1

Overview of Heterogeneous Networks

Geng Wu,[1] Qian (Clara) Li,[1] Rose Qingyang Hu,[2] and Yi Qian[3]
[1] Intel Corporation, USA
[2] Utah State University, USA
[3] University of Nebraska – Lincoln, USA

We are living in a rapidly changing world. Every two days now we create as much information as we did from the dawn of civilization up until 2003 [1]. Users want to communicate with each other at any time, anywhere and through any media, including instant messages, email, voice and video. Users want to share their personal life experience, ideas and news with friends through social networking, and use their intelligent mobile devices to produce and to consume content generated by users or by commercial media. In the meantime, mobile internet is rapidly evolving towards embedded internet, expanding its reach from people to machines [2]. In fact, the wireless industry now expects 50 billion machine-type devices connected to the global network by 2020 [3], truly forming an internet of everything.

The advancement of a number of fundamental technologies powers the rapid market growth. Moore's Law continues to provide more transistors and power budget, enabling the semiconductor industry to deliver more powerful signal processing capabilities at lower power consumption and lower cost. Application developers continue to innovate and maximize the benefits of the signal processing technology, with user interface evolving from keypad to touch to gesture, and applications from voice to video to augmented reality. As our society enters the age of 'Big Data' [4], our communication infrastructure also needs to evolve to meet the overwhelming demands for capacity and bandwidth. The migration from homogenous to heterogeneous network architecture is therefore essential to support a broad range of connectivity and to deliver unprecedented user experience. The future is coming today.

As one of the main pillars and the future trends of mobile communication technology, heterogeneous networks have received a lot of attention in the wireless industry and in the

Heterogeneous Cellular Networks, First Edition. Edited by Rose Qingyang Hu and Yi Qian.

academic research communities. This chapter is intended to provide a technology and business overview of heterogeneous networks, the state of the art in technology development, the main challenges and tradeoffs, and the future research and development directions. However, we are still at an early stage of development of heterogeneous network technology. As you will find throughout this chapter, there are many more questions than answers at this time, and many questions may have more than one valid answer, depending on the market, the target applications and the exact deployment scenarios and competitive environment. We expect that heterogeneous network technology will continue to evolve along with the convergence of information technology and telecommunication, and increasingly intelligent mobile devices.

1.1 Motivations for Heterogeneous Networks

There are significant economic and technological reasons for the rapid development of heterogeneous networks. The outcomes of this technological development are expected to have profound impacts on the future of telecommunications.

1.1.1 Explosive Growth of Data Capacity Demands

In recent years, mobile internet has witnessed an explosive growth in demand for data capacity [5]. This is largely fuelled by the proliferation of more intelligent mobile devices. Market studies have shown that the data traffic volume is a direct function of the device's screen size, the user-friendliness of its operating system and the responsiveness of wireless network that the device is connected to. For example, a 3G smartphone on average consumes about 30 times the system capacity of a 2G voice phone, and a tablet consumes five times the system capacity of a smartphone. As the mobile devices continue to increase in screen size, image resolution and battery life, and as the network infrastructures continue to improve in peak data rate and network latency, the growth in data capacity demand will continue.

In addition to this organic growth in capacity, demand from the improved mobile devices and communication infrastructure, user-generated content and social networking add significant additional burden to the network. In fact, mobile devices are an ideal platform for social networking applications such as Facebook since they offer ubiquitous coverage with its always-on and always-connected connectivity. Social networking and other similar applications usually produce small but frequent data transmissions. A network may have to frequently set up and tear down the radio links to conserve precious radio resources in order to accommodate a large number of users. This often results in an excessive amount of control messages over the control plane. On the other hand, as watching YouTube videos on mobile devices gains popularity, the capacity demand on the data plane is also growing rapidly, and often in an asymmetric fashion between the uplink and the downlink. Finally, depending on how cloud and client partition the signal processing load, cloud-based services may further accelerate demand, as information is shipped between the mobile devices to the cloud for cloud computing and network storage. One such example is Apple's Siri voice reorganization application software. Since the popularity of mobile applications is often difficult to predict, we start to see drastically different capacity demands between the control plane and the data plane, between the uplink and the downlink. We also start to see network congestion expanding from the access network (the traditional capacity bottleneck) to the core network and even to the backbone network and connections.

Machine-type communications add yet another complexity to the future generations of wireless networks. With mobile internet evolving towards embedded internet, future networks need to scale up in size and complexity in order to accommodate an unprecedented number of connected devices with vastly different traffic characteristics, usage models and security requirements. The capacity demands from these machine-type devices range from very low traffic volume monthly meter reading to high speed real-time video surveillance. In addition, securely managing billions of such connected devices across many different types of networks and operating environments adds to the complexity of capacity planning.

The combined capacity demands from organic traffic growth, user-generated contents, social networking and machine-type connected devices require orders of magnitude capacity increase in future wireless networks. This heterogeneous data traffic growth also mandates a paradigm shift in network architecture design and provisioning.

1.1.2 From Spectral Efficiency to Network Efficiency

The wireless industry has several options for meeting the explosive data traffic growth. After decades of relentless air interface innovations, today we are practically reaching the theoretical limit of radio channel capacity, commonly known as the Shannon limit. Although air interface improvement will continue to maximize the benefits of advanced wireless communication research and take full advantage of advanced signal processing technologies for an even higher spectral efficiency, we need several orders of magnitude greater system capacity than what the air interface spectral efficiency improvement can offer. The future capacity increases therefore need to come from a combination of technology solutions, including, in particular, maximizing the overall network efficiency instead of solely relying on the spectral efficiency improvement at the radio link level (Figure 1.1). Heterogeneous networks are a fundamental technology behind most of these solutions.

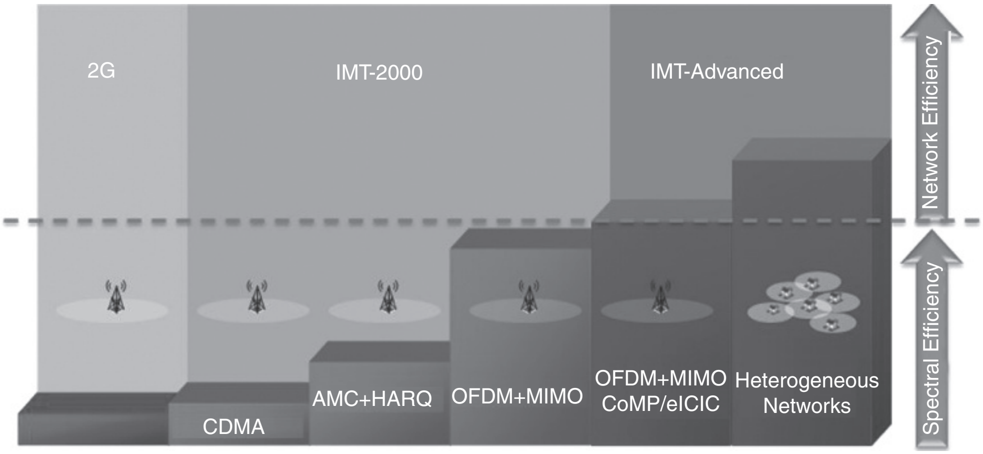


Figure 1.1 Wireless technology evolution.

In the near term, mobile network operators are looking at limiting the monthly data usage of each subscriber over the wireless wide areas networks (WWAN), and throttling the data rate of heavy usage users when necessary. However, limiting usage or throttling capacity demand is in general only a temporary fix to the immediate network overloading problems. We need more proactive solutions to encourage and enable future sustained data traffic growth, and to provide mobile broadband access to all users, and to enrich every person's life on earth.

One such solution that mobile network operators are looking at is the data offloading strategy. This includes (but is not limited to) facilitating and encouraging subscribers to offload their traffic from macro base stations to the alternative small-cell networks, essentially forming a basic heterogeneous network. Since the capacity bottleneck varies from market to market and from network to network, there are many flavours and technical options for offloading strategy, including macrocell network and small-cell network of the same air interface technology, between networks of different air interface technologies, or between mobile operator core network and public internet. There is no single answer to the mobile data offloading question. These options are complementary, and all of them will continue to develop to meet the ever-increasing capacity demands.

Another obvious answer to the growing demand in data capacity is to add more spectrum. The wireless industry and regulators are working together to investigate the possibility of adding more frequency bands, both licensed and unlicensed, for mobile internet applications. However, since there is a limited supply of spectrum, and there is the strong desire for globally harmonized frequency allocation to maximize the economy of scale, the progress in new frequency allocation has been slow. As many densely populated markets are already on the verge of running out of spectrum, we see increased pressure to re-farm the existing frequency bands and for the rapid deployment of small cells for high spatial frequency reuse. In addition, the wireless industry has also started to look at high frequency bands such as millimeter wave for mobile internet applications. Since these bands have very different radio propagation characteristics from the traditional lower frequency bands (usually below 3 GHz) used for high mobility cellular networks, the technology, design and operation of these networks are expected to be very different from traditional cellular networks. Therefore, heterogeneous networks consisting of layers of networks operating at different frequency bands become the main venue for achieving higher system capacity.

In addition to obtaining additional spectrum allocation and developing new technologies for the higher frequency bands, the wireless industry and the research community are also looking at innovative ways for more flexible spectrum utilization, including spectrum sharing, dynamic spectrum access and cognitive radio with opportunistic network access. One such example is the experimental use of TV white space spectrum for wireless communication in the US market. This new type of spectrum access requires additional network entities such as databases that administrate the alternative radio transmitters to operate in the broadcast television spectrum when that spectrum is not used by the licensed service. Since the network coverage and service availability are different from those of the traditional wireless mobile networks due to the dynamic nature of the spectrum availability, the industry is still investigating suitable network architecture and business models to achieve viable return on investment. From a telecommunication infrastructure viewpoint, such new types of networks are expected to become part of the global heterogeneous networks.

1.1.3 Challenges in Service Revenue and Capacity Investment

In recent years, mobile service revenue growth has shifted from circuit-switched voice and short message service (SMS) to data services. This shift adds significant pressure to mobile network operators' profitability for three main reasons. First, mobile data in general yields a lower revenue per bit compared to the traditional voice services and SMS. Secondly, the highly profitable operator walled-garden mobile applications are facing stiff competition from over-the-top mobile applications. Finally, as mobile data traffic explodes, operators need extensive capital investment in new network capacity to meet the demand. Since mobile network operators are instrumental in investing, operating and maintaining global mobile internet infrastructure, it is crucial for the wireless industry and the academic research communities to develop new networking technologies that allow operators to remain profitable and competitive so that they can continue to invest in capacity and new services. Heterogeneous networking is considered one of the most important technologies that not only deliver tens- to thousands-fold system capacity increase but also enable new generations of services to replace the revenue from traditional but diminishing voice-centric telecom services.

To summarize, while the demand for data capacity is exploding and the improvement in spectral efficiency in homogeneous networks is slowing down due to the approaching Shannon limit, it becomes essential that the future focus of wireless technology shifts from further increasing the spectral efficiency of the radio link to improving the overall network efficiency through heterogeneous network architecture and related signal processing technologies. We need heterogeneous networks to deliver a higher system capacity to meet the higher traffic density. We want to leverage heterogeneous network architectures to expand network coverage, to improve service quality and fairness throughout the network coverage areas, in particular at the cell edge. We also want to use heterogeneous networks as a platform for future technological innovations, including the integration of new types of networks, new types of connectivity and new types of connected devices and applications.

1.2 Definitions of Heterogeneous Networks

Heterogeneous networking is one of the most widely used but most loosely defined terms in today's wireless communications industry. Some people consider the overlay of macro base station network and small cell network (e.g., micro, pico and femtocells) of the same air interface technology as heterogeneous networks. Others consider cellular network plus WiFi network as a main use case. There are also those who consider the inclusion of new network topologies and connectivity as part of the heterogeneous networks vision, such as personal hotspot, relay, peer-to-peer, device-to-device, near field communication (NFC) and traffic aggregation for machine type devices. In fact, as flexible sharing and dynamic access of spectrum become part of the network infrastructure, we can expect heterogeneous networks to also include cognitive radios.

Despite these diverse definitions and understandings, heterogeneous network research and deployment have made significant progress in the past several years, in particular in the area of data offloading through small cells (including WiFi access points). In practice, heterogeneous deployments are defined as mixed deployments consisting of macro, pico, femto and relay nodes. To the authors, heterogeneous networking is about a set of essential technologies and capabilities that deliver unprecedented large system capacity through the integration of heterogeneous architectures from WAN to LAN to PAN, provide always-on and always-best-connected connectivity for compute continuum, and offer innovative services and significantly better user experiences through the introduction of improved network efficiency.

In general, a heterogeneous network consists of multiple tiers (or layers) of networks of different cell sizes/footprints and/or of multiple radio access technologies [6]. An LTE macro base station network overlaying an LTE pico base station network is a good example of a multi-tier heterogeneous network. In this case since the same LTE air interface technology is used across different layers/tiers of networks, 3GPP (the standards body that created LTE) has developed solutions and design provisions to facilitate the interaction and the integration of such a heterogeneous network, including an extensive performance evaluation methodology that models a variety of deployment scenarios. On the other hand, a heterogeneous network consisting of a macrocellular LTE network and a WiFi network is a good example of multi-tier and multi-air-interface networks. Since the air interfaces were developed by different standards bodies (in this case 3GPP for LTE and IEEE802.11/Wi-Fi Alliance for WiFi), collaboration between standards development bodies is necessary to make the heterogeneous network work. The Hotspot 2.0 specification developed by Wi-Fi Alliance is one such example.

In addition to small cells, future heterogeneous networks may also include super big base stations. Cloud-RAN is one example [7]. Through high-speed optical fibre connections, a cloud-RAN base station relocates all or most of the baseband signals from tens to hundreds of traditional stations to a centralized server platform for massive signal processing. This architecture may significantly reduce the network's energy consumption since air conditioning at each cell site may no longer be required. Furthermore, due to its large size, a super base station can dynamically allocate its signal processing resource to adapt to the varying traffic loading within its geographical coverage during a day, a phenomenon often referred to as the 'tidal effect'. This further reduces hardware requirements and energy consumption. These super base stations may facilitate tighter couplings between different types of base stations including macro and small cells. They may also serve as a platform for cost-effective implementation of advanced air interface and network features to significantly increase network efficiency, which we will discuss in more details in future sections.

Although the sizes of base stations in each tier or layer of a hybrid network may differ significantly, ranging from femto station to picocell, microcell, macrocell and cloud-RAN, as shown in Figure 1.2, and although the radio access technology used in each tier may be the same or different, there is a common set of challenges and techniques to integrate them together to form a high performance heterogeneous network. This chapter will discuss a number of fundamental issues and solutions. It should be noted that due to the broad technical scope and highly complex economic tradeoffs, the designs of a heterogeneous network may have different flavours and focuses, depending on the existing installed network equipment, the choices of network transport, the availability of the required multi-mode devices, and the main set of applications expected to be supported by a particular heterogeneous network. As the requirements from users and mobile network operators continue to evolve, the definitions and the technical focuses of heterogeneous networks are expected to also evolve with time.

1.3 Economics of Heterogeneous Networks

There are various aspects to the economics of heterogeneous networks: the total cost of ownership to mobile network operators, the performance and cost benefits to end users and the

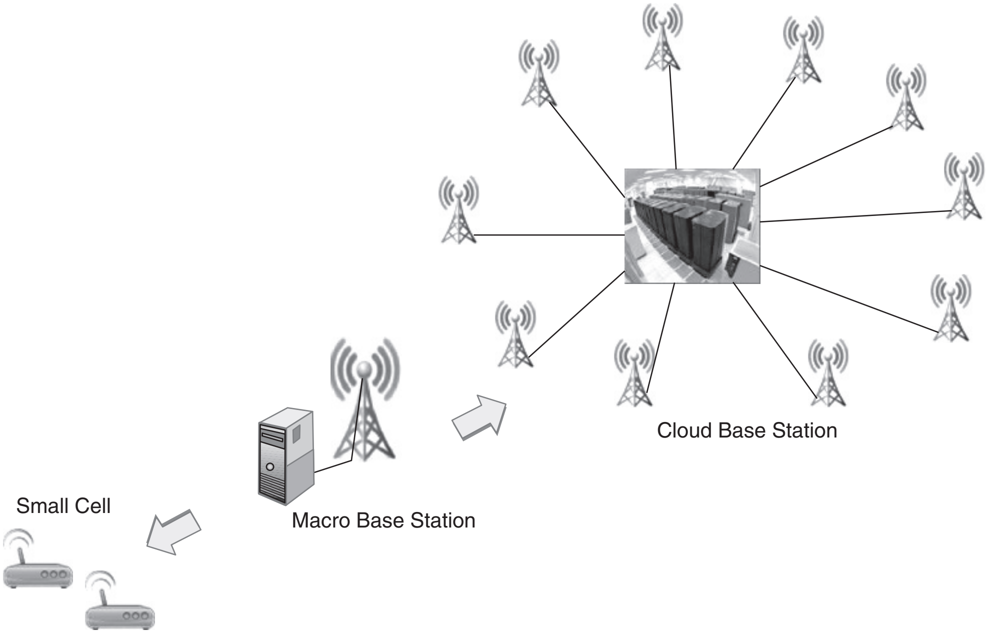


Figure 1.2 Base stations are becoming both bigger and smaller.

increase of the size of both total and addressable markets for telecommunication equipment manufacturers.

1.3.1 Total Cost of Ownership

The most important elements in the total cost of ownership (TCO) to a mobile wireless network operator are the capital expenditure (CAPEX) for network construction and the operating expenditure (OPEX) for network operation.

The cost structure for a traditional wireless network is relatively well understood [8]. For a typical cellular network, the CAPEX usually includes the cost of the radio access network (the base stations and the radio network controllers), the mobile core network (the gateways and the IMS platforms), the backhaul infrastructure and site acquisition, construction, engineering and integration. The radio access network represents about 60% of CAPEX, followed by the core network at about 15%, backhaul at about 5% and site acquisition, construction and engineering at 10–20%. There are of course large variations from deployment to deployment; for example, the radio access network cost can be reduced when an operator can upgrade its existing base station equipment to support a new air interface technology, or the site acquisition cost can be avoided if an operator can overlay a new network on its existing cell sites. The OPEX is mainly associated with network operation and management, including the expense of site rental, backhaul transmission, operation and maintenance of the network, and electric power. Given a 7-year depreciation period of base station equipment, OPEX may account for up to 60% of the TCO [7]. The cost of acquiring wireless licences is normally excluded from CAPEX or OPEX. This is because such acquisitions are infrequent and sometimes extremely expensive, ranging from several billions to tens of billions of dollars. In certain markets, the spectrum cost can be a significant portion of the TCO.

The cost structure for a heterogeneous network is often very different. Although the cost of each small cell is often an order of magnitude less than that of a macrocell base station, there are usually a large number of them in a heterogeneous deployment. As the number of cell sites increases, the backhaul cost may also significantly increase. Since many small cells are installed indoors on the wall or outdoors on utility poles, the cost structure for site acquisition, installation and site leasing may be very different. The operation and maintenance cost for a large number of small cells also poses new challenges to the operators of heterogeneous networks. Self-organizing and self-optimizing techniques become essential to reduce the overall network cost. The situation may be further complicated with consumer-installed femtocells. In this case, although there may be little or no cost to a mobile operator for site acquisition, equipment installation and backhaul connectivity, they may incur a significantly higher cost in customer technical support. Finally, a heterogeneous network may consist of network layers that operate on unlicensed bands such as WiFi. Although unlicensed bands do not incur a licensing cost as a traditional cellular network, the uncontrolled radio environment may be challenging to manage and operate, particularly in very dense hotspot deployments.

There have been a number of studies on the impact of the changes of cost structure on heterogeneous networks. A good example is given in [9, 10], which proposed a methodology for analyzing the total cost and performance of a heterogeneous network composed of multiple base station classes and radio access technologies with different cost and technical characteristics. It also demonstrated the impacts of the shift in cost structure and business model. Another good example is given in [7], which highlighted the fact that electricity cost at macrocell sites is about 41% of the OPEX per year, of which 48% is consumed by the air conditioning. As the industry continues the push for higher energy efficiency, there is a strong desire to significantly reduce the energy consumption through advanced base station design and network architectures, including fan-less cooling small cells and centralized baseband processing of large-scale cloud base stations.

It should be noted that the evolution towards heterogeneous networks presents both opportunities and challenges in terms of TCO for the mobile wireless industry. This is partly due to the significant shift of cost structure on base stations, backhaul, network installation and maintenance, energy cost and radio spectrum cost, and partly due to the significant shift of usage model from outdoor to indoor/hotspot, from high system capacity to both high system capacity and high data rate.

1.3.2 Heterogeneous Networks Use Scenarios

Heterogeneous networks have many architectural flavours and implementation variations to meet different market requirements and cost considerations [11]. However, the goals are similar. For consumers, heterogeneous networks need to provide ubiquitous coverage, secure, high data rate, high capacity, always-on, and always-connected-to-the-best-network user experience. For mobile operators, heterogeneous networks need to provide fast time-to-market, optimal network utilization, and operator control and network manageability.

The most classical heterogeneous network deployment is home femtocells. They use mobile operator licensed spectrum and are primarily deployed by end users for network coverage extension in an indoor environment or remote rural areas where outdoor macrocells have difficulty in providing coverage. As a cellular coverage extension, they are primarily used for voice services. These home femtocells are typically connected to the mobile core network through the consumer's own internet service such as DSL, and therefore there is usually no backhaul transmission cost to the mobile operator. A home femtocell often limits its access to a 'closed subscriber group', which makes economic sense to an end user since he/she pays for the backhaul transmission cost, and there is little reason to share it with others. One unique characteristic of home femtocells is that they are mostly installed by the consumers who do not necessarily have adequate knowledge about radio technologies. As a result, the RF interference from/to a home femtocell may be difficult to manage. A popular solution is to assign home femtocells to a different carrier frequency than the macrocellular network, provided that the mobile operator has sufficient spectrum for a standalone RF carrier. Since home femtocells are considered consumer products, the average selling price is in the order of hundreds of dollars.

Another class of small cells is the so-called picocell, sometimes also referred to as an enterprise femtocell or metro femtocell. They use the same air interface technology over mobile operator's licensed spectrum, and serve as an extension of the macrocellular network. This class of small cells usually has a larger subscriber capacity compared to home femtocells and provides voice and data services in office environments, indoor coverage in places like shopping centres or outdoor hotspot coverage such as a busy shopping street or sports stadium. They are often environmentally hardened in particular for outdoor deployment, professionally installed with more advanced antennas, with service open to all qualified subscribers instead of only to the members of a closed subscriber group. It should be noted that a home femtocell and a picocell may not be very different in terms of subscriber capacity and transmission power. Table 1.1 shows a typical example. The main differences are actually related to how they are connected to mobile operator's core network, an important issue that we will discuss in a later section. The average cost of a picocell is in the order of one to several thousand dollars since it often needs carrier-grade equipment.

In the past few years, data traffic offloaded from cellular networks to WiFi has gained significant momentum. This type of heterogeneous network operates in both licensed (for

Table 1.1 A comparison of home femto and public picocell key features

	Femtocell	Picocell
TX power	low power, <250 mW	higher power, 250 mW to 2 W
capacity	low capacity, <8 users	higher capacity, 16–32 users
backhaul	consumer grade, paid for by user	carrier grade, paid for by operator
equipment	owned by consumer	owned by operator
cell 'site'	consumer installation	operator professional installation
deployment	unplanned, consumer deployment	planned, deployment by operator
user access	closed, restricted access	open to all qualified subscribers
handover	loose coupling at network layer	tight coupling, intra-network handover
security	not trusted by operator network	trusted by operator network
enclosure	consumer electronics	often environmentally hardened

2G/3G/4G cellular) and unlicensed band (for WiFi). Since cellular networks and WiFi are very different in user/device credentials, authentication, air interface characteristics, network architectures, interference environments and subscriber management and billing systems, the integration of WiFi and cellular network requires careful economic consideration and therefore varies from market to market. Some mobile operators choose to deploy their own WiFi networks, while others offer services through third-party WiFi service providers either under mobile operator's own brand or under a third-party brand. The size of the WiFi networks also varies, from several thousand access points (APs) for selected hotspot coverage at airports to more than 2 million APs for major city coverage across a country. Due to such wide variations in network equipment ownership, in backhaul transmission provider and in cell site real estate arrangement, the cost structure tends to be complex. The business model is further complicated by the exact data offloading strategy, since some operators choose to offload the radio link only and bring the traffic over WiFi to mobile operator's core network, while others choose to offload both the radio access and the core networks, with the WiFi traffic directly going into the public internet.

1.3.3 General Tends in Heterogeneous Networks Development

There are several general trends in heterogeneous network development, including the integration of WiFi and cellular air interfaces in the same small-cell platform, the techniques for dense deployment of small cells in extremely heavy loading conditions such as a sports stadium, the convergence of Hotspot 2.0 and Access Network Discovery and Selection Function (ANDSF) for network interoperability, the evolution from providing coverage or capacity to offering value-added services by integrating location and proximity services into the integrated small cells, and the possibility of developing super control nodes that coordinate the radio resources across different layers of networks. The debate on femto station versus WiFi access point is practically settled. It is now widely recognized that they are complementary.

Ultimately, a practical heterogeneous network must deliver three things to realize its economic potential: technically, it needs to fill the capacity gap created by the data tsunami; business-wise it must help mobile operators with in-service differentiation and new revenue opportunities, while minimizing CAPEX and OPEX; and to end users, it must offer superior user experience through always-on and always-best-connected. To achieve these goals, a heterogeneous network solution needs to leverage existing technologies and deployment, to enrich and to expand existing applications, and to enable future service innovations [12].

1.4 Aspects of Heterogeneous Network Technology

In this section we discuss different technical aspects of heterogeneous network technology, the challenges, the tradeoffs and the future technology directions.

1.4.1 RF Interference

When deploying 3G femtocells within the coverage area of a 3G macrocell network, these two layers of network may share the same carrier frequency, or they can be deployed on two separate carrier frequencies.

Sharing the same frequency has the advantage of minimum spectrum usage, which is particularly important to operators with a limited spectrum supply. However, the interference could be severe between these two networks if they are not properly engineered. For example, a UE connected to a macro base station may produce excessive amount of interference to nearby femtocells, in particular when the UE is at the cell edge transmitting at close to its maximum power level. One practical solution to address this problem is to desensitize the femto station receiver to avoid RF saturation. In general, it is always desirable to place the femto stations in locations that provide certain natural RF signal isolation with the macro network, and to carefully engineer the handover triggers for reliable UE mobility between these two networks.

When macro stations and femto stations are deployed on separate carrier frequencies, there are still challenges for the UEs to associate with the most appropriate network. Traditional cellular networks were designed for homogeneous network deployment, where the pilot strength measured at a UE is a good indication of its distance from the base station. However, this is generally no longer true for a heterogeneous deployment, where the pilot signal from the macro station can overwhelm the much weaker pilot signal from a femto station, even when the UE is already well within the coverage area of the femto station. A much better measurement in this case should be the actual path loss between the UE and the base station instead of the pilot strength. However, since such change would require some fundamental modifications to the existing standards and the deployed equipment, the industry adopted the ‘range extension’ technique which essentially adds a bias value to compensate for the weaker femto pilot, and therefore ‘extend’ the coverage area of a femto station [13].

In actual consumer femtocell deployment, the closed-subscriber-group (CSG) poses yet another challenge to RF interference management [14]. Since consumers usually use their own internet subscription for femtocell backhaul connection, the access to the femto station is restricted to a predefined group of users, usually the family members or house guests. A UE will not be able to access a consumer-deployed femto station even in close proximity unless it is part of the CSG. This can produce an excessive amount of interference to the networks, in particular in a dense femtocell deployment environment such as within an apartment building. Fortunately this is less of a concern for enterprise environments, since all users usually have access to all femto stations in the office space. There have been discussions on open subscriber group for femtocells, but it is more of a business model issue than a technical solution.

The situation with WiFi networks is very different. On the one hand, there is no ‘macro’ WiFi station to worry about, but on the other hand, the interference among WiFi access points and devices is more unpredictable due to the unlicensed-band nature of the network. As capacity demand continues to grow, more and more hot spot areas are served by multiple WiFi networks, which can cause excessive interference to each other. The wireless industry is looking at network sharing (e.g., several service providers sharing one access point by broadcasting several SSIDs over the air), local WiFi channel coordination and new deployment in the higher unlicensed spectrum including the 5 GHz band.

Future heterogeneous networks are also expected to support new types of connectivity and network architectures. These include layer 2 and layer 3 relay (including in-band backhaul application similar to a mess network), mobile relay (e.g., a relay station installed in a high speed train to provide relay service between the passengers in the train and the base station network along the trackside [15]), D2D (device-to-device direct communication [16]), traffic aggregation point for machine-type communications (a small radio station that provides short-range connectivity such as ZigBee to local sensors and relays the aggregated traffic to the network through a long-range cellular connection) and cognitive radio for spectrum sharing (e.g., in TV white space bands). Although the RF interference issues and challenges are very different in each case, there is a common set of requirements that a heterogeneous network RF engineer needs to consider, such as frequency planning, co-existence, transmission power level and receiver sensitivity within a heterogeneous network and within a neighbouring homogeneous or heterogeneous network operated by a different mobile network operator. At this moment, most heterogeneous network studies are focused on developing solutions for a specific combination of network layers. There is a need for longer-term vision and framework to scale up the heterogeneous network design and deployment through an optimal distribution of functions among devices, access points and cloud.

1.4.2 Radio System Configuration

Traditional cellular networks follow strict managed-network design principles. The cell sites are individually selected and base stations are manually engineered with sophisticated RF planning tools that model the actual deployment environment. Mobile devices provide real-time channel quality, interference level and resource request as inputs, but the network makes final decisions on radio resource management. This network-centric design has been effective in achieving optimal balance between coverage and capacity in homogeneous networks. The homogenous nature also simplifies mobile reporting, since the pilot to interference ratio can be directly used to estimate the electronic distance between the base station and a mobile device.

This traditional approach faces many challenges in a heterogeneous network environment. The first is site selection. Small cells are deployed in more versatile environments, ranging from residential, office, shopping centres and sports stadiums, both indoors and outdoors. The number of sites and environments are too costly to apply traditional deployment methodology. The second is carrier frequency coordination. Cellular operators have control over small-cell deployment due to its ownership of the licensed band. For unlicensed systems such as WiFi, the frequency planning needs to not only coordinate the channel allocation among access points, but also to adapt to the local interference environment. This is further complicated by consumer self-installation or professional installation without any network optimization.

The self-organizing network (SON) technology is therefore widely used in small cells [17]. SON requires interference sensing capability at the small-cell stations. For 3G/4G small cells operating in the FDD bands, this means the addition of an interference sensor to monitor its transmit frequency. For WiFi access points, this capability is inherited from the time division duplex(TDD) nature of its air interface. Based on the sensed interference information, the network can self-configure its channel allocation, scheduling algorithm and even its transmit pattern. The latter is particularly useful in places such as an apartment building where two femto stations are unintentionally installed next to each other across a building wall [18]. SON can be implemented in a distributed or a more centralized manner. Centralized SON coordination usually yields better performance, in particular in the dense deployment environments.

SON technology will further evolve in the future with the proliferation of cloud-RAN and small cells. SON technology needs to extend its capability to optimize across different layers of networks, and also across different base stations in a geographic location. We expect mobile devices to take a bigger role, for two reasons. First, multi-mode devices have the visibility of the local RF environment across layers of networks. Secondly, mobile devices may add features to help to reduce SON complexity. We also expect SON to heavily leverage cloud-RAN technology for optimum performance in a heterogeneous network environment. One possible solution is to have the cloud provide policy and general information on the RF environment, based on which each network layer can configure its radio system, and move more decision-making to devices. Recent development in connection manager software at certain mobile devices can be considered an early experiment of this technology concept.

1.4.3 Network Coupling

Heterogeneous networking is about multiple layers of networks interconnecting and working together. The relationship between networks of different layers can be loosely described as a network coupling issue. Although there are no precise definitions, loose network coupling usually refers to two networks generally maintaining independence of each other when forming a heterogeneous network, while tight coupling usually means a greater level of integration between two networks.

To illustrate the concept of network coupling, let's use 3GPP-WiFi interworking as a concrete example. According to the 3GPP specification [19], an 'un-trusted' WiFi network can connect to a 3G cellular network through an 'enhanced packet data gateway' (ePDG), which provides backhaul security through IPSec tunnels, real-time packet processing, and subscribers, service and applications monitoring. We consider these two networks loosely coupled when the WiFi access network and 3GPP core network share only the AAA server for authentication. In this loose-coupling case, each network operates independently, and neither of the networks needs to change its architecture or protocol stack. However, since there is no support for service continuity during handovers, a user may experience longer handover latency and a certain amount of packet loss. In a tight-coupling case, a WiFi data flows through the 3GPP core network from ePDG to the 3GPP packet data network gateway (PDN-GW), in the same way as cellular data does. In addition to 3GPP authentication as in the loose-coupling case, WiFi users can also access 3GPP services with the potential support of guaranteed QoS and seamless mobility. Other 3GPP services may be enabled for WiFi users as well through tight coupling.

In a broader sense, network coupling is about the level of integration between two networks. Femto station is another good example of tight coupling. In addition to the tight coupling at the core network level as we just described, a heterogeneous network may also create a tight coupling at lower layers between two radio access networks, and allow one radio access network to have certain visibility and radio resource management control over the other network. A relevant example is the CDMA-LTE handover design through tight coupling at the radio link layer, which potentially offers the same level of fast handover performance as a homogeneous network [20]. In LTE-Advanced and beyond, this type of 'tighter coupling' can be further extended to air interface PHY and MAC layers to allow the implementation of joint radio resource management and traffic scheduling. There are already suggestions to add an X2-like interface between a macro base station and a femto station, and add the aggregation of carriers of different radio access technologies, which we will discuss later.

Network coupling is a complex technical issue, but in real life deployment, it is also very often a critical business decision driven by the tradeoffs of complexity, cost, security and performance. A wireless network operator may choose loose coupling over tight coupling despite its potential higher performance. For example, if a cellular operator uses a third-party

Table 1.2 Types of network coupling

Type of network coupling	Example design characteristics	Benefits and complexity
no coupling	• two networks operate independently • mobile device connection manager coordinates wireless connectivity.	• no change to existing networks, suitable for roaming • operator may perform offline billings consolidation for single bill
loose coupling	• two networks share user credential and AAA • data traffic goes through separate core networks	• common user/device credential for two networks; potentially eliminate user intervention for WiFi access • core network traffic offload to internet
tight coupling	• user traffic goes through mobile operator core network, (e.g., home femtocells connected through a femto gateway)	• opportunity for operator to offer value-added services • opportunity to offer service continuity or even seamless handover
very tight coupling	• two access networks are directly interconnected • real-time radio resource management across networks, (e.g., picocell connected through X2 interface)	• opportunity to implement carrier aggregation and interference coordination • requires significant design and implement efforts

WiFi network to form a heterogeneous network, loose coupling may offer the simplicity to both networks for both business relationship and network management. This may also be the case for operators who own both networks but do not want to make significant changes to the already installed equipment. The network coupling decision is also significantly driven by the expected use scenarios of an operator. Certain operators may choose to use the networks for different purposes, for example, one for voice calls and smartphone data access, and the other for data offloading to the internet to avoid core network congestion. In this case, the benefits of high performance service continuity of a tight couple may be secondary. Table 1.2 provides a summary of types of network coupling, the general design characteristics and the associated benefits and complexity.

From a standards development viewpoint, the industry tends towards tightly integrated solutions with secure data offloading and service continuity to offer operators a complete range of technology choices. These solutions will continue to evolve for different markets including home, enterprise and wireless carriers.

1.4.4 User and Device Credential

A key technology element in a heterogeneous network is the user or device credential used for authentication, authorization and accounting. The most common user/device credentials include SIM card in mobile phones, embedded SIM and soft SIM in certain machine type communication devices, certificate in laptop computers, and username/password for WiFi access. These fragmented user/device credential solutions were mainly due to the legacy of the telecom and the information technology industries, and they are now increasingly becoming a challenge to the implementation and deployment of heterogeneous networks.

Taking cellular and WiFi interworking as an example, while most mobile smartphones use SIM cards, most WiFi-only devices currently use username/password pair or certificate for access authentication. A unified user/device credential solution is highly desirable for seamless network access, improved user privacy and information protection, uniform service availability, consistent quality of experience and simplified device management. Unfortunately, it may take time to implement the required changes to the portable device platform design, and more importantly to operators; accounting and billing system. In tightly coupling heterogeneous networks, we see stronger motivations to unified user/device credential to facilitate fast network transition and seamless service continuity across networks. In loosely coupling cases, we anticipate slower convergence of solutions, or no convergence at all.

This situation is further complicated by the proliferation of machine-type communication devices. Compared to consumer electronics such as smartphones, some machine-type communication devices require a much longer life cycle (e.g., seven or more years) and need to be environmentally hardened and able to survive deployment with minimum physical protection for the devices. Therefore new forms of device credentials may be required.

Nevertheless, the wireless industry has taken steps to address the issues, and there are several initiatives currently at different stages of development [21]. The convergence of telecom and information technology may also help to accelerate the process.

1.4.5 Interworking

Interworking provides a degree of seamlessness to multi-mode devices capable of transiting between different networks. There are two forms of transiting. The first is between networks that are geographically separated, commonly known as roaming. The second is between networks that are co-located, such as handover between different layers of networks within a heterogeneous network. For heterogeneous networks and their associated multi-mode devices, roaming and handover support is one of the fundamental capabilities that enable end users to access basic services seamlessly.

Interworking between networks is often achieved through an interworking function node or the access gateway [19]. In a heterogeneous network, since different layers of the network may have different network architecture and protocol stack design, such a node provides protocol conversion or translation on the control plane and intermediate protocol termination on the data plane. This node may also serve as the security demarcation point between two networks, in particular when the ownership of two networks is different. The home NodeB gateway defined in 3GPP runs an IPsec tunnel to the home NodeB (commonly known as a femto station) over the public internet to provide information security and an interworking function for handover. Since some of the networks such as WiFi do not have core network architecture and the associated protocol stack, an interworking node (in this case the ePDG) provides WiFi devices the access to the AAA server in the 3GPP core network for common authentication and charging. Future development will also add unified QoS capability across different layers of the network to provide more seamless experience to end users.

The main challenge in interworking is that the design and implementation complexity goes up exponentially when we scale up the number of layers in a heterogeneous network. Since each access technology usually comes with its own control plane design, making all layers talk to each other requires extensive engineering efforts, in particular for tight coupling. We expect the challenge to further increase in the future due to the introduction of new types of connectivity (e.g., ZigBee for machine-type devices) and new network topologies (e.g., mobile relay, device-to-device communications).

The wireless industry is looking at two technology directions to address the interworking complexity issue. One is to put more intelligence in the mobile devices, therefore offloading some of the interworking complexities from the network infrastructure to the devices. One example is to design and deploy sophisticated connection manager software in the mobile devices. The other one is to improve data plane processing, moving some of the complexities from the control plane to the data plane. Since the device has detailed knowledge of the applications it is running, enhanced data plane processing is in particularly attractive for applications that require consistent QoS support throughout layers of a heterogeneous network.

1.4.6 Handover

Once networks can interwork with each other, the next technical requirement is the ability to handover a mobile device among base stations and/or networks. As in any homogenous network, the reliability and the performance of handover directly affect user experience in a heterogeneous network and are therefore the key performance metrics.

Before we discuss the details of handover techniques in heterogeneous networks, we first look at the differences. In a homogeneous network, handover is primarily triggered by a mobile device moving out of the coverage area of a serving base station or network. Occasionally a wireless operator may also choose to direct a device to handover to a specific base station for network maintenance or load balancing purposes. In a heterogeneous network, a high-power macro base station network usually provides ubiquitous network coverage and high-speed mobility support, while the low-power small cells offer data offload in hotspot areas at a lower cost.

For intra-system handovers within the same layer of a heterogeneous network, the handover techniques and procedures are similar to that of a homogeneous network. For inter-system handovers across different layers, however, the handover decision is much more involved. In addition to the traditional coverage-triggered handovers, many decisions in a heterogeneous network are also based on the service delivery cost to an operator, the data pricing to users, the QoS requirements of specific applications, and the non-ubiquitous nature of the hotspot network layer [22]. For example, for data offloading purposes it is highly desirable to handover a mobile device from the macro base station network to small cells as soon as it moves within the hotspot or home network coverage area. For real-time and conversational applications such as voice calls, it may be desirable to avoid inter-layer handover due to service continuity and quality of service requirements. The degree of network coupling also has a significant impact on the handover algorithm design. Due to its smaller interruption to services, a heterogeneous network with tight coupling between layers can afford more aggressive inter-layer handover decisions than a similar network with loose coupling. The handover algorithm design in a heterogeneous network environment is therefore significantly more complex and challenging than that of a homogeneous network.

Similar to the traditional designs widely used in homogeneous networks, a handover process in heterogeneous networks also involves five stages: (1) handover preparation, (2) trigger for handover, (3) handover decision, (4) handover execution and session transfer and (5) handover completion. Handover preparation includes the mobile device looking for possible handover targets according to certain pre-optimized mobile-network association rules [23]. To facilitate this searching process and potentially save the battery consumption of the mobile device, the serving network may offer certain information about the candidate networks, such as the type of air interface technology and the basic radio configuration information including carrier frequency. The serving network may also provide certain handover-related parameters to the mobile device to optimize performance. In addition, a mobile device may pre-register itself to the target networks to save time in authentication and service authorization when handover actually happens. However, all these come at a cost to a heterogeneous network, including different layers of the network having to have sufficient information about each other, and having to maintain consistency of registration from a mobile device. This complexity goes up rapidly as the number of layers increases or when the deployment environment is complex (e.g., a mix of indoor and outdoor networks, or different network topologies). Fortunately, the industry and university research teams are working on advanced network self-configuration techniques. The efforts in unifying user and device credentials may also reduce the complexity in managing device pre-registration.

The next step is to determine the need for handover when certain preset conditions are met, commonly known as handover triggers. Handover triggers are mostly generated at the mobile device, based on a set of rules and parameters that are pre-installed by the operator or downloaded from the network. A network may also trigger handovers for network maintenance or performance optimization purposes. It is probably one of the most important technologies in a wireless network. In a homogeneous network, wrong handover triggers result in radio link failure (e.g., call drops) and/or inefficient usage of radio resources. In a heterogeneous network, since small-cell coverage may not be ubiquitous, knowing where and when to switch from one network to another affects both network performance and user experience. The design is further complicated since in addition to radio link conditions the handover trigger algorithm for a heterogeneous network also needs to consider operator policy, subscriber usage cost and application quality of service requirements. These requirements may vary in time and/or may be location-dependent.

After handover is triggered, an entity in the system needs to make a handover decision. In traditional cellular system designs, the infrastructure network is responsible for making the final handover decision. As soon as the handover conditions are met, the mobile device is responsible for sending a handover request to the network, but the network can decide either to issue a handover command or to take no action. This network-centric handover decision process has shortcomings in heterogeneous network environments. For example, the serving layer of the heterogeneous network may not have the full information or capability to make a handover decision to other layers. This is particularly true when networks are loosely coupled. We expect mobile devices to take a more active role in handover decision-making in future evolved heterogeneous networks, and as a result, some of the handover message flows will change accordingly.

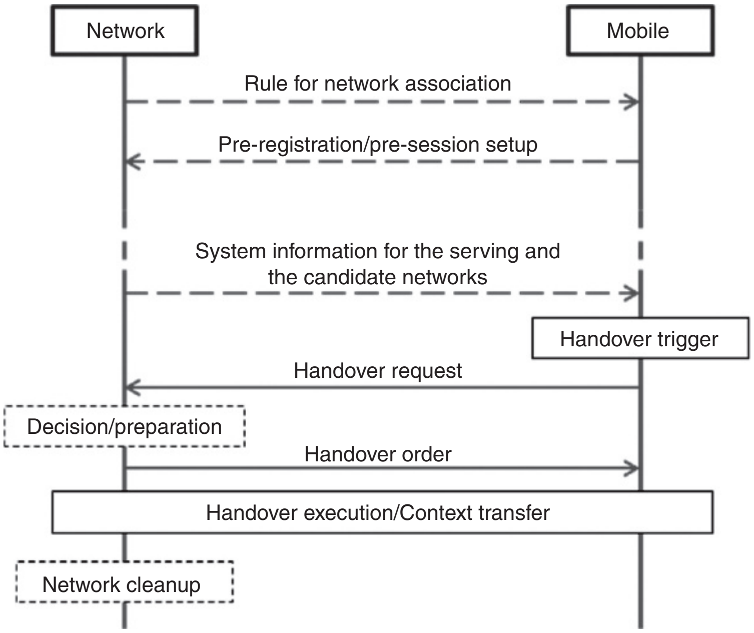

Network
Mobile
Rule for network association
Pre-registration/pre-session setup
System information for the serving and the candidate networks
Handover trigger
Handover request
Decision/preparation
Handover order
Handover execution/Context transfer
Network cleanup

Figure 1.3 An example of a mobile-triggered network-decided generic handover process.

Heterogeneous networks can apply the same techniques used in homogeneous networks for fast handover execution and minimum service interruption. In a similar way to traditional handovers, the source network may prepare the target network(s) to pre-fetch the required context information before actual handover occurs. During handover execution, depending on the degree of network coupling, session information including the remaining user data in the transmit buffer may be transferred from the source to the target network to minimize data loss and to achieve session continuity [24].

Upon completion of the handover process, the heterogeneous network needs to perform certain clean-up to maintain its overall resource utilization efficiency and software robustness. An example of a mobile-triggered network-decided generic handover process is shown in Figure 1.3.

1.4.7 Data Routing

While handover is mostly a link layer issue, data packet routing is mostly a network layer function. There are two main aspects to data packet routing in a heterogeneous network: (1) the radio link selection and/or aggregation over the air and (2) IP packet routine in network. The actual design of data routing is based both on business and technical considerations.

Unlike a homogeneous network where a mobile device usually connects to the network through a single radio link, a heterogeneous network can allow multiple air interface technologies to operate concurrently, therefore potentially creating multiple radio links between a mobile device and the heterogeneous network, for example, a smartphone with both 3G and

WiFi connections turned on at the same time. One simple way of using this capability is to aggregate them, in order to achieve a higher data rate. The other way is to intelligently map user application flows to different radio links, according to their respective QoS requirements, operator policy and user preference. For example, the mobile device can use the cellular connection for voice calls and the WiFi link for high-speed best-effort internet access. This capability is supported in 3GPP IFOM (IP flow mobility) feature [25].

In addition to radio link selection and aggregation over the air, a heterogeneous network may also support different routing options within the infrastructure network to different layers. This flexibility becomes increasingly important as the industry is starting to experience more and more core network overloading issues as the air interface capacity rapidly improves due to the deployment of small cells and also due to more spectrally efficient air interfaces such as LTE and LTE-Advanced. 3GPP is working on a number of solutions including LIPA (Local IP Access, commonly known as local breakout for 3GPP femto stations) and SIPTO (Selective IP Traffic Offloading) to offload an operator's core network or to selectively route traffic to the public internet or to the core network according to operator business considerations [26].

The combination of multi-radio-link capability and the network routing flexibility is an area of intensive research and technology development in industry. It is one of the unique tools in a heterogeneous network to achieve an optimum balance between user experience and service revenue, and to implement an operator's data offloading strategy.

1.4.8 Quality of Service

Providing consistent quality of service to end users is extremely important to any wireless networks. Since it also offers opportunities for product differentiation and incremental revenue for network operators, it is one of the most mission-critical technology elements in future wireless networks.

Due to its end-to-end nature, delivering quality of service is complex and challenging, even in a traditional homogeneous network. First of all, the network needs to have the capability of establishing and managing service policy, and have the associated accounting and charging infrastructure. When a subscriber requests a service, the network authenticates the user and authorizes the service according to its subscription record. The network then configures each network node along the data path of this specific application. Once network nodes start performing the desired QoS treatment for this particular application IP flow, the network needs to have the capability of monitoring the actual service quality delivered to the end user. When the application session is ended, the network needs to collect all relevant accounting information and to perform system clean up.

This rather complex process faces several challenges in real life. First of all, quality of service is meaningless to a user unless it is delivered end to end, but a wireless network only has the span of control over the radio link and within the operator's own core network. This has not been a major issue, since traditionally the radio link was almost always the bottleneck for service quality, but recently the backbone networks have also started to experience congestion due to the proliferation of video and other media-rich contents. The second challenge is the lack of consistency in QoS implementation. Different segments of the network may have different interpretations of QoS marking, and may use different parameters in traffic scheduling algorithms. This is, in particular, an issue when traffic travels through networks of different operators. The third challenge is that QoS scheduling may not always produce as significant a result as users expect. This is partly due to the significant increase in air interface data rate and backhaul bandwidth in the past decade, and partly due to the fact that many popular contents and applications are the so-called ‘over-the-top’ traffic where operators have little real control of service quality (over-the-top traffic refers to contents or applications that are delivered directly from the provider to the customer using an open internet connection, independently of wireless operator network). Finally, end users see little reason to pay extra for QoS unless the service becomes practically unusable without it. As a result, current focus in the wireless industry is limited to VoLTE (VoIP over LTE), where operators own both the network and the service, and QoS support is essential for VoLTE service reliability to approach that of circuit switched voice [27].

QoS support in heterogeneous networks incurs additional complexity. There are multiple layers of networks involved, each of which may employ different air interface technology and network architecture, with different QoS mechanisms. Even for a heterogeneous network consisting of the same air interface technology – for example 3G macro stations with 3G femto stations – the backhaul transport over public internet for femto stations may affect the QoS management and delivery. Network coupling brings further variations, and some deployments may face limitations in delivering consistent QoS across loosely coupled networks. For heterogeneous networks with data routing capability among layers, the routing change itself may introduce additional latency in the data path. Finally, the RF interference condition in unlicensed band is often difficult to predict and changes with time, particularly in dense hotspots where QoS happens to be most needed.

As discussed earlier, due to the proliferation of mobile internet traffic, wireless operators have to invest heavily to meet the capacity demand while their revenues from traditional voice and SMS services are decreasing. As a result, QoS is increasingly essential to maintain revenue stream and to differentiate in the highly competitive market place. We expect research and development of QoS for heterogeneous networks to intensify in the next few years, and the results will become increasingly critical and rewarding with time.

Since a significant portion of the mobile internet traffic is over-the-top with origination outside of operator’s network, such as Skype and YouTube, a heterogeneous network today mostly serves as best-effort data pipe. There are two basic approaches to introduce QoS support in this case. The first one is to expand the traditional QoS model we discussed earlier in this sector into the third-party application service providers. The MOSAP development in 3GPP (interworking between Mobile Operators using the evolved packet System and data Application Providers) can be used as a framework [28]. We can call this a control-plane-centric approach since it requires the control entities of the mobile operator network and the data application provider system to exchange application requirements and policy information with each other so that the mobile operator can properly set up the data pipe for QoS support. However, since many over-the-top applications are already deployed and are working well in the field, many data application providers may not be willing to make additional investment in this capability and negotiate a business agreement with mobile network operators. In addition, the required interoperability testing may be challenging to scale up the solution. Finally, the question of how to handle QoS within a heterogeneous network remains the same.

Another very different approach is to introduce advanced signal processing on the data plane, thereby shifting some of the complexity from the control plane to the data plane. A classic example is the application of deep packet inspect (DPI) at both the base station network and the mobile device. By performing DPI, the data plane integrates an autonomous QoS handling capability, which can remain intact when a radio link is handed over between different layers of networks in a heterogeneous network. For the base station network, the data plane packet processing can be integrated with other advanced content distribution technologies including caching, cross-layer optimization and adaptive media reformatting to address the network overloading conditions. With the convergence of information technology and the telecom in the cloud network, data plane processing is becoming increasingly feasible and attractive. For mobile devices, shifting complexity from the control plane to the data plane has the added potential benefits of minimum additional signal processing. This is because both the applications and the complete protocol stack are implemented within the same physical device, so in practice there is little need to perform packet-by-packet inspection since such information can be easily and directly obtained from the API for each application. In a heterogeneous network environment the mobile devices also have the advantage of knowing its RF environment and the availability of each network. It is natural for the mobile devices to take a greater role in QoS management and decision making.

We expect the wireless industry and the research community to continue to investigate the traditional QoS mechanisms and to look for new technology options in the next few years in the context of IT and telecom convergence and media content distribution for heterogeneous networks. We expect cloud network architecture and multi-communication device to play a major role in QoS over heterogeneous networks. From an economic viewpoint, it is also one of the most important capabilities for mobile network operators since it can offer unprecedented user experience in terms of service quality and consistency to the end users in a highly complex and sometimes confusing heterogeneous network environment, and can also provide a major revenue opportunity for operators who are facing the challenge of becoming a featureless transparent-data-pipe utility provider.

1.4.9 Security and Privacy

Security and device interconnectivity engineered in a heterogeneous network and within the protocol stack of each layer of networks enable consistent and compelling experience for consumers and expand the network applications for machine-type connected devices in a future information society. As devices like smartphones, tablets and smart TVs become more popular, consumers are clamouring for access to applications and data everywhere, but are faced with security barriers of each individual system and therefore disparate user experiences. For machine-type communications such as smart grid and smart transportation, security is paramount in protecting the integrity of strategic infrastructure and for ensuring the normal functioning of an intelligent society and the underlying future embedded internet [2].

As mobile internet penetrates into every aspect of an individual’s daily life and becomes the fabric of our modern society, the concerns over security further increase. With the smartphone operating systems becoming more concentrated around four options (the combined market share of Android, Symbian, iOS and Blackberry accounts for more than 95% globally [29]), any major security breach would potentially affect a massive number of users around the world. In addition to the large-scale potential security impacts, the drive towards better user convenience also requires more security safeguards. Features such as single-sign-on (SSO) and unified billing completely rely on secure communications among different systems. Ultimately, convenience does not mean anything unless it is secure. The industry is taking steps to fortify security including embedding security functions directly in the silicon chips [30].

With the proliferation of mobile internet there is also an increased awareness of, or concerns over, privacy. In the heterogeneous network environment in particular, since layers of network need to exchange essential information to ensure the compute continuum, potentially among different operators that own and operate each layer of the network, protecting privacy is an even more critical issue. Since mobile network operators have a strong desire to offer context-aware and location-based services instead of simply providing a featureless data pipe, the collection – and more importantly the exchange – of user and application information sometimes become critical. Taking device-to-device (D2D) for social networking applications as an example, if everyone is concerned with privacy and security, therefore turns off the D2D function, this feature will never become useful. For location-based services, privacy has always been a concern, in particular in a heterogeneous network environment involving a mixed network ownership. We expect mobile network operators to play a bigger role in managing and maintaining security and privacy, and to heavily leverage their subscriber ownership and the cloud infrastructure.

1.4.10 Capacity and Performance Evaluation

The capacity and performance evaluation of heterogeneous networks poses a unique set of challenges and issues. On the one hand, the cellular industry has established an elaborate system capacity modelling and evaluation methodology for traditional homogeneous cellular networks. The focus has been on ubiquitous network coverage (e.g., 95%), overall cell throughput and spectral efficiency, and tail user throughput (e.g., 95% point on the CDF curve). On the other hand, the IT industry uses its own network modelling and evaluation methodology, with focus on enterprise and hotspot traffic density, and peak data rate. For heterogeneous networks, the cellular community has extended its traditional model to cover femto/small-cell deployment scenarios. This includes adding median user throughput, average macrocell area spectral efficiency, and the percentage of total throughput carried by low-power cells. In addition, latency-based metrics were also considered for the case of bursty-traffic evaluations. The cellular model has been further extended to include new types of wireless connectivity and use scenarios, such as closed subscriber group (CSG) and relay, including in-band self-backhauling use cases [31].

Nevertheless, significant research and development are still required for heterogeneous network capacity and performance evaluations. Key focus areas include more accurate modelling of hotspot and hot-area user distribution, including: the impact of network coupling and interworking performance on subscriber distribution among layers of networks; the inclusion of new types of connectivity in the models such as mobile relay and device-to-device communications; cross-network-layer interference coordination and radio resource management; a unified or hybrid cellular and WiFi evaluation methodology that enables more TCO modelling; and a more accurate modelling of unlicensed band radio environment.

1.5 Future Heterogeneous Network Applications

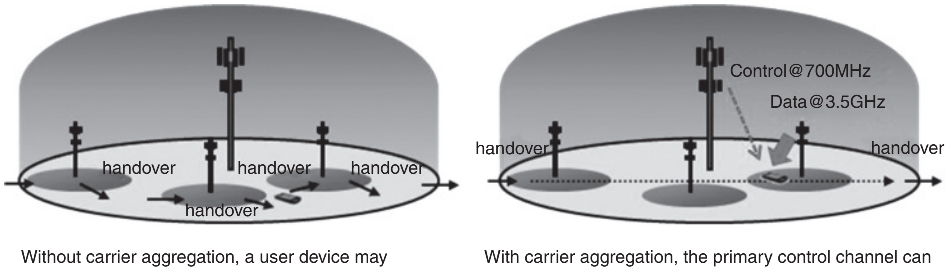
The initial deployments of heterogeneous networks have been for network capacity and traffic offloading [32]. Future technology development needs to support new capabilities for innovative applications that offer unprecedented new user experience and to deliver a true world of compute continuum.

We can expect that traffic offloading solutions continue to evolve and improve, to address congestion in both access networks and core network. This is achieved through the deployment of both IP flow mobility and local breakout solutions such as LIPA and SIPTO [26]. As backhaul transport becomes a major cost factor, we will see an increasing number of existing and new hybrid solutions using optical fiber, copper cable, microwave, in-band relay, multi-hope relay and mesh networks. Network security and user privacy need to continue to improve as user devices increasingly roam between layers of networks and between coverage areas of different mobile network operators. Finally, we expect the wireless industry to expend significant research and development effort in addressing very dense heterogeneous network design and deployment issues for public venues such as sports stadium and mass public transportation hubs.

Collaborative communication will become a major driving force for heterogeneous network innovations to support new types of connectivity and the associated new business models [33, 34], which include device-to-device communication, client relay and mobile relay, and sensor hub for machine type communication devices. Cognitive radio may also be added to this framework since it may offer an additional connectivity option for collaborative communication.

Future heterogeneous networks are also expected to implement and to take full advantage of the next generation air interface features. A good example is the implementation of carrier aggregation over a heterogeneous network. For a deployment scenario shown on the left side of Figure 1.4, a mobile device may need to handover frequently between macro and small networks without carrier aggregation, which disrupts services and add signalling traffic to the network. With carrier aggregation, it is possible to anchor the primary control channel at the macrocell, preferably operating at a lower frequency band for reliability and coverage, while operating the secondary data channels from small cells at a higher frequency band to deliver high data rate and network capacity in an opportunistic fashion. There is no handover required as long as the mobile device is under the coverage area of the macrocell. This is shown on the right side of Figure 1.4.

Multiple carriers also enable interference management between different power class cells as well as open access and closed subscriber group (CSG) cells. Long-term resource partitioning can be carried out by exclusively dedicating carriers to certain power class cells, while dynamic radio resource management and network load balancing can be applied by sharing those carriers



Without carrier aggregation, a user device may need to handover frequently between two layers of networks, disrupting servics and adding significant amount of signaling traffic

With carrier aggregation, the primary control channel can be anchored at the macro-cell, while the secondary data traffic channels can operate from the small cells for high data rate and system capacity

Figure 1.4 Carrier aggregation in a heterogeneous network.

among cells through cross-network-layer coordinated multi-point transmission and beamforming, application scheduling, power control, inter-cell and inter-network-layer interference coordination and alignment, examples of which include fractional frequency reuse (FFR) and time-domain resource partitioning.

As the wireless technology focus shifts from spectral efficiency to network efficiency, a major network performance improvement is expected to come from media optimization through the application of information and communication technology (ICT) [35]. In particular, caching enables more efficient content distribution for video, local navigation mapping information, social networking content uploading and distribution. Virtualization technology allows a unified signal processing platform for both control and data planes. It also enables dynamic workload balancing between signalling and application processing. The introduction of ICT brings many new possibilities for future heterogeneous networks, including the co-existence of both super-sized cloud base station and small cells, the possible introduction of a super control node that coordinates the operation of different network layers, the proliferation of data plane signal processing for cross-network QoS support, and the further flattening of the core network which facilitates the implementation of cross-layer optimization [36].

Finally, future heterogeneous networks also require advanced multi-mode and multi-band user devices. Consider in the following a rather extreme use case in China: a sales representative drives his car to visit his customer. He is making a GSM call using his Bluetooth headset, while downloading a movie via TD-LTE on his new smartphone, while his smartphone is searching for a WiFi hot spot for offloading the cellular network, while he is listening to the latest MP3 hits, streamed from his smartphone to the car radio via FM radio, while a GPS navigation application on the smartphone is running. And, by the way, the car is running at all only due to the fact the sales representative could identify himself as its owner by the NFC ID feature in his nice smartphone. As the world goes increasingly wireless and increasingly connected, it is no longer uncommon to find a user device with seven or more antennas. The devices themselves become an integral part of the future heterogeneous networks. The tunable RF circuit becomes essential to achieve compact design at low cost. Co-existence technologies based on TDM, FDM, SDM domains become increasingly critical. Together with advanced mobile devices, future heterogeneous networks will bring user experience to an unprecedented new level.

References

1. Panel discussion by Google CEO Eric Schmidt at Techonomy conference in Lake Tahoe, CA, August 2010.
2. Geng Wu, Shilpa Talwar, Kerstin Johnsson, Nageen Himayat and Kevin D. Johnson, ‘M2M: From Mobile to Embedded Internet’, IEEE Communications Magazine, April 2011.
3. Ericsson, ‘More than 50 billion connected devices’, February 2011.
4. McKinsey Global Institute, ‘Big data: The next frontier for innovation, competition, and productivity’, May 2011.
5. Cisco, ‘Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010–2015’, 2011.
6. Shu-Ping Yeh, Shilpa Talwar, Geng Wu, Nageen Himayat and Kerstin Johnsson, ‘Capacity and Coverage Enhancement in Heterogeneous Networks’, IEEE Wireless Communications, June 2011.
7. China Mobile, ‘C-RAN, The Road Towards Green RAN’, Version 2.5, October 2011.
8. ABI, Mobile Operator CAPEX Market Data, 4/5/2012.
9. K. Johansson, A. Furuskär and C. Bergljung, ‘A Methodology for Estimating Cost and Performance of Heterogeneous Wireless Access Networks’, PIMRC ’07.

10. K. Johansson, J. Zander, and A. Furuskär, 'Modeling the cost of heterogeneous wireless access networks', *Int. J. Mobile Network Design and Innovation*, Vol. 2, No. 1, pp. 58–66.
11. Caroline Chan and Geng Wu, 'Pivotal Role of Heterogeneous Networks in 4G Deployment', ZTE Technologies, Issue 1, 2010.
12. Jennifer Pigg, 'The Operator as Innovator: Smartphones, Smart Apps and Smart Pipes', Yankee Group Research, February 2011.
13. Aleksandar Damnjanovic et al., 'A Survey On 3GPP Heterogeneous Networks', IEEE Wireless Communications, June 2011.
14. Sayandev Mukherjee, 'UE Coverage in LTE Macro Network with mixed CSG and Open Access Femto Overlay.' 2011 IEEE International Conference on Communications (ICC), Kyoto, Japan, 4 June 2011.
15. R. Balakrishnan, X. Yang, M. Venkatachalam Ian F. Akyildiz, 'Mobile Relay and Group Mobility for 4G WiMAX Networks', 2011 IEEE Wireless Communications and Networking Conference (WCNC), Cancun, Mexico, 28–31 March 2011.
16. Klaus Doppler, Cássio B. Ribeiro and Jarkko Kneckt, 'Advances in D2D Communications: Energy Efficient Service and Device Discovery Radio', 2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), 28 February to 3 March, Chennai, India.
17. Seppo Hämäläinen, Henning Sanneck and Cinzia Sartori, *LTE Self-Organizing Networks (SON): Network Management Automation for Operational Efficiency*, John Wiley & Sons, ISBN 1119970679, 7 February 2012.
18. Small Cell Forum, 'Femto Forum Summary Report: Interference Management in UMTS Femtocells', February 2010.
19. 3GPP TS 23.402, 'Architecture enhancements for non-3GPP accesses'.
20. 3GPP TR 36.938, 'Improved Network Controlled Mobility between E-UTRAN and 3GPP2/Mobile WiMAX Radio Technologies'.
21. 'GSMA, WBA Collaborate on wifi Roaming', http://wirelessweek.com/News/2012/03/gsma-wba-collaborate-on-wifi-roaming/.
22. Julio Puschel, 'Learning from the Femtocell and Wi-Fi Pioneers', Webinar, Informa, 18 May 2011.
23. Qian (Clara) Li, Rose Qingyang Hu, Geng Wu and Yi Qian, 'On the Optimal Mobile Association in Heterogeneous Wireless Relay Networks', IEEE INFOCOM 2012, 25–30 March 2012, Orlando, Florida USA.
24. Rose Qingyang Hu, David Paranchych, Mo-Han Fong and Geng Wu, 'On the Evolution of Handoff Management and Network Architecture in WiMAX', IEEE Mobile WiMAX Symposium, Orlando, USA, pp. 144–149, March 2007.
25. 3GPP TS 23.261, 'IP flow mobility and seamless Wireless Local Area Network (WLAN) offload'.
26. 3GPP TR 23.829, 'Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO).
27. Miikka Poikselkä, Harri Holma, Jukka Hongisto and Juha Kallio, *Voice over LTE (VoLTE)*, John Wiley & Sons, ISBN 1119951682, March 13, 2012.
28. 3GPP TR 23.862, 'EPC enhancements to support Interworking with data application providers'.
29. Gartner, 2011 Q3 Global Smartphones OS Market Share.
30. McAfee, 'McAfee Deep Defender – security beyond the OS to expose and eliminate covert threats', Data Sheet, 2011.
31. 3GPP TR 36.814, 'Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects'.
32. ABI Research, 'Mobile Network Offloading', December 2010.
33. Qian (Clara) Li, Rose Qingyang Hu, Yi Qian and Geng Wu, 'Cooperative Communications for Wireless Networks: Techniques and Applications in LTE-Advanced Systems', IEEE Wireless Communications April 2012.
34. Aamod Khandekar, Naga Bhushan, Ji Tingfang and Vieri Vanghi, 'LTE-Advanced: Heterogeneous Networks', 2010 European Wireless Conference, 12–15 April 2010.
35. Geng Wu, 'Virtualized Next Generation Wireless Network in the Cloud', International Mobile Internet Conference 2010, 13–14 December 2010, Beijing, China.
36. Nokia Siemens Networks, 'Liquid Radio – Let traffic waves flow most efficiently', White Paper, 2011.