS subscriber identity module (USIM), network elements, radio sites)

	Single authentication (independent of the access network) Flexible billing capabilities Access type selection optimizing service delivery
MANUFACTURER / APPLICATION DEVELOPER	Reduced cost of terminals and network equipment based on global economies of scale Access to a global marketplace Open physical and logical interfaces between modular and integrated subsystems Programmable platforms that enable fast and low-cost development

The services that users will want, and the rising number of users, will place increasing demands on access networks. These demands aren't met by the enhancement of IMT-2000 radio access systems (in terms of peak bit rate to a user, aggregate throughput, and greater flexibility to support many different types of service simultaneously). It is therefore anticipated that there will be a requirement for a new radio access technology, as IMT-Advanced, or technologies at some point in the future to satisfy the anticipated demands for higher bandwidth services.

Nowadays and further ITU-R Recommendations will develop these concepts in more detail. Other new Recommendations will address spectrum requirements for systems beyond IMT-2000 (IMT-Advanced), which frequency bands might be suitable, and in what time-frame such spectrum would be needed, with a view to accommodating emerging broadband services and applications. It is expected that new spectrum requirements documented in these Recommendations will be addressed at a future World Radiocommunication Conference.

IMT-Advanced 4G systems will support a wide range of symmetrical, asymmetrical, and unidirectional services. They will also provide management of different quality of service levels to realize the underlying objective of efficient transport of packet based services. In parallel, there will be an increased penetration of nomadic and mobile wireless access multimedia services.

The technologies, applications and services associated with IMT-Advanced 4G systems could well be radically different from the present, challenging the perceptions of what may be considered viable by today's standards and going beyond what can be achieved by the future enhancement of IMT-2000 working with other radio systems.

The new radio access interface(s) are envisaged to handle a wide range of supported data rates according to economic and service demands in multi-user environments with target peak data rates of up to approximately 100 Mbit/s for high mobility such as mobile access and up to approximately 1 Gbit/s for low mobility such as nomadic/local wireless access (Figure1.1). These data rates are targets for research and investigation. They should not be taken as the definitive requirements for 4G systems.

Moreover, these data rates will be shared between active users. The achievable (peak or sustained) throughput for any individual user depends on many parameters, including the number of active users, traffic characteristics, service parameters, deployment scenarios, spectrum availability, and propagation and interference conditions. These data rates are the maximum value of the sum of the data rate for all of the active users on a radio resource; it is possible that the peak data rate needed in the upstream direction will be different from the downstream direction. The transport data rates may need to be higher due to overheads, such as signalling and coding. Depending on the services for which the technology(or technologies, such: LTE, LTE-Advanced, 802.11n, 802.16m) will be used, continuous radio coverage may not be needed in order to meet the service requirements.

Long story short, here are summarized the key features (which cover the 4G requirements) of IMT-Advanced systems:

- ✓ a high degree of commonality of functionality worldwide while retaining the flexibility to support a wide range of services and applications in a cost efficient manner;
- ✓ compatibility of services within IMT and with fixed networks;
- ✓ capability of interworking with other radio access systems;
- ✓ high-quality mobile services;
- ✓ user equipment suitable for worldwide use;
- ✓ user-friendly applications, services and equipment;
- ✓ worldwide roaming capability;
- ✓ enhanced peak data rates to support advanced services and applications (100 Mbit/s for high and 1 Gbit/s for low mobility were established as targets for research).

These features enable IMT-Advanced to address evolving user needs. The capabilities of IMT-Advanced systems are being continuously enhanced in line with user trends and technology developments.

1.2. 3GPP LTE and LTE-Advanced technologies

Wireless operators are under intense pressure to offer compelling next generation mobile services to provide a personalized user experience for mobile users. They first introduced GPRS/EDGE services that introduced packet switching and connected the mobile device to the Internet over IP. The pursuit for higher data rate, higher capacity, higher throughput, lower delay, better spectrum efficiency and flexibility, high level of QoS provisioning, diversified mobile speed and greater coverage over cellular, resulted in GPRS/EDGE (2.5/2.75G) evolving to UMTS (3G) to HSPA (3.5G) to HSPA+ (3.75G). Compared to a data rate of 180 kbps in EDGE, HSPA+ promises data rates of 42 Mbps downlink and 22 Mbps uplink. Clearly the trend indicates that the mobile phone will soon support broadband speeds. Now, with the advent of LTE (3.9G or Super 3G) and LTE Advanced (4G), mobile broadband just got broader.

3GPP has not only evolved beyond addressing the Universal Terrestrial Radio Access Network (UTRAN) requirements to providing bandwidth intensive services. It has also put in a significant effort to evolve and simplify the packet core network. Branded as System Architecture Evolution (SAE), 3GPP has proposed a framework to evolve the 3GPP system to a higher data rate, lower latency, packet-optimized packet core system (Evolved Packet Core) that supports multiple access technologies, including 3GPP Internet Protocol Connectivity Access Network (IP CANs) like GSM EDGE Radio Access Network (GERAN), UTRAN and Evolved UTRAN (E-UTRAN) and non-3GPP IP CANs like WiFi, WiMAX and even wired technologies. This access independent evolution of the packet core system architecture is the first major step towards the realization of an All-IP Network and reaching the point where LTE-Advanced meeting IMT-Advanced. The timeline of LTE releases and LTE-Advanced development is shown in Figure 1.4.

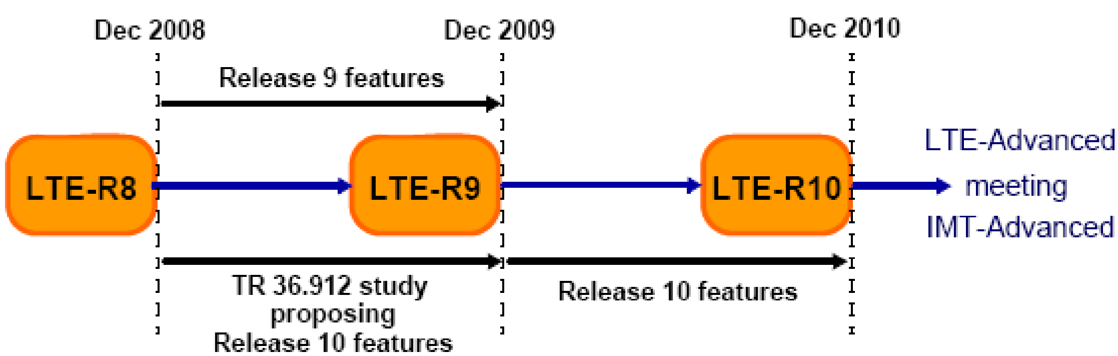


Figure 1.4. 3GPP LTE releases timeline.

Although there are major step changes between LTE and its 3G predecessors, it is nevertheless looked upon as an evolution of the UMTS/3GPP 3G standards. Despite it uses a different form of radio interface, using OFDMA/SC-FDMA instead of CDMA, there are many similarities with the earlier forms of 3G architecture and there is scope for much re-use. Moreover, 3GPP LTE can be seen for provide a further evolution of functionality, increased speeds and general improved performance. In addition to this, LTE is an all IP based network, supporting both IPv4 and IPv6. There is also no basic provision for voice, although this can be carried as VoIP.

Moreover, LTE has introduced a number of new technologies when compared to the previous cellular systems. They enable LTE to be able to operate more efficiently with respect to the use of spectrum, and also to provide the much higher data rates that are being required.

- OFDM (Orthogonal Frequency Division Multiplex): OFDM technology has been incorporated into LTE because it enables high data bandwidths to be transmitted efficiently while still providing a high degree of resilience to reflections and interference. The access schemes differ between the uplink and downlink: OFDMA (Orthogonal Frequency Division Multiple Access is used in the downlink; while SC-FDMA(Single Carrier - Frequency Division Multiple Access) is used in the uplink. SC-FDMA is used in view of the fact that its peak to average power ratio is small and the more constant power enables high RF power amplifier efficiency in the mobile handsets - an important factor for battery power equipment.
- MIMO (Multiple Input Multiple Output): One of the main problems that previous telecommunications systems has encountered is that of multiple signals arising from the many reflections that are encountered. By using MIMO, these additional signal paths can be used to advantage and are able to be used to increase the throughput. When using MIMO, it is necessary to use multiple antennas to enable the different paths to be distinguished. Accordingly schemes using 2 x 2, 4 x 2, or 4 x 4 antenna matrices can be used. While it is relatively easy to add further antennas to a base station, the same is not true of mobile handsets, where the dimensions of the user equipment limit the number of antennas which should be place at least a half wavelength apart.
- SAE (System Architecture Evolution): as a part of the Evolved Packet Core (EPC). With the very high data rate and low latency requirements for 3G LTE, it is necessary to evolve the system architecture to enable the improved performance to be achieved. One change is that a number of the functions previously handled by the core network have been transferred out to the periphery. Essentially this provides a much "flatter" form of network architecture. In this way latency times can be reduced and data can be routed more directly to its destination.

In Figure 1.5. we can see the functional decomposition of the Evolved Packet Core for 3GPP and non-3GPP IP Core Access Network. The EPC architecture is guided by the principle of logical separation of the signalling and data transport networks. The fact that some EPC functions reside in the same equipment as some transport functions, does not make the transport functions a part of the EPC. It is also possible that one physical network element in the EPC implements multiple logical nodes. More details about EPC and SAE will be provided in the Module 2 (Section 2.3).

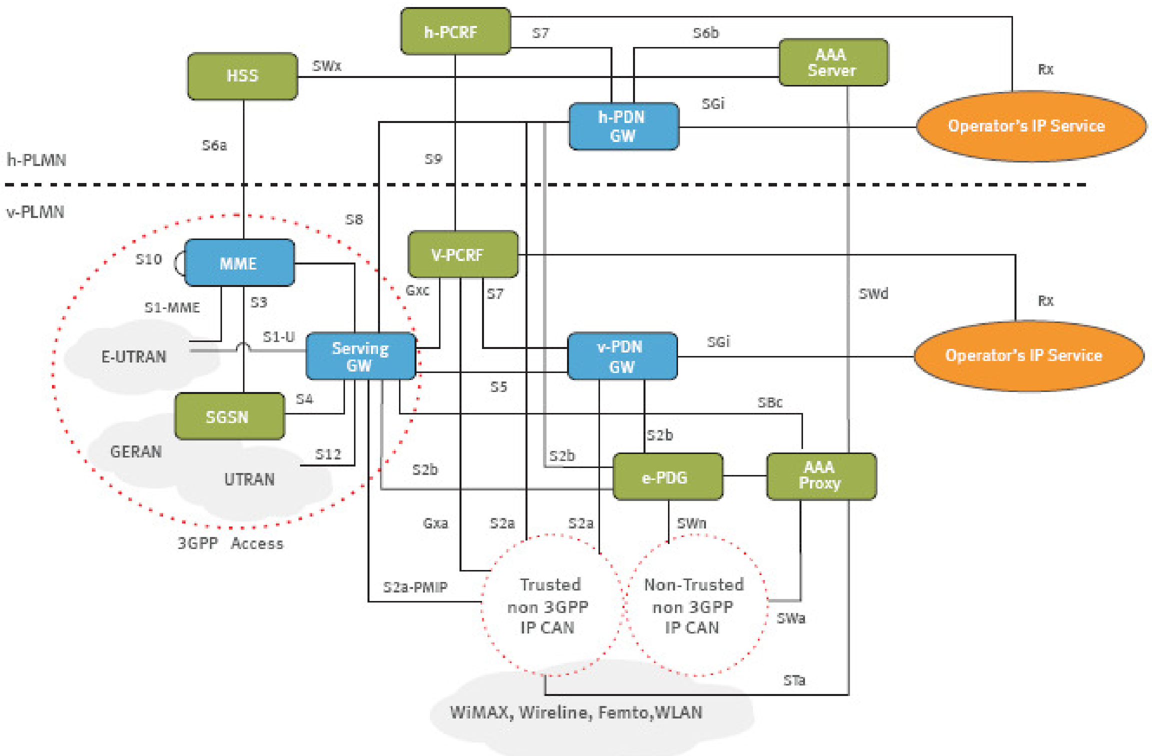


Figure 1.5. 3GPP Evolved Packet Core System Architecture

It is worth summarizing the key parameters of the 3G LTE specification. In view of the fact that there are a number of differences between the operation of the uplink and downlink, these naturally differ in the performance they can offer. In Table 1.2 the 3GPP LTE highlight specifications are summarized. Those specifications give an overall view of the performance that 3G LTE is offering. It meets the requirements of industry for high data download speeds as well as reduced latency - a factor important for many applications from VoIP to gaming and interactive use of data. It also provides significant improvements in the use of the available spectrum.

Table 1.2. 3GPP LTE highlight specifications

Parameter	Details
Peak downlink speed 64QAM (Mbps)	100 (SISO), 172 (2x2 MIMO), 326 (4x4 MIMO)
Peak uplink speeds (Mbps)	50 (QPSK), 57 (16QAM), 86 (64QAM)
Data type	All packet switched data (voice and data). No circuit switched.
Channel bandwidths (MHz)	1.4, 3, 5, 10, 15, 20
Duplex schemes	FDD and TDD
Mobility	0 - 15 km/h (optimised), 15 - 120 km/h (high performance)
Latency	Idle to active less than 100ms Small packets ~10 ms
Spectral efficiency	Downlink: 3 - 4 times Rel 6 HSDPA Uplink: 2 -3 x Rel 6 HSUPA
Access schemes	OFDMA (Downlink) SC-FDMA (Uplink)
Modulation types supported	QPSK, 16QAM, 64QAM (Uplink and downlink)

With the standards definitions now available for LTE, the Long Term Evolution of the 3G services, eyes are now turning towards the next development, that of the truly 4G technology named IMT-Advanced. The new technology being developed under the auspices of 3GPP to meet these requirements is often termed LTE-Advanced (LTE Release 10 & beyond).

In order that the cellular telecommunications technology is able to keep pace with technologies that may compete, it is necessary to ensure that the next generation mobile and wireless technologies are being formulated and developed. This is the reasoning behind starting the development of the new LTE Advanced systems, proving the technology and developing the LTE Advanced standards. In order that the correct solution is adopted for the 4G system, as we mention in the first section, the ITU-R has started its evaluation process to develop the recommendations for the terrestrial components of the IMT Advanced radio interface. And as one of the main approved technology for this is the LTE Advanced.

If we look back in the past, in 2008 3GPP held two workshops on IMT Advanced, where the "Requirements for Further Advancements for E-UTRA" were gathered. The resulting Technical Report 36.913 was then published in June 2008 and submitted to the ITU-R defining the LTE-Advanced system as their proposal for IMT-Advanced. The development of LTE Advanced/IMT Advanced can be seen to follow and evolution from the 3G services that were developed using UMTS/W-CDMA technology. However, the LTE Advanced is not the only candidate technology for 4G (and in the same time for IMT-Advanced). WiMAX is also there, offering very high data rates and high levels of mobility. Consequently, as it was mention, WiMAX 802.16m together with the LTE

Advanced are adopted as main 4G technologies. In the following table (Table 1.3) the main comparisons between existing 3G (3,5 G) and LTE Advanced technologies are summarized.

Table 1.3. 3G and 4 G comparisons

	WCDMA (UMTS)	HSPA HSDPA / HSUPA	HSPA+	LTE	LTE Advanced (IMT Advanced)
Max downlink speed bps	384 k	14 M	28 M	100 M	1 G
Max uplink speed bps	128 k	5.7 M	11 M	50 M	500 M
Latency round trip time approx	150 ms	100 ms	50ms (max)	~10 ms	less than 5 ms
3GPP releases	Rel 99/4	Rel 5 / 6	Rel 7	Rel 8	Rel 10 & beyond
Approx years of initial roll out	2003 / 4	2005 / 6 HSDPA 2007 / 8 HSUPA	2008 / 9	2009 / 10	2010 / 11
Access methodology	CDMA	CDMA	CDMA	OFDMA / SC-FDMA	OFDMA / SC-FDMA

The performance and capabilities of 4G LTE will be unmatched in the marketplace, allowing customers to do things never before possible in a wireless and mobile environment. Although not fixed yet in the specifications, there are many high level aims for the new LTE Advanced specification. These will need to be verified and much work remains to be undertaken in the specifications before these are all fixed. Currently some of the main that wireless and mobile implementation of 4G LTE will provide are the following:

- ✓ Peak data rates: downlink - 1 Gbps; uplink - 500 Mbps.
- ✓ Spectrum efficiency: 3 times greater than LTE. LTE-Advanced shall operate in spectrum allocations of different sizes including wider spectrum allocations than those of LTE Release 8. The main focus for bandwidth solutions wider than 20MHz should be on consecutive spectrum. However, aggregation of the spectrum for LTE-Advanced should take into account reasonable user equipment (UE) complexity. Frequency division duplex (FDD) and time division duplex (TDD) should be supported for existing paired and unpaired frequency bands, respectively.
- ✓ Peak spectrum efficiency: downlink - 30 bps/Hz; uplink - 15 bps/Hz.

- ✓ Spectrum use: the ability to support scalable bandwidth use and spectrum aggregation where non-contiguous spectrum needs to be used.
- ✓ Latency: from Idle to Connected in less than 50 ms and then shorter than 5 ms one way for individual packet transmission.
- ✓ Cell edge user throughput to be twice that of LTE.
- ✓ Average user throughput to be 3 times that of LTE.
- ✓ Simultaneous user support: LTE provides the ability to perform two-dimensional resource scheduling (in time and frequency), allowing support of multiple users in a time slot, resulting in a much better always-on experience while enabling the proliferation of embedded wireless applications/systems (in contrast, existing 3G technology performs one-dimensional scheduling, which limits service to one user for each timeslot).
- ✓ Mobility: Same as that in LTE: System shall support mobility across the cellular network for various mobile speeds up to 350km/h (or even up to 500km/h depending on the frequency band). In comparison to LTE Release 8, the system performance shall be enhanced for 0 up to 10 km/h.
- ✓ Compatibility: LTE Advanced shall be capable of interworking with LTE and 3GPP legacy systems.
- ✓ Security: LTE provides enhanced security through the implementation of Universal Integrated Circuit Card (UICC) Subscriber Identity Module (SIM) and the associated robust and non-invasive key storage and symmetric key authentication using 128-bit private keys. LTE additionally incorporates strong mutual authentication, user identity confidentiality, integrity protection of all signaling messages between UE and Mobility Management Entity (MME) and optional multi-level bearer data encryption.
- ✓ Simplified Worldwide Roaming: the widely adopted next-generation 3GPP standard, will provide the greatest opportunities for seamless international roaming.
- ✓ Mass Deployment: LTE's inherent support for Internet Protocol version 6 (IPV6) addressing and International Mobile Subscriber Identity (IMSI)-based identifiers makes mass deployments of machine-to-machine applications over LTE-Advanced possible.

These are many of the development aims for LTE Advanced. Their actual figures and the actual implementation of them will need to be worked out during the specification stage of the LTE Advanced system.

However, these two 3GPP technologies are addressed in much greater detail in the following Modules (Module 2 and Module 3).

1.3. IEEE wireless technologies (802.11n, 802.16m)

In order the data packets to travel safety and regular from one computer terminal to other terminal they should also follow set of rules and regulations. One such set of rules for the networking traffic to follow is IEEE802 standards. Its developed by IEEE (Institute of Electrical and Electronics Engineers, Inc.) The IEEE is the world's leading professional association for the advancement of technology. It's a non-profit organization offering its members immense benefits. The standards such as IEEE 802 helps industry provide advantages such as, interoperability, low product cost, and easy to manage standards. IEEE standards deal with only Local Area Networks (LAN) and Metropolitan Area Networks (MAN). See in the figure below (Figure 1.6), to know where exactly the IEEE802 standards are used in a classical OSI layer.

<table>
<tr><td colspan="4">Application layer (OSI 7)</td></tr>
<tr><td colspan="4">Presentation layer (OSI 6)</td></tr>
<tr><td colspan="4">Session layer (OSI 5)</td></tr>
<tr><td colspan="4">Transport layer (OSI 4)</td></tr>
<tr><td colspan="4">Network layer (OSI 3)</td></tr>
<tr><td colspan="4">MAC layer (OSI 2)
IEEE 802.2 (LLC)</td></tr>
<tr><td>IEEE 802.11
WLAN</td><td>IEEE 802.3
CSMA/CD</td><td>IEEE 802.4
Token Bus</td><td>IEEE 802.5
Token Ring</td></tr>
</table>

Figure 1.6. Illustration of the position of IEEE 802 standards (fields marked in blue) in an OSI layer.

The IEEE 802 standards are further divided into many parts.
Such as:

IEEE 802.1 Bridging (networking) and Network Management
IEEE 802.2 Logical link control (upper part of data link layer)
IEEE 802.3 Ethernet (CSMA/CD)
IEEE 802.4 Token bus (disbanded)
IEEE 802.5 Defines the MAC layer for a Token Ring (inactive)
IEEE 802.6 Metropolitan Area Networks (disbanded)
IEEE 802.7 Broadband LAN using Coaxial Cable (disbanded)
IEEE 802.8 Fiber Optic TAG (disbanded)
IEEE 802.9 Integrated Services LAN (disbanded)
IEEE 802.10 Interoperable LAN Security (disbanded)
IEEE 802.11 Wireless LAN & Mesh (Wi-Fi certification)
IEEE 802.12 demand priority (disbanded)
IEEE 802.13 Not Used

IEEE 802.14 Cable modems (disbanded)
IEEE 802.15 Wireless PAN
IEEE 802.15.1 (Bluetooth certification)
IEEE 802.15.4 (ZigBee certification)
IEEE 802.16 Broadband Wireless Access (WiMAX certification)
IEEE 802.16e (Mobile) Broadband Wireless Access
IEEE 802.16m (Mobile WiMAX release 2.0)
IEEE 802.17 Resilient packet ring
IEEE 802.18 Radio Regulatory TAG
IEEE 802.19 Coexistence TAG
IEEE 802.20 Mobile Broadband Wireless Access
IEEE 802.21 Media Independent Handoff
IEEE 802.22 Wireless Regional Area Network

Here we will focus on the most popular IEEE wireless standard: **IEEE 802.11n** and **IEEE 802.16m** (Mobile WiMAX release 2.0, a.k.a. WirelessMAN-Advanced), key standards accepted within the 4G mobile network paradigm (e.g. included within IMT-Advanced umbrella).

1.3.1. IEEE 802.11n standard

A wireless communication standard like IEEE802.11 or WLAN (or HiperLAN/2) is always envisaged to provide ubiquitous high-speed access to data or information. The greed for a higher data rate communication system has challenged the wireless researchers all over the world. Thus, the data rate offered by the current wireless communication systems has increased tremendously. The current wireless local area networks (WLAN) such as IEEE 802.11a/g standard offers data rate at physical (PHY) layer up to 54 Mbps (see Figure 1.7). The advent of new applications such as video transmission to high definition TV (HDTV), video-streaming, video phones, etc., demand very high data rates. The latest digital communication techniques can be used to increase the data rate. The basic ways to increase the data rate are larger channel bandwidth with utilization of a higher number of data subcarriers, larger size of constellation, higher coding rate and use of multiple antennas (MIMO).

Moreover, the IEEE **802.11n** amendment is the latest addition under development for the IEEE 802.11 standard providing a marked increase in throughput (from **20 Mbps to around 200 Mbps**, in practice) as well as range of reception (through reducing signal fading) over the IEEE 802.11a/g standards currently in use. Multiple antennas, or MIMO (Multiple-Input, Multiple-Output), is the key innovation used to obtain these benefits. The current draft for the IEEE 802.11n amendment supports the use of MIMO features such as spatial-division multiplexing (SDM), space-time block coding (STBC) and transmitter beamforming. In addition, there are provisions for the use of advanced coding with LDPC (low-density parity check codes), as well as a 40 MHz bandwidth mode (known as channel bonding). The above features allow the IEEE 802.11n amendment to specify data rates up to **600 Mbps**, a more than ten-fold increase over the maximum data rate with the 11a/g standards.

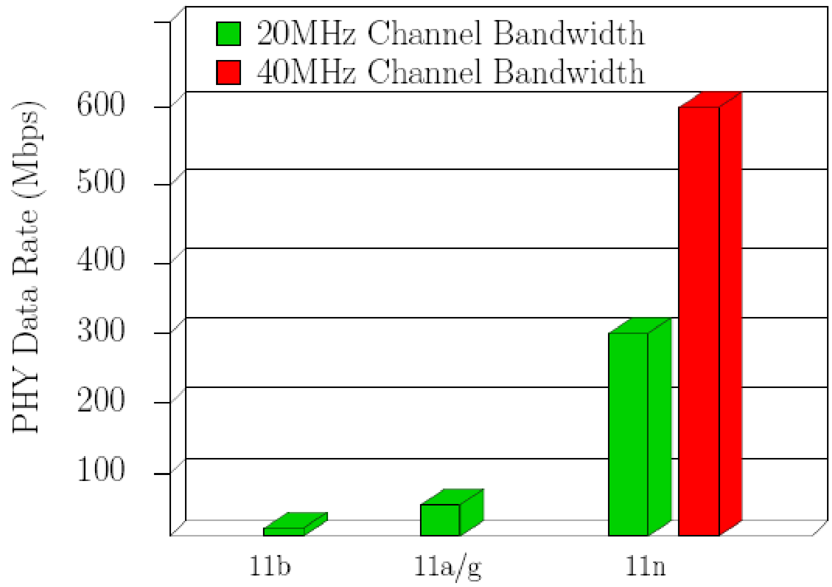


Figure 1.7. Comparison of data rates of IEEE 802.11 WLAN standards.

Moreover. the greatest impact of 802.11n is in the technology used for the physical layer (OSI 1). The ability to create low-cost radios in CMOS now allows the simultaneous use of multiple radios and antennas in every client. Advanced signal processing enables 802.11n to integrate multiple radios and a number of other PHY improvements to effectively increase the burst transmission speed and the total system capacity by a factor of ten when all of the new enhancements are used.

In the following are summarized the main advantages of IEEE 802.11n wireless technology:

- ✓ Improved coding and modulation: 802.11n updates the OFDM methods pioneered by 802.11a for bit and frame encoding. The more efficient OFDM used in 802.11n produces a maximum data rate of 65 Mbps per stream compared to 54 Mbps for 802.11a and 802.11g systems. These changes result in 20% more burst transmission speed. However, the **600 Mbit/s** data rates are achieved only with the maximum of **four** spatial streams using a **40 MHz**-wide channel. Various modulation schemes and coding rates are defined by the standard and are represented by a Modulation and Coding Scheme (MCS) index value. The table below (Table 1.4) shows the relationships between the variables that allow for the maximum data rate.

Table 1.4. Illustration of IEEE 802.11n data rates.

MCS index	Spatial streams	Modulation type	Coding rate	Data rate (Mbit/s)			
				20 MHz channel		40 MHz channel	
				800 ns GI	400 ns GI	800 ns GI	400 ns GI
0	1	BPSK	1/2	6.5	7.2	13.5	15
1	1	QPSK	1/2	13	14.4	27	30
2	1	QPSK	3/4	19.5	21.7	40.5	45
3	1	16-QAM	1/2	26	28.9	54	60
4	1	16-QAM	3/4	39	43.3	81	90
5	1	64-QAM	2/3	52	57.8	108	120
6	1	64-QAM	3/4	58.5	65	121.5	135
7	1	64-QAM	5/6	65	72.2	135	150
8	2	BPSK	1/2	13	14.4	27	30
9	2	QPSK	1/2	26	28.9	54	60
10	2	QPSK	3/4	39	43.3	81	90
11	2	16-QAM	1/2	52	57.8	108	120
12	2	16-QAM	3/4	78	86.7	162	180
13	2	64-QAM	2/3	104	115.6	216	240
14	2	64-QAM	3/4	117	130	243	270
15	2	64-QAM	5/6	130	144.4	270	300
16	3	BPSK	1/2	19.5	21.7	40.5	45
17	3	QPSK	1/2	39	43.3	81	90
18	3	QPSK	3/4	58.5	65	121.5	135
19	3	16-QAM	1/2	78	86.7	162	180
20	3	16-QAM	3/4	117	130.7	243	270
21	3	64-QAM	2/3	156	173.3	324	360
22	3	64-QAM	3/4	175.5	195	364.5	405
23	3	64-QAM	5/6	195	216.7	405	450
24	4	BPSK	1/2	26	28.8	54	60
25	4	QPSK	1/2	52	57.6	108	120
26	4	QPSK	3/4	78	86.8	162	180
27	4	16-QAM	1/2	104	115.6	216	240
28	4	16-QAM	3/4	156	173.2	324	360
29	4	64-QAM	2/3	208	231.2	432	480
30	4	64-QAM	3/4	234	260	486	540
31	4	64-QAM	5/6	260	288.8	540	600

- ✓ Shorter Guard Interval: 802.11n reduces the Guard Interval (GI) from 800 nanoseconds to 400 nanoseconds (see Table 1.4). This small change increases the symbol rate by 10 percent.

- ✓ MIMO and Spatial Multiplexing: 802.11n introduces smart radio technology to dramatically increase the quality and speed of the physical layer by creating multiple simultaneous data streams. Previous wireless systems have modestly used multiple antennas at the receiver to take advantage of the fact that a transmission from a source to a destination may take multiple paths based on reflections from obstructions in the direct path between source and destination. Historically, multipath has been viewed as a signal impairment that degrades the quality of the radio transmission. 802.11n exploits multipath to enhance the delivery of multiple spatial streams. 802.11n systems employ MIMO (multiple input, multiple output) which defines multiple transmit and receive radios – each with its own antenna – that combine to deliver multiple streams of data between stations on the same channel. By digitally controlling simultaneous transmissions and reception, 802.11n stations can effectively multiply the data rate by the number of simultaneous spatial streams they support. 802.11n defines up to 4 spatial streams. The 802.11n allows up to **4x4:4** (the first number (4) is the maximum number of transmit antennas or RF chains that can be used by the radio, the second number (4) is the maximum number of receive antennas or RF chains that can be used by the radio, and the third number (4) is the maximum number of data spatial streams the radio can use.). Common configurations of 802.11n devices are **2x2:2, 2x3:2,** and **3x3:2**. All three configurations have the same maximum throughputs and features, and differ only in the amount of diversity the antenna systems provide. In addition, a fourth configuration, 3x3:3 is becoming common, which has a higher throughput, due to the additional data stream. Moreover, the Wi-Fi Alliance 802.11n Draft 2 certification requires that systems support at least 2 spatial streams. Spatial multiplexing can increase the burst transmission rate up to 4 times.

- ✓ Channel Bonding: 802.11n allows channel aggregation that bonds two adjacent 20 MHz channels into a single 40 MHz channel in both the 2.4 GHz and 5 GHz bands. Channel bonding doubles the burst transmission rate. The overall system capacity for large systems with many access points will not increase, since more spectrum is consumed. However, for smaller networks with one or two access points, it can increase total capacity as well. Channel bonding is supported in both the 2.4 GHz and 5 GHz bands. Only one independent 40 MHz channel is possible in the 2.4 GHz band. Channel bonding will be most useful in the 5 GHz band for green field 802.11n networks.

- ✓ Beamforming: is an optional feature of 802.11n. It is a natural extension of the physical layer that has multiple radios and antennas in each station. By controlling the transmit power and phase of the collection of transmission antennas, it is possible to shape the effective gain of the antennas to create a pattern that points towards the receiving station - a beam. Beamforming will be used to extend the effective range and create more robust coverage with 802.11n systems.

- ✓ Available Spectrum: 802.11n supports both 2.4 GHz and 5 GHz bands. It has a single MAC that operates with a multiple frequency physical layers. This is really a configuration benefit rather than a Physical layer's feature. 802.11n makes use of the legacy 2.4 GHz band and constructs three (3) largely non-interfering 20 MHz channels or 1 20 MHz channel and 1.40 MHz channel. It is backward compatible with 802.11b/g stations and channelization. 802.11n makes use of the existing 802.11a channel set in the 5 GHz band at (5.15-5.25, 5.25-5.35, and 5.75-5.85 GHz) to construct 12 non-overlapping 20 MHz channels or as many as 6 nonoverlapping 40 MHz channels. 802.11n also takes advantage of new worldwide regulatory changes making the 5.47-5.75 GHz band available for unlicensed WLAN use. Existing primary users of this band (largely radars) have heretofore limited its use, but the Dynamic Frequency Selection and power control features defined by the companion standard 802.11h and new regulations open up this band to use.

- ✓ MAC Layer Enhancements: the 802.11n MAC allows successive frames to be combined and supports frames up to 64 k bytes with the Aggregated MAC Protocol Data Unit. It makes 802.11n MAC protocol more efficient than the other 802.11 legacy MAC protocols. Moreover, there are Block ACK, which enables multiple frames to be transmitted and then acknowledged with a single ACK frame. These and other MAC improvements reduce protocol overhead and improve the efficiency of the MAC protocol resulting in higher throughput. The new mechanisms also allow new 802.11n equipment and legacy 802.11 a/b/g equipment to share the airwaves more fairly in a mixed network. The impact of MAC performance improvements will be more subtle than the PHY improvements. The amount of performance gain will vary according to the type of application traffic and the mix of legacy and new stations in the network. With all of the enhancements and the most optimistic set of conditions, these improvements could double the effective throughput.

On the other hand, the additional capacity of 802.11n exacts its own price on the wired network infrastructure that supports WLAN. Large-scale WLAN deployments in particular are usually based on a wired Ethernet network that supplies both backbone trunking between WLAN access points and power over the network wiring through Power Over Ethernet (POE). Enterprise class 802.11n access points will often be capable of delivering more than 100 Mbps of useful capacity each. The first wave of Draft 2 802.11n APs will support at least two spatial streams with a short GI. That yields 145 Mbps of raw capacity. With the new MAC enhancements, 100 Mbps of IP throughput per AP is possible. Channel bonding and more spatial streams will increase capacity per AP even further. Hundreds of APs with these capabilities will strain the wired infrastructure of many enterprises. In some networks, the wired core network will become the bottleneck rather than the wireless LAN at the edge. When deploying an 802.11n wireless LAN, enterprises will have to consider upgrading their wired infrastructure as well. Newer 802.11n multi-radio access points will consume more electrical power than legacy 802.11a/b/g access points and more than the current POE standard delivers. This will require an upgrade of the wired infrastructure to the next generation POE standard or provisioning two Ethernet drops per AP location.

Moreover, the dominant Wireless LAN architecture for enterprises has been a central Wireless LAN controller that manages and secures a set of “thin” APs. The WLAN controller provides management, security and deployment tools for the entire wireless network. All of the wireless traffic passes through the WLAN Controller. The improvements of 802.11n can be difficult for a centralized WLAN Controller architecture to handle and have prompted a new round of innovation at the system level. 802.11n APs are getting “fatter” anyway – they have more radios and are more complex than legacy a/b/g APs. The significant increase in data rate and capacity brought by 802.11n challenges the wisdom of aggregating all of the wireless traffic back through a single node in the network.

Also, a new system level architectures are being proposed for 802.11n in the enterprise:

- Scaled up WLAN controllers – 802.11n APs deliver ten times the capacity, so create a new bigger WLAN controller that can handle that increase and leave the network architecture the same.
- Distributed clusters of controllers – each WLAN controller is smaller and manages a subset of APs. The clusters communicate with each other to manage system wide functions.
- Hierarchical network of controllers – define the notion of edge controllers and master controller. Data and management traffic are split and routed directly from the edge controllers. Only management traffic flows to the master controller.
- No controller at all – back to the future. New APs are “fat APs” capable of independent operation without a controller. For large deployments there is a management appliance.

Moreover, one of the biggest and least appreciated challenges of large-scale enterprise WLAN deployments is the problem of interference. In particular, the self-interference between nearby WLAN cells served by different access points in the same system. The vast majority of enterprise WLAN networks today are constructed using only the three non-overlapping channels of the 2.4 GHz band to create a microcellular architecture. Three channels are not sufficient for adequate isolation between micro-cells on the same channel. The interference range of WLAN is much greater than the effective communication range. A substantial amount of system capacity is lost to interference from micro-cells operating on the same channel – often quite some physical distance away.

802.11n improves this situation by opening up the additional channels in the 5 GHz band thus introducing many more new channels that can mitigate this self-interference challenge. However, while more channels help – they do not eliminate the problem entirely. The use of smart antennas in 802.11n, particularly with beam forming, dramatically increases range but not necessarily in predictable or manageable patterns. While the range increase is welcome, it also brings with it increased interference at distance. These issues suggest that though the problem of interference is reduced with the advances of 802.11n; it will continue to be a key challenge in the design of large scale, high performance 802.11n WLAN networks.

Furthermore, the range of 802.11n networks is much better that its predecessors, but the way that range is achieved is very different and much more sensitive to the physical environment. Local multi-path conditions allow greater range in some directions preferentially to others resulting in unpredictable coverage. These new coverage patterns will interact with “adjacent” access points in ways that existing network planning and site survey tools do not anticipate. New versions of these tools will be required to help manage the deployment of 802.11n networks in the enterprise. Conventional wisdom about how to deploy WLANs in an office environment.

Coverage and range will also vary depending on the type of client using the 802.11n infrastructure. Like any 802 standard, 802.11n includes the provision of supporting earlier version of the standard. So there will be 802.11b, 802.11g and 802.11a clients operating on 802.11n infrastructure. These devices will have very different range than 802.11n clients. 802.11n introduces the possibility of a wider variety for client devices – from low-power single radio, single antenna clients to multi-radio clients supporting up to four spatial streams and beam forming. The different client capabilities will make designing 802.11n networks for coverage much more complicated.

Moreover, the consumer and small business markets are the first to adopt 802.11n, and will ultimately speed the process for other markets – just as consumer adoption of 802.11g did in 2003. As early as 2006, so-called “pre-N” non-standard products have been sold to this market. These early products delivered many of the benefits (increased range and performance) of 802.11n in non-standard products. But the lack of interoperability across different vendors stalled wide-scale deployment of these early products. With the adoption of a firm draft 802.11n standard, this limitation is effectively removed. Apple quietly converted all of its wireless products to be native 802.11n in early 2007. We expect to see most other computer and consumer electronic manufacturers move to 802.11n in their next product cycle – Lenovo, Dell and HP have already done so in recent notebook products incorporating dual-band 802.11n capability. Consumer products featuring wireless distribution of high definition video streams through 802.11n will be coming soon. The low-cost availability of built-in dual band 802.11n clients in a variety of mobile devices will accelerate the deployment of 802.11n in the enterprise and metropolitan networks.

As a conclusion of this sub-section for IEEE 802.11n, we can say that 802.11n are substantially increaseing the performance and ubiquitous wireless access of laptops, desktops, smart phones and entertainment devices nowadays and over the next several years. Without cost increase, just like the migration from 802.11 cousin 10 Mbps Ethernet to gigabit Ethernet, 802.11n will first appear in client devices and begin to be pervasively deployed in enterprises, homes and eventually metro networks. While their will be teething problems with this new 4G technology, there is no doubt of its pervasive and inevitable deployment.

1.3.2. IEEE 802.16m standard

With the tremendous development of the mobile devices such as smart phones, PDAs, iPhones and etc., there is now a huge demand for wireless networks which both support high data rates and high mobility. WiMAX (IEEE 802.16) networks offer a response to this demand. However, with the expectation of constantly increasing data rates comes the need to continually upgrade the standards behind these networks. The next major revision undoable is WiMAX Release 2.0, which will be based on the 802.16m standard. In Figure 1.8 is plotted the latest IEEE 802.16 technology roadmaps.

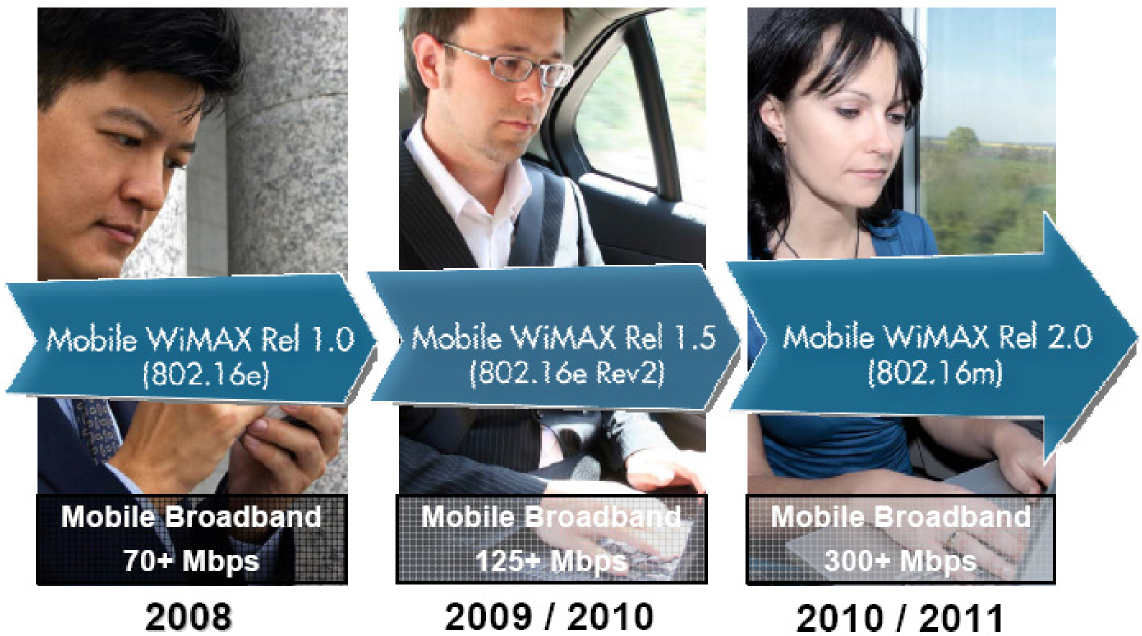


Figure 1.8. Mobile WiMAX 802.16 Roadmap

How we can see from the figure above, the WiMAX Rel 2.0 is the third significant release of WiMAX (IEEE 802.16). The first, Rel 1.0 (IEEE 802.16e), targeted International Telecommunications Union's IMT-2000 standard by implementing a subset of the 802.16e + Cor2 standard. The final network specification was released in the second quarter of 2007. It employed Time Division Duplexing (TDD), and provided data rates of around 70 Mbps. Rel 1.5 (IEEE 802.16e Rev2) followed with a final release of the network specification in the fourth quarter of 2008. It employed Frequency Division Duplexing (FDD) and allowed speeds of around 125 Mbps. The Rel 2.0 (IEEE 802.16m) network specification is released in 2010. However, this release will be additionally update (with a new draft and recommendations), and will feature data rates around 300 Mbps, employ a combination of FDD and TDD, and is being designed to target the International Telecommunications Union's IMT-Advanced specification. In the Figure 1.9, just for comparison, the timeline of IMT-Advanced and IEEE 802.16m is given.

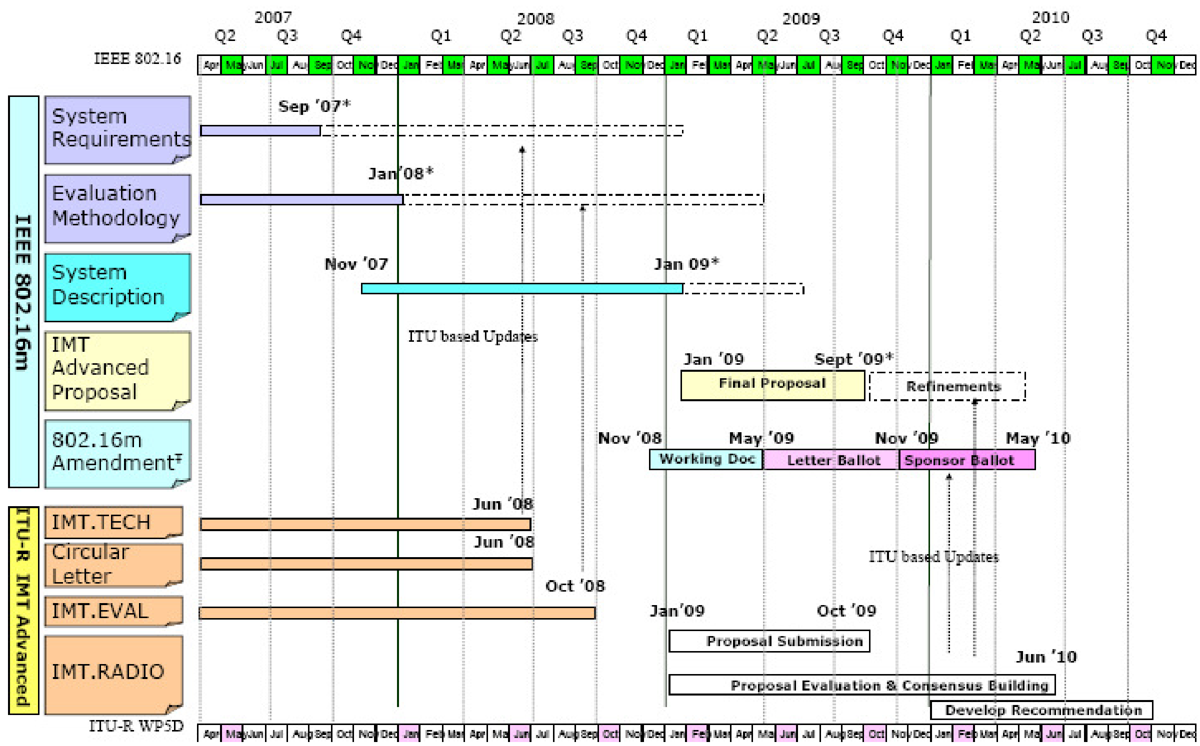


Figure 1.9. WiMAX 802.16m and ITU IMT-Advanced Roadmaps.

The ITU-R IMT-Advanced specification which 802.16m is being designed to target is the successor to IMT-2000. The primary improvements include adding support for many new service classes (especially telecommunications services), increased mobility, and better Quality of Service (QoS) guarantees. The stated goals of the 802.16m amendments are “to provide an advanced air interface for operation in licensed bands”. It maintains support for legacy devices based on earlier variations of the 802.16 standard, while providing significant performance increases for new devices. Devices, which operate on an 802.16m network, fall into one of several categories. Advanced Mobile Stations (AMS), Advanced Base Stations (ABS), and Advanced Relay Stations (ARS) support 802.16m operation, while R1MSs and R1BSs (Revision 1 mobile and base stations, respectively) are legacy devices.

Moreover, here are summarized the advanced features in IEEE 802.16m technology:

- ✓ New subframe-based frame structure that allows faster air-link transmissions/retransmissions, resulting in significantly shorter user-plane and control plane latencies.
- ✓ New subchannelization schemes and more efficient pilot structures in the downlink and uplink to reduce L1 overhead and to increase spectral efficiency.
- ✓ New and improved control channel structures in the downlink and uplink to increase efficiency and reduce latency of resource allocation and transmission as well as system entry/re-entry.

- ✓ Multi-carrier operation using a single MAC instantiation to enable operation in contiguous/non-contiguous RF bands in excess of 20 MHz.
- ✓ Extended and improved MIMO modes in the downlink and uplink.
- ✓ Enhanced Multicast and Broadcast Services using new E-MBS control channels and subchannelization.
- ✓ Enhanced GPS-based and Non-GPS-based Location Based Services.
- ✓ Support of Femtocells and Self-Organization and Optimization features.
- ✓ Increased VoIP capacity though use of new control structure, frame structure, faster HARQ retransmissions, persistent scheduling, group scheduling, and reduced MAC overhead.
- ✓ Improved and increased control channel and data channel coverage and link budget through use of transmit diversity schemes as well as more robust transmission formats and link adaptation.
- ✓ Support for multi-hop relay operation with unified access and relay links.
- ✓ Support for advanced interference mitigation techniques including Multi-BS MIMO, Fractional Frequency Reuse, Closed-loop and Open-loop power control schemes.
- ✓ Improved intra-RAT and inter-RAT handover schemes with shorter handoff interruption times.
- ✓ Improved sleep and idle mode operations.
- ✓ Improved QoS provisioning.

In the following we discussed shortly the key advanced features that IEEE 802.16m technology incorporate within.

Moreover, IEEE 802.16m is designed to operate in a variety of **frequency bands** in a listened spectrum. The specific bands supported are 450-470 MHz, 698-960 MHz (also R1.0 target), 1710-2025 MHz, 2110-2200 MHz, 2300-2400 MHz (also R1.0 target), 2500-2690 MHz (also R1.0 target), 3400-3600 MHz (also R1.0 target). The 450-470 MHz, 1710-2025 MHz, and 2110-2200 MHz bands are not supported in previous revisions of the 802.16 specification. Details in regard to channel sizes appear to be unchanged from previous releases and are therefore not discussed here.

A goal of the 802.16m standard is backwards compatibility with previous releases. This is accomplished by employing Time Division Duplexing (TDD) between a legacy mode and a 802.16m mode of operation. To emphasize that TDD is used to separate these operation modes regardless of how the Duplexing of the connection as a whole is setup. This means that if TDD is employed for the general network, each uplink or downlink time slot is divided into legacy and 802.16m sections. Likewise, if FDD is employed for the general network, the uplink and downlink channels are each partitioned into legacy and 802.16m sections. Time is divided into Time Zones represented as an integer larger than zero that represents several consecutive subframes. A given Time Zone is either an LZone if legacy mode is employed during it, or an MZone if 802.16m mode is employed during it. This means that generally an R1MS will communicate during the LZone and an AMS will communicate during the MZone. It is possible for an AMS to communicate during the LZone, but it will operate in legacy mode and will not experience any of the benefits of using 802.16m.

For more clear presentation in the following two figures (Figure 1.10 and Figure 1.11) the IEEE 802.16m Reference Model and Protocol Structure is illustrated.

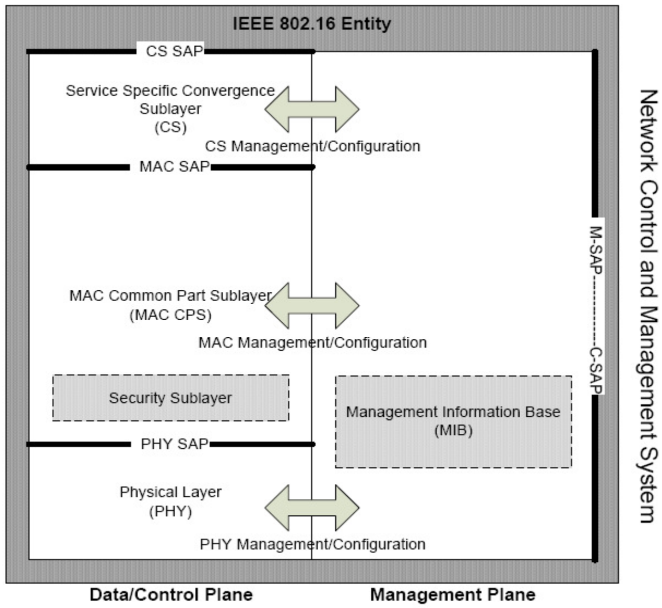


Figure 1.10. IEEE 802.16m Reference Model.

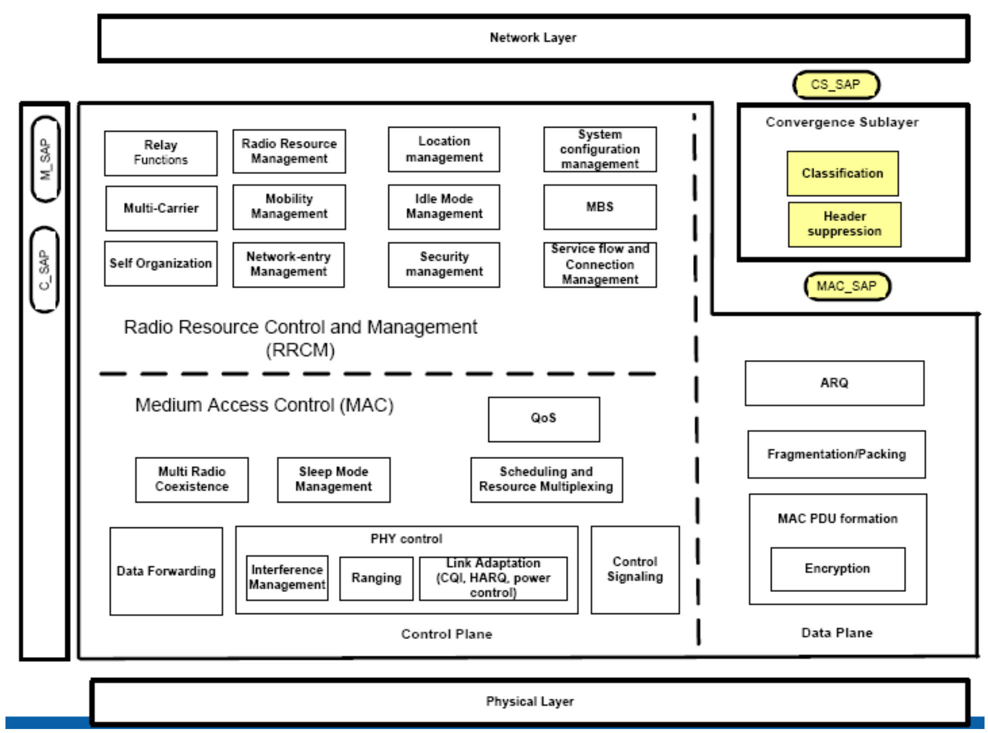


Figure 1.11. IEEE 802.16m Protocol Structure.

The general (basic) frame structure (TDD/FDD) is illustrated in Figure 1.12. Data is organized into a hierarchy of Superframes, Frames, subframes, and OFDM symbols. Superframes last 20 ms and contain four 5ms frames, each of which contains 8 subframes unless the channel size is 7 MHZ in which case frames contain 6 subframes, or 8.75 MHZ, in which case frames contain 7 subframes. Subframes fall into one of four categories, three of which are employed as part of 802.16m and a fourth that is included for legacy operation with 802.16-2009 devices operating on 8.75 MHz channels.

These types are:

- Type 1: 6 OFDMA symbols
- Type 2: 7 OFDMA symbols
- Type 3: 5 OFDMA symbols
- Type 4: 9 OFDMA symbols

For better understanding of the IEEE 802.16m OFDMA numerology that is explain above, see Figure 1.13.

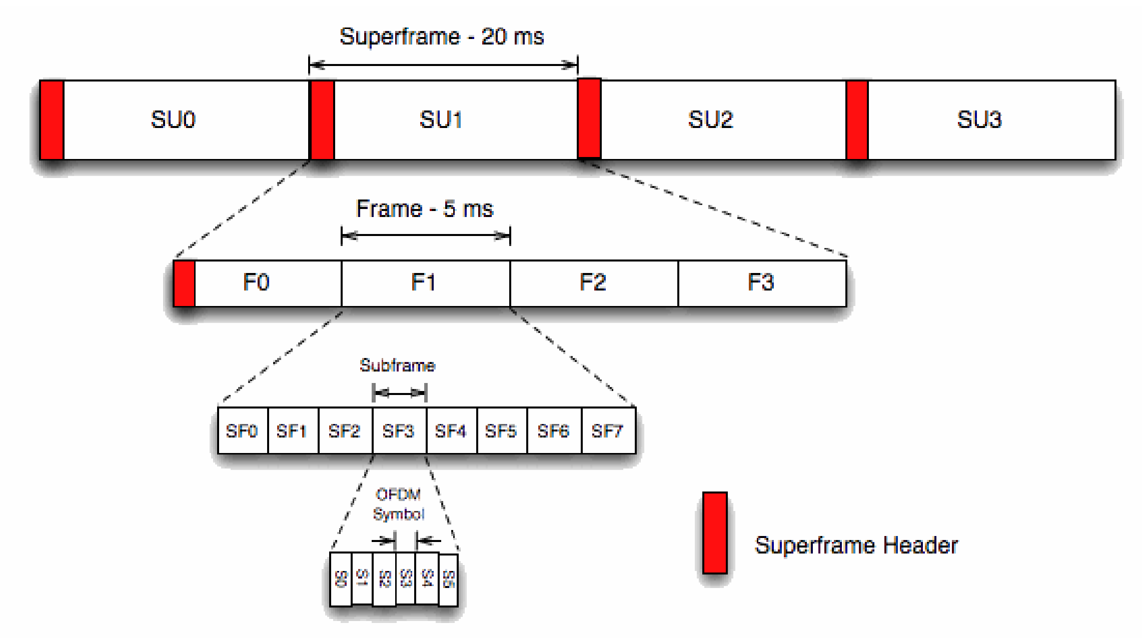


Figure 1.12. Basic Frame Structure (FDD/TDD).

IEEE 802.16m uses OFDMA in both uplink and downlink as the multiple access scheme and in the same time supports other bandwidths between 5MHz and 20MHz than listed by dropping edge tones from 10MHz or 20MHz. To emphasize that 802.16m supports both FDD and TDD modes. TDD must obviously be half duplex, but FDD operation may be either half or full duplex. ABSs must support both half and full duplex modes for AMSs operating with FDD. This means that some AMSs may be operating in half duplex FDD mode at the same time that others are operating in full duplex FDD mode on the same ABS, and still more may be operating in TDD mode.

		Nominal channel bandwidth (MHz)	5	7	8.75	10	20
		Sampling factor	28/25	8/7	8/7	28/25	28/25
		Sampling frequency (MHz)	5.6	8	10	11.2	22.4
		FFT size	512	1024	1024	1024	2048
		Sub-carrier spacing (kHz)	10.94	7.81	9.76	10.94	10.94
		Useful symbol time T_u (μs)	91.429	128	102.4	91.429	91.429
CP T_g=1/8 T_u		Symbol time T_s (μs)	102.857	144	115.2	102.857	102.857
	FDD	Number of OFDM symbols per 5ms frame	48	34	43	48	48
		Idle time (μs)	62.857	104	46.40	62.857	62.857
	TDD	Number of OFDM symbols per 5ms frame	47	33	42	47	47
		TTG + RTG (μs)	165.714	248	161.6	165.714	165.714
CP T_g=1/16 T_u		Symbol time T_s (μs)	97.143	136	108.8	97.143	97.143
	FDD	Number of OFDM symbols per 5ms frame	51	36	45	51	51
		Idle time (μs)	45.71	104	104	45.71	45.71
	TDD	Number of OFDM symbols per 5ms frame	50	35	44	50	50
		TTG + RTG (μs)	142.853	240	212.8	142.853	142.853
CP T_g=1/4 T_u		Symbol Time T_s (μs)	114.286	160	128	114.286	114.286
	FDD	Number of OFDM symbols per 5ms frame	43	31	39	43	43
		Idle time (μs)	85.694	40	8	85.694	85.694
	TDD	Number of OFDM symbols per 5ms frame	42	30	37	42	42
		TTG + RTG (μs)	199.98	200	264	199.98	199.98

Figure 1.13. Illustration of the IEEE 802.16m OFDMA numerology.

Using FDD has the clear advantage of allowing full duplex mode, but this ability comes at the cost of needing paired channels. While TDD is limited to half duplex mode, it requires only a single channel. TDD also has the advantage that it is very easy to adjust the downlink/uplink ratio by simply altering the amount of time given to each. If the circumstances require it, this could even be adjusted to the extreme case of unidirectional communication.

Multiple-Input-Multiple-Output (**MIMO**) involves using several antennas, and using processing on the differences in what is received by each antenna to allow for faster data rates. Multiple Antennas are required for 802.16m operation (though legacy modes may be operational with single antenna setups).

ABSs must support at least 2x2 (two transmitting antennas and two receiving antennas) MIMO, and may have two, four, or eight transmit antennas. AMSs must support at least 1x2 MIMO.

There are two forms of MIMO defined in the 802.16m standard. Single User MIMO (SU-MIMO) and Multiple User MIMO (MU-MIMO). These modes employ the multiple antennas in different ways, for MU-MIMO it is primarily to help differentiate between users, while in SU-MIMO it is primarily to increase data rates. Moreover in the following are summarized the key advantages and shames of DL and UL 802.16m MIMO.

- ➢ Key features of IEEE 802.16m DL MIMO:
  - ✓ Single-BS and Multi-BS MIMO
  - ✓ Single-User MIMO (SU-MIMO) and Multi-User MIMO (MU-MIMO)
    - ▪ Vertical encoding for SU-MIMO and Horizontal encoding for MU-MIMO.

- ✓ Adaptive-precoding (closed loop) and non-adaptive (open loop) MIMO precoding
- ✓ Codebook and sounding based precoding:
  - Short and long term adaptive precoding as well as Dedicated (precoded) pilots for MIMO operation
- ✓ Enhanced codebook design:
  - Enhanced base codebook, Transformed codebook, Differential codebook.

Moreover, in Figure 1.14 the IEEE 802.16m DL MIMO classification into Single and Multi BS-MIMO is illustrated.

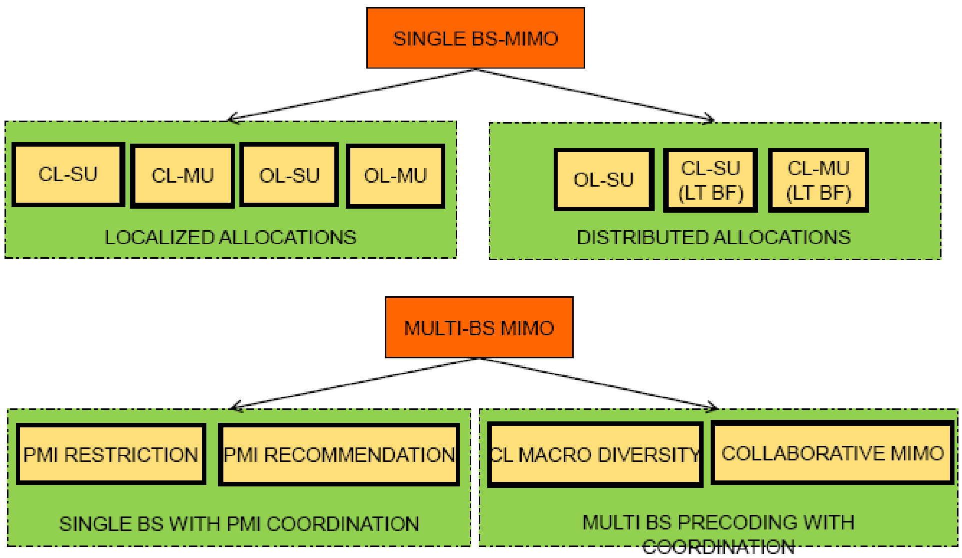
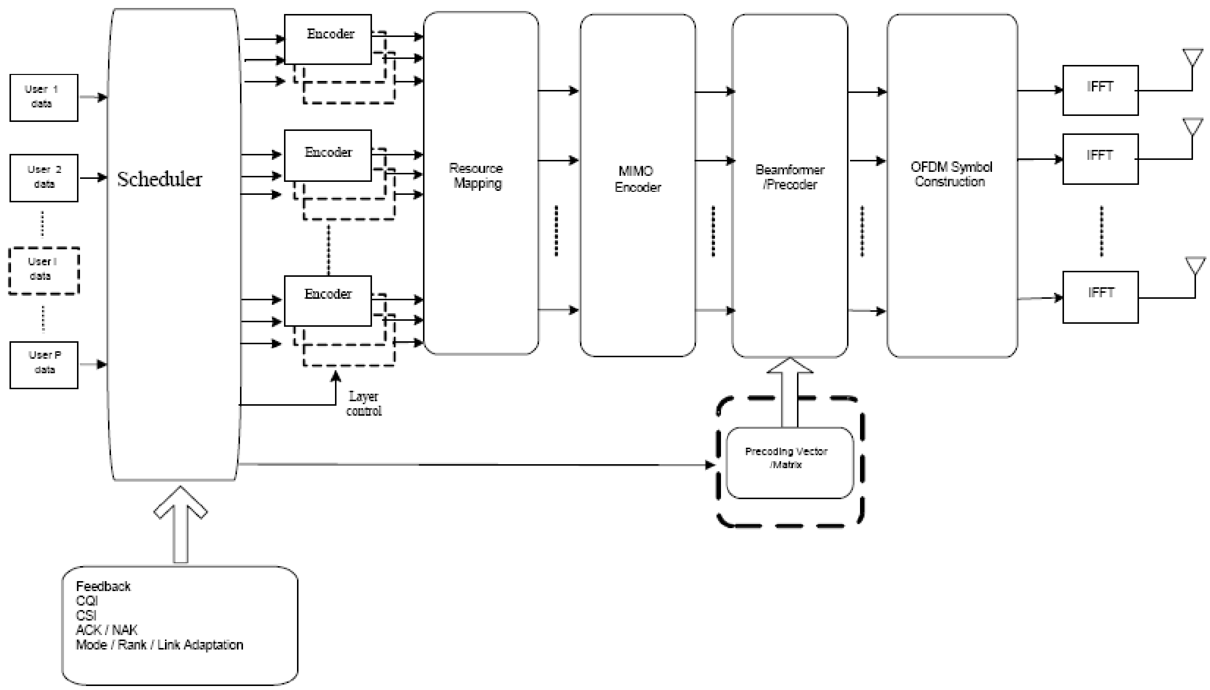


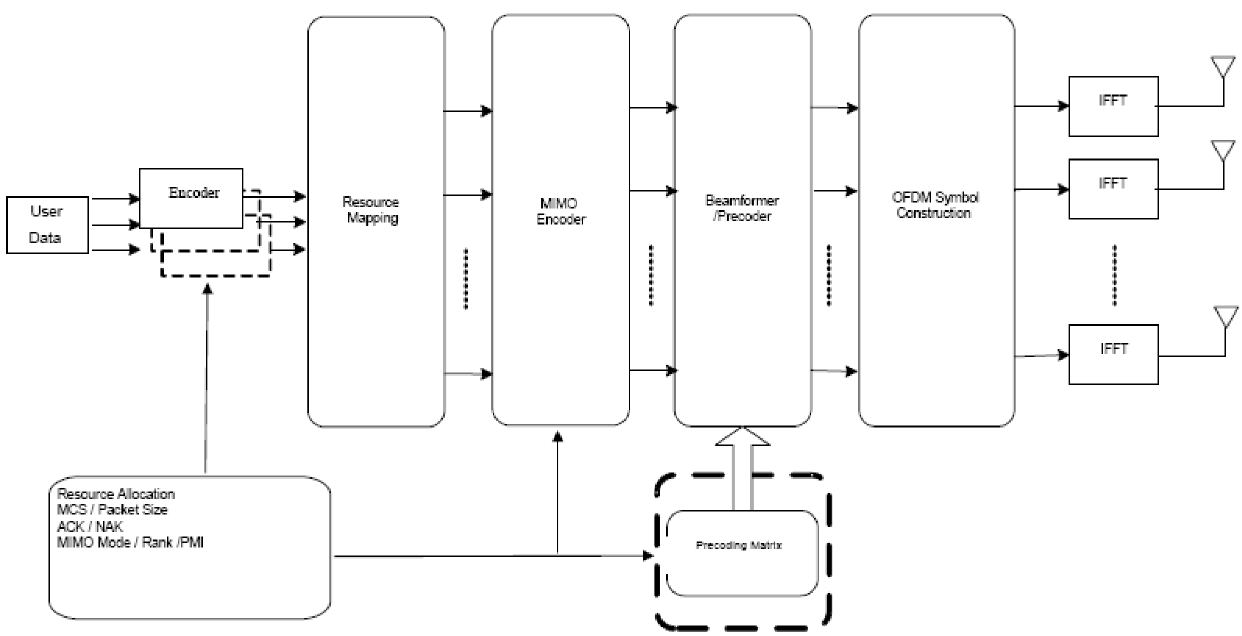
Figure 1.14. IEEE 802.16m downling MIMO classification.

- ➢ Key features of IEEE 802.16m UL MIMO:
  - ✓ Single-User MIMO (SU-MIMO) and Collaborative Spatial Multiplexing (CSM)
  - ✓ Vertical encoding for SU-MIMO and CSM
  - ✓ Open Loop and Closed Loop MIMO operation
  - ✓ Codebook based and vendor specific precoding:
    - Short and Long term precoding as well as Precoded (dedicated) pilots for MIMO operation.
  - ✓ Enhanced codebook design:
    - Enhanced base codebook for both correlated and uncorrelated channel.
    - Antenna selection codewords to reduce MS power consumption.

Furthermore, the Downlink/Uplink MIMO architectures are illustrated in Figure 1.15, and the Downlink and Uplink MIMO modes are given in Table 1.5 and Table 1.6, respectively.



a) DL MIMO architecture



b) UL MIMO architecture

Figure 1.15. Illustration of the DL and UL MIMO architectures.

Table 1.5. DL MIMO Modes.

MODE INDEX	DESCRIPTION	MIMO ENCODING FORMAT (MEF)	MIMO PRECODING
MODE 0	OL SU-MIMO	SFBC	NON-ADAPTIVE
MODE 1	OL SU-MIMO (SM)	VERTICAL ENCODING	NON-ADAPTIVE
MODE 2	CL SU-MIMO (SM)	VERTICAL ENCODING	ADAPTIVE
MODE 3	OL MU-MIMO (SM)	HORIZONTAL ENCODING	NON-ADAPTIVE
MODE 4	CL MU-MIMO (SM)	HORIZONTAL ENCODING	ADAPTIVE
MODE 5	OL SU-MIMO (TX DIVERSITY)	CONJUGATE DATA REPETITION (CDR)	NON-ADAPTIVE

	# OF TX ANTENNAS	STC RATE PER LAYER	# OF STREAMS	# OF SUBCARRIERS	# OF LAYERS
MIMO MODE 0	2	1	2	2	1
	4	1	2	2	1
	8	1	2	2	1
MIMO MODE 1 AND MIMO MODE 2	2	1	1	1	1
	2	2	2	1	1
	4	1	1	1	1
	4	2	2	1	1
	4	3	3	1	1
	4	4	4	1	1
	8	1	1	1	1
	8	2	2	1	1
	8	3	3	1	1
	8	4	4	1	1
	8	5	5	1	1
	8	6	6	1	1
	8	7	7	1	1
	8	8	8	1	1
MIMO MODE 3 AND MIMO MODE 4	2	1	2	1	2
	4	1	2	1	2
	4	1	3	1	3
	4	1	4	1	4
	8	1	2	1	2
	8	1	3	1	3
	8	1	4	1	4
MIMO MODE 5	2	1/2	1	2	1
	4	1/2	1	2	1
	8	1/2	1	2	1

Table 1.6. UL MIMO Modes.

MODE INDEX	DESCRIPTION	MIMO ENCODING FORMAT	MIMO PRECODING
MODE 0	OL SU-MIMO	SFBC	NON-ADAPTIVE
MODE 1	OL SU-MIMO (SM)	VERTICAL ENCODING	NON-ADAPTIVE
MODE 2	CL SU-MIMO (SM)	VERTICAL ENCODING	ADAPTIVE
MODE 3	OL MU-MIMO (COLLABORATIVE SM)	VERTICAL ENCODING	NON-ADAPTIVE
MODE 4	CL MU-MIMO (COLLABORATIVE SM)	VERTICAL ENCODING	ADAPTIVE

	NUMBER OF TRANSMIT ANTENNAS	STC RATE PER LAYER	NUMBER OF STREAMS	NUMBER OF SUBCARRIERS	NUMBER OF LAYERS
MIMO MODE 0	2	1	2	2	1
	4	1	2	2	1
MIMO MODE 1 AND MIMO MODE 2	2	1	1	1	1
	2	2	2	1	1
	4	1	1	1	1
	4	2	2	1	1
	4	3	3	1	1
	4	4	4	1	1
MIMO MODE 3 AND MIMO MODE 4	2	1	1	1	1
	4	1	1	1	1
	4	2	2	1	1
	4	3	3	1	1

Moreover, in a physical layer in IEEE 802.16m there is adaptive **modulation and coding**. Also, there is a Convolutional Turbo Code (CTC) with code rate 1/3, which incliding:

–FEC block sizes ranging from 48 to 4800.

–Bit grouping: solve the 64QAM degradation problem.

–FEC CRC and burst CRC.

Moreover, there are several Control channels (DL: SFH and A-A-MAP; UL: SFBCH and BW-REQ) where FEC is based on TBCC. To emphasize that here in 802.16m there is a minimal code rate of 1/4 for DL and 1/5 for UL, also it use a random puncturing with sub-block interleaver and rate-matching.

Furthermore, in order to ensure all packets to be transmitted and correctly received in 802.16m use a Hybrid Automatic Repeat Request (HARQ) on a MAC layer. While there are several variations of ARQ, 802.16m uses one primarily based on stop-and-wait. This means that when each frame is sent, the sender waits until it received an ACK (acknowledgement) before sending the next frame. Multiple HARQ channels can run in parallel (up to 16), mitigating the performance hit of waiting for an ACK before sending more data. Each of these channels has a unique identifier that is determined differently for UL and DL traffic. For DL traffic, it is simply the HARQ Channel ID (ACID). For UL traffic this identifier is a combination of the ACID and the index of the subframe containing the HARQ data.

Legacy systems used a different version of HARQ, which is supported in 802.16m as a special case. The old version is called Chase Combining, and involves retransmitting exactly the same data in the event that an ACK is not received. 802.16m uses a variant of HARQ known as Incremental Redundancy (IR). IR retransmits the same data with a (potentially) different encoding. The idea behind using different coding is that if there is some interference preventing the data from being received correctly, an encoding with more redundancy or simply a different bit-pattern may either work better, or different portions may be received allowing the incorrectly received data to be combined to create a correct copy. Chase Combining is simply the special case where the new encoding is the same as the old one.

On the other side, the MAC layer of an 802.16m network is based on the concept of connections, which are conceptualized as unidirectional data flows (each of which will generally be paired with a data flow in the opposite direction). Each flow is assigned a four-bit Flow ID (FID). The FID can be combined with a 12-bit Station ID (STID) to generate a network-unique 16-bit identifier for that flow. The separation of FIDs and STIDs is useful due to the fact that in Handovers the FIDs do not need to change. This allows all connections to be very quickly reestablished by simply changing the STID to the new value assigned by the new ABS.

There is a downside in that the legacy model allowed many more than the 16 connections per MS limit dictated by the four bit FID. Legacy systems used the full 16-bit connection ID for each connection, which allowed up to 216 connections per station. However, each of these connection IDs had to be reassigned on handover, which created significant overhead. Each flow in

802.16m may have **QoS** service parameters, which are negotiated between the ABS and AMS when the flow is setup. These parameters are the same as the parameters assigned to connections in legacy systems. An important part of QoS is bandwidth allocation. The bandwidth request protocol has been reworked for 802.16m. In legacy systems, a five-message request was needed, which specified the bandwidth grant side explicitly each time. In 802.16m there is a shorter 3-message grant request available that will automatically assume some default size and allow two messages to be skipped, thereby lowering latency.

When we mention **handover processes** in 802.16m, there are four cases for the handover of an AMS from the currently Serving-BS (S-BS) to a new Target-BS (T-BS). These cases are the following:

1. R1BS → R1BS
2. ABS → R1BS
3. R1BS → ABS
4. ABS → ABS

The first case simply uses the legacy handover protocol, and the second uses the legacy handover protocol in the LZone of the S-ABS. The third uses the legacy handover protocol in the LZone of the T-ABS, but after handover is complete if the MS is an AMS, it can transition to the MZone of the ABS it is now associated with. The fourth case, in which an AMS is handed over from one ABS to another ABS uses a new protocol comprised of three phases: HO-Initiation, HO-Preparation, and HO-Execution. The HO-Initiation phase is only necessary if the Handover is being started by the AMS, and is comprised only of the AMS sending a signal to the S-ABS requesting a handover.

The HO-Preparation step involves the S-ABS sending authentication and identification information about the AMS to one or more potential T-ABS via the network backbone. Each potential T-ABS then performs ranging on the AMS, and the results are returned to the AMS, which then selects which of the potential T-ABSs will actually be targeted. Note that this step can be skipped if there is only one potential T-ABS. Finally control information is sent from the S-ABS to the T-ABS giving information about whether the handover will be soft or hard, information about the data flows on the AMS (and any other information necessary to optimize the handover), and the time at which the handover will complete and the AMS will lose contact with the S-ABS.

The HO-Execution Phase involves the AMS and T-ABS going through the network reentry protocol at the appropriate time (designated at the end of the last phase). The AMS disconnects from the S-ABS either before or after this completes (as appropriate). If it occurs afterwards (the handover is soft) then the AMS employs TDD to maintain communication with both ABSs during the handover. The Reentry procedure is largely unchanged from earlier releases, except that some extra ranging and channel data may be specified to avoid interference. A Handover will always result in some downtime in network communications, even if only to allow for control data to be sent. Table 1.7 provides a breakdown of the maximum acceptable downtime for different types of handovers between 802.16 BSs.

Table 1.7. 802.16m Handover Downtime

Handover Type		Max Downtime (ms)
Intra-Frequency		27.5
Inter-Frequency	Within a spectrum band	40
	Between spectrum bands	60

Moreover, IEEE 802.16m network can also deal with handovers of some MS to a different Radio Access Technology (RAT). The IEEE 802.21 standard for Media Independent Handover is supported to accomplish this. The 802.16m standard defines specific handover procedures for several other RATs, including 802.11, GSM/EDGE, UTRA, E-UTRA, 3GPP2, and CDMA2000.

As one interesting advantage of IEEE 802.16m technology is the capability of supporting **Femtocells**. A Femtocell is a short range ABS with very low transmit power compared to a normal (macro) ABS. They are generally set up either for home/small office use or as a hotspot and connect to a conventional wired network backbone to provide network connectivity. There are three varieties of Femto ABS, which primarily relate to policies surrounding Closed Subscriber Groups (CSG). The three varieties are: CSG-Closed, CSG-Open, and OSG (Open Subscriber Group). CSG-Closed only allows members of the CSG to associate with the Femto ABS (and emergency access, as required by law). CSG-Open allows anyone to connect, but gives members of the CSG priority (and will not allow non-CSG members to connect if it would damage QoS to members). It also allows anyone to connect for emergency services. An OSG allows anyone to connect with the same priority.

Femtocells are identified in the SA-Preamble, which both identifies an ABS as a Femto-ABS and identifies the type of Femto-ABS. AMSs can maintain a MAC-Address based list of Femto-ABSs for which they are a member of the CSG . Also, there are several special cases regarding handover to a Femto ABS. If the ABS is CSG-Closed and the AMS is not in the CSG, handover is only attempted in the case of emergency. If the ABS is CSG-Open and the AMS is not in the CSG, handover is only attempted in the case of emergency or if that is the only way to maintain a stable connection. The network will attempt to perform load balancing on Femto-ABSs and the overlapping macro ABS. It is actually done by Femto ABSs initiating handovers when their load is too high.

Because the low range of Femto-ABSs can result in relatively frequent handover between the Femto-ABS and the overlapping macro ABS, Femto ABSs and AMSs associated with them may cache significant amounts of information to speed up handovers. While this is also possible strictly between macro AMSs, it is not worth it due to the large range making handovers rare.

Femto ABSs have a low duty mode similar to power-save mode in a MS. The goal of this mode is not to conserve power (as they will generally not be reliant on batteries), but rather to avoid excessive noise in the frequency band that could cause interference with neighboring Femto ABSs or communication between an AMS and a macro ABS. A Femto-ABS can only enter this low duty mode if no AMSs are associated with it.

As a conclusion for IEEE 802.16m, we can say that 802.16m provides a number of improvements over legacy systems while maintaining backwards compatibility. The WiMAX forum is currently in the process of updating the new WiMAX Rel 2.0 based on the 802.16m standard. The PHY layer has been changed in several ways. More frequency bands are usable than in previous revisions. The Frame structure has been updated to provide more variations on subframe structure. Better support for MIMO and more Duplexing options were added (now both TDD and FDD are supported). A system was also put in place to ensure backwards compatibility with older releases.

The MAC layer also has a number of changes. The HARQ system was redesigned to use incremental redundancy instead of chase combining. Significant changes to how connections are processed (and QoS updates to work with the new format) were made with the introduction of the concept of data flows. The Handover procedure received a significant amount of work, especially in defining handover to and from legacy BSs. Support for Femtocells was made significantly more robust.

All of above changes resulted in significant improvements to performance, with a throughput gain of at least 2x and a decrease in latency in all cases, undoubtedly place IEEE 802.16m in the IMT-Advanced family of networks.

1.4. ITU NGN model and mobile networks

Nowadays we are all witnesses of the tremendous grow in the world of telecommunication and information technologies and services. In this grow the next generation networks (NGN) provide one new reality by many factors such as: the need to converge and optimise the operating networks and the extraordinary expansion of digital traffic (i.e., increasing demand for new multimedia services, increasing demand for mobility, All-IP concept, and etc. While different services converge at the level of digital transmission, the separation of distinct network “layers” in the NGN structure (transport, control, service and applications functions – see Figure 1.16) provides support for competition and innovation at each horizontal level. At the same time NGNs also create strong commercial incentives for network operators to bundle, and therefore increase vertical and horizontal integration, leveraging their market power across these layers. This may bring about the need for closer regulatory and policy monitoring, in order to prevent the restriction of potential development of competition and innovation in a next generation environment, and therefore the risk of reducing benefits for consumers and the potential of new networks for economic growth and for providing multimedia services with high level of QoS provisioning.

The NGN standardization work started in 2003 within ITU-T, and is worldwide today in various major telecom standardization bodies. The most active NGN relevant standardization bodies are ITU, ETSI, ATIS, CJK and TMF. The Next Generation Mobile Networks (NGMN) initiative is a major body for mobile-specific NGN activities, which are important contributors to the 3GPP specification for NGMN. Although there is a significant amount of work underway in standardisation forums on NGN, at the policy level, there is a still not complete agreement on a specific definition of “NGNs”. The term is generally used to depict the shift to higher network speeds using broadband, the migration from the PSTN to an IP-network, and a greater integration of services on a single network, and often is representative of a vision and a market concept. According to ITU-T Recommendation Y.2001 (12/2004) General overview of NGN is described with the following definition:

“**A Next Generation Networks (NGN) is a packet-based network able to provide Telecommunication Services to users and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent of the underlying transport-related technologies. It enables unfettered access for users to networks and to competing service providers and services of their choice. It supports generalised mobility which will allow consistent and ubiquitous provision of services to users.**”

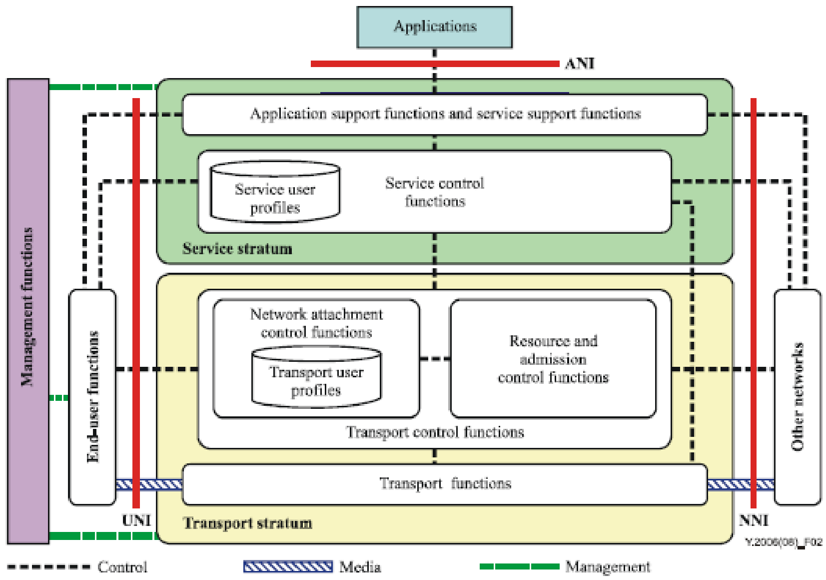


Figure 1.16. Separation of the NGN functional plans in two stratums.

If we go back in history of NGN, we can say that its beginning was in late 2003. Moreover, in 2003 year, under the name JRG-NGN (Joint Rapporteur Group on NGN), the NGN pioneer work was initiated. The key study topics are:

- ✓ NGN requirements;
- ✓ the general reference model;
- ✓ functional requirements and architecture of the NGN;
- ✓ evolution to NGN.

Moreover, the two fundamental recommendations on NGN are:

- ➢ Y.2001: 'General overview of NGN'.
- ➢ Y.2011: 'General principles and general reference model for next-generation networks'.

These two documents comprise the basic concept and definition of NGN. In May 2004, the FG-NGN (Focus Group on Next Generation Networks) was established in order to continue and accelerate NGN activities initiated by the JRG-NGN. FG-NGN addressed the urgent need for an initial suite of global standards for NGN. The NGN standardization work was launched and mandated to FG-NGN.

On 18 November 2005, the ITU-T published its NGN specification Release 1, which is the first global standard of NGN and marked a milestone in ITU's work on NGN. The NGN specification Release 1, with 30 documents, specified the NGN Framework, including the key features, functional architecture, component view, network evolution, etc. The Figure 1.17 illustrates one key part of this NGN Release 1: configuration of transport and service layer. Lacking protocol specifications, the ITU NGN Release 1 is not at an implementable stage; however, it is clear enough to guide the evolution of today's telecom networks. With the release of NGN Release 1, the FG-NGN has fulfilled its mission and closed.

Following FG-NGN, the ITU-T NGN standardization work continues under the name GSI-NGN (NGN Global Standards Initiative) in order to maintain and develop the FG-NGN momentum. In parallel with the FG-NGN, there are two other groups working on the NGN relevant issues. They are the NGN-MFG (NGN Management Focus Group) and the OCAF-FG (Open Communication Architecture Forum Focus Group), directly contributing to the GSI-NGN.

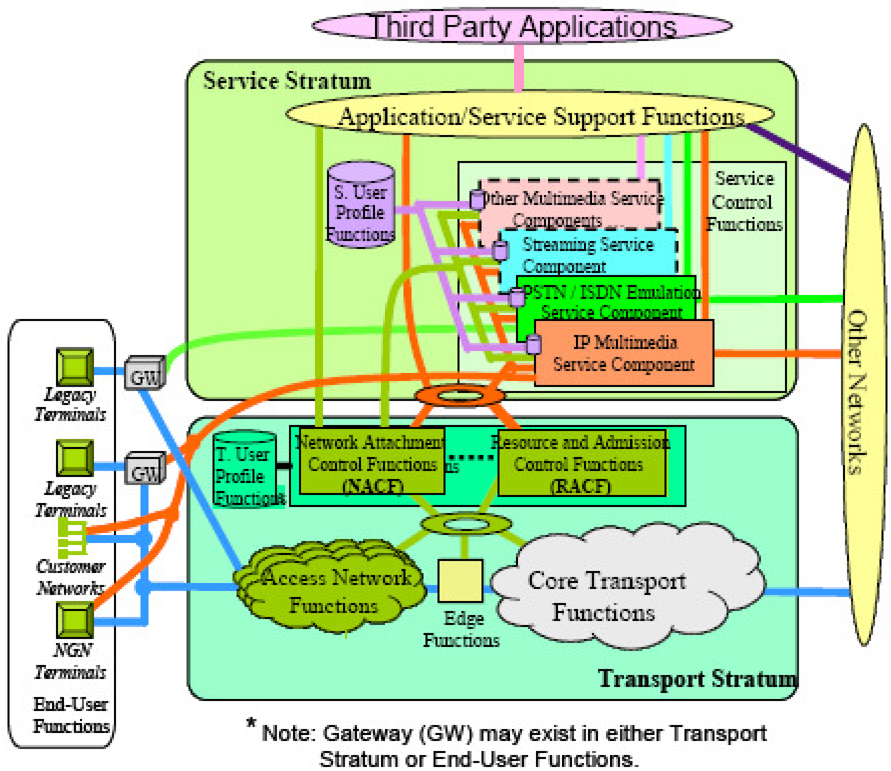


Figure 1.17. Transport and service configuration of NGN.

As we mention before, ITU coordinates the global efforts (including governments, regional and national SDOs, industry forums, vendors, operators, etc.) in developing the ITU recommendations. Moreover, the ITU takes a three-stage approach as follows to develop the NGN standards:

(a) Stage 1: identify service requirements;

(b) Stage 2: describe network architecture and functions to map service requirements into network capabilities;

and (c) Stage 3 : define protocol capabilities to support the services.

All multimedia services and capabilities have to be specified to stage 3 to ensure that the standards are implementable. ITU's NGN specifications are mainly contained in the Y-series and Q-series recommendations. Some of the NGN related recommendations are listed in Table 1.8.

Table 1.8. NGN related ITU recommendations

Recommendation	Short title
Y.2001	General overview of NGN
Y.2002	Overview of ubiquitous networking and of its support in NGN
Y.2006	Description of capability set 1 of NGN release 1
Y.2007	NGN capability set 2
Y.2011	General principles and general reference model for Next Generation Networks
Y.2012	Functional requirements and architecture of the NGN
Y.2013	Converged services framework functional requirements and architecture
Y.2014	Network attachment control functions in next generation networks
Y.2015	General requirements for ID/locator separation in NGN
Y.2016	Functional requirements and architecture of the NGN for applications and services using tag-based identification
Y.2017	Multicast functions in next generation networks
Y.2018	Mobility management and control framework and architecture within the NGN transport stratum
Y.2019	Content delivery functional architecture in NGN
Y.2021	IMS for Next Generation Networks
Y.2031	PSTN/ISDN emulation architecture
Y.2051	General overview of IPv6-based NGN
Y.2052	Framework of multi-homing in IPv6-based NGN
Y.2053	Functional requirements for IPv6 migration in NGN
Y.2054	Framework to support signalling for IPv6-based NGN
Y.2091	Terms and definitions for Next Generation Networks
Y.2111	Resource and admission control functions in Next Generation Networks
Y.2112	A QoS control architecture for Ethernet-based IP access network
Y.2113	Ethernet QoS control for next generation networks
Y.2121	Requirements for the support of flow state aware transport technology in an NGN
Y.2122	low aggregate information exchange functions in NGN
Y.2171	Admission control priority levels in Next Generation Networks
Y.2172	Service restoration priority levels in Next Generation Networks
Y.2173	Management of performance measurement for NGN
Y.2174	Distributed RACF architecture for MPLS networks
Y.2175	Centralized RACF architecture for MPLS core networks
Y.2201	NGN release 1 requirements
Y.2205	Next Generation Networks - Emergency telecommunications – Technical considerations
Y.2206	Requirements for distributed service network (DSN)
Y.2211	IMS-based real time conversational multimedia services over NGN
Y.2212	Requirements of managed delivery services
Y.2213	NGN service requirements and capabilities for network aspects of applications and services using tag-based identification
Y.2214	Functional model for customized multimedia ring service
Y.2215	Requirements and framework for the support of VPN services in NGN including mobile environment
Y.2216	NGN capability requirements to support multimedia communication centre (MCC) service
Y.2221	Requirements for support of ubiquitous sensor network (USN) applications and services in the NGN environment
Y.2232	NGN convergence service model and scenario using Web Services
Y.2233	Requirements and framework allowing accounting and charging capabilities in NGN
Y.2234	Open service environment capabilities for NGN

Y.2235	Converged web-browsing service scenarios in NGN
Y.2236	Framework for NGN support of multicast-based services
Y.2237	Functional model, service scenarios and use cases for QoS enabled mobile VoIP service
Y.2261	PSTN/ISDN evolution to NGN
Y.2262	PSTN/ISDN emulation and simulation
Y.2271	Call server based PSTN/ISDN emulation
Y.2401	Principles for the Management of the Next Generation Networks
Y.2601	Fundamental characteristics and requirements of future packet based networks
Y.2611	High level architecture of future packet based networks
Y.2612	Generic requirements and framework of FPBN addressing, routing and forwarding
Y.2613	The general technical architecture for public packet telecommunication data network (PTDN)
Y.2701	Security requirements for NGN release 1
Y.2702	Authentication and authorization requirements for NGN release 1
Y.2703	The application of AAA service in NGN
Y.2704	Security mechanisms and procedures for NGN
Y.2720	NGN identity management framework
Y.2721	NGN identity management requirements and use cases
Y.2801	Mobility management requirements for NGN
Y.2802	Fixed-mobile convergence general requirements
Y.2803	FMC service using legacy PSTN or ISDN as the fixed access network for mobile network users
Y.2804	Generic framework of mobility management for next generation networks
Y.2805	Framework of location management for NGN
Y.2806	Framework of handover control for NGN
Y.2807	MPLS-based mobility capabilities in NGN
Y.2808	Fixed mobile convergence with a common IMS session control domain
Y.2901	The carrier grade open environment reference model
Y.2902	Carrier grade open environment components

Moreover, despite of the definition for NGN (given within the ITU Recommendation Y.2001), NGN is also defined as “broadband managed IP networks”, includes next generation “*core*” networks, which evolve towards a converged IP infrastructure capable of carrying and convergating a multitude of services, such as voice, video and data services, and next generation “*access*” networks (wireless and wire-line), *i.e.* the development of high-speed local loop networks that will guarantee the delivery of innovative services.

In the following, the NGN fundamental aspects are summarized:

- ✓ Packet-based transfer
- ✓ Separation of control functions among bearer capabilities, call/session, and application/service
- ✓ Decoupling of service provision from transport, and provision of open interfaces
- ✓ Support for a wide range of services, applications and mechanisms based on service building blocks (including real time/streaming/non-real time services and multi-media)
- ✓ Broadband capabilities with end-to-end QoS and transparency
- ✓ Interworking with legacy networks via open interfaces

- ✓ Generalised mobility
- ✓ Unfettered access by users to different service providers
- ✓ A variety of identification schemes which can be resolved to IP addresses for the purposes of routing in IP networks
- ✓ Unified service characteristics for the same service as perceived by the user
- ✓ Converged services between Fixed and Mobile networks
- ✓ Independence of service-related functions from underlying transport technologies
- ✓ Support of multiple last mile technologies
- ✓ Compliant with all Regulatory requirements, for example concerning emergency communications and security/privacy, etc.

And moreover, the six key ITU NGN concepts (six criteria) are the following:

- ✓ Packet-oriented network;
- ✓ Integration of existing infrastructure;
- ✓ Support broad variety of services;
- ✓ Openness and flexibility regarding new services;
- ✓ Application focused-access independent;
- ✓ Separation into different layers using open interfaces.

The definition of next generation ***core*** **networks** are defined on the basis of their underlying technological "components" that include – as mentioned in the ITU definition – packet-based networks, with the service layer separated by the transport layer, which transforms them into a platform of converged infrastructure for a range of previously distinct networks and related services. These features may have an impact on traditional business models and market structure, as well as on regulation:

- ➢ ***IP-based network*****:** "Next generation core networks" generally cover the migration from multiple legacy core networks to IP-based networks for the provision of all services. This means that all information is transmitted via packets. Packets can take different routes to the same destination, and therefore do not require the establishment of an end-to-end dedicated path as is the case for PSTN-based communications. The core will be IP Multimedia Subsystem (IMS) based. For IMS in more details can be found in section 2.4 in the following module.

- ➢ ***Packet-based, multi-purpose*****:** While traditionally separate networks are used to provide voice, data and video applications, each requiring separate access devices, with NGN different kinds of applications can be transformed into packets, labelled accordingly and delivered simultaneously over a number of different transport technologies, allowing a shift from single-purpose networks (one network, one service), to multi-purpose networks (one network, many services). Interworking between the NGN and existing networks such as PSTN, ISDN, cable, and mobile networks can be provided by means of media gateways.

- ***Separation of transport and service layer***: This constitutes the key common factor between NGN and convergence, bringing about the radical change in relationship between network "layers" (transport infrastructure, transport services and control, content services and applications). In next generation networks service-related functions are independent from underlying transport-related technologies (see Figure 1.18). The uncoupling of applications and networks allow applications to be defined directly at the service level and provided seamlessly over different platforms, allowing for market entry by multiple service providers on a non-discriminatory basis.

These features may foster the development and provision of new services and constitute a new opportunity for innovation, allowing different market players to create value at the separate functional levels of access, transport, control and services. However, while initially it was a common assumption that this layered structure would lead to a market model where services could be increasingly provided across the value chain, in a more decentralised manner, today it appears that the network provider will decide whether the "horizontal" model will prevail, or whether they will simply (commercially) vertically integrate transport and services across functional levels, offering bundled services.

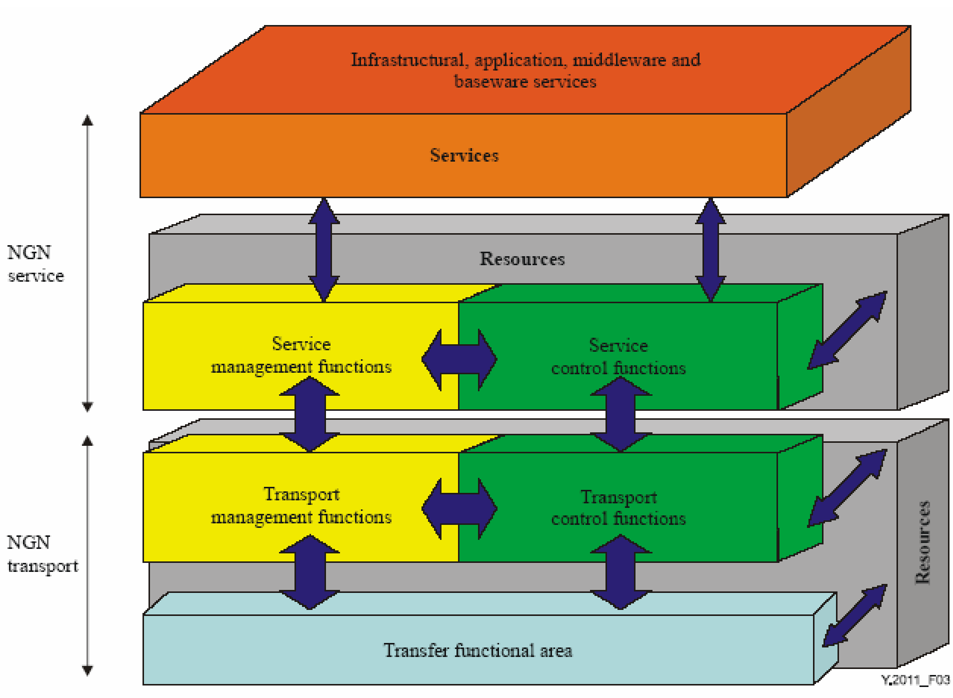


Figure 1.18. General NGN functional model.

Currently, bundling of a variety of services is a key trend in the sector, bringing greater competition between formerly distinct sectors. Bundles include all sorts of combinations of fixed and mobile voice calls, Internet access and media/entertainment services. With services and transport commercially integrated at the vertical level, customers are somehow “locked-in” in a vertical relationship with a single operator. This is not negative in itself, as packages are often more convenient, or easier to use, at the same time it is important to maintain the possibility for users to choose which services they want to purchase, and to have clear information about the cost and characteristics of these. The risk would be to create a situation in which the network provider may limit the possibility of users to access IP-based services and applications provided by third parties. Considering the economic drivers behind the shift towards next generation networks, there is an incentive for the network provider to also become an integrated market player, in order to maintain/extend their user base or benefit from a privileged relationship with subscribers. This raises questions regarding obligations for access to networks by service providers and issues of traffic prioritisation. In this context access plays an important role for all service providers to be able to provide their content, services and applications to end users.

One essential feature of next generation mobile networks is the capability to support “**generalised mobility which will allow consistent and ubiquitous provision of services to users**”. Although core next generation networks tend to be on a fixed infrastructure, the possibility to improve interconnection with mobile networks is being explored, and standardisation organisations as well as operator and manufacturers associations are working to the development of appropriate standards.

In addition, the deployment of wireless infrastructures facilitates access to IP networks, and the adoption of increasingly sophisticated devices and handsets will allow an easy access to IP services from anywhere. The migration process towards IP-NGN potentially entails several structural changes in the core network topology, such as the rearrangement of core network nodes and changes in the number of network hierarchy levels. As a result, an overall reduction in the number of points of interconnection will take place, especially with regard to interconnection points at the lowest level. This could negatively affect alternative operators whose previous interconnection investment may become stranded.

On the other side, the definition of next generation ***access*** **networks** is usually specific to investment in fibre in the local loop, *i.e*. fibre replacing copper local loops, able to deliver next generation access services – *i.e*. an array of innovative services, including those requiring high bandwidth (voice, high-speed data, TV and video). In general, this is the definition used in a number of national initiatives in examining NGN. However, while next generation access networks tend to refer to a specific technological deployment, there are other technologies which can compete in providing some of the services which it is envisaged will be provided by NGNs. There are also other technologies which may not be able to fully compete with NGN access networks in terms of capacity and the plethora of bundled offers which NGNs can provide, but may be perfectly suitable for users who do not have the need for high capacity access.

The different technologies (nowadays available) include **wire-line access**: existing copper networks upgraded to DSL, coaxial cable networks, powerline communications, optical networks and ect., and **high speed wireless and mobile networks** (LTE, LTE-Advanced, WiMAX 802.16m, HSPA and ect.), or **hybrid deployments** of these technologies. Although fibre, in particular point-to-point fibre development, is often described as the most "future proof" of network technologies to deliver next generation access, there are likely to be a number of alternative and complementary options for deployment of access infrastructures by incumbent telecommunications operators, and new entrants. In Figure 1.19 the different levels of NGN architecture together with the NGN access networks are given.

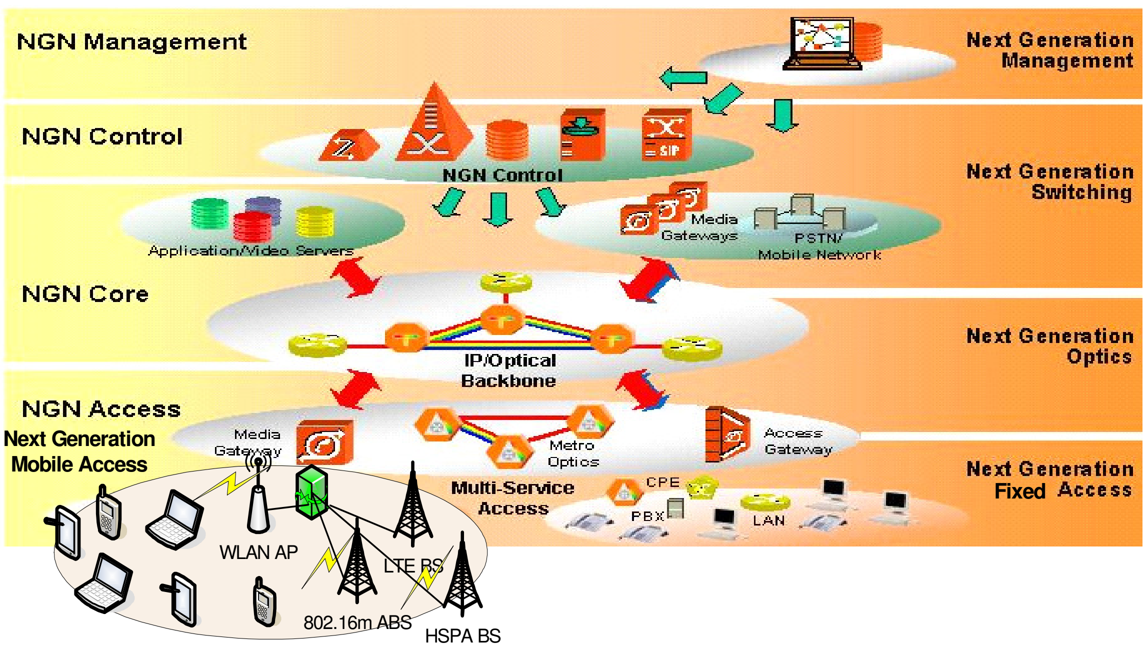


Figure 1.19. Illustration of different levels of NGN architecture.

Due to the aim of this section, furthermore, we will focus on mobile and wireless NGN access networks, (the wireline networks will be discussed in some another place).

Broadband Wireless Access (BWA)

BWA technologies aim at providing high speed wireless access over a wide area. Certain early fixedwireless access technologies, such as Local Multipoint Distribution Service (LMDS) and Multichannel Multipoint Distribution Service (MMDS), never gained widespread market adoption. WiMAX technologies, – the IEEE 802.16 set of standards that are the foundation of WiMAX certification, and similar wireless broadband technologies, are expected to address some of these shortcomings, and fill market gaps left by wired networks, or compete with wired access providers. The WiMAX Forum has estimated that new WiMAX 802.16m (which belong to 4G mobile technology) equipment will be capable of sending high-speed data over long distances (a theoretical 300 Mbit/s over 3 to 10 kilometres, in a line-of-sight fixed environment). When users are connected at the same time, capacity sharing will significantly reduce speeds for individual users sharing the same capacity.

WLAN (or wireless fidelity) refers to wireless local area networks that use one of several standards in the 802.11 family. Previously we discussed about the capability of WLAN 802.11n to reach 600 Mbps (using 4x4:4 MIMO). Moreover any type of WLAN can be deployed without cabling for client devices, typically reducing the costs of network deployment and expansion. Due to its affordability, scalability and versatility, its popularity has spread to rural and urban area. WLAN range is usually limited to about 45 m indoor and 90 m outdoors, however WLAN technologies can also be configured into point-to-point and point-to-multipoint networks in order to improve their range and provide last mile fixed wireless broadband access. One way to serve remote areas which cannot be reached with the above-mentioned technologies is with wireless “mesh” solutions. They often include a satellite backhaul connection through Very Small Aperture Terminals, usually coupled with wireless technologies such as WLAN. This combination allows access to telecommunication and data services even to more remote areas, albeit with limited (and expensive) bandwidth. Terrestrial wireless services offer the opportunity to deploy competing access infrastructure. However, they may offer different service characteristics to fixed-line services in terms of coverage, symmetry and speeds. These networks may be less suitable to deliver sustained high bandwidth connections for larger numbers of users, or for high bandwidth applications, such as High Definition TV on demand. In addition, wireless service deployments are constrained by spectrum availability. At the same time, the economics of their deployment is often relatively scalable, which means that they have lower economic barriers to entry compared to fibre deployments. Therefore while not a complete substitute, they can complement wireline networks and be an alternative provider in certain areas or for specific services.

3G mobile networks

The term next generation networks frequently encompasses some kind of fixed-mobile convergence (FMC), as it allows the transition from separate network infrastructures into a unified network for electronic communications based on IP, which facilitate affordable multiple play business models, seamlessly integrating voice, data and video. The introduction of 3G technology supports the transmission of high-speed data with speeds theoretically reaching 2/4 Mbit/s, and third-generation handsets give users access to the Internet and multimedia content on the go. In addition, new handsets in countries such as Japan, Korea, Italy or the United States allow users to access innovative, dedicated terrestrial and in some cases satellite television networks. Operators are expanding their 3G networks across the OECD and this will provide higher data speeds to users, who will be able to access innovative networks dedicated to providing mobile video or television programming. In 2005, 11% of all OECD mobile subscribers were on a 3G network, which offered a broader “blanket” data coverage to users. However, existing 3G technologies will need to be upgraded in order to support very high bandwidth or extensive concurrent usage that may be demanded by users in the future.

4G mobile networks

As it was mentioned before, the next generation wireless mobile communications will be shifted from today's traditional transmission-specific radio technology to an interface-based technology in order to be more compatible with computer system architecture. The future mobile device will therefore be first and foremost a computer, then an open wireless architecture (OWA) low-power terminal. This OWA technology offers an optimal solution to open up the wireless platform for complete openness and simplicity and would support the service-oriented architecture and infrastructure that is necessary for future mobile phone development and advancement.

Though different regions have diversified approaches towards the next generation mobile communication technology (4G Mobile Networks), the future trend can be described shortly with the following processes: Convergence among fixed, mobile and wireless communications, All-IP concept, seamless vertical handovers, high speed data transfer, omnipresence, advance mobility and ubiquitous provision of any services to users (any-place, any-time, any-where on any devices). No single wireless radio transmission technology (RTT) can do both broadband high-speed data-rate and seamless mobility, and therefore we need multiple RTTs to complement each other in any optimal way to ensure the information is delivered to the mobile user in a more cost-effective way and in a more spectrum-efficient way.

The future evolution of mobile networks for example using LTE technology and LTE-Advanced (4G technology) or/and IEEE 802.16m – a next generation mobile technology – may significantly increase speeds, enabling high peak data rates of 300 Mbit/s downlink and 100Mbit/s uplink. However, deployment of IMT-Advanced technologies may not begin, at the earliest, before 2011.

Satellite

Satellite services are typically dedicated to direct-to-home television and video services, satellite radio, and specialised mobile telephony uses. More recently technological advances – such as spot beam technology and data compression algorithms – increased technical efficiency in satellite communications, enabling more efficient use of spectrum, and reducing redundancy, thus increasing effective data density and reducing required transmission bandwidth. Satellite broadband is usually provided to the customer via geosynchronous satellite. Ground-based infrastructure includes remote equipment consisting of a small antenna and an indoor unit. Gateways connect the satellite network to the terrestrial network. Except for gateway locations, satellite broadband is independent of terrestrial infrastructure such as conduits and towers, allowing it to provide coverage also to remote areas.

In this context, several operators began to offer broadband satellite service to residential consumers, especially to those in areas not otherwise reached by broadband networks, at affordable prices, and at speeds comparable to those offered by some wired broadband services.

For example, Wildblue in the United States offers broadband connectivity (512Kbps downstream and 128Kbps upstream) for about USD 50/month. In Europe, Eutelsat and Viasat jointly launched consumer broadband satellite service, targeting underserved markets in European countries. While technological developments allowed satellite services to offer significantly higher capacity and improved performance, there are still some challenges to users of satellite connectivity. In particular, limited upstream capacity will restrain the possibility of users to benefit from new Web 2.0 opportunities, while latency issues will continue to limit the usability of satellite for certain broadband services and applications (*e.g.* voice and video conferencing) and speeds are expected to be significantly lower than can be offered by fibre networks.

1.5. All-IP concept for 4G

The mobile communication generations has traversed a long way through different phases of evolution since its birth (with appearing of 1G mobile networks) early in the 1970s. The incredible global boom in the number of mobile users each year has periodically spurned the development of more and more sophisticated technologies trying to strike the right chord primarily in terms of provision of seamless global roaming, quality services and high data rate.

Nowadays numerous different generation technologies with their individual pros and cons are existing globally. The coming era of **4G** systems is foreseeing a potential smooth merger of all these heterogeneous technologies with a natural progression to support seamless cost-effective high data rate global roaming, efficient personalized services, typical user-centric integrated service model, high Qos and overall stable system performance with All-IP concept within. However, every step in such technological advancements presents huge research challenges. This section aims to focus upon some of these potential challenges along with different proposed feasible and non-feasible concepts in the areas of mobile terminals and users, mobile services, mobile and wireless access networks, and communication, in order to give an indepth view of the next-generation communication systems. Moreover, ITU IMT-Advanced announced that current versions of LTE-Advanced, WiMax 802.16m and other evolved 3G technologies that do not fulfill "IMT-Advanced" requirements could be considered "4G", provided they represent forerunners to IMT-Advanced and "a substantial level of improvement in performance and capabilities with respect to the initial third generation systems now deployed."

The term 4G is used broadly to include several types of broadband wireless access communication systems, not only cellular telephone systems. One of the terms used to describe 4G is **MAGIC**—**M**obile multimedia, **a**nytime anywhere, **G**lobal mobility support, **i**ntegrated wireless solution, and **c**ustomized personal service. As a promise for the future, 4G systems, that is, cellular broadband wireless access systems have been attracting much interest in the mobile communication arena. The 4G systems not only will support the next generation of mobile service, but also will support the fixed wireless networks. This obviously invites new challenges on every step and researchers worldwide face an uphill task of designing suitable solutions. Figure 1.20, shows such a 4G vision. Moreover, the features of 4G systems might be summarized with one word- **Integration**. The 4G systems are about seamlessly integrating terminals, networks, and applications to satisfy increasing user demands. The continuous expansion of mobile communication and wireless networks shows evidence of exceptional growth in the areas of mobile subscriber, wireless network access, mobile services, and applications. An estimate of 1 billion users by the end of 2013 justifies the study and research for 4G systems.

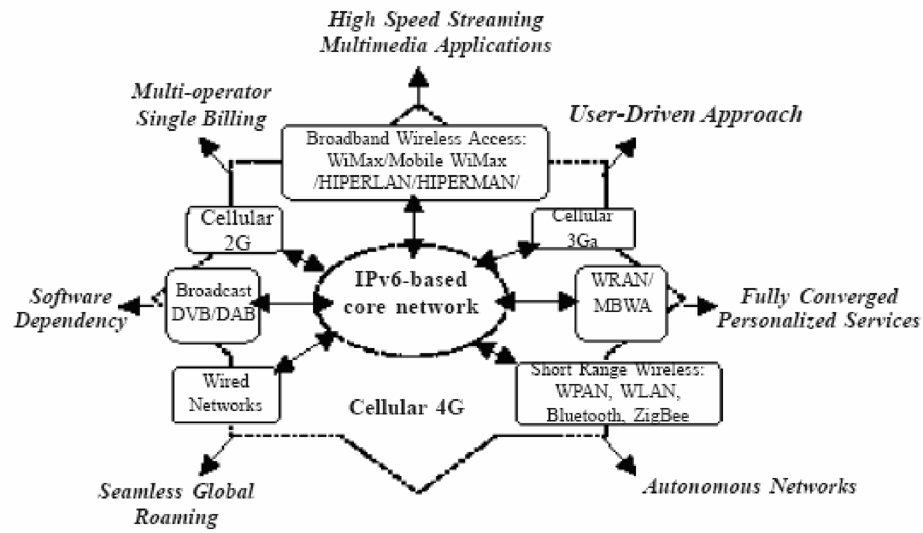


Figure 1.20. Illustration of one 4G vision.

The third-generation (3G) access networks, like WCDMA (wideband code division multiple access) and cdma2000, have a complicated network structure and define many protocols to cover the system structure. Accordingly, it is expected that 4G networks will have a simple structure based on **all-IP** where Internet Protocol (IP) packets traverse an access network and a backbone network without any protocol conversion. Since 3G networks basically evolved from a circuit-switched cellular network, they have their own gateways to interpret IP from the backbone network. They also have their own protocols and interfaces for the communication within themselves. To overcome these problems, 4G networks are expected to become an **all-IP based packet-switched system**, similar to the IP backbone network.

4G networks have two different visions: revolution — developing an innovative system and evolution — interworking with existing systems. The interworking model takes an approach that integrates cellular networks, wireless metropolitan area networks (WMAN), wireless local area networks (WLAN), and wireless personal area networks (WPAN). This model covers a future scenario of ubiquitous networking where anyone can access a network anytime, anywhere, and anyway. The IEEE 802.11 WLAN achieves system throughput up to 54 Mb/s while the service area is limited to two or three hundred meters. In contrast, a current cellular network such as cdma2000 1x EV-DO (evolution-data only) covers several kilometers, but its cell throughput is at most, 2 Mb/s. Therefore, it is essential to develop an innovative system with high throughput and wide coverage. Moreover, as it was mentioned before, the new system is expected to employ novel techniques such as orthogonal frequency division multiplexing (OFDM) and multiple input multiple output (MIMO) antennas. If the system considers various conditions such as high speed mobile or nomadic users, data or voice traffic, and cell center or boundary conditions, it may be required to exploit hybrid multiple access techniques.

As a candidate for the 4G network, the IEEE 802.16 standard sets a goal of WMAN/WLAN based on OFDM or orthogonal frequency division multiple access (OFDMA). The IEEE 802.16e and IEEE 802.16m standards supplemented it for mobility support such as IEEE 802.20. In Korea, a new service for the 2.3 GHz band based on IEEE 802.16 OFDMA, named WiBro, started in the second quarter of 2006. It is designed to support the maximum mobility of 60 km/h. To support high mobility, there is a hybrid multiple access scheme combining OFDMA and frequency hopping (FH)-OFDMA, where fast-moving users access the network via FH-OFDMA. Decoupling multiple access techniques for the hierarchical cell structure, we consider a new wireless network, comprised of OFDMA microcells and FHOFDMA macrocells. Although innovative 4G systems — medium access control (MAC) layer and physical (PHY) layer — are under development, there exists little work that considers the network architecture. Moreover, there are many frameworks and proposals which expand cross layer techniques, which are dedicated to the cooperation between MAC (L2) and PHY (L1) layers, to cover the network layer (L3) together.

Existing cellular networks, based on circuitswitching, consist of base stations (or base transceiver stations), base station controllers, switching centers, gateways, and so on. The base station (BS) plays the role of physical transmission with fast power control and wireless scheduling. The base station controller (BSC) executes the largest part of the radio resource management. Whenever a mobile terminal (MT) moves into another cell, it requires handoff to another base station.

In contrast to that, the 4G network has a simple structure where each BS must function intelligently to perform radio resource management as well as physical transmission. This makes the BS act the role of an access router (AR). This architecture is shown in Figure 1.21.

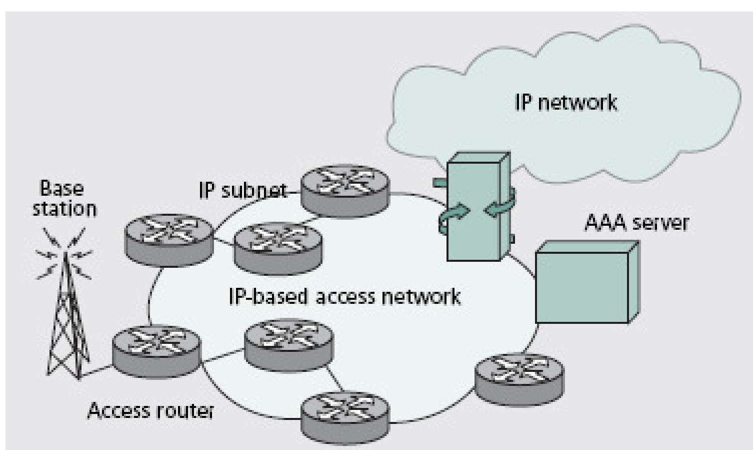


Figure 1.21. The pure all-IP 4G network.

It incurs high overhead, however, especially when an MT configures mobile IP (MIP) addresses for handoff. As we know, it takes several seconds to run the MIP handoff, and MIP hinders an MT from performing smooth handoff. In addition, the 4G network is expected to have a small cell radius due to its use of high frequency band, which possibly results in short cell residence time. For this matter, reducing the latency is still a challenging issue. In particular, there are a fast handoff schemes which proposed to decrease the address resolution delay by pre-configuration. To solve the fundamental problem of all-IP cellular networks, the functionality of an AR can be separated from that of an access point (AP) so that each undertakes L3 and L2 protocols, respectively. Figure 1.22 shows an example of a simple network where an AR manages several AP.

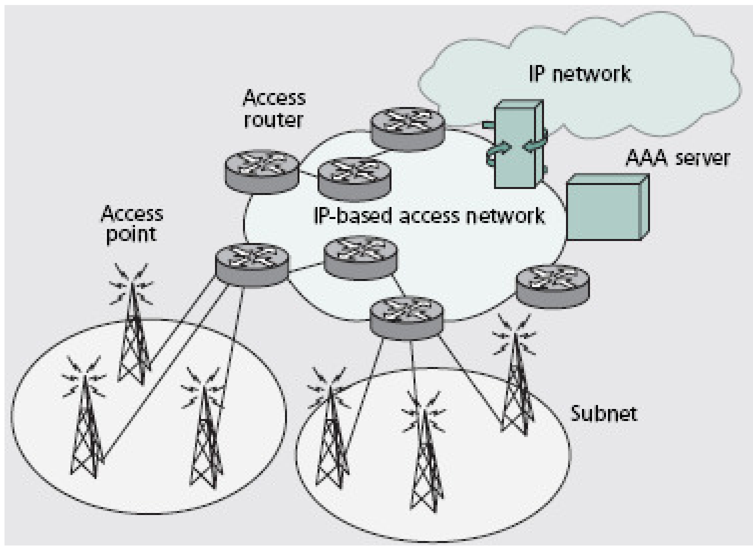


Figure 1.22. The subnet-based 4G network.

This relation is similar to that between BSC and BS in existing cellular networks. A subnet consisting of an AR and several AP can be configured by Gigabit Ethernet. Then, an MT moving within the subnet (i.e., changing AP) performs L2 handoff without changing MIP attachment. When the MT moves into another AR area, it experiences L3 handoff. Table 1.9 compares the network architecture of pure and subnet-based all-IP cellular networks.

Table 1.9. Comparison of two network types.

	Pure all-IP network	Subnet-based all-IP network
Access network components	AR	AR+AP
Operation type	Decentralized	Centralized
Coordination among cells	Complex but flexible	Simple but fixed
Handoff overhead	High	Low
Handoff protocol	L3	L2+L3
Cost	Low	High

A main difference is that the former is decentralized, and the latter is centralized. Since the pure all-IP network incurs L3 protocol in the end access link, it requires long handoff latency and high signaling overhead. However, the architecture is simple and cost-efficient for implementation. On the other hand, the subnet-based all-IP network implements hierarchical architecture, so it is possible to do efficient resource management in spite of its inflexibility.

Moreover, some of the key technologies required for 4G are briefly described below:

- ✓ OFDMA: it is possible to exploit the time domain, the space domain, the frequency domain and even the code domain to optimize radio channel usage. It ensures very robust transmission in multi-path environments with reduced receiver complexity. OFDM also provides a frequency diversity gain, improving the physical layer performance .It is also compatible with other enhancement Technologies, such as smart antennas and MIMO radar antenna.
- ✓ MIMO: uses signal multiplexing between multiple transmitting antennas (space multiplex) and time or frequency. It is well suited to OFDM, as it is possible to process independent time symbols as soon as the OFDM waveform is correctly designed for the channel. This aspect of OFDM greatly simplifies processing. The signal transmitted by m antennas is received by n antennas. Processing of the received signals may deliver several performance improvements range, quality of received signal and spectrum efficiency.
- ✓ Software Defined Radio (SDR): benefits from today's high processing power to develop multi-band, multi-standard base stations and terminals. Although in future the terminals will adapt the air interface to the available radio access technology, at present this is done by the infrastructure. Several infrastructure gains are expected from SDR. For example, to increase network capacity at a specific time (e.g. during a sports event), an operator will reconfigure its network adding several modems at a given Base Transceiver Station (BTS). SDR makes this reconfiguration easy. In the context of 4G systems, SDR will become an enabler for the aggregation of multi-standard pico/micro cells. For a manufacturer, this can be a powerful aid to providing multi-standard, multi-band equipment with reduced development effort and costs through simultaneous multi-channel processing.
- ✓ Handover and mobility: Handover technologies based on mobile IP technology have been considered for data and voice. Mobile IPv6 techniques are slow but can be accelerated with classical methods (hierarchical, fast mobile IP). These methods are applicable to data and probably also voice. In single-frequency networks, it is necessary to reconsider the handover methods. Several techniques can be used when the carrier to interference ratio is negative (e.g. Variable Spreading Factor Orthogonal Frequency and code Division Multiplexing (VSFOFDM), bit repetition), but the drawback of these techniques is capacity. In OFDM, the same alternative exists as in CDMA, which is to use macro-diversity. In the case of OFDM, MIMO allows macro-diversity processing with performance gains. However, the implementation of macro-diversity implies that MIMO processing is centralized and transmissions are

synchronous. This is not as complex as in CDMA, but such a technique should only be used in situations where spectrum is very scarce.

- ✓ QoS: provisioning of high level of QoS for any given service in 4 G network is a main objective. For more details about QoS in 4G see the following section.

As the history of mobile communications shows, attempts have been made to reduce a number of technologies to a single global standard. Projected 4G systems offer this promise of a standard that can be embraced worldwide through its key concept of integration. Future wireless networks will need to support diverse IP multimedia applications to allow sharing of resources among multiple users. There must be a low complexity of implementation and an efficient means of negotiation between the end users and the wireless infrastructure. The fourth generation promises to fulfill the goal of PCC (Personal Computing and Communication) a vision that affordably provides high data rates everywhere over a wireless network. In few countries like South Korea and Japan 4G was launched in 2010 and the world is looking forward for the most intelligent technology that would connect the entire globe. In India, Mukesh Ambani's Reliance Communications conducted trial for 4G in India, got 80 Mbps Download Speed. With the plethora of promising features 4G is truly moving towards getting universally accepted as the ideal next generation communication system.

1.6. QoS and mobility management concepts for 4G

The vision of IMT-Advanced and moreover the paradigm of 4 G mobile networks is an all-IP network supporting heterogeneous wireless access technologies to accommodate a variety of services and traffic types and to allow the mobile user to roam within the service area or across the different networks without degrading the QoS provided. In general, one wireless and mobile access network can manage QoS independently of the IP network because it becomes a bottleneck for providing end-to-end QoS. QoS control is possible by using access and scheduling methods. Recently, the QoS of IEEE 802.11 WLAN system was supplemented by the IEEE 802.11e and IEEE 802.11n standards. The advantages of IEEE 802.11n is defined in section 2, here we can say briefly a few words about IEEE 802.11e. It defines extended distributed contention access (EDCA), which assigns a small backoff number to high priority traffic and hybrid coordination function (HCF), which improves the conventional polling scheme of point coordination function (PCF). Also, cdma2000 1x EV-DO and WCDMA-HSDPA (high speed downlink packet access) adopted opportunistic scheduling to exploit channel fluctuation. The opportunistic scheduling is a hot issue in designing various scheduling algorithms for QoS.

The Third Generation Partnership Projects (3GPP and 3GPP2) define four traffic classes and their related parameters for QoS provisioning. There exist gateways between IP backbone and access networks that perform protocol conversion and QoS mapping between IP and access networks. However, direct translation is difficult since access networks have their own QoS attributes that require strict QoS mechanisms within them. The QoS mechanisms, including resource reservation (signaling), admission control and traffic control, allow multimedia applications to get certain quality guarantee e.g., on bandwidth and delay for its packets delivery.

Meanwhile, the importance of unified QoS management grows in 4G networks as QoS management for both access networks and IP networks becomes cumbersome in all-IP networks. If each network has an individual QoS model, it requires a rule that integrates the QoS models to ensure end-to-end QoS. Furthermore, let us see what QoS in 4G networks are providing to the end users:

- ✓ Traffic generated by the different services will not only increase traffic loads on the networks, but will also require different quality of service (QoS) requirements (e.g., cell loss rate, delay, and jitter) for different streams (e.g., video, voice, data).
- ✓ Providing QoS guarantees in 4G networks is a non-trivial issue where both QoS signaling across different networks and service differentiation between mobile flows will have to be addressed.

- ✓ One of the most difficult problems that are to be solved, when it comes to IP mobility, is how to insure the constant QoS level during the handover.
- ✓ Depending on whether the new access router is in the same or some other subnetwork, we recognize the horizontal and vertical handover.
- ✓ However, the mobile terminal cannot receive IP packets while the process of handover is finished. This time is called the handover latency.
- ✓ Handover latency has a great influence on the flow of multimedia applications in real-time.
- ✓ Mobile IPv6 has been proposed to reduce the handover latency and the number of lost packets.
- ✓ The field "Traffic Class" and "Flow Label" in IPv6 header enables the routers to secure the special QoS for specific packet series with marked priority.

For unified QoS management, there are many proposes, adaptive QoS frameworks, cross-layer designs and etc, that considers L1, L2, and L3 together. Undoabatly, one appropriate QoS framework for 4G technology must to conceder all those layers (and Application layer too) together in same time in order to provide the best QoS performances for any given service. Moreover, the L1 and L2 allocate radio resources and logical channels, respectively, according to the QoS indication of L3, and application layer gives the information for the Type of Service (ToS). Moreover, the QoS frameworks for 4G concerns the information about IntServ and DiffServ for the resource management of L1 and L2. When IntServ establishes a real-time session, MAC reserves a dedicated channel. In contrast, when DiffServ is used for low mobility users, MAC can exploit either a dedicated or a shared channel. If the shared channel is allocated for DiffServ, the wireless scheduler runs a scheduling algorithm for QoS provisioning. In contrast, the dedicated channel allocation requires an admission control that enables a limited number of users into the network for QoS support. Therefore, IP QoS information helps MAC and PHY manage the following resources in a flexible manner: Cell type — microcell or macrocell, Multiple access — OFDMA or FH-OFDMA, MAC channel — dedicated or shared, PHY scheduling — priority or fairness.

IntServ is easy to involve in radio resource management because wireless access is usually accompanied by signaling. When a mobile terminal (MT) requests a real-time service in a 4G network, the corresponding AR can initiate IntServ and allocate a dedicated channel. For a downlink call, the AR can adjust the bandwidth of a dedicated channel with the aid of RSVP. As real-time traffic usually requires a constant data rate, the dedicated channel is recommended to use power control rather than AMC. In this aspect, FH-OFDMA and CDMA may be more suitable than OFDMA for real-time services.

Regarding DiffServ in 4G networks, it is sufficient for an MT to set the DS field properly for uplink packets, because the AP controls radio resources before transferring them to the AR. For downlink traffic, the AR classifies packets according to the DS field and chooses a multiple access method, and accordingly, the AP allocates a dedicated or shared channel. The dedicated channel has the advantage of simple management, while the shared channel works well with DiffServ because both require scheduling. In contrast to scheduling in routers, which must handle several flows, wireless scheduling handles fewer connections, which enables it to use a per-user buffer. Therefore, the wireless scheduler can exploit an algorithm with high granularity of radio resources. Figure 4 and Table 2 summarize tightly coupled resource management among three layers through a unified QoS strategy.

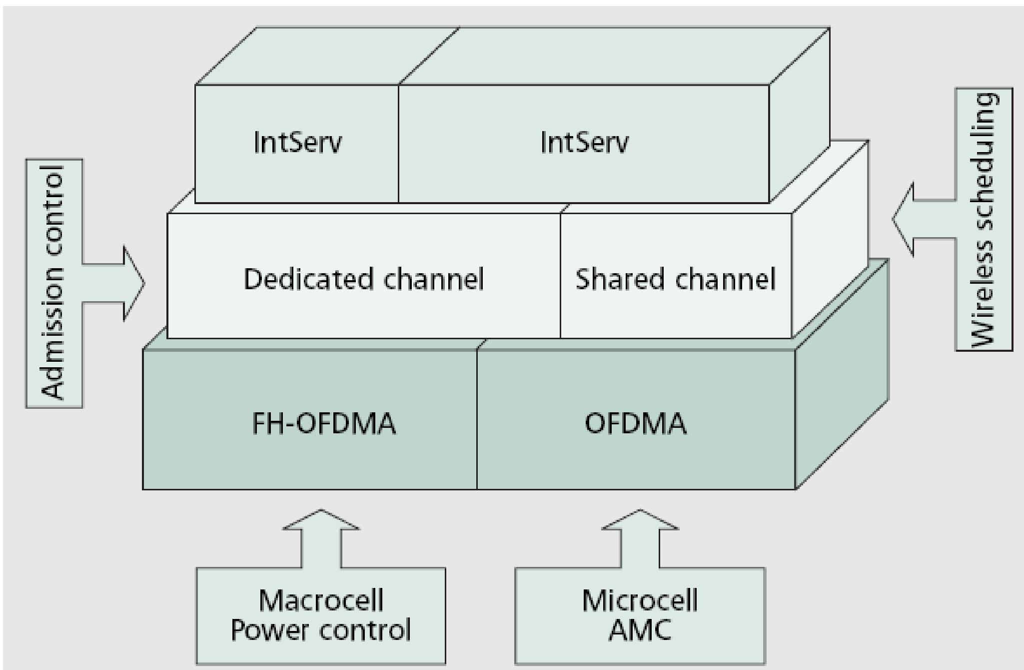


Figure 1.23. Illustration of the coupled layering for resource management.

Table 1.10. Example of unified 4G QoS strategy.

Traffic class	Mobility	IP QoS	Logical channel	Multiple access
Real-time	High	IntServ/DiffServ	Dedicated	FH-OFDMA
	Low	IntServ	Dedicated	FH-OFDMA/ OFDMA
		DiffServ	Dedicated/shared	
Non-real-time	High	DiffServ	Shared	FH-OFDMA
	Low	DiffServ	Shared	OFDMA

However, providing QoS provisioning for any given service in 4G networks is non-trivial issue and requires further study. Both QoS signaling across different networks and service differentiation between mobile flows will have to be addressed. On the other hand, before providing network access and allocating resources for an Mobile Terminals, the network needs to authenticate the MT's (or the mobile user's) credential. Furthermore, a security association needs to be established between the MT and the network to ensure data integrity and encryption. Thus, in order to achieve seamless handover, mobility, QoS and security technologies must be integrated.

While resource management functions are primarily handled by BSC in the existing cellular networks, more functions will be imposed on AP in 4G networks. Basically AR will be responsible for IP QoS, and AP will play the primary role of resource management. Another challenge is the application of unified QoS framework to the hierarchical network. In this case, a coordinator is required to decide whether an incoming session is served by a macrocell or a microcell. It also will have the capability of load balancing by triggering vertical handoff.

Moreover, beside QoS, there are also the security and mobility which together with the QoS can be viewed as three different, indispensable aspects in 4G networks; however all are related to network nodes involving the controlling or the processing of IP packets for end-to-end flows between an MT and the core network. In Figure 1.24 one basic view of the 4G network infrastructure is shown.

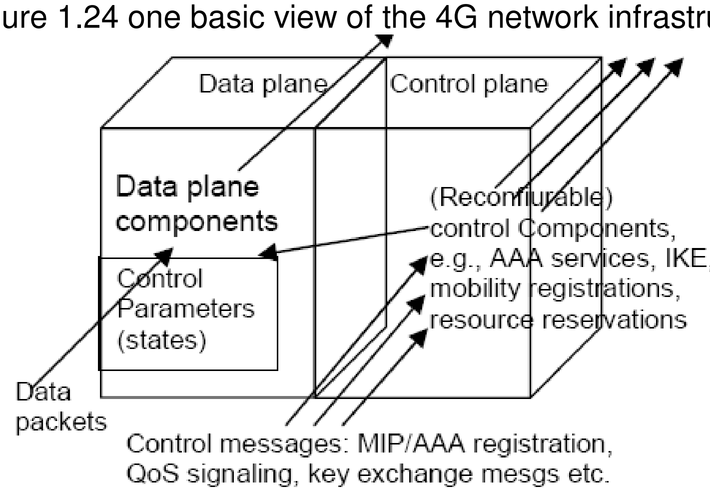


Figure 1.24. The decomposition of control plane and data plane Functionalities in 4G networks.

We argue the separation and coordination of control plane and data plane is critical for seamless mobility with QoS and security support in 4G networks, with the reasons as follows. Per-flow or per-user level actions occur much less frequent than per-packet actions, while per-packet actions are part of critical forwarding behavior, which involves very few control actions (which are typically simply to read and enforce according the install state during forwarding data). Actually, this separation concept is not new – routing protocols have the similar abstraction together used with the traditional IP packet delivery, this abstraction is recently being investigated in the IETF ForCES working group. However, we emphasize the three critical dimensions of future 4G networks: mobility, QoS and security, as well as other new emerging or replacement components might appear, integrated into a unified framework and allowing more extensibility for 4G networks design.

Moreover, the mobility issue is also one of the main aspects in 4G, and furthermore we will elaborate the mobility and mobility management in 4G mobile networks. As we know that mobility involves both control plane and data plane. The control plane is mainly involved with pathdecoupled, end-to-end way of mobility registrations, while data plane concerns mobility-enabled routing for data flows into and from an MT while it moves between different locations. The data plane behavior is achieved by installing/changing certain binding caches upon certain control plane information exchange (e.g., the binding update/acknowledge procedure in MobileIP). In fact, although MobileIP does not change the traditional IP routing table, when the MT is away from home and changes its location, associated with its fixed home address information, routing information is added in certain data processing and/or forwarding entities such as mobility agents (e.g., home agent and foreign agent) and systems themselves upon successful MobileIP registrations. Localized mobility solutions such as fast handover for Mobile IPv6 (FMIPv6) and Hierarchical Mobile IPv6 (HMIPv6) make this a little bit more complicated. However, more details about various MobileIPv6 and MobileIPv4 cases you can find in the following section (section 1.7).

On the other hand, let we see what is beyond "mobility management in 4G" which is very crucial concept in 4G. Beyond mobility management concepts in 4G, there is not only user mobility, but also there is application-layer mobility, which included itself the following types of mobility:

- Terminal mobility: one terminal, multiple network addresses;
- Personal mobility: one person, multiple terminals;
- Session mobility: one user, multiple terminals in sequence or in parallel;
- Service mobility: services move with user.

Behind the Terminal mobility there is a user's ability to use his/her terminal to move across (heterogeneous) networks while having access to the same set of subscribed services. Furthermore the Terminal Mobility can be provided in Session Initiation Protocol (SIP) through the use of the SIP Registrar and SIP Redirect Server. As the terminal moves across heterogeneous networks new temporary identifiers (IP addresses) are assigned to the terminal. These are updated with the SIP Registrar by using the REGISTER method. The current location of the device is always up-to-date so that messages can be redirected successfully, in a time-efficient manner. Although SIP has been chosen by the 3GPP for call signalling it also could be used for Application-Layer mobility management although no standardized specifications for this purpose have been worked out. In principle, SIP is an Application-Layer multimedia signalling protocol. Terminal, personal, session and service mobility are all largely in the application level as well. Therefore, with the help of SIP infrastructure, SIP provides a framework and has the potential capabilities to support advanced high-level mobility management by augmented signalling. To provide personal mobility, the infrastructure can obtain the current device the user is using in the visited network. If the device changes its current IP address during a session due to the user's roaming, theoretically what the device needs to do is simply notify the corresponding host by the versatile SIP INVITE message (notification) so that the session can resume. However, the basic SIP has some difficulties in dealing with mid-session mobility since the IP issue belongs to Network Layer and is beyond SIP. For session mobility, SIP can INVITE another device to join the session and hand it over. However, simultaneous handoffs of the two communicating parties have some problems in the complete implementation. To sum up, SIP is able to facilitate an advanced mobility management scenario but currently it is more of a conceptual model rather than a mature solution, although **extensions of SIP** (or SIPbased solutions) are a promising approach. **Alternative approaches** such as ICERGERG (Internet Core Network Architecture for Integrated Communications) and IPMoA (Integrated Personal Mobile Architecture) prefer to tackle part of this problem from their own points of view and take different angles, which could contribute to building parts of a complete MM solution.

For sure, one feature of 4G should be its heterogeneous communication environment, where different generation cellular systems coexist and various wireless access networks and corresponding services are used. Advanced mobility management (MM) is thus needed to fulfill seamless global roaming in terms of not only the traditional terminal mobility, but the other high-level personal, session and service mobility. From a general point of view, three kind of terminal mobility can be considered:

- Terminal Mobility with respect to user communications that can be divided into discrete and continuous mobility. Discrete mobility takes place when movement of the terminal has to be managed only when the user is not in communication, whereas continuous mobility would require the maintaining of ongoing communications sessions while the user is moving with its terminal.
- Terminal mobility with respect to change in network access. This includes mobility within a single access network (i.e. same access technology) as well as mobility between access networks of different type (vertical handover).
- Terminal Mobility with respect to administrative domains. Movement from one administrative domain to another (e.g. roaming between networks of two operators) will require so extra network functions (e.g. AAA) to be handle in a distributed manner among the domains.

In general, two Terminal mobility contexts according to network hierarchy can be identified namely, **macro-mobility** and **micro-mobility**. Shortly, micro-mobility takes place when moving between access points that are close to each other according to network hierarchy (e.g. between Node-Bs attached to the same RNC in UTRAN) while macro-mobility management handles mobility between distant access points according to networks hierarchy (e.g. between different RNCs in UTRAN, between distant administrative domain, between different operator’s networks, etc.). In global domain terminal mobility is supported by means of a global mobility protocol (GMP), such as Mobile IPv6 (MIPv6) or Host Identity Protocol (HIP). On the other hand, terminal mobility within a local domain is handled via local protocol operators, local mobility protocols (LMPs), which are transparent to the core network and independent of the GMP. In this case, when a mobile node moves within a local domain, only the LMP used in that domain operates; when the node moves across domains, only GMP operates.

Furthermore, Personal Mobility refers to the user's ability to access mobility services from anywhere, at anytime, using any terminal. Similar to terminal mobility, in personal mobility using SIP can be provided transparent support of name mapping and redirection services through the use of the SIP Registrar and Location Services Server. These components keep track of all the possible terminals associated with a user, and the permanent and temporary addresses of each of those terminals, at any given time. This provides SIP with inherent personal mobility support - users can use a single personal identifier in all occasions, regardless of the terminal(s) used or their network locations. Any incoming messages or calls will be redirected to the appropriate active terminal(s). In the case where a single user is associated with multiple terminals, SIP can provide personal mobility support through the use of a SIP Registrar and Location Services Server, and a simple call forking procedure. The actual address of the active terminal is determined dynamically, as a session initiation request is received. This disassociates the user from any one particular terminal or network location, providing true personal mobility.

The Session Mobility refers to the user's ability to maintain an active session while switching between terminals. SIP can successfully implement session mobility (or mid-call mobility) through the use of the re-INVITE method. This is an INVITE method sent while a multimedia session is in progress. While maintaining the existing session alive, new terminal(s) can be added to the session and existing ones can be removed. Changes could be made to the parameters of the session in progress in order to match the capabilities of the newly added terminal(s).

One example of session mobility system architecture is shown in Figure 1.25. This architecture is based on three standardized protocol: Service Location Protocol (SLP) for device discovery, Session Initiation Protocol (SIP) and its extensions for signalling in order to support session transfer, Real-Time Transport Protocol (RTP) for all media transport such as audio and video.

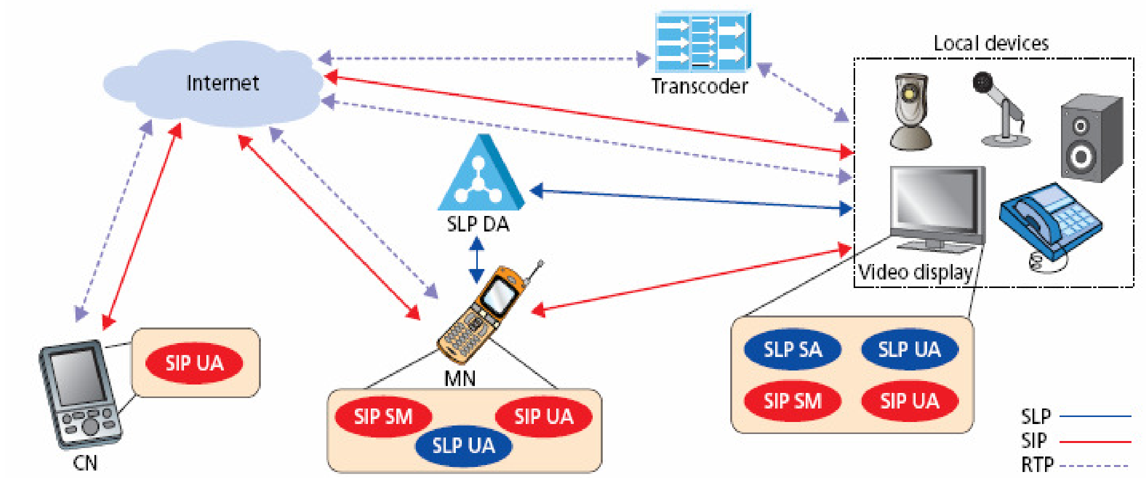


Figure 1.25. The session mobility system architecture.

In Figure 1.25, the mobile node (MN) that invokes a session transfer is a SIP-enabled mobile device. It implements standard SIP and the described SIP Session Mobility (SM) extension, which handles SIP signalling for session transfer. Local Devices, which are discovered by the MN, may be standard SIP-enabled devices or they may include a session mobility extension as well. At least one local device must have this extension, which is, for example, the video display in Figure 1.25. The Correspondent Node (CN) is a basic SIP device implementing a SIP User Agent (UA), capable of setting up SIP calls. A transcoder is only used when the reconciliation of different capabilities during the session transfer is needed. For example, transcoding is needed when there is no codec of the CN available on a transfer target local device. We have chosen to implement device discovery using SLP, as it is a standardized protocol and allows discovery using SLP, as it is a standardized protocol and allows discovery in different granularities, such as in a specific room or on a floor. In SLP, the SLP Service Agent (SA) is responsible for advertising services, while the SLP User Agent (UA) queries for services. The SLP Directory Agent (DA) is the centralized directory keeping track of devices based on location and capabilities. Nevertheless, this architecture is not dependent on any specific discovery protocol.

Furthermore, this architecture (Figure 1.25) supports various options of session transfer as follows:

- ✓ Transfer and retrieval: after a session is transferred to local devices, a user may want to have his session back on the original device before the transfer happens.
- ✓ Complete Session Transfer (CST) or Splitting Session Transfer (SST): a user has a choice to either transfer his/her ongoing session consisting of several media types (e.g., audio and video media in a video call) completely to a single local device (CST) or split it across local devices (SST).
- ✓ Session Control Retention or Relinquishment: a user may or may not want to keep the control on the source device initiating the

session transfer. Our architecture provides two different models: Mobile Node Control (MNC) mode and Session Handoff (SH) mode.

And the last, the Service Mobility is defined as the ability of the network to consistently provide personalized services to the user, with the expected QoS, regardless of the user's location. It "allows users to maintain access to their services even while moving or changing devices and network service providers". With respect to service mobility two aspects must be considered: maintaining adequate QoS for the duration of a session regardless of the network changes, and ensuring that the users have access to all their subscribed services regardless of the point of attachment to the network.

A possible SIP approach to QoS support is through appropriate resource allocation during session initiation and hand-offs. In order to support access to subscribed services, SIP uses a combination of AAA and SIP REGISTRAR functions. This can be implemented in two ways:

- Centralized registration through the home network;
- Distributed registration through the visited network.

The Home Network Control approach provides easier call control and accounting functionality, as well as better security, at the expense of increased hand-off delays and latencies for ongoing multimedia sessions. The Visited Network Control approach reduces the domain hand-off delays and latencies, but it requires a more intelligent and powerful terminal, as well as more complex accounting and call control functionality. An appropriate choice can be made after careful consideration of the resources available and the performance-complexity trade-offs.

As it is known, the services are composed of many components. Most of the components are service specific, but two components are required by the service architecture for every service, as discussed in this section. First component is a service factory that is responsible for creating and assembling all the components of the service. It is also capable of maintaining a repository of service instances of that service type. A service provider has one service factory for each service type it offers. Another component is a service handle that performs two main functions. One, it provides a service interface to the outside world. Two, it also connects all the components of the service. When one service component wants to interact with another, it gets the component through the handle.

Services are mobile when they are accessible through different network domains. There are **three** main approaches to make a service mobile. These approaches are not mutually serving network operator establishes a communication channel between the service access point in its domain and the service provider domain. The service access point is typically the user terminal connected to the serving network. The service data moves across the channel using some form of request-response protocol or RPC. The terminal must be capable of receiving and presenting the service data. If there are multiple users accessing the same service simultaneously, the network operator must support each remote connection to prevent denial of service.

Another way of accessing a foreign service is to deploy a service proxy in the serving network. The service proxy typically serves a single user. It can be implemented as a mobile agent that mediates the requests between the terminal and the service running in the service provider domain. It may have limited storage capacity for caching service data. It simplifies the terminal software design by managing the connections to the service and hiding the details of service access, and in some cases even by massaging the service data. This can be implemented using the Client/Server/Agent model of mobile computing. And the last way is with the cloning and moving the service components. So, each service may have its own strategy for the service mobility.

Which components are moved to the serving network depends upon several factors. Some are internal factors related to the nature of the service and its mobility policy. For example, typically database components of a service are unlikely to move. Other factors are external related to service management and logistics. For example, load condition of the servers at the serving network, the service hosting policies of the serving network, and the contract between the serving network and the service provider are some external factors that play a role in the decision of the final mobility of the service components. The network support is required to move services across provider domains. Before moving a service some sort of agreements must be in place between the network operator and the service provider. A handshake protocol is described below to move a service dynamically. The network operators should have some kind of service platform that is capable of hosting mobile services.

The request to move a service comes from a network operator to the service. A service level agreement (SLA) between the network operator and the service provider must be in place prior to establishing a connection for the service. The service level agreement contains some agreed upon service parameters between the network and the provider. It may also contain some service management information, such as whether the network operator can offer new subscriptions for that service, time to lease the service if it is on lease, etc. The network operator sets up one of its servers to receive the service components before sending a *request for a service* message to the service provider. After receiving the request, the service provider creates a mobile instance of the service based on the SLA, and *moves the service* components to the server assigned by the network operator for receiving the service. Once all the mobile components of the service are moved to the new network and assembled to form the service, the service provider receives a signal from the service to mark the end of the service to the ready to serve state. At this point the network operator sends a *received the service* message to the service provider to conclude the service mobility handshake. All above is illustrated in Figure 1.26.

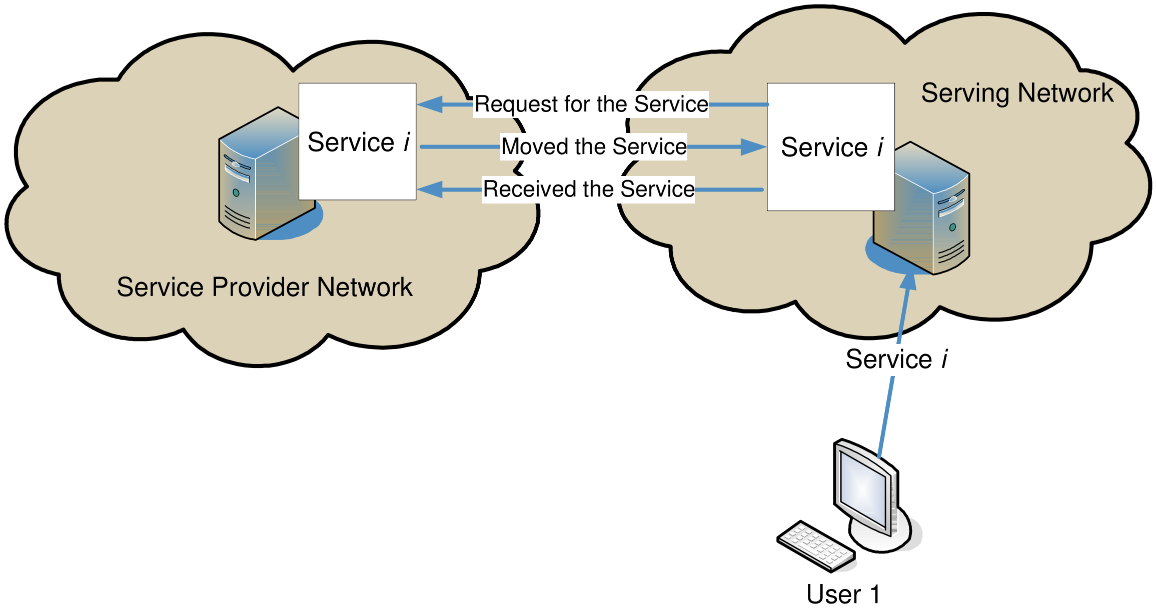


Figure 1.26. Service Mobility Protocol.

A mobile service can be accessed from different types of terminals in different network environments. The service presentation needs to be adjusted for different terminal types without affecting the basic service functionality. Hence, the presentation logic must be kept separate from the service core logic. The service presentation logic or *service user interface* (service UI) resides at the terminal, which is used to access the service. For full-scale service mobility, both the service core logic (or the *service*) and the service UI must be mobile. In the architecture presented in this section both sub-systems are mobile, but they are discussed separately. Terminals only the front end resides there. The service UI can be downloaded from the service provider on the user's request. It can also be a value-added service obtained from a third party value-added service provider. For example, the directory service provider can provide a directory service, but to access the directory from a specific terminal, say an SMS terminal, an independent SMS gateway service provider may provide SMS gateway service.

Overall, we can conclude that the mobility is a critical aspect of 4G. In the following the three main issues regarding mobility management in 4G networks are summarized:

- The first issue deals with optimal choice of access technology, or how to be best connected. Given that a user may be offered connectivity from more than one technology at any one time, one has to consider how the terminal and an overlay network choose the radio access technology suitable for services the user is accessing. There are several network technologies available today, which can be viewed as complementary. For example, WLAN (IEEE 802.11n and IEEE 802.11e) is best suited for high data rate indoor coverage. GPRS, UMTS, LTE, WiMAX 802.16m and LTE-Advanced, on the other hand, are best suited for nation wide coverage and can be regarded as wide area networks, providing a higher degree of mobility. Thus a user of the mobile terminal or the network needs to make the optimal choice of radio access technology among all those available. A handover algorithm should both determine which network to connect to as well as when to perform a handover between the different networks. Ideally, the handover algorithm would assure that the best overall wireless link is chosen. The network selection strategy should take into consideration the type of application being run by the user at the time of handover. This ensures stability as well as optimal bandwidth for interactive and background services.

- The second issue regards the design of a mobility enabled IP networking architecture, which contains the functionality to deal with mobility between access technologies. This includes fast, seamless vertical (between heterogeneous technologies) handovers (IP micro-mobility), quality of service (QoS), security and accounting. Real-time applications in the future will require fast/seamless handovers for smooth operation. Mobility in IPv6 is not optimised to take advantage of specific mechanisms that may be deployed in different administrative domains. Instead, IPv6 provides mobility in a manner that resembles only simple portability. To enhance Mobility in IPv6, 'micro-mobility' protocols (such as Hawaii, Cellular IP and Hierarchical Mobile IPv6 (see in the next Section)) have been developed for seamless handovers i.e. handovers that result in minimal handover delay, minimal packet loss, and minimal loss of communication state.
- And the third final issue concerns the adaptation of multimedia transmission across 4G networks. Indeed multimedia will be a main service feature of 4G networks, and changing radio access networks may in particular result in drastic changes in the network condition. Thus the framework for multimedia transmission must be adaptive.

1.7. Control and signaling protocols for 4G

As IMT-Advanced (4G) networks scale, carriers are finding one of the biggest challenges to keeping networks up and running is **scaling the control and signaling plane**.

The first hurdle to overcome when making a multimedia seession (e.g., VoIP call, video-conference, ftp session, sending sms, web browsing and etc.) is to establish a connection between the parties involved. In legacy telephony, this is done by switching circuits until a physical wire is established between locations. The Internet Protocol on the other hand is connectionless by nature. IP packets have a tendency to take whatever route they find first, and end up in whatever order they arrive. For time sensitive applications such as voice and video (real-time services) this is unacceptable. Steps must be taken to establish a point to point connection and to keep it open for the duration of the call. Similar to the handshake of the DHCP protocol, the VoIP signaling protocols use TCP to set up, manage and tear down the VoIP phone call. Signaling protocols are not concerned with the actual media stream of voice or video, and could care less about QoS and traffic engineering. Their basic functions are to first initiate a session, then to find common ground for communication between the parties involved, and to terminate the session at calls end. In the following the most used control and signaling protocols for 4G networks (SIP, DIAMETER and Mobile IP) are presented.

1.7.1. SIP

Session Initiation Protocol (SIP) is the core networking protocol used within the IP Multimedia Subsystem (IMS) within the 4G networks and is one of the more widely known examples of a session control protocol. Session control refers to the process used to create, modify and terminate IP-based communication sessions, and a session can include two-way voice communication, multimedia (text, audio or video) conference collaboration, instant messaging, application sharing and other contemplated but not yet fully specified services. Session control is accomplished through signaling between various network elements and endpoints using a session control protocol. Overall, SIP is capable of providing support for not only terminal mobility but also for session mobility, personal mobility and service mobility.

However, the performance of SIP, operating at the highest layer of the protocol stack, is only as good as the performance of the underlying transport layers in such a heterogeneous 4G environment. Although SIP is the most widely known session control protocol, SIP has a major limitation that is of great importance to any GSM-UMTS operator. It does not provide any method of directly inter-working with the Public Switched Telephone Network (PSTN) because it was not created with the intention of it being fully backwards-compatible with legacy PSTN signaling mechanisms.

The IETF has been developing a competing, but potentially complementary, architecture for multiparty, multimedia conferencing on the Internet as is shown in Figure 1.27. Undoubtedly, SIP is a component of this architecture and provides the basic session control mechanism used within it. The SIP protocol has gained a substantial following within the industry by offering the potential for an easily implemented method of establishing and controlling basic voice calls. From its very lightweight inception, SIP has been developed to address the challenges of being used outside basic point-to-point voice calls and an overly simplistic direct-mode signaling model. Moreover, in Figure 1.28, the simple call flows using SIP is given.

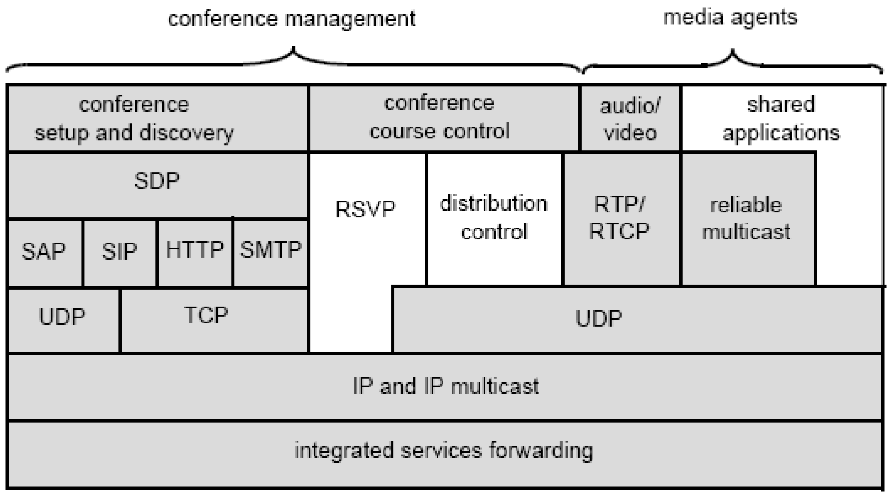


Figure 1.27. IETF multimedia conferencing architecture.

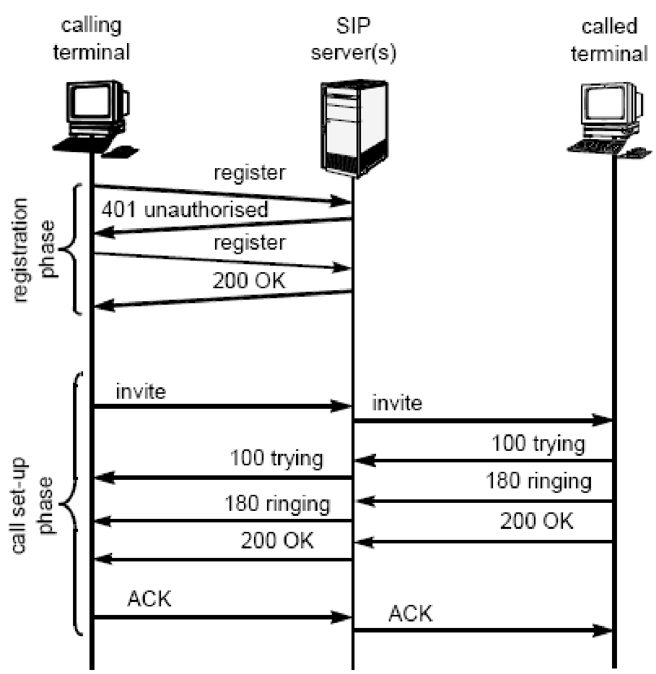


Figure 1.28. Typical registration and call set-up using SIP.

Furthermore, a SIP network is made up of end points, a proxy and/or redirect server, location server, and registrar. A diagram for basic SIP architecture is provided in Figure 1.29. In the SIP model, a user is not bound to a specific host, moreover, the user initially reports their location to a registrar, which may be integrated into a proxy or redirect server. This information is in turn stored in the external location server.

Also, the messages from endpoints must be routed through either a proxy or redirect server. The proxy server intercepts messages from endpoints or other services, inspects their "To:" field, contacts the location server to resolve the username into an address and forwards the message along to the appropriate end point or another server. Redirect servers perform the same resolution functionality, but the onus is placed on the end points to perform the actual transmission. That is, Redirect servers obtain the actual address of the destination from the location server and return this information to the original sender, which then must send its message directly to this resolved address.

It is more then obviously that SIP protocol itself is modeled on the three-way handshake method implemented in TCP (see Figure 1.28).

The main advantages of SIP consists of its offering an easily implemented, powerful, control environment capable of scaling to very large networks due to its simple message request/response format. This, combined with its relative immaturity compared with H.323, encouraged its adoption in the access segment of 4G networks, since this affords the opportunity to incorporate any mobile-specific elements that were subsequently identified.

On the other hand, both protocols can be extended to manage new capabilities. The argument has been advanced that H.323 is more stable because of its maturity but SIP provides better support for some functionality and is easier to implement. Fortunately the ITU and the IETF are now co-operating in developing standards in this area.

Because of the transparency to the lower layer characteristics, application-layer mobility management protocol like the SIP has been considered as the right candidate for handling mobility in the heterogeneous 4G wireless networks.

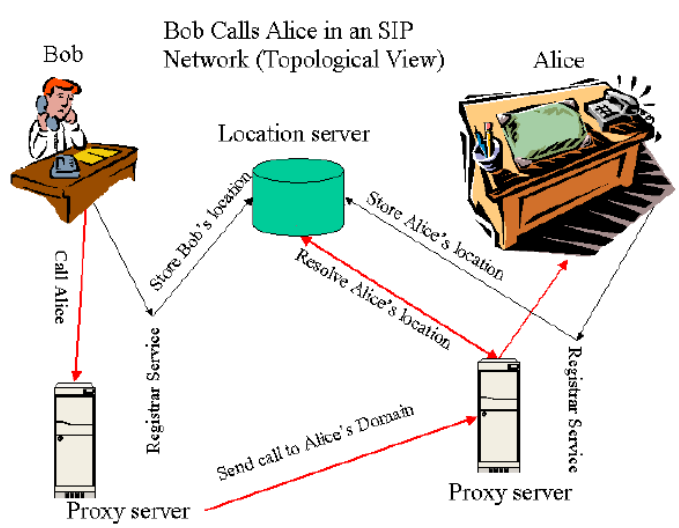


Figure 1.29. Typical SIP Architecture.

SIP sessions utilize up to four major entities: SIP User Agents, SIP Registrar Servers, SIP Proxy Servers and SIP Redirect Servers. Together, these systems deliver messages embedded with the SDP protocol defining their content and characteristics to complete a SIP session. Below is a high-level description of each SIP component and the role it plays in this process.

SIP User Agents (UAs) are the end-user devices, such as cell phones, multimedia handsets, PCs, PDAs, etc. used to create and manage a SIP session. The User Agent Client initiates the message. The User Agent Server responds to it.

SIP Registrar Servers are databases that contain the location of all User Agents within a domain. In SIP messaging, these servers retrieve and send participants' IP addresses and other pertinent information to the SIP Proxy Server.

SIP Proxy Servers accept session requests made by a SIP UA and query the SIP Registrar Server to obtain the recipient UA's addressing information. It then forwards the session invitation directly to the recipient UA if it is located in the same domain or to a Proxy Server if the UA resides in another domain.

SIP Redirect Servers allow SIP Proxy Servers to direct SIP session invitations to external domains. SIP Redirect Servers may reside in the same hardware as SIP Registrar Severs and SIP Proxy Servers.

The following scenarios demonstrate how SIP components work in harmony to establish SIP sessions between UAs in nowadays and future 4G networks. In Figure 1.30 a simplified illustration of a call between VoIP SIP phones within the same SIP IP telephony network is given. When calls are made within a single SIP IP telephony network, the process typically involves the origination and destination phones and a single proxy server.

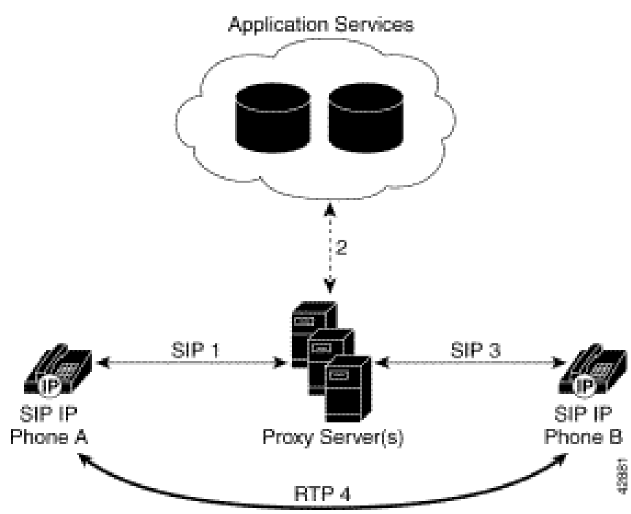


Figure 1.30. Calls Within a Single SIP VoIP Network

In this illustration, the following sequence occurs:
1. SIP IP phone A initiates a call by sending an INVITE message to the SIP proxy server. (There can be more than one proxy server for redundancy.)
2. The SIP proxy server interacts with the location server and possibly with application services to determine user addressing, location, or features.
3. The SIP proxy server then proxies the INVITE message to the destination phone.
4. Responses and acknowledgments are exchanged, and an RTP session is established between SIP IP phones A and B.

When calls are made between SIP VoIP networks, the process typically involves the origination and destination phones as well as two or more SIP proxy servers. Figure 1.31 is a simplified illustration of a call between SIP VoIP phones in different SIP VoIP networks.

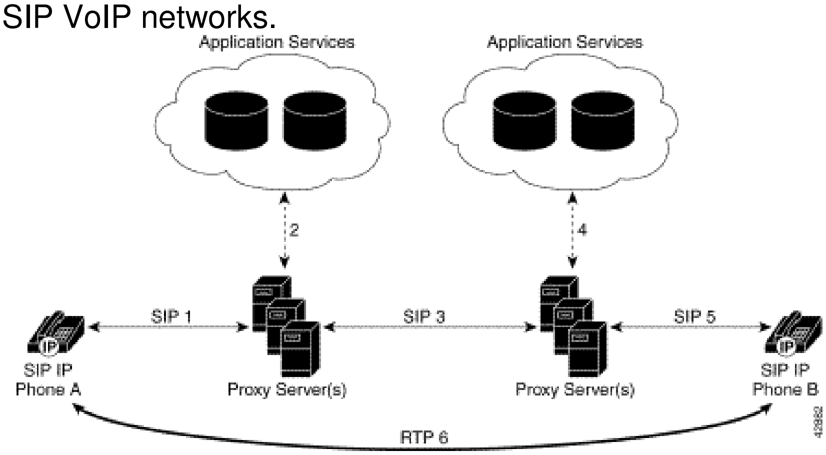


Figure 1.31 Calls Between SIP IP Telephony Networks

In this illustration, the following sequence occurs:

1. SIP IP phone A initiates a call by sending an INVITE to the SIP proxy server. (There can be more than one proxy server for redundancy.)
2. The SIP proxy server might interact with application services such as RADIUS to obtain additional information.
3. The SIP proxy server in phone A's network contacts the SIP proxy server in phone B's network. The local proxy uses the domain name system (DNS) domain to determine if it should handle the call or route it to another proxy. The remote proxy is contacted based on the domain of the destination device.
4. The SIP proxy server in phone B's network might interact with application services to obtain additional information.
5. The SIP proxy server in phone B's network contacts the destination phone (SIP IP phone B).
6. Responses and acknowledgments are exchanged, and an RTP session is established between SIP IP phones A and B.

Moreover, SIP 200 OK, 180 Ringing, and 183 Session Progress messages pass through the same set of proxies, for they are in the same call sequence. SIP CANCEL or BYE requests sent by a terminating user agent might or might not pass through the same set of proxies.

Furthermore, when calls are made between a SIP VoIP network and a traditional telephony network, the process typically involves the origination phone, one or more proxy servers, a gateway, and a PBX or PSTN device. Figure 1.32 is a simplified illustration of a call between a SIP IP phone and a traditional phone in a traditional PSTN.

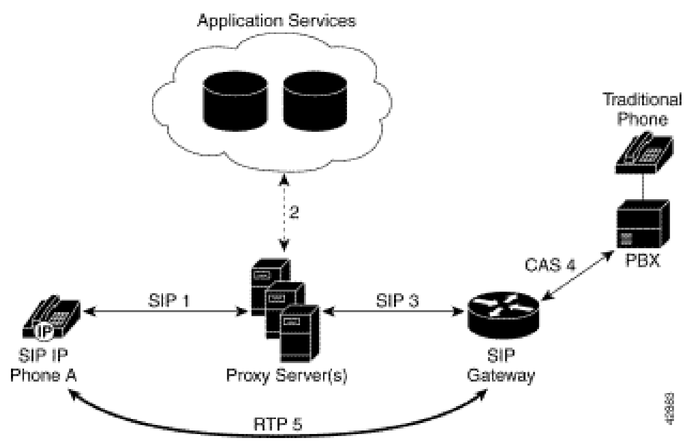


Figure 1.32 Calls Between a SIP VoIP Network and a Traditional Telephony Network

In this illustration, the following occurs:

1. SIP IP phone A initiates a call by sending an INVITE to the SIP proxy server. (There can be more than one proxy server for redundancy.)
2. The SIP proxy server might interact with application services such as RADIUS to obtain additional information.
3. The SIP proxy server proxies the INVITE to the SIP gateway.
4. The SIP gateway establishes communication with the traditional telephony network, in this case a PBX.
5. Responses and acknowledgments are exchanged, and an RTP session is established between SIP IP phone and the SIP gateway. The signaling on the plain-old-telephone-service (POTS) side of the gateway is translated into SIP messages on the IP network to provide proper ringback signaling to the end-user phones.

In addition to SIP, other examples of session control protocols include BICC, SIP-I and SIP-T. BICC, or Bearer Independent Call Control, is the protocol standardized in the 3GPP Release 4 architecture and deployed in some networks today. BICC, however, is not an optimal choice for ongoing evolution because it has been limited to, and is predicted to remain limited to, operation within a GSM-UMTS context. BICC does not address domains beyond GSM-UMTS such as LTE; as a result, it does not automatically offer the future level of flexibility of continued development and evolution that would accompany the SIP with ISUP encapsulation variants (i.e. either SIP-T, SIP for Telephones or SIP-I, SIP with ISUP encapsulation).

With a technical analysis of capabilities existing within the two SIP technologies with ISUP encapsulation variants, 4G Americas recommends SIP-I as the direction for evolution. There are four areas where SIP-I is better suited for a GSM-UMTS environment than SIP-T: Assumptions regarding the trust and security environment, encapsulation procedures & message mapping, support of Request for Comments (RFCs) and user plane interoperability.

To conclude why SIP is one of the main signalling protocol used in 4G signalling processes? Because of the following SIP advantages:

- ✓ Readability- SIP is expressed in form of text. Its syntax is easy to understand and the semantics is close to human natural language. Hence, as a communication signaling protocol, SIP is very conducive to communication whether for development, debugging, or interoperability test.
- ✓ Scalability- SIP protocol structure boasts of strong scalability.So far, there are hundreds of extension protocol based on the SIP core of Protocol RFC3261, among which are documents defined by the IETF, OMA, GSMA and 3GPP Organizations. And thus various types of business expand. SIP has no restriction in communication carrier and can build any communication session from voice to video.
- ✓ Collaboration- SIP can work together with many other protocols to provide more powerful business capability. Its protocols include SDP, DHCP, HTTP, DNS, RADIUS, XCAP, TBCP, MSRP, etc. What's more, it can even work with the protocols that you defined.
- ✓ Epidemic– there are more and more technology using SIP. Apart from the traditional application of VoIP, the main protocol in IMS, the application system of IP domain in 3G and 4G LTE in the network operator area,also adopts SIP. SIP protocol is broadly supported by the Internet and IT manufacturers, among which are the famous Microsoft, Apple and AOL.

1.7.2. DIAMETER

The Diameter protocol was derived from the RADIUS protocol with a lot of improvements in different aspects, and is generally believed to be the next generation Authentication, Authorization, and Accounting (AAA) protocol used in 4G control and signalling plane. The Diameter protocol was widely used in the IMS architecture for IMS entities to exchange AAA-related information. Because the IMS system is the next big thing in the telecom industry as part of the 4 G core network architecture. A clear understanding of the Diameter protocol is necessary for understanding the essence of the IMS architecture. This article offers an overview of Diameter and how it works. For developers interested in how AAA in IMS works, or who want to implement Diameter applications, this section is a good starting page.

With the emergence of new technologies and applications such as wireless networks and Mobile IPs, the requirements for authentication and authorization have greatly increased, and access control mechanisms are more complex than ever. The existing RADIUS (Remote Authentication Dial-In User Service) protocol can be insufficient to cope with these new requirements; what's needed is a new protocol that is capable of fulfilling new access control features while keeping the flexibility for further extension. This is where the Diameter protocol comes into play.

To emphasise that this section provides an overview of Diameter and does not cover all the protocol details. If you want to go further and implement the Diameter base protocol, refer to RFC3588 in Resources for more details. So, because this section mainly addresses the base protocol, Diameter will refer to the Diameter Base Protocol. Before immersing ourselves in protocol details, let's see what drives the requirement for AAA protocol.

In the past, people tried to dial into their Internet Service providers (ISPs) by providing their ID and password to an access server, which then authenticated the user before granting Internet access. In most cases, a user's credential information is not stored directly in the access server, but in a more secure location such as a Lightweight Directory Access Protocol (LDAP) server behind a boundary firewall. Therefore, a standardized protocol is required between the access server and the user information repository in order to exchange authentication-, authorization-, and accounting-related information. The RADIUS protocol was designed to provide a simple, but efficient, way to deliver such AAA capability. As with the evolution of network applications and protocols, new requirements and mechanisms are required to authenticate users. These requirements are summarized in RFC2989 (see References), which includes such topics as failover, security, and audit ability. Although there are some subsidiary protocols intended to extend the capability of the RADIUS protocol, a more extensible and general protocol was expected. The Diameter protocol was then derived from that of RADIUS, and designed to be a general framework for future AAA applications.

The Diameter protocol is not a brand-new one for AAA, but rather, as its name implies, is an enhanced version of the RADIUS protocol. It includes numerous enhancements in all aspects, such as error handling and message delivery reliability. It extracts the essence of the AAA protocol from RADIUS and defines a set of messages that are general enough to be the core of the Diameter Base protocol. The various applications that require AAA functions can define their own extensions on top of the Diameter base protocol, and can benefit from the general capabilities provided by the Diameter base protocol. Figure 1.33 illustrates the relationship between the Diameter base protocol and various Diameter applications.

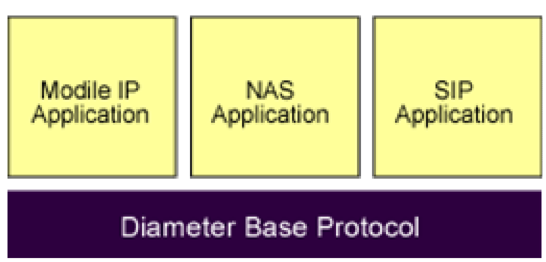
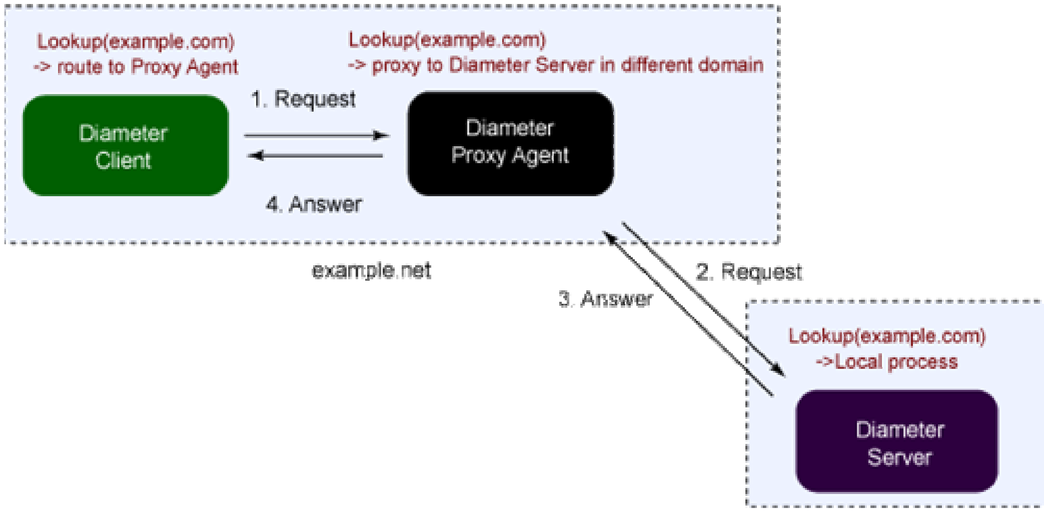


Figure 1.33. The relationship of the Diameter base protocol and Diameter applications.

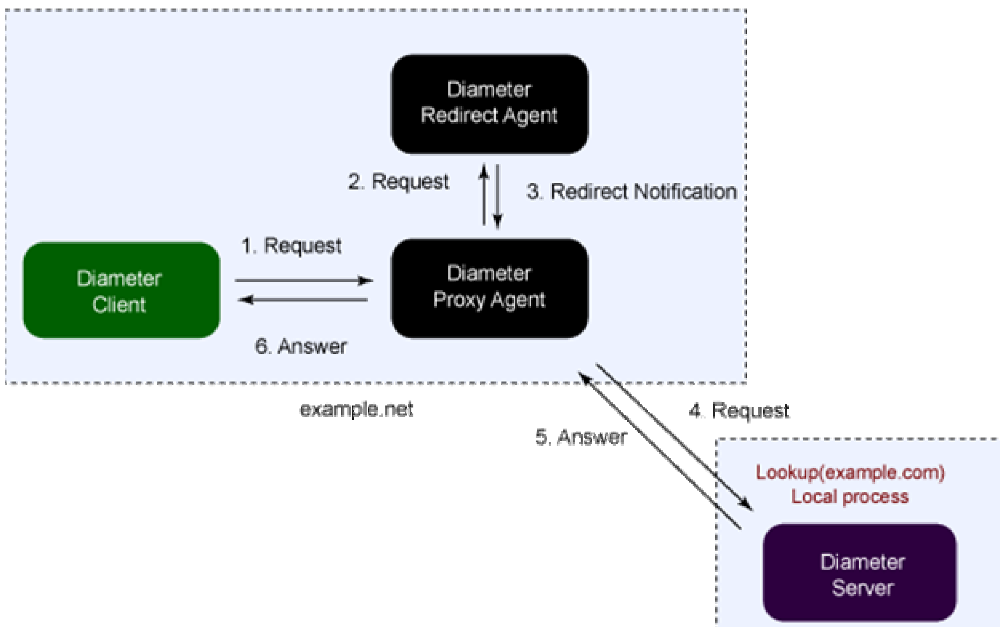
Moreover, the Diameter is designed as a Peer-To-Peer architecture, and every host who implements the Diameter protocol can act as either a client or a server depending on network deployment. So the term Diameter node is used to refer to a **Diameter client, a Diameter server, or a Diameter agent**, which we will introduce later. The Diameter node that receives the user connection request will act as the Diameter client. In most cases, a Diameter client will be a Network Access Server. After collecting user credentials, such as username and password, it will send an access request message to one Diameter node serving the request. For simplicity, we assume it is the Diameter server. The Diameter server authenticates the user based on the information provided. If the authentication process succeeds, the user's access privileges are included in the response message and sent back to the corresponding Diameter client. Otherwise, an access reject message is sent.

Although the architecture just described looks like a traditional client-server architecture, a node acting as the Diameter server for some requests might actually act as a Diameter client in some situations; the Diameter protocol is actually peer-to-peer-based architecture in a more generic sense. Besides, a special Diameter node called Diameter agent is clearly defined in Diameter. Typically, there are three kinds of Diameter agents:

- Relay Agent: A Relay Agent is used to forward a message to the appropriate destination, depending on the information contained in the message. The Relay Agent is advantageous because it can aggregate requests from different realms (or regions) to a specific realm, which eliminates the burdensome configurations of network access servers for every Diameter server change.
- Proxy Agent: A Proxy Agent can also be used to forward messages, but unlike a Relay Agent, a Proxy Agent can modify the message content and, therefore, provide value-added services, enforce rules on different messages, or perform administrative tasks for a specific realm. Figure 1.34 shows how a Proxy Agent is used to forward a message to another domain. If the Proxy Agent will not modify the content of an original request, a Relay Agent in this scenario would be sufficient.
- Redirect Agent: A Redirect Agent acts as a centralized configuration repository for other Diameter nodes. When it receives a message, it checks its routing table, and returns a response message along with redirection information to its original sender. This would be very useful for other Diameter nodes because they won't need to keep a list routing entries locally and can look up a Redirect Agent when needed. Figure 1.35 illustrates how a Redirect Agent works. The scenario in Figure 1.35 below is basically identical to the one in Figure 1.34, but this time the Proxy Agent is not aware of the address of the contacting Diameter node within example.com. Therefore, it looks up the information in the Redirect Agent of its own realm to get the address.



1.34. The Diameter Proxy Agent



1.35. The Diameter Redirect Agent

In addition to these agents, there is a special agent called **Translation Agent**. The responsibility of this agent, as you might have guessed, is to convert a message from one AAA protocol to another. The Translation Agent is helpful for a company or a service provider to integrate the user database of two application domains, while keeping their original AAA protocols. Another situation is that a company wants to migrate to Diameter protocol, but the migration consists of many phases. The Translation Agent could provide the backward capability for a smooth migration. Figure 1.36 shows how one agent translates the RADIUS protocol into the Diameter protocol, but, of course, other kinds of protocol translation (for example, Diameter to RADIUS, Diameter to TACACS+) are also possible.

Figure 1.36. The Diameter Translation Agent

A Diameter message is the base unit to send a command or deliver a notification to other Diameter nodes. For different purposes, Diameter protocol has defined several types of Diameter messages, which are identified by their command code. For example, an Accounting-Request message recognizes that the message carries accounting-related information, while a Capability-Exchange-Request message recognizes that the message carries capability information of the Diameter node sending the message. Because the message exchange style of Diameter is synchronous, each message has its corresponding counterpart, which shares the same command code. In both previous examples, the receiver of an Accounting-Request message prepares an Account-Answer message and sends it to the original sender. The command code is used to identify the intention of a message, but the actual data is carried by a set of Attribute-Value-Pairs (AVPs). The Diameter protocol has predefined a set of common attributes and imposes each attribute with a corresponding semantic. These AVPs carry the detail of AAA as well as routing, security, and capability information between two Diameter nodes. In addition, each AVP is associated with an AVP Data Format, which is defined within the Diameter protocol (for example, OctetString, Integer32), so the value of each attribute must follow the data format. In Figure 1.37 the relationship between Diameter messages and their AVPs is given. For more details on each part of Figure 1.37, see Chapters 3 and Chapter 4 of the RFC3588 Diameter base protocol.

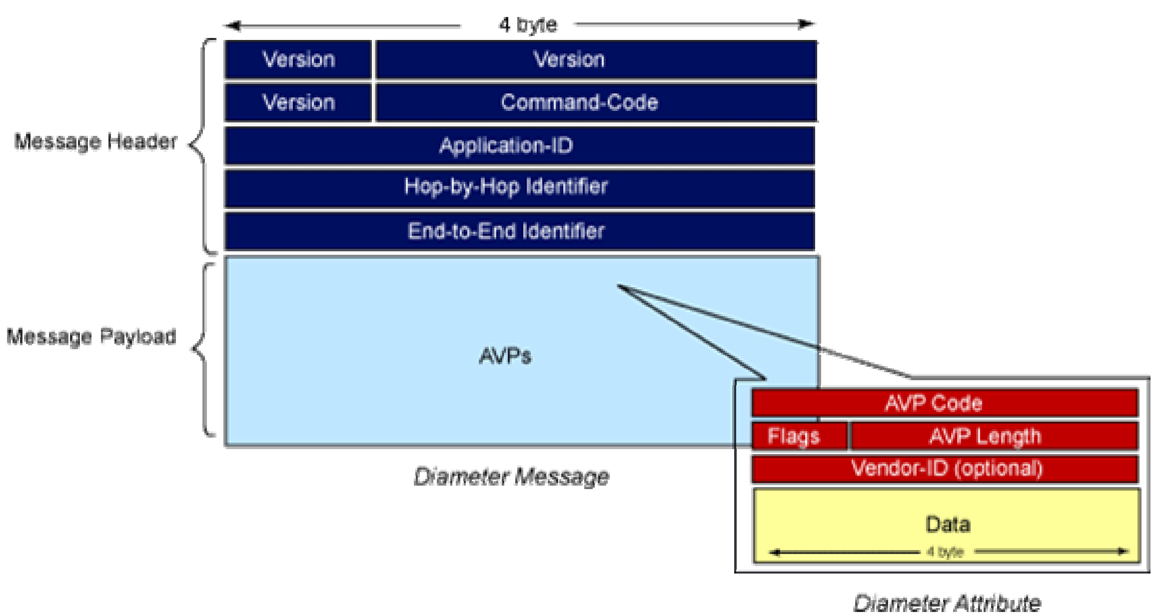


Figure 1.37. The Diameter Packet Format.

As we mentioned, the Diameter protocol isn't bound to a specific application running on top of it. It focuses on general message exchanging features. Because authentication and authorization mechanisms vary among applications, the Diameter base protocol doesn't define command codes and AVPs specific to authentication and authorization. It is the responsibility of Diameter applications to define their own messages and corresponding attributes based on the application's characteristics. For example, the AA-Request message is used to carry authentication and authorization information in the NAS application, while in the SIP application the message is called User-Authorization-Request.

Unlike authentication and authorization, the behavior and message to be exchanged for accounting is clearly defined. Accounting in Diameter essentially follows a server directed model, which means the device that generates accounting records follows the direction of an authorization server.

Based on the user profile or any business condition, a Diameter server informs the corresponding Diameter client as to what behavior is expected, such as how often the accounting record should be sent from client to server, or if the accounting record should be generated continuously within an accounting session.

Generally speaking, depending on the service to be provided, there are two kinds of accounting records: For one-time invocation-based services, the EVENT_RECORD is used. However, if the service will be provided in a measurable period, the accounting record types START_RECORD, INTERIM_RECORD, and STOP_RECORD could be used to mark the start, update, and end of a session. In order to prevent duplicated accounting records, each accounting message is associated with a Session-Id AVP along with an Accounting-Record-Number AVP. As this combination can uniquely identify an accounting record, a Diameter node acting as a Diameter agent can use this information to detect duplicated accounting messages being sent to the Diameter server, thereby avoiding unnecessary processing for the Diameter server. This situation might come from temporary network problems or client shutdowns. Also, it is required that the Diameter client keep a local cache of outgoing accounting messages until an acknowledgement message arrives.

In addition to SIP, Diameter is the other core protocol used in the IMS architecture for 4G, in the service and control plane. IMS defines a set of reference points between different IMS entities and some of them use Diameter as the underlying protocol to exchange subscription, presence, and billing-related messages. For example, the Sh reference point in IMS defined a set of Diameter messages for subscription and notification purposes.

As IMS 4G core continues to evolve, we believe there will be more Diameter applications to come, as well as Diameter-related implementations.

1.7.3. Mobile IP

The Mobile IP is officially referred to as "Internet Protocol Mobility Support." It is an area under rapid development and one of the factors driving the requirements to redevelop the Internet Protocol as IPv6. Generally, Mobile IP can be thought of as the cooperation of **three major subsystems**. First, there is a discovery mechanism defined so that mobile terminals can determine their new attachment points (new IP addresses) as they move from place to place within the Internet. Second, once the mobile terminal knows the IP address at its new attachment point, it registers with an agent representing it at its home network. Lastly, mobile IP defines simple mechanisms to deliver datagrams to the mobile node when it is away from its home network.

In the beginning of this sub-section, it is a good idea to frame the discussion by setting some terminology, adapted from the mobile IP specification. Mobile IP introduces the following new functional entities:

- **Mobile node** - A host or router that changes its point of attachment from one network or subnetwork to another, without changing its IP address. A mobile node can continue to communicate with other Internet nodes at any location using its (constant) IP address.
- **Home agent** - A router on a mobile node's home network which delivers datagrams to departed mobile nodes, and maintains current location information for each.
- **Foreign agent** - A router on a mobile node's visited network which cooperates with the home agent to complete the delivery of datagrams to the mobile node while it is away from home.

A mobile node has a home address, which is a long-term IP address on its home network. When away from its home network, a care-of address is associated with the mobile node and reflects the mobile node's current point of attachment. The mobile node uses its home address as the source address of all IP datagrams it sends, except where otherwise required for certain registration request datagrams.

The following terms are frequently used in connection with Mobile IP:

- **Agent advertisement** - Foreign agents advertise their presence by using a special message, which is constructed by attaching a special extension to a router advertisement, as described in the next section.
- **Care-of-address** - The termination point of a tunnel toward a mobile node, for datagrams forwarded to the mobile node while it is away from home. There are two different types of care-of-address (CoA): a foreign agent care-of address is an address of a foreign agent with which the mobile node is registered; a collocated care-of address is an externally obtained local address which the mobile node has associated with one of its own network interfaces.
- **Correspondent node** - A peer with which a mobile node is communicating. A correspondent node may be either mobile or stationary.
- **Foreign network** - Any network other than the mobile node's home network.
- **Home address** - An IP address that is assigned for an extended period of time to a mobile node. It remains unchanged regardless of where the node is attached to the Internet.

- **Home network** - A network, possibly virtual, having a network prefix matching that of a mobile node's home address. Note that standard IP routing mechanisms will deliver datagrams destined to a mobile node's home address to the mobile node's home network.
- **Link** - A facility or medium over which nodes can communicate at the link layer. A link underlies the network layer.
- **Link-layer address** - The address used to identify an endpoint of some communication over a physical link. Typically, the link-layer address is an interface's media access control (MAC) address.
- **Mobility agent** - Either a home agent or a foreign agent.
- **Mobility binding** - The association of a home address with a care-of address, along with the remaining lifetime of that association.
- **Mobility security association** - A collection of security contexts between a pair of nodes which may be applied to mobile IP protocol messages exchanged between them. Each context indicates an authentication algorithm and mode (as described in the fourth section), a secret (a shared key, or appropriate publiciprivate key pair), and a style of replay protection in use.
- **Nonce**- A randomly chosen value, different from previous choices, inserted in a message to protect against replays. Security parameters index (SPI) - An index identifying a security context between a pair of nodes among the contexts available in the mobility security association.
- **Tunnel** - The path followed by a datagram while it is encapsulated. The model is that, while encapsulated, a data gram is routed to a knowledgable agent, which decapsulates the datagram and then forwards it along to its ultimate destination.
- **Virtual network** - A network with no physical instantiation beyond its router (with a physical network interface on another network). The router (e.g., a home agent) generally advertises reachability to the virtual network using conventional routing protocols.
- **Visited network** - A network other than a mobile node's home network to which the mobile node is currently connected.
- **Visitor list** - The list of mobile nodes visiting a foreign agent.

To emphasize that nowadays we have two types of Mobile IP protocol versions: Mobile IPv4 (MIPv4) and Mobile IPv6 (MIPv6). MIPv4 is a popular mobility protocol used in the current IPv4 networks, but with the next generation networks emerging developments, there are the IPv6 networks, and the MIPv6 protocol. MIPv6 is design to deal with mobility and to overcome some problems suffered by MIPv4. Although MIPv6 shares many features with MIPv4, there are some differences between them (discussed later in this sub-section). The most significant difference between MIPv4 and MIPv6 is that MIPv6 is integrated into the base IPv6 protocol and not an add-on feature, as is the case with IPv4 and MIPv4. Because most Internet devices will soon be mobile, it is important that all devices are inherently designed to be mobile and IPv6/MIPv6 allows for this.

Furthermore, we will elaborate the three general functions in Mobile IPv4. Those related functions are:

- ✓ **Agent Discovery** - Mobility agents advertise their availability on each link for which they provide service.
- ✓ **Registration** - When the mobile node is away from home, it registers its CoA with its home agent.
- ✓ **Tunneling** - In order for datagrams to be delivered to the mobile node when it is away from home, the home agent has to tunnel the datagrams to the care-of address.

The following will give a rough outline of operation of the Mobile IPv4 protocol, making use of the above-mentioned operations. Figure 1.38 may be used to help envision the roles played by the entities.

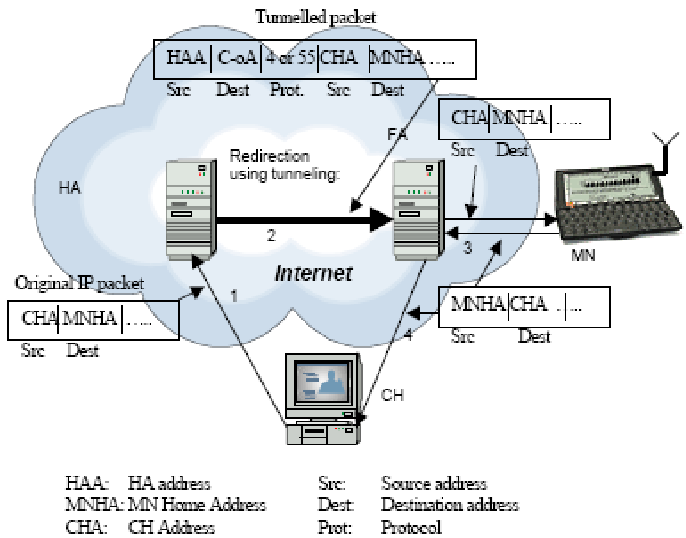


Figure 1.38. Mobile IP packet flow.

Mobility agents make themselves known by sending agent advertisement messages. An impatient mobile node may optionally solicit an agent advertisement message. After receiving an agent advertisement, a mobile node determines whether it is on its home network or a foreign network. A mobile node basically works like any other node on its home network when it is at home.

When a mobile node moves away from its home network, it obtains a CoA on the foreign network, for instance, by soliciting or listening for agent advertisements, or contacting DHCP or Point-to-Point Protocol (PPP). While away from home, the mobile node registers each new CoA with its home agent, possibly by way of a foreign agent. Figure 1.39 illustrated the Mobile IP registration process. IP packets sent to the mobile node's home address are intercepted by its home agent, tunneled by its home agent to the CoA, received at the tunnel endpoint (at either a foreign agent or the mobile node itself), and finally delivered to the mobile node. In the reverse direction, datagrams sent by the mobile node are generally delivered to their destination using standard IP routing mechanisms, not necessarily passing through the home agent.

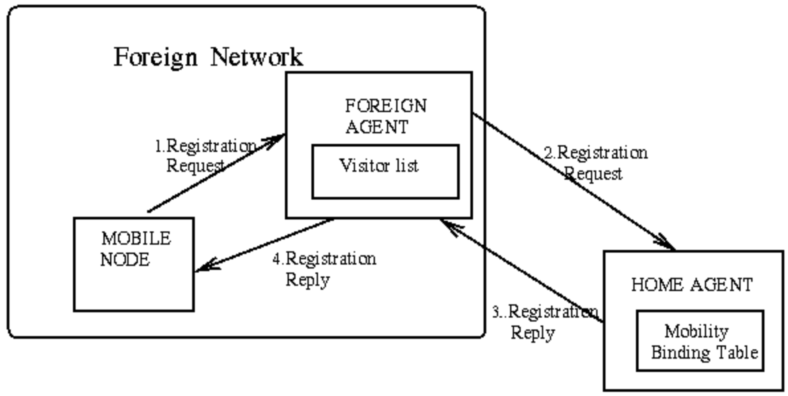


Figure 1.39. Mobile IP registration process.

When the home agent tunnels a datagram to the CoA, the inner IP header destination (i.e., the mobile node's home address) is effectively shielded from intervening routers between its home network and its current location. At the CoA, the original datagram exits from the tunnel and is delivered to the mobile node. It is the job of every home agent to attract and intercept datagrams that are destined to the home address of any of its registered mobile nodes. In Figure 1.40, the tunneling process in Mobile IPv4 is plotted.

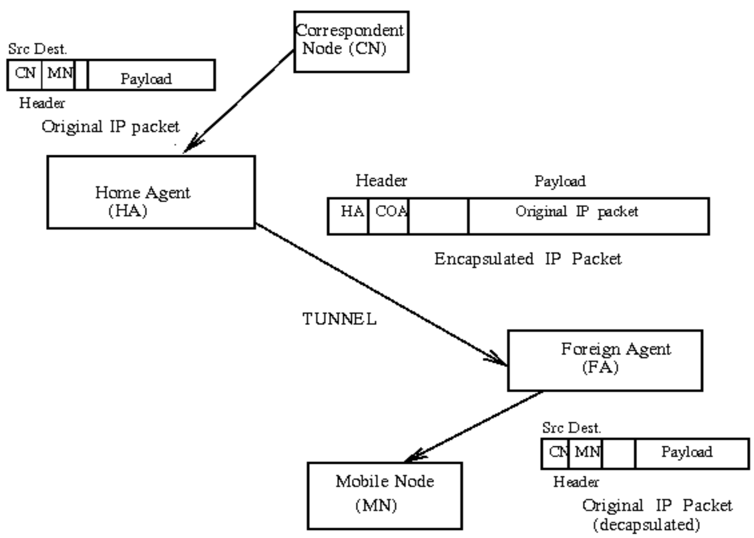


Figure 1.40. Mobile IP tunneling process.

The home agent basically does this by using a minor variation on proxy Address Resolution Protocol (ARP), and to do so in the natural model it has to have a network interface on the link indicated by the mobile node's home address. However, the latter requirement is not part of the mobile IP specification. When foreign agents are in use, similarly, the natural model of operation suggests that the mobile node be able to establish a link its foreign agent. Other configurations are possible, however, using protocol operations not defined by (and invisible to) mobile IP. Notice that, if the home agent is the only router advertising reachability to the home network, but there is no physical link instantiating the home network, then all datagrams transmitted to mobile nodes addressed on that home network will naturally reach the home agent without any special link operations.

On the other hand, Mobile IPv6 is the next generation mobile protocol and in the near future, all nodes/routers are going to become more faster and the new technologies are going to reduce the Internet delay and will provide advanced mobility management. IETF (Internet Engineering Task Force) expects that the IPv6 protocol will replace the IPv4 protocol in the near future. Although space does not permit a full exposition of the details of the proposed MIPv6, some overall discussion is certainly in order.

The Mobile IPv6 uses the experiences gained from the design and development of Mobile IPv4 together with the new IPv6 protocol features. Mobile IPv6 shares many features with Mobile IPv4, but the protocol is now fully integrated into IPv6 and provides many improvements over Mobile IPv4. The major differences between Mobile IPv4 and Mobile IPv6 are:

- ✓ Support for "Route Optimisation": This feature is now built in as a fundamental part of the Mobile IPv6 protocol. In Mobile Ipv4 the route optimisation feature is being added on as an optional set of extensions that may not be supported by all IP nodes.
- ✓ In Mobile IPv6 (also integrated in the IPv6) a new feature is specified that allows Mobile Nodes and Mobile IP to coexist efficiently with routers that perform "ingress filtering" [RFC2267]. The packets sent by a Mobile Node can pass normally through ingress filtering routers. This can be accomplished due to the fact that the CoA is used as the Source Address in each packet's IP header. Moreover, the Mobile Node's home address is carried in the packet in a Home Address destination option. This allows the use of the care-of address in the packet to be transparent above the IP layer, e.g., TCP.
- ✓ By using the CoA as the Source Address in each packet's IP header the routing of multicast packets sent by a Mobile Node is simplified. In Mobile IPv6 the Mobile Node will not anymore have to tunnel multicast packets, as specified in Mobile IPv4, to its Home Agent. Moreover, the use of the Home Address option allows the home address to be used but still be compatible with multicast routing that is based in part, on the packet's Source Address.
- ✓ In Mobile IPv6 the functionality of the Foreign Agents can be accomplished by IPv6 enhanced features, such as Neighbour Discovery and Address Autoconfiguration [RFC1971]. Therefore, there is no need to deploy Foreign Agents in Mobile IPv6.
- ✓ The Mobile IPv6, unlike Mobile IPv4, uses IPsec for all security requirements such as sender authentication, data integrity protection, and replay protection for Binding Updates (which serve the role of both registration and Route Optimisation in Mobile IPv4). In Mobile IPv4 the security requirements are provided by its own security mechanisms for each function, based on statically configured mobility security associations.

- ✓ In mobile IPv6 a mechanism is provided to support bidirectional (i.e., packets that the router sends are reaching the Mobile Node, and packets that the Mobile Node sends are reaching the router) confirmation of a Mobile Node's ability to communicate with its default router in its current location. This bidirectional confirmation can be used to detect the “black hole” situation, where the link to the router does not work equally well in both directions. In contrast, Mobile IPv4 does not support bidirectional confirmation, but only the forward direction (packets from the router are reaching the Mobile Node) is confirmed, and therefore the black hole situation may not be detected.
- ✓ Mobile IPv6 and IPv6 use the source routing feature. This feature makes it possible for a Correspondent Host to send packets to a Mobile Node while it is away from its home network using an IPv6 Routing header rather than IP encapsulation, whereas Mobile IPv4 must use encapsulation for all packets. However, in Mobile IPv6 the Home Agents are allowed to use encapsulation for tunnelling. This is required, during the initiation phase of the binding update procedure.
- ✓ In Mobile IPv6 the packets which arrive at the home network and are destined for a Mobile Node that is away from home, are intercepted by the Mobile Node’s Home Agent using IPv6 Neighbour Discovery [RFC1970] rather than ARP [RFC826] as is used in Mobile IPv4.
- ✓ The source routing (routing header) feature in Mobile IPv6 removes the need to manage "tunnel soft state", which was required in Mobile IPv4 due to limitations in ICMP error procedure for IPv4. In Mobile IPv4 an ICMP error message that is created due to a failure of delivering an IP packet to the Care-of Address, will be returned to the home network, but will may not contain the IP address of the original source of the tunnelled IP packet. This is solved in the Home Agent by storing the tunneling information, i.e., which IP packets have been tunnelled to which Care-of Address, called tunneling soft state.
- ✓ In IPv6 a new routing procedure is defined called anycast. This feature is used in Mobile IPv6 for the dynamic Home Agent address discovery mechanism. This mechanism returns one single reply to the Mobile Node, rather than the corresponding Mobile IPv4 mechanism that used IPv4 directed broadcast and returned a separate reply from each Home Agent on the Mobile Node's home subnetwork. The Mobile IPv6 mechanism is more efficient and more reliable. This is due to the fact that only one packet need to be replied to the Mobile Node.
- ✓ In Mobile IPv6 an Advertisement Interval option on Router Advertisements (equivalent to Agent Advertisements in Mobile IPv4) is defined, that allows a Mobile Node to decide for itself how many Router Advertisements (Agent Advertisements) it is tolerating to miss before declaring its current router unreachable.
- ✓ All Mobile IPv6 control traffic can be piggybacked on any existing IPv6 packets. This can be accomplished by using the IPv6 destination options. In contrary, for Mobile IPv4 and its Route Optimisation

extensions, separate UDP packets were required for each control message.

- ✓ In Mobile IPv6 supports Hierarchical Mobile IPv6 (HIPv6) plus Fast Handovers for Mobile IPv6 (FHIPv6). HMIPv6 is a localized mobility management proposal that aims to reduce the signaling load due to user mobility. The mobility management inside the local domain is handled by a Mobility Anchor Point (MAP). Mobility between separate MAP domains is handled by MIPv6. Moreover, the HMIPv6 presents the following advantages: it includes a mechanism to reduce the signaling load in case of handoffs within the same domain and may improve handoff performance reducing handoff latency and packet losses since intra-domain handoffs are performed locally. However, since the periodic BUs are not reduced but the ones due to handoffs, the gain depends on the mobility of the mobile nodes. For more details on HMIPv6 see the reference [33]. On the other hand, FHIPv6 protocol enables an mobile nodes to quickly detect that it has moved to a new subnet by providing the new access point and the associated subnet prefix information when the mobile node is still connected to its current subnet. For instance, a mobile node may discover available access points using link-layer specific mechanisms (i.e., a "scan" in WLAN) and then request subnet information corresponding to one or more of those discovered access points. The mobile node may do this after performing router discovery or at any time while connected to its current router. For more details on FMIPv6 see the reference [34].

Moreover, in comparison to Mobile IPv4 protocol, Mobile IPv6 protocol can provide mobility support that combines the experience gained in the design of Mobile IPv4 and the new features of the IPv6 protocol. Some of the Mobile IPv4 open issues, i.e., Triangle routing, Mobility routing crossings in an Intranet, RSVP operation over IP tunnels, Inefficient maintenance of simultaneous bindings, Ingress filtering, Minimize the number of required trusted entities and Authentication are partially solved. Most of the solutions provided in Mobile IPv6 are mainly generated for Mobile IPv4. However, it is expected that some of these (above mentioned) solutions, after some minor modifications, can also be applied for Mobile IPv6.

The goal for Mobile IPv6 is to provide provides seamless mobility for next generation mobile services and applications and across several access technologies such as WCDMA, WiMAX 802.16m, WLAN 802.11m, LTE-Advanced and other 4G access networks. Undoubtedly, Mobile IPv6, along with fast-handoffs and context transfer mechanisms will be essential for the large scale deployment of real-time services (such as VoIP) and broadcast services in 4G networks.

References

[1] ITU-R Rec. M.1645, 'Framework and overall objectives of the future development of IMT 2000 and systems beyond IMT 2000'.

[2] ITU-R M.2072, 'World mobile telecommunication market forecast'.

[3] ITU-R M.2134, 'Requirements related to technical performance for IMT-Advanced radio interface(s)'

[4] 3GPP LTE : http://www.radio-electronics.com/info/cellulartelecomms/lte-long-term-evolution/3g-lte-basics.php

[5] 3 GPP LTE-Advanced: http://www.radio-electronics.com/info/cellulartelecomms/lte-long-term-evolution/3gpp-4g-imt-lte-advanced-tutorial.php

[6] http://www.3gpp.org/

[7] http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/

[8] IEEE Computer Society and the IEEE Microwave Theory and Techniques Society: "IEEE Std 802.16m-2011 Standard for local and metropolitan area networks", 2011.

[9] Chris Thomas, Raj Jain, "802.16m and WiMAX Release 2.0", 2010.

[10] Yaghoobi, Hassan, "Mobile WiMAX Update and IEEE 802.16m," IEEE, March 2009.

[11] IEEE 802.11n-2009—Amendment 5: Enhancements for Higher Throughput. IEEE-SA. 29 October 2009.

[12] Broadcom corporation, "802.11n: Next-Generation Wireless LAN Technology", April 2006.

[13] Thomas Paul and Tokunbo Ogunfunmi , "Wireless LAN Comes of Age: Understanding the IEEE 802.11n Amendment", IEEE Circuits and Systems Magazine Vol. 8(1), pp. 28-54, 2008.

[14] A. Bacioccola, C. Cicconetti, C. Eklund, L. Lenzini, Z. Li, E. Mingozzi, "IEEE 802.16: History, Status and Future Trends", Computer Communications, Volume 33, Issue 2, Pages 113-123, 15 February 2010.

[15] Srinivasan, Roshni (ed), Hamiti, Shkumbin (ed), "IEEE 802.16m System Description Document (SDD)," IEEE 802.16 Task Group m, September 2009.

[16] 3GPP TR 23.882: '3GPP system architecture evolution, report on technical options and conclusions'.

[17] ETSI: 'Long term evolution of the 3GPP radio technology' and 'System architecture evolution'.

[18] ITU-T Rec. Y.2216, ‘NGN capability requirements to support the multimedia communication centre service’.

[19] ITU-T Rec. Y.2205, ‘Next Generation Networks – Emergency telecommunications – Technical considerations’.

[20] ITU-T NGN – GSI Release 1 ‘NGN_FG-book II’.

[21] ITU-T Rec. Y.2001, ‘General overview of NGN’, 12/2004.

[22] ITU-T Rec. Y.2211, ‘IMS-based real-time conversational multimedia services over NGN’.

[23] Sayan Kumar Ray, “Fourth Generation (4G) Networks: Roadmap- Migration to the Future”, IETE Technical Review Vol 23, No 4, pp 253-265, July-August 2006.

[24] Young-June Choi, Kwang Bok Lee, Saewoong Bahk, “All-IP 4G Network architecture for efficient mobility and resource management”, Wireless Communications, IEEE, Volume: 14 Issue:2, pp.: 42 – 46, April 2007.

[25] Savo G Glisic, “Advanced Wireless Networks: 4G Technology”, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, 2006.

[26] RFC 2989 - Criteria for Evaluating AAA Protocols for Network Access.

[27] RFC3588 – DIAMETER base protocol.

[28] RFC4005: Diameter Network Access Server Application.

[29] RFC2002: IP Mobility Support.

[30] Perkins, C., E., “Mobile IP”, IEEE Communications Magazine, May 1997.

[31] RFC 3344: IP Mobility Support for IPv4.

[32] RFC 3775: Mobility Support in IPv6.

[33] RFC 4140: Hierarchical Mobile IPv6 Mobility Management (HMIPv6).

[34] RFC 4068: Fast Handovers for Mobile IPv6.