12

Software Defined Networks for Mobile Application Services

Ram Gopal Lakshmi Narayanan
Verizon, San Jose, CA, USA

12.1 Overview

Cloud, virtualization, and software defined networking (SDN) are emerging IT technologies, which revolutionizes business model and technical realization. Enterprise and network operators are consolidating their network and data center resources using virtualization technologies into service architecture. First, it is better to understand essential requirements that are driving these technologies so that we can realize how these technologies are achieving the end goals. The driving requirements are:

- Virtualization: Implement network function in software, and decouple the hardware dependency. Then run the network functions anywhere without the need to know physical location and how it is organized.
- Programmability: Topology should be flexible and able to change the behavior of the network on demand.
- Orchestration: Ability to manage and control different devices and software uniformly with simple and fewer operations.
- Scaling: System should be scalable up or down based on the usage of the network.
- Automation: System should provide automatic operations to lower operational expense. It must support troubleshooting, reduced downtime, easy life cycle management of infrastructure resource, and load usage.
- Performance: System must provide features to understand the network insights and take actions to optimize network device utilization such as capacity optimization, load balancing, etc.

Software Defined Mobile Networks (SDMN): Beyond LTE Network Architecture, First Edition.
Edited by Madhusanka Liyanage, Andrei Gurtov, and Mika Ylianttila.

- Multitenancy: Tenants need complete control over their addresses, topology, routing, and security.
- Service integration: Various middleboxes such as firewall, security gateway, load balancers, video optimizer, TCP optimizer, intrusion detection systems (IDS), and application-level optimizer must be provisioned on demand and placed appropriately on the traffic path.
- Open interface: Allow multiple equipment suppliers to be part of the topology and open control functions to control them.

The SDN goals are being envisioned by providing (i) separation of control and user plane traffic, (ii) centralized control of network functions and policies, (iii) open interfaces to hardware and software that need control, and (iv) control of traffic flows and programmability from external application. As SDN is a bigger concept, there are various organizations working toward standard approach including network function virtualization (NFV) from the European Telecommunication Standards Institute (ETSI), OpenFlow from the Open Network Forum (ONF), Interface to the Routing Systems (I2RS) from the Internet Engineering Task Force (IETF) [1–5].

Mobile network is expanding in all directions including the number of base station to accommodate traffic growth, increase in number of connected devices to provide machine-to-machine (M2M) services, and number of applications being developed for users. Operators are challenged on how can they manage and control such sudden growth in mobile network and, in parallel, allow growth in mobile networks. IT infrastructure has already consolidated their data center using cloud and SDN-based architecture when they faced with similar challenges. Therefore, telecommunication operators are following similar evolution on wireless networks. In this regard, 3GPP standard organization has started study group activity on SDN for Long-Term Evolution (LTE) wireless architectures. The goal of this chapter is to provide an overview of mobile network and then describe how NFV and SDN-based mechanisms are applied to wireless architecture. Next, we describe various application-level use cases and how SDN can be applied to improve the operation of network. Finally, we conclude with list of open research problems for future study.

12.2 Overview of 3GPP Network Architecture

Mobile broadband access network consists of packet core network (CN), radio access network (RAN), and transport backhaul network (TN). Simplified 3G and 4G mobile broadband network architecture is shown in Figure 12.1. Third-generation RAN shown in Figure 12.1a consists of NodeBs and radio network controller (RNC). The functions of RAN include radio resource management (RRM), radio transmission and reception, channel coding and decoding, and multiplexing and demultiplexing. Layer 2 radio network protocol messages are used to carry both control and user plane traffic from RAN to user terminal. RNC identifies signaling and user plan messages and forwards layer 3 mobility management messages toward CN HSS for authentication and authorization of users. Uplink user plane traffic from UE is received at RNC, and RNC performs GPRS Tunneling Protocol (GTP) operation and forwards the IP packet toward GGSN. Similarly, downlink packets for UE are received at RNC, and GTP packets are terminated and inner IP packet is forwarded to UE.

3G packet core consists of Gateway GPRS Support Node (GGSN) and Serving GPRS Support Node (SGSN). The functions of packet core include IP session management,

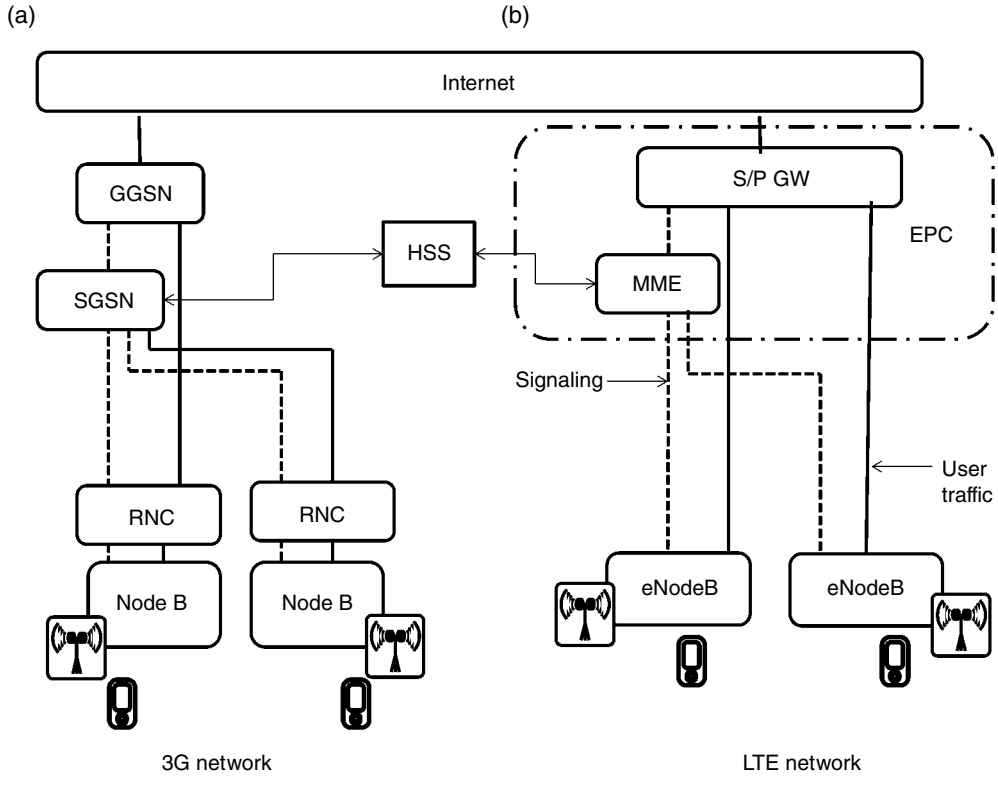

Figure 12.1 Mobile broadband networks.

legal interception functions, policy-based routing and charging functions, etc. User mobility and its associated sessions are handled by SGSN node. SGSN also acts as anchor point for ciphering and authentication of session between UE and wireless access networks.

Fourth-generation LTE radio network is shown in Figure 12.1b and it consists of eNodeB network element [6]. LTE design is based on flat architecture with reduced number of network elements. The RAN functions are consolidated into single network element eNodeB. 3G and 4G RAN networks are not backward compatible as their physical layer technologies and radio network protocols are different. 4G uses orthogonal frequency division multiplexing (OFDM), whereas 3G uses wideband code division multiple access (WCDMA)-based technologies for wireless physical layer processing.

Similar to 3G, there is a logical separation of networks such as RAN, packet core, and transport network in 4G networks. Mobility management entity (MME) and serving and packet gateway (S/P-GW) are part of Evolved Packet Core (EPC) in LTE. The functions of MME include radio signaling functions and mobility management sessions maintenance. The EPC functions are similar to that of 3G GGSN and contain improved IP mobility management functions.

12.3 Wireless Network Architecture Evolution toward NFV and SDN

12.3.1 NFV in Packet Core

ETSI NFV ISG standard organization is defining NFV standards, and most of contributing members are from telecom operators and network equipment suppliers. Both SDN and NFV share same goals and NFV came out of SDN concept. Therefore, it is worth to investigate NFV goals and architecture and its applicability to LTE and beyond architectures.

Implementation of each network function as software implementation and running them in virtual environment is called NFV. Today, 10 gigabit/s links are commonly used in most of switches and routers, and general-purpose computers are becoming cheaper and are capable of processing of most of switching and routing functions in software itself. NFV concepts were based upon SDN and were complementary to each other, and both can exist independently. Figure 12.2 describes the simplified view of ETSI NFV ISG architecture. One could approach this activity in two steps: firstly, move network functions from proprietary appliances to generic virtualized IT HW and SW stacks, and secondly, implement software functions as virtualized software functions (VNF), for example, IMS or MME each running inside a virtual machine (VM). This allows the network operator to use standardized computer infrastructure without HW vendor lock-in, and based on the network conditions, they can dynamically instantiate software and enable flexible service innovation. To manage with high degree of flexibility, management and orchestration (MANO) functions include service and network orchestration layer, and it interacts with operators' business and operation support systems (BSS and OSS). ETSI NFV ISG has divided their activities into architecture of the virtualization infrastructure, MANO, software architecture, reliability and availability, performance and portability, and security working groups to define the standardized interfaces and solutions.

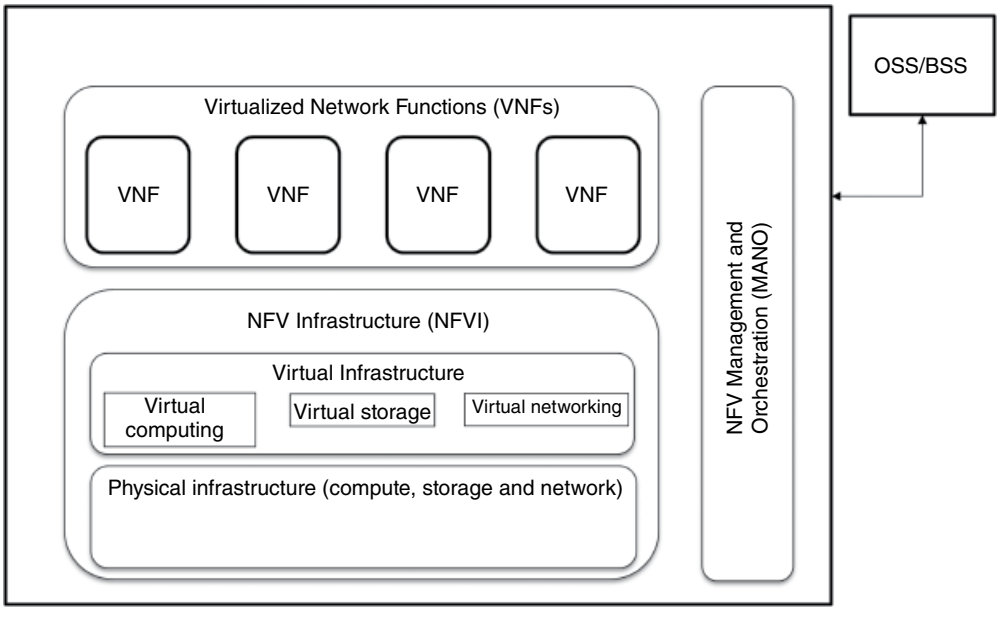

Figure 12.2 Simplified ETSI NFV ISG architecture.

The relationship between NFV and SDN is as follows:

- NFV came out from SDN, and SDN and NFV are complementary as both serve same goals.
- NFV can be implemented without SDN.
- According to NFV, virtualization of network function alone is sufficient to solve most of the problems, and this is used in current data centers.
- SDN relies on separation of control and data plane forwarding, defining additional control and interfaces.
- NFV and SDN can be combined to create potentially greater value.

12.3.2 SDN in Packet Core

SDN has become apparent as it provides simple interface to vendor equipment and allows dynamic network topology changes and easy integration of new services and reduces operating cost. The goal of SDN is to provide (i) separation of control and user plane traffic, (ii) centralized control of network functions and policies, (iii) open interfaces to hardware and software that needs control, and (iv) control of traffic flows and programmability from external application.

In 3GPP architecture, there exists separation of user plane and control plane in interface and in protocols. However, all these interfaces are implemented in proprietary hardware and it is difficult to decouple and introduce new services. Processing elements such as MME and policy charging and routing function (PCRF) process only control plane messages, and it contributes toward proper handling of user sessions. Elements such as eNodeB and EPC P/S gateway process both control plane and user plane traffic, and for performance reasons, each vendor implements it in dedicated hardware. To have clear separation between control plane and user plane, the following could be one of the possible approaches. First, virtualize the network function so that they can be made available on any hardware or on the same hardware. Secondly, centralize the control functions via separation of the control plane processing and user plane processing via appropriate SDN protocols. Both these steps are described in the following sections.

12.3.2.1 Virtualized CN Elements

Figure 12.3 describes the virtualized network functions of existing network interfaces. This step is similar to the proposal proposed by NFV standards. But centralize all network functions in single or less hardware so that higher degree of control and simplification to the internal communication can be achieved. As we could see that SDN, virtualization, and cloud are related and essential concepts:

- Removing hardware and software dependency: As the network element is no longer a collection of integrated hardware and software entities, the evolution of both is independent of each other. This enables the software to progress separately from the hardware and vice versa.
- Flexible network function deployment: As network function is virtualized, it improves the resource utilization of the hardware. For example, MME and PCRF could be running on one

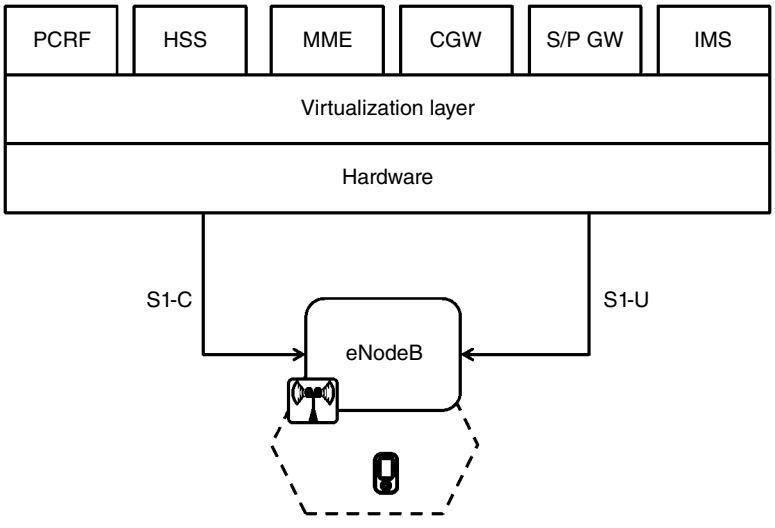

Figure 12.3 Virtualized 3GPP network functions.

hardware and EPC core on different target hardware, or if the network is serving smaller number of users, all of them can be made to run on the same hardware platform. We bring the concept of hardware pools and via virtualization any of the network function can be made to run on that hardware. Such dynamic start, stop, and movement of network functions bring different cloud and network topologies.

- Dynamic operation: When network functions can be virtualized and moved within the hardware pool, we can further extend and allow instantiated software to run with certain configuration and do dynamic operations.

There are several approaches to LTE wireless network element functions. Today, high-end commercial off-the-shelf (COTS) hardware and operating system are capable of handling higher traffic and can be configured as layer 2 switch or router by software configuration. Given this to the advantage, virtualize the 3GPP network elements into virtual environment without changing the 3GPP interfaces. This is one of form separation and gives more flexibility to operator who currently runs dedicated hardware for each network function. Virtualization gives improved resource utilization and ability to consolidate or distribute the processing functions at any time based on load and enables to optimize the resources. Figure 12.3 describes a scenario wherein all 3GPP control plane applications and EPC core functions are moved into virtualized network environment. For next-generation wireless architecture, virtualization becomes prerequisite for cloud; therefore, we must consider SDN on virtualized network environment. Pure control plane functions like MME, SGSN can run as applications in the cloud without impacts to other nodes as long as 3GPP-defined interfaces do not change. The initial step in the evolution process could involve virtualization of EPC nodes. The second step could involve moving whole P-GW/S-GW, other EPC functions like MME and PCRF to cloud or virtualized environment.

12.3.2.2 SDN Mechanism on Virtualized CN Elements

To achieve separation between control and user plane traffic, network element functions such as P-GW, S-GW must be logically separated and be able to communicate via open interface protocols. Identification of forwarding and control functions requires careful investigation of how and where each feature of S-GW and P-GW resides. The goal of this exercise is to achieve total separation of control and user plane functions. Figure 12.4 describes the control and user plane separated EPC core and running of SDN protocol as open interface to control the network functions. There are many protocol candidates such as OpenFlow and I2RS proposed by various standard organizations as SDN protocol. The purpose of this protocol is to support mechanism to configure, control, and manage the network functions and sessions seamlessly.

When control plane function of the gateway is virtualized (running on a VM) as shown in Figure 12.3 and the user plane function of the gateway application protocol (e.g., GTP-U) is not virtualized (running on dedicated hardware) as shown in Figure 12.4, the 3GPP specific user plane control and reporting functionalities shall be supported by the control protocol between the virtualized S/P-GW-C and the nonvirtualized S/P-GW-U.

12.4 NFV/SDN Service Chaining

12.4.1 Service Chaining at Packet Core

Network provides diverse functions. User gets service based on subscriber or type of application traffic types. Depending upon type of service, traffic passes through one or more middlebox functions deployed in the networks. Operator typically deploys several middlebox functions such as carrier-grade network address translation (CG-NAT), firewall, video optimizer, TCP optimizer, caching server, etc. Figure 12.5 shows deployment scenario of middlebox functions and how different packets go through series of network interfaces. Middleboxes are deployed

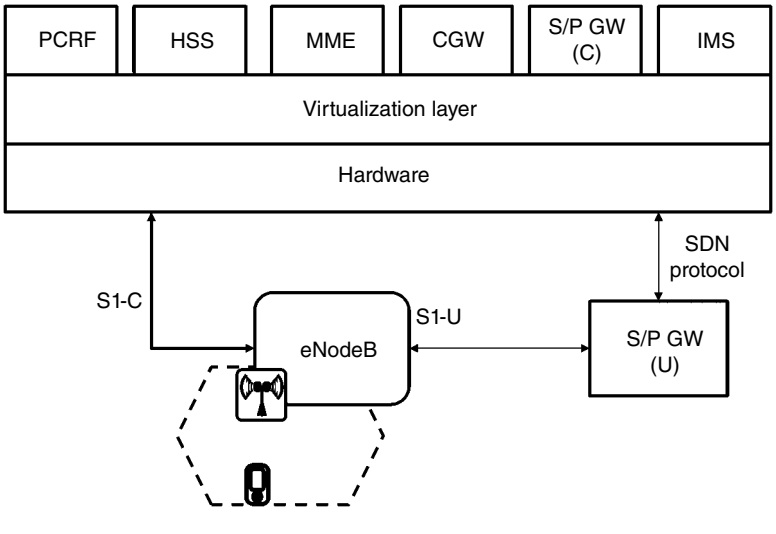

Figure 12.4 Control and user plane split architecture.

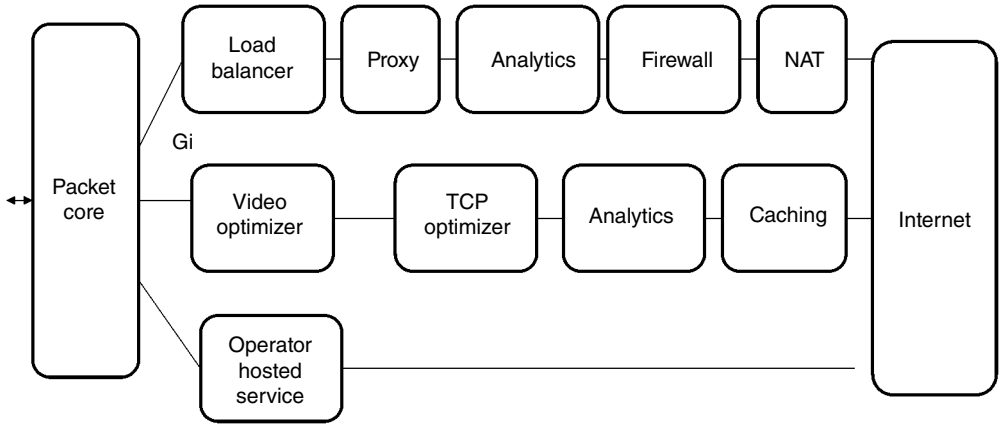

Figure 12.5 Traditional Gi LAN service chaining.

at SGi interface that connects data network and P-GW. When packet traverse through different path and treated by middleboxes on the path, it is called service chaining [7, 8].

Service chaining is currently operated via static policy and is not flexible. It has the following shortcomings:

- Network equipment supplier provides middlebox functions on dedicated hardware and configuration in SGi. LAN interface is hardwired to network interfaces and it is not easy to add or change the topology.
- Network operator does not have sufficient insights about how the traffic on their network is flowing and how it is impacting their service. Most of the services are offered by third-party applications and it is difficult to achieve standardized way of monitoring and control. Due to these reasons, often, operators preprovision systems to maximum capacity. Consequently, it is more difficult to enforce policy, manage traffic, and differentiate services dynamically.
- Static policy enforcement does not give flexibility and does not allow the network to scale when the traffic grows dynamically.
- Operator has to purchase dedicated hardware to run each middlebox functions such as caching server, firewall, load balancer switch, analytics engine, video optimizer, etc. Each such independent unit brings complexity and duplicated functionality. For example, a function such as DPI functionality is available in EPC and also in certain content filter or optimizers and it is difficult to disable and enable selectively. This often results in poor treatment and extra load on the network.

The goal of service chaining is to address the current problems:

- Dynamic addition and modification of services to chains.
- Packets are getting treated only once, and the same service can be applied to other chains.
- Packets may take different routes based on dynamic policy being enforced (e.g., based on subscriber profile, application type, network condition, etc.).
- Avoid unnecessary hardware, and enable virtualized software instance and form graphs and connect links to treat different packet flows.

This nature of dynamic service changing has been in discussion with various standardization bodies including IETF and NFV [7, 8].

Combining SDN and service chaining addresses dynamic control of each middlebox with central policies. SDN enables open interfaces and centralized control point for the middlebox and allows service creation functions such as bandwidth management, traffic steering, etc.

Virtualized network element function is a key enabler for dynamic service chaining. When network functions are virtualized and are running in same hardware, it is easy to perform different chain structure and create services easily. Each packet could be treated differently based on the service by making the packet to go through different next hop network functions and form a graph. As the packet has to go through series of middleboxes as part of service, the next hop vectors can be easily provisioned via flow label protocol such as OpenFlow. Figure 12.6 describes the service chain mechanism, wherein only necessary chains are selected based on the dynamic policy from orchestration layer. The orchestration entity has SDN controller functions and aggregates network state information and delivers service aware policy information dynamically. There is no preprovisioning involved network scale dynamically.

12.4.2 Traffic Optimization inside Mobile Networks

In 3GPP architecture, all user plane control and policy functions are centralized and reside at packet core. This creates difficulty to introduce any new service easily. To illustrate this, consider two users, say, U1 and U2 as shown in Figure 12.7a, are using voice over IP (VoIP) or P2P-type applications. All uplink traffic from user U1 goes through radio, transports, and reaches packet CN. At the packet core, GTP tunnel functions, charging, and NAT functions

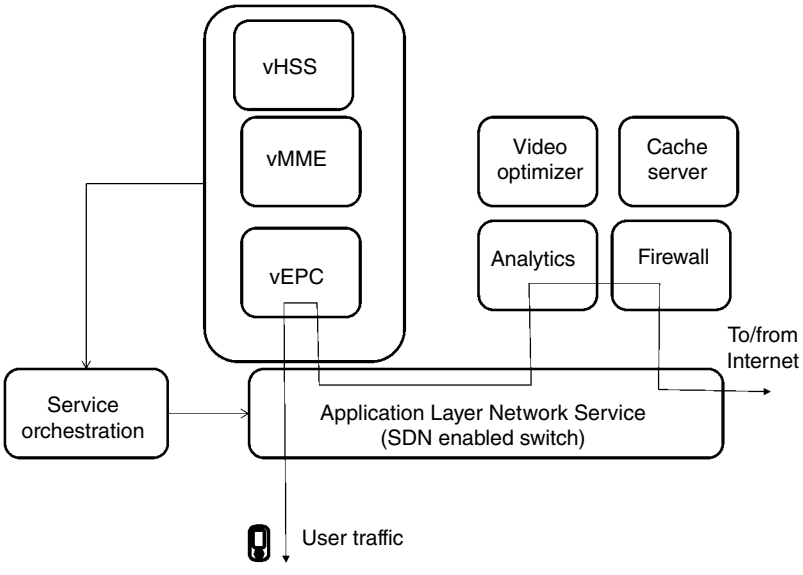

Figure 12.6 Dynamic service chaining.

are applied, and packet is sent back to U2 as downlink traffic. The traffic flow is north–south as shown in Figure 12.7a. It would be optimal to keep the traffic between the two users inside the RAN network (east–west traffic flow), and this will save bandwidth in transport network and reduce processing in EPC.

Concept of separation of signaling and switching is there in 2G mobile switching systems (MSC) and in plain old telephone systems (POTS) architectures. Traditionally, MSC is capable of handling both call flow signaling and switching function. Later, signaling and switching functions got separated into two different processing entities and are called soft switch. Operator will have one signaling server and deploy many switching servers closer to BTS or BSC in 2G architecture, thus allowing only signaling to be handled in the CN, and user traffic stays within the radio access networks.

Using SDN, it is possible to covert north–east to east–west traffic inside mobile networks. Design behind this is to have dynamic control of traffic flows, and traffic flow must be controlled based on the network state and internal network functions. Two key steps while designing SDN enabled network function are identification and exposure of internal state of network function and when to configure, monitor, and manage those states based on the traffic or network conditions. By having such network function, we could achieve traffic optimization for each application or flow level.

Figure 12.8 describes how to achieve traffic steering functions using SDN architecture. In this architecture, eNodeB is integrated with SDN layer 3 switch, and all traditional GTP tunnel endpoint functions are performed in this switch. Also, at the packet core, EPC is virtualized; control plane and user plane functions are separated. To illustrate the concept, we have shown only charging gateway (CGW), carrier-grade network address translation (CG-NAT or NAT) network functions. The following are the sequence of message that happen during a VoIP traffic stay within RAN network:

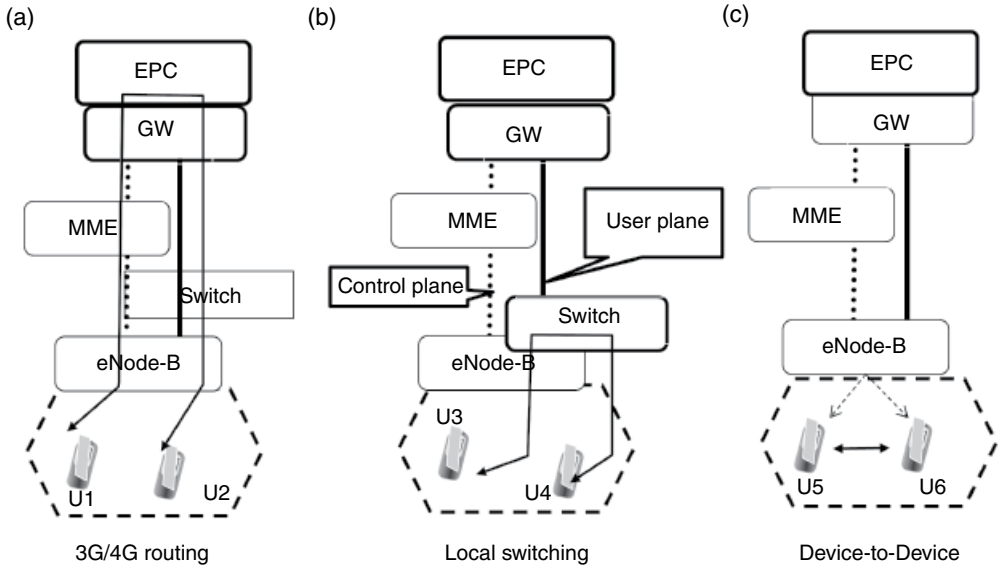

Figure 12.7 Traffic trend inside mobile networks.

1. SDN controller could be part of service and network orchestration layer function, and it provides initial configuration to CG-NAT. The CG-NAT is SDN enabled, and it will expose all its internal NAT table entries for management. Apart from traditional NAT, it includes extra control so that it will send a trigger message when two users behind CG-NAT are communicating. For example, when U1 traffic is seen first time on the NAT, it creates an outbound NAT table entry with U1 source IP address, destination address (public IP address of NAT), protocol, and port numbers. When NAT sees the traffic of U2, it will create an outbound traffic similar to U1. It compares internally when an entry is created with all other entries in the table to make sure that two users are not behind the NAT. SDN controller configures the CG-NAT, and CG-NAT will act on the user plane traffic and wait for the trigger conditions.
2. Either U1 or U2 is initiating VoIP session. VoIP application performs sequence of steps that includes peer endpoint IP address detection using simple traversal of UDP over NAT (STUN) or other suitable protocols and then exchange of signaling messages and starts application data transfer. As of now, IPv4 is still dominating and operators are using CG-NAT and deploy NAT function for all traffic toward the Internet. CG-NAT maintains mapping table in each direction.
3. CG-NAT detects the match on traffic flow between U1 and U2, and sends the traffic flow information trigger to SDN controller.
4. SDN controller instructs respective SDN-enabled eNodeB to apply traffic policies locally. Typically, policy would be to forward packet between U1 and U2 and start counting bytes in each direction. To have this functionality to be separately programmed, eNodeB's internal state processing must be identified and separated for external control. EnodeB typically performs GTP processing, and IP QoS functions toward packet core. Careful state identification and exposure is required so that higher degree of control is achieved. Timers such as dead-peer detection logics are part of OpenFlow protocols and those are also supplied along with the control messages. eNodeB SDSN switch function will hairpin the traffic locally for that session, thus saving the transport backhaul traffic. When traffic is being routed locally with eNodeB, the CG-NAT timer could expire; to avoid timer expiry, either SDN controller or port control protocol mechanism can be used explicitly to keep the NAT association active till the session is completed.
5. When either U1 or U2 or both terminate their sessions, the eNodeB SDN interface function will send IP flow statistics information pertaining to that flow including the duration and number of bytes transferred to SDN controller.
6. SDN controller could then send explicit command to purge the NAT table entry and also send the appropriate byte count information to CGW server.

To see the benefits of SDN, each network function when they are introducing control plane and user plane separation, they must identify each network internal state and expose them carefully. When control and user plane are separated properly along with internal state operations, SDN adds higher degree of freedom to service chaining and traffic steering.

Similarly, we can apply SDN to device-to-device (D2D) communication. 3GPP is standardizing D2D communication as part of Release 12 and Release 13 LTE features. These are going to be part of LTE-Advanced implementations. The goal of D2D is to combine location and proximity to enable direct D2D communication. There are two ways to use D2D, namely, infrastructure mode and ad hoc mode communication. Operator-assisted D2D communications are useful under following scenarios:

- Operators have limited spectrum available; infrastructure-based D2D communication will give relief for operator. By using D2D, operator could accommodate more devices to communicate and scale the network.
- When two users who wish to communicate among themselves are in cell edge of the network, they may not get sufficient bandwidth for their communication. In that situation, by using D2D network, they will be able to establish communications.
- A group of users want to share images or files among themselves. They share common profiles, and when D2D communication is present, operator could establish and manage their wireless connectivity to have proximity-based communications.

Without putting additional requirements on the user and application, we need to establish D2D communication when possible. To achieve this, SDN-based signaling can be used. For example, we described earlier a scenario of VoIP wherein two users U1 and U2 are behind the same cell or the same BTS. Now, BTS in the D2D communication is more flexible to lease frequencies to couple of mobile devices for D2D communication. If we extend that concept further, if two users are behind the same BTS (or cell) and are in close proximity, we could apply SDN principle to have D2D communication. Combining application endpoint information and location proximity information, we could create configuration parameters that will then eventually be used to establish D2D communication. Following are the possible additional procedures that must be incorporated in Figure 12.7.

Referring to sequence of message described earlier for VoIP application, there is no change to steps 1–3 of Figure 12.8. Additional functions are implemented in eNodeB. At step 3, eNodeB needs to check whether two devices are in close proximity, and they are allowed to communicate directly via D2D channel. eNodeB can verify the location and proximity information either with the help of MME or location-based service network elements. Assume that two UEs are in close proximity; then they will establish network-assisted D2D session, and control information will go to eNodeB.

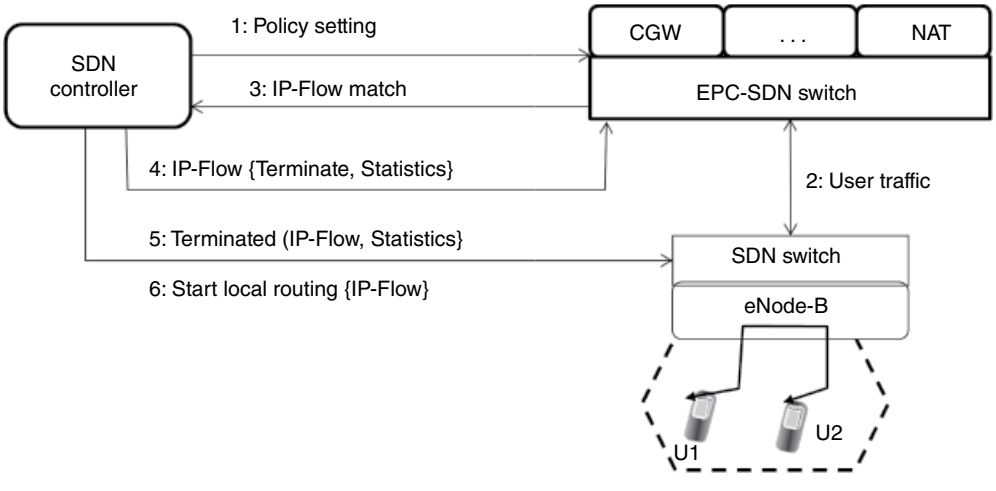

Figure 12.8 Traffic steering inside mobile networks.

12.4.3 Metadata Export from RAN to Packet CN

In the previous section, we described using SDN how to extract and transfer packet CN state information to radio network and achieve traffic optimization. There are situations wherein radio network states information needs to be extracted and transferred to packet core or to SGi LAN to achieve service differentiation. We will illustrate this use case by considering video streaming application.

Video traffic contributes toward a major portion of Internet traffic. Due to increase in penetration of smartphones and tablets, video traffic will grow by many folds. User-generated content (UGC) such as YouTube and premium traffic such as Netflix are major contributors in the United States [9, 10]. Such a sudden explosion of video traffic makes network unmanageable. Lessons learned from deploying fourth-generation mobile network such as LTE and third-generation mobile network such as WCDMA technologies have shown that video streaming does not perform well over mobile broadband networks. Video streaming session has a long duration and demands latency and bandwidth throughout the video play. Hence, network resource must be made available throughout the video play. To satisfy these requirements, mobile operators have increased their server and network capacity. However, wireless signal strength varies with respect to location, time, and environment, and delivering guaranteed bandwidth to video streaming application in such nonuniform wireless network condition becomes a challenge.

Three popular mechanisms exist in the Internet to deliver video content using standardized protocols. They are as follows: (i) user could download a video file from server using FTP service and watch video locally in his device at later time, (ii) video could be streamed as video on demand (VOD) from a server, and (iii) video could be delivered in real time. VOD and live video streaming are most popularly used today. Pseudo streaming and adaptive bit rate (ABR) streaming are two popular streaming mechanisms and both run on HTTP [11]. As of today, ABR is adopted by most of the content distributors. In ABR streaming, a video data may be coded into different quality levels such as high, medium, or low resolution. Each of those coded video data is further divided into chunks and kept as small files in the video content distribution server. Each chunk contains 3–8 s of video data. Based on the available bandwidth, client selects video quality and then requests respective chunk file for play. When the requested chunk is being downloaded by client, client computes several parameters including round-trip time, total download time for the chunk, etc. and keeps this history information. The stored information is then used to decide next chunk request. As the wireless network is time varying, TCP congestion control algorithms are not well suited for such burst video transmission and results in retransmission and video stalls. Application service providers are exploring the possibility to get additional information about radio conditions so that they can adjust their video transmission accordingly. As 3GPP standards suggest full inspection of control and data packets to solve RAN congestion [12], it is not possible when application service providers are enabling SSL-based encryption on user plane data. Therefore, one of the possible approaches is to have cooperative solution between application service provider and network operator. Figure 12.9 describes a solution wherein SDN controller is part of PCRF function and interacts with video service function such as caching server or video optimizer or TCP optimizer hosted on SGi LAN. The traffic from the UE is encrypted and is not possible to apply DPI and perform RAN traffic management functions.

There are two approaches to trigger the information export from eNodeB. (i) In the first approach, SDN controller receives periodic flow information that requires radio state information from eNodeB. (ii) In the second method, video server IP address is known to SDN

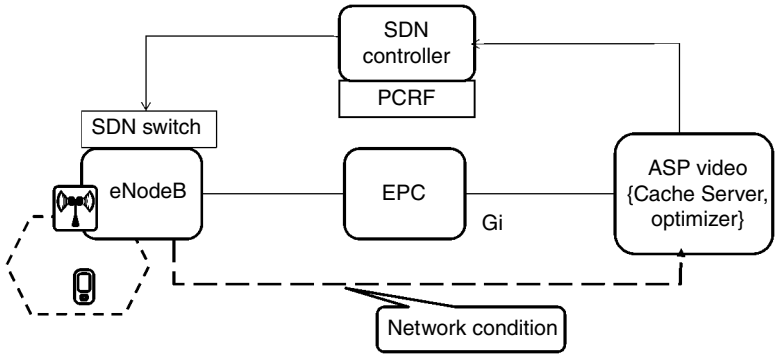

Figure 12.9 User plane RAN optimization.

controller, and it could instruct eNodeB to provide radio information when uplink traffic matches the video server IP address. In both approaches, eNodeB will export its internal radio state information such as cell load condition, traffic congestion, etc. EnodeB can piggyback this information on uplink traffic from UE toward video server, or alternatively, it could send at IP layer option information. This metainformation is then passed through EPC core and reaches video server on SGi LAN.

The role of SDN controller is to communicate the required flow configuration information to eNodeB. Then, eNodeB will act upon the IP traffic flow; eNodeB will generate metadata information when a match is found in IP traffic against the configuration. For example, whenever UE sends uplink traffic toward video server, as the server's IP address is put as part of eNodeB configuration, the eNodeB will provide the metadata and sends along with the UE uplink packet. The metadata may include radio load information, cell-level information, UE location information, etc. Existing eNodeB does not export this information at this moment. This is not yet defined in standards. When we are enabling SDN for each network function, there must be flexibility in configuring each network function, and hence, respective actions can be managed.

12.5 Open Research and Further Study

We described application scenarios and how it can be improved by combing service chaining and SDN principles. We emphasize that SDN provides separation of control and user plane and enhance higher degree of control and give the notion of programmable networks. As part of SDN design, careful design choice must be made on how to expose internal network and network element states. What we presented is just a beginning; as part of this exercise in wireless network, we need to consider several aspects. We believe that the following are still exploratory activities that are still open to both research and innovation:

- Legal interception gateway is performed in packet CN. Lawful agencies apply policies on user traffic and collect the user data [13]. When traffic optimizations as described in Figure 12.8 are performed, the traffic will not reach LIG. Therefore, we must distribute LI functions carefully to RAN network or disable the localized traffic routing for sessions requiring LI.
- Separation and exposing network state on each network function is a complex task. As EPC supports more than 100s of protocols, we need to ensure that state information can be manipulated or exported to outside world as programmable API.

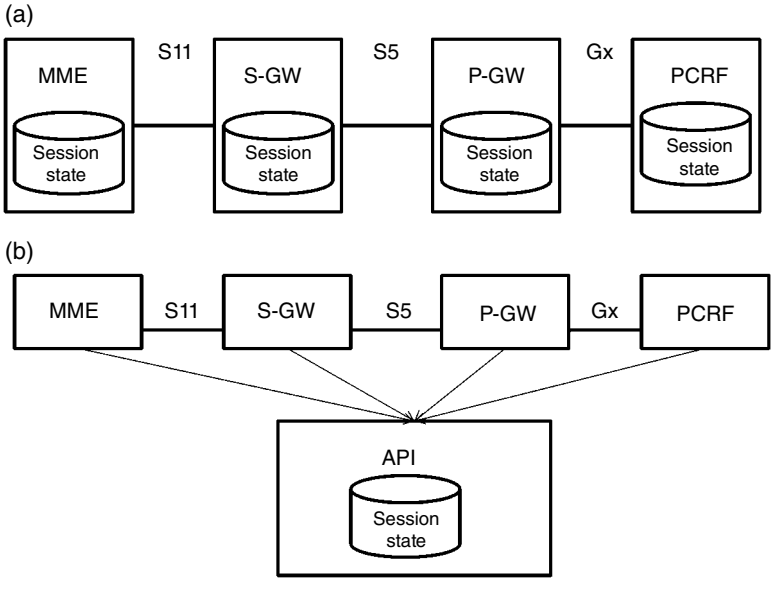

Figure 12.10 State consolidation and SDN API exposure.

- MME, S-GW, and P-GW maintain UE state information internally. Network elements such as MME process control plane information, S-GW acts as mobility anchor point, and PDN terminates traffic at SGi interfaces and acts as policy enforcement point. As the functionality is different, they need to maintain redundant information of each UE. As described in Figure 12.9, when a UE-attached procedure is performed, MME, S-GW, and P-GW creates state information for each UE and has lot of redundant information. When network functions are virtualized and going to run on a single hardware platform, we can revisit dedicated box approach to share virtual environment and improve the network design by making network element stateless as much as possible. Figure 12.10b describes one of the approaches to consolidate the state information in the CN elements and allow the database to be exposed with an API.

Acknowledgments

This work started while the author was working at Nokia. At the time of publication, the author works at Verizon. This work is supported both by Nokia and Verizon. Opinions, findings or recommendations expressed in this chapter is from the author and does not reflect the views of Nokia and Verizon.

References

[1] Atlas, A., Nadeau, T.D., and Ward, D. (editors) (2014) Interface to the Routing System Problem Statement, IETF, (work-in-progress) draft-ietf-i2rs-problem-statement-03, June 2014. Available at http://tools.ietf.org/id/draft-ietf-i2rs-problem-statement-03.txt (accessed February 17, 2015).

[2] ETSI, NFV (2013, October) Network Functions Virtualisation—Update White Paper. Available at http://portal.etsi.org/NFV/NFV_White_Paper2.pdf (accessed January 24, 2015).

[3] ETSI, NFV (2013, October) NFV Virtualization Requirements. Available at http://www.etsi.org/deliver/etsi_gs/NFV/001_099/004/01.01.01_60/gs_NFV004v010101p.pdf (accessed January 24, 2015).
[4] ETSI, Network Functions Virtualization (2014). Available at http://www.etsi.org/technologies-clusters/technologies/nfv (accessed January 24, 2015).
[5] Open Networking Foundation (ONF) (2012) Software-Defined Networking: The New Norm for Networks. Available at https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf (accessed January 24, 2015).
[6] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description. Available at http://www.3gpp.org/dynareport/36300.htm (accessed February 17, 2015).
[7] Quinn, P., and Nadeau, T. (editor) (2014) Service Function Chaining Problem Statement, IETF, (work-in-progress) draft-ietf-sfc-problem-statement-05.txt. Available at https://tools.ietf.org/html/draft-ietf-sfc-problem-statement-05 (accessed February 17, 2015).
[8] ETSI GS NFV 002 V1.1.1 (2013, October) Network Functions Virtualisation (NFV); Architectural Framework. Available at http://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.01.01_60/gs_NFV002v010101p.pdf (accessed January 24, 2015).
[9] About YouTube, Available at https://www.youtube.com/yt/about/ (accessed April 9, 2015).
[10] Netflix. How does Netflix work?, Available at https://help.netflix.com/en/node/412 (accessed April 9, 2015).
[11] Stockhammer, T. (2011) Dynamic Adaptive Streaming Over HTTP: Standards and Design Principles. In: Proceedings of the second annual ACM conference on Multimedia systems, pp. 133–144. ACM, 2011. San Jose, CA, USA.
[12] 3GPP TR 23.705, System Enhancements for User Plane Congestion Management, Release, draft 0.11.0. Available at http://www.3gpp.org/DynaReport/23705.htm (accessed February 17, 2015).
[13] 3GPP TS 33.107, Lawful Interception Architecture and Functions. Available at http://www.3gpp.org/ftp/Specs/html-info/33107.htm (accessed January 24, 2015).