# 1 Ultra-dense Networks and Sliced Network Services

Anwer Al-Dulaimi and Chih-Lin I

The network densification is one of the prominent solutions in fifth-generation (5G) networks to utilize spectrum resources through intensive deployment of small cells. However, the traffic management in dense networks may become a serious challenge for underlying infrastructure supporting the virtual core network. Moreover, 5G will employ different types of communication frameworks: ultrareliable low-latency communication (URLLC), enhanced mobile broadband (eMBB), and massive Internet of Things (mIoT). Each identify standard slice types (STTs) that have different performance requirements and enabling technologies. However, network developers will need to provide a concise modeling on how those logic networks would be administrated on top of physical network. Therefore, this chapter investigates the 5G sliced networks and studies virtual networking options to meet the performance requirements of service-based architecture.

## 1.1 Introduction

The tremendous increase in mobile connectivity and downloaded applications motivates mobile operators to employ massive numbers of access points that offload traffic locally. This dense deployment of various radio interface technologies aims to leverage the utilization of radio resources through intensive segmentation of spectra into small contagious domains. In addition, 5G relies on spectrum extensions using multi-radio interfaces accessing multiple bands to increase throughput and meet the rapidly growing numbers of connected appliances/applications. The concept for this resource segregation seems to be logical and requires radio access modifications and enhanced radio interface technologies to achieve higher spectrum accessibility. However, this network densification exhausts the underlying network infrastructure and overloads physical ports of conventional physical network. This inconsistency between various network segment capacities may scale further into sequential network failure. This requires network designers to expand their vision and look at the end-to-end network infrastructure capacity to provide new models for operations management.

The explosion of data services is envisioned to exceed 1,000 times the capacity of current cellular networks post 2020. This incredible increase in traffic volumes forces to devise radio access technologies (RATs) along with upgrading legacy network systems into highly efficient ultra-dense networks (UDN). The massive multiple-input



https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

and multiple-output (MIMO), New Radio (NR), and license-assisted access (LAA) are just examples of new RATs that support 5G throughput demands [1]. Tracking and predicting the exact position of connected mobile users is a key factor for all aforementioned RATs to maintain cost-effective and reliable radio links. This can also be reflected in proactive radio resource management (RRM) and mobility management that defines the different mobile equipment status prior to transmission actions [2]. A key factor for 5G RATs is to facilitate multiuser transmissions using current radio frame structures to reduce the impact of deploying new technologies and maintain interoperability with other legacy long-term evolution (LTE) releases. Therefore, transmissions should still be bounded to power ranges of current LTE while round-trip time (RTT) latency is reduced to below 1 ms [3]. From an architectural perspective, the trade-off between available free spectrum resources and UDN-deployed units is another key factor for network optimization to be considered [4]. However, this is an area that is still not well investigated since resources management is utilized separately at various components or layers. This also reflects the lack of deploying an efficient controlling mechanism that manages the network resources from end-to-end (E2E). However, network planners need to develop a long-term vision for an adaptive network that can maximize capacity and also commits to the intensive performance requirements of 5G.

Developing a sustainable network architecture that complies with 5G ultra-dense deployments is well investigated in literature. In [5], authors provide multiple solutions to evolve current networks for the 5G era. The main focus is to develop a new waveform and adopt a millimeter wave at the base station level. The work does not have any real proposals for backbone development. The authors in [6] provide topology analysis for deploying small-cell access points (SAPs) considering *centralized* and *distributed* network densification models. The network function virtualization (NFV) is the additive solution that allows networks to meet the dynamic traffic growth as a direct impact for network densification. The authors mention that network virtualization can be used to manage wireless traffic since physical fronthaul is sharing the same virtual backbone resources. However, there is no reference to enabling connectivity technologies and what features need to be supported for virtual and physical networking competency. The network architecture given in [7] consists of macrocells, microcells, picocells, and femtocells that are connected to a cloud radio access network (C-RAN) as the simplest way to migrate current LTE infrastructures into cloud-based networks. Considering traffic challenges, the paper looks into improving the radio accessibility through enhanced RRM. The 5G ultra-dense cloud small-cell network (UDCSNet) architecture was presented in [8] to investigate the small cell deployments at enterprise locations and how to manage resource allocations in master to small-cell modes. The authors presented optical fiber as the backhaul linking technology between the core network and the virtual baseband processing (vBBP) pool. This seems to be limited to the mobile-edge computing (MEC) side and does not really provide any solutions to handle large-scale networks that employ different landline technologies mostly from third-party providers. Similar limited solutions that lean toward RAN can still be seen in other papers [9–11], without any specified technologies.

The NFV/SDN offers magnificent opportunities to manage traffic on the backbone side through traffic-forwarding features. Although such use cases are well understood

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

from the industrial perspective, they have not been investigated enough in academia. In [12], the authors highlight the role of three-tiered SDN architecture in managing dense networks provided with two additive features: femtocell IP access (FIPA) and selective local controller traffic offloading (SLCTO) technique. The goal is to offload traffic at the network edge using OpenFlow (OF) switches that are connected directly to the Internet without requiring further processing from core network. Overall, this is a very interesting solution for managing user plane data at the network edge, but it makes the whole network vulnerable to the external world besides the additional expenses incurred from accessing external data repositories through different providers. A hierarchical software-defined networking (SDN) control plane is used in [13] to manage various domain resources. The paper investigates the transport layer without direct investigation for the SDN functionality in editing the traffic directions. The SDN presented in [14] can identify users momentarily requesting the same video downloads to initiate a multicast-based content distribution scheme for that video. This model is useful in reducing the load of processing similar requests at various locations, but it does not provide a comprehensive solution that explores traffic forwarding between involved network access points. Software-defined small-cell networking (SDSCN) is another framework that defines traffic by originating clusters and uses energy-efficient metrics to deploy/enable small cells considering traffic status [15]. The SDN is used here as a controller that activates small cells for specific RAN operations.

In the following sections, we will show the networking options at the core network side and potential options to meet the 5G performance requirements. Our goal is to explore thoroughly the features of hypervisor platforms and how the SDN is integrated with them to manage traffic forwarding. We will also explore how to leverage traffic editing in sliced networks where the network infrastructure is logically separated in service-based architecture.

### 1.1.1 5G Network Slicing

Network slicing is one of the main new concepts supported by 5G to meet the performance requirements of supported services. The network slicing refers to separate logical networks that are responsible for handling the communications with different types of connected users. A network slice is a dedicated virtual network supporting a specific service or customer over a common network infrastructure. This separation of services and users allows the accommodation of numerous and varied services as requested in 5G. Conventional LTE networks employ random access (RA) to connect services and users due to the limited set of preambles. This is one of the major cases for system complexity when supporting large numbers of users and is also more vulnerable to congestion [16]. Therefore, 5G uses the NFV/SDN features to utilize network resources across different segments to reduce operational costs; see Figure 1.1. The network architecture may not have all the technical features required to support every use case of connectivity. Therefore, service-oriented separation helps to define the necessary physical infrastructure for each service type and also reduces the impact of processing different types of requests using the same components [17]. This separation deepens

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

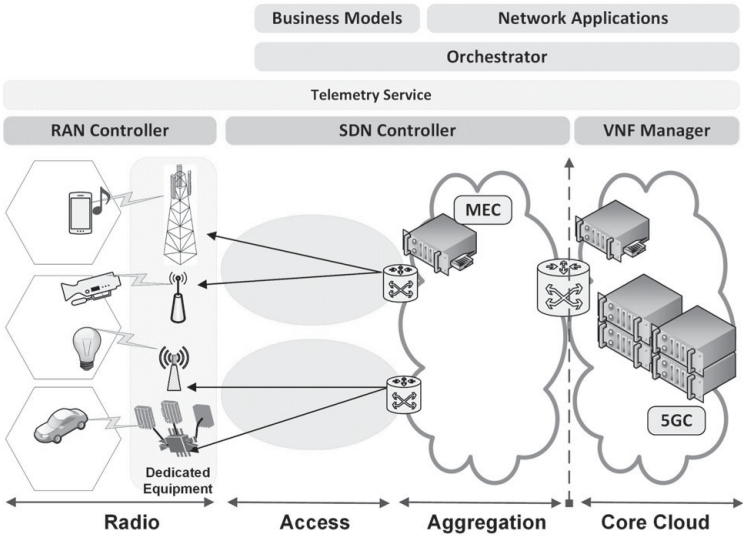


**Figure 1.1** 5G system: architecture, segments, and slices

further into the core network, creating E2E separation between various network slices. Each slice appears to be an E2E virtualized network that spans from the core to the RAN. From the RAN perspective, different RAT technologies that operate over different quality of service (QoS) bands are considered to be independent systems from each other. On the core side, the slicing refers to compressed computational resources of cores, storage, and accelerator cards that provide the computational resources 5G core (5GC) network components. The relevant service-based architecture (SBA) cores are also instantiated as separate virtual network functions (VNFs) from each other. Although this separation is necessary to improve processing different types of 5G services, it also raises many challenges for network operators on how this can be managed on a large scale and what could be the consequences for network physical infrastructure when overwhelmed with those logical networks.

One of the potential solutions is to limit the number of logical networks to avoid increasing network complexity. It is also necessary to understand that creating those networks may not be possible at all times due to the lack of physical computational resources and ports capacity. In fact, the rapid creation of multiple networks may increase the systematic time delays of platforms, causing a degradation in performance over many network segments. Therefore, we think that operators will choose to create permanent SBA cores that process main 5G slice services to reduce the technical burden of launching and removing such cores on demand. However, operators may accept to allow the a creation of a certain number of additional SBA cores that handle certain service functionalities that require additional computational processing speed or access to specific data resources. The final decisions on such scenarios will be subject to physical network status and services provided by each operator. In general, we can see that network slicing is a multitask challenge

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

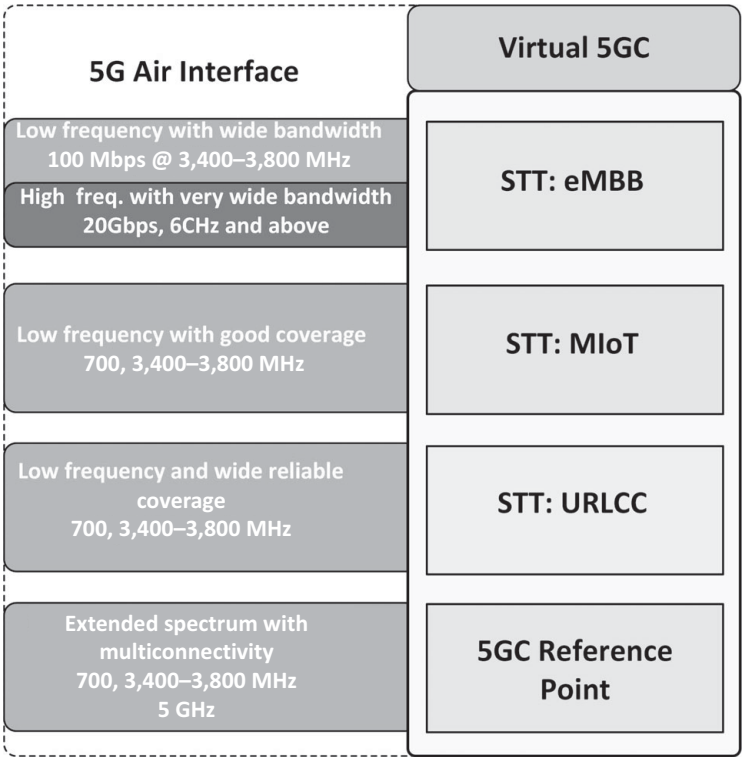


**Figure 1.2** The frequency bands for each service-based architecture

for service providers that involve technology evaluation, use cases, and also the business model.

In the following subsections, we summarize the slicing challenges at both RAN and core sides.

## RAN Slicing

The 5G network is intended to support multitype services include: eMBB, URLLC, and mIoT. Those types of services will have their own assigned RAT interfaces, as shown in Figure 1.2. Although various RATs and SBAs seem to be isolated from each other, this is far from reality. All networks share the same physical backbone and resources, and they are just logically existed to ease the processing of relevant services. Therefore, any multiple types of RATs can be used to create multiconnectivity to any end user to achieve a multigigabyte 5G wireless link with preestimated time delays to support a chosen service type. This raises many questions about service management and how to distribute wireless resources between multiple service types. The challenge can be further amplified as which service has the priority in a shared RAT infrastructure. Therefore, it is important to employ controllers that manage RAT sharing with predefined policies for prioritizing service delivery and user

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

connectivity requests. The MEC data centers can be the host of such controllers since they support multiple RATs located in one domain. This challenge becomes more complicated when a multi-RAT transmission employs third-party RATs such as an open-source Wi-Fi or a customer's own Wi-Fi. Overall, it seems that the multiconnectivity models are at the very early stages of feasibility since there was no urgency to have such solutions in fourth-generation (4G) networks.

The New Radio (NR) is accepted as the next generation of radio interface technology, which supports higher utilization of the spectrum compared to any previous technology. However, deploying the NR on top of an existing 4G network structure may not be the best path to take for long-term 5G deployments. The NR, supported by LAA features, should motivate operators to restructure the whole underlaying infrastructure into new models that support telemetry services. The latter will help operators to predict the maximum load of any network segment/domain along with the computational resources for the data center that serves that domain locally. It is also important to understand what slice service types may only instantiated when there is a need to support real use cases of connected users. This makes the RAT deployment completely dependent on use case and local services within domain. This actually reduces the cost of operation but also requires a lot of measurements during the deployment phases.

### 5GC Slicing

The 5GC is a new generation of core network that is designed to run over the cloud as software packages. Therefore, the 5GC entities will be deployed in the form of virtual entities over hypervisors. This virtualization allows instantiating, scaling, and interconnecting 5GC components subject to the arrival of user requests and the processed traffic. The 3rd Generation Partnership Project (3GPP) adopts SBA to separate various services from each other [18]. This means that not all 5GC components are necessary to support a logical network, and some components may not be needed in the case of certain services. The required SBA core can be instantiated at smaller local data centers once a service request is received with the service slice type (SST) value, as shown in Table 1.1. The SST also helps to divert the arrival traffic to relevant existing SBA when the request is processed by a network centrally. Regardless of whether a data center is used at MEC or a core network, they all need to employ hypervisors that support the scalability feature of *elasticity* to allow virtual 5GC entities to scale, subject to arrival traffic.

Since SBA is designed for a cloud environment, the Diameter protocol, and the Stream Control Transmission Protocol (SCTP) were removed from 5GC and replaced with the Hypertext Transfer Protocol (HTTP2) as the application layer protocol for service-based interfaces (SBIs) [19]. This was a major change for mobile operators that were used to proprietary hardware, and now they needed to dive deeply into the world of software development (DevOps). It also meant that mobile infrastructure was adapting to the world of cyber-physical systems (CPSs), where physical systems such as RATs are actually driven, controlled, and monitored by computer-based algorithms.

The 5GC comprises various network functions (NFs) that can be defined as authentication server function (AUSF), access and mobility management function (AMF), data network (DN), unstructured data storage function (UDSF), network exposure function

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

**Table 1.1** Standard slice type (STT) values for 5GC SBA

| Slice/service type | SST value | Characteristics |
|---|---|---|
| **eMBB (enhanced mobile broadband)** | 1 | This slice supports 5G-enhanced mobile broadband. It is also useful for general mobile broadband applications, including streaming of high-quality videos, fast large-file transfers, etc. This SST is anticipated to be the main login network supporting high data rates and high traffic densities. |
| **URLLC (ultra reliable low-latency communications)** | 2 | This slice is intended to support ultra reliable low-latency communications for applications, including industrial automation, remote control systems, vehicle-to-everything (V2X), etc. |
| **mIoT (massive IoT)** | 3 | This slice is intended to support a large number and high density of IoT devices efficiently and cost-effectively. |

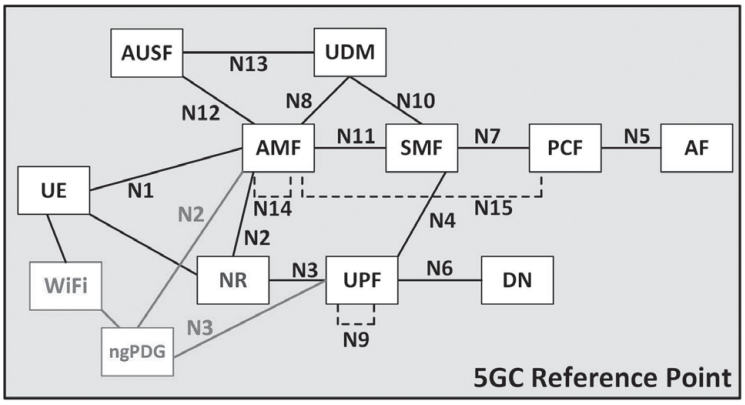


**Figure 1.3** The entities and interfaces of the 5GC reference point

(NEF), network repository function (NRF), network slice selection function (NSSF), policy control function (PCF), session management function (SMF), unified data management (UDM), unified data repository (UDR), user plane function (UPF), application function (AF), user equipment (UE), (Radio) access network ((R)AN), 5G equipment identity register (5G-EIR), security edge protection proxy (SEPP), and network data analytics function (NWDAF) [18].

A normal phone call will be handled by reference point 5GC, which is shown in Figure 1.3. It can be seen that the new interface model, consisting of (N1, . . . , N15), is provided with this new architecture that differs completely from the 4G evolved-packet core (EPC). 5GC provides the necessary interface mechanism to connect various RAT technologies (e.g., Wi-Fi) and also maintain a logic interface (N1) to the mobile user. However, it is necessary to keep in mind that incorporating different RATs using backbone does not enable multiconnectivity schemes immediately. In fact, the core interfaces do not solve the technical challenge of having different QoS transmissions. This

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

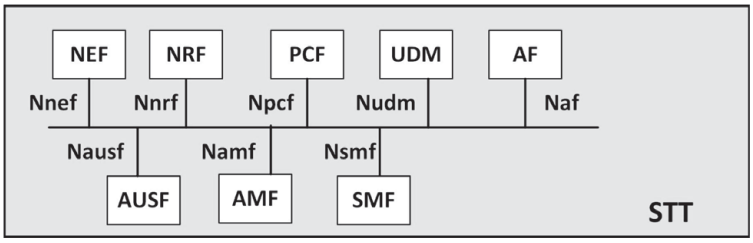


**Figure 1.4** The entities and interfaces of 5GC service-based architecture

is considered to be one of the challenges for 5G and is investigated by standardization projects such as IEEE 1932.1 [20].

During the interim transition from 4G to 5G, *nonstandalone mode* will be adopted to deploy the next Generation NodeB (gNB). In this mode, the control plane for gNBs will be interconnected to EPC, and the user plane will be obtained through LTE eNBs [21]. Once the 5GC is deployed on data centers, the gNBs will be connected to the new core and eNBs will be accessing the 5GC through the control plane, while the user plane will be provided by gNBs. This mode of core connectivity is referred to be as the *standalone mode*, and it is anticipated to be arriving at the final stages of 5G network deployments.

The 5GC SBA employs different core entities and interfaces, as shown in Figure 1.4. This type of 5GC modeling allows the support of requirements of URLLC, eMBB, and IoT services. SBA types have the ability to process a specific service type with direct access to relevant service providers, reducing the time for the search and download of requested services. Although the architecture is well defined, there is still a lack of definition for the sequence of processing operations and what entities will be coordinating requests. The 5GC SBI framework is using a representational state transfer (REST) application programming interface (API) that takes advantage of HTTP methodologies defined by the RFC 2616 protocol [22]. This design is fully compatible with 5G goals of achieving a fully virtual 5GC that can be deployed over the cloud. Moreover, it allows the scaling of various network functions, independently assuming that each may run as a virtual machine (VM) or even as an NFV. However, there are still many challenges to be addressed:

- The deployment of 5GC will be based on open-source software (OSS). Therefore, operators need to develop their own security schemes to protect their systems against offensive maneuvers that target information, infrastructures, and user data or devices. The security threats become seriously challenging in small deployments, when a G5GC could be deployed on small data centers to manage the calls at local sites, making network functions more vulnerable to cyber attacks.
- The 5GC entities may be deployed at various servers in a large cloud, causing performance challenges to meet slicing service requirements considering the software and hardware specifications of underlying platforms. Operators need to consider the computational resources and networking required to support expected

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

5G performance figures. This implies deploying new data centers that offer high processing speeds and larger-capacity ports at the hardware level as well as fully tested software stacks that reflect compatibility with hardware and harmony with other operator platforms.

It is important to understand that achieving gigabyte downloading speeds below 1 ms may not always be possible even with dedicated computational resources to 5GC. A sliced network is a logic network that spans from the RAN to application servers on the core network side. Therefore, sliced networks need to verify that all network components E2E meet the relevant slice performance measures. This might be hard to achieve since the basic infrastructure is laid down and shared by different operators. Additionally, most of the landline networking is provided by third parties that provide services to different types of customers. Assuming the access network will be upgraded and will have no congestion joints that may delay transmissions throughout the network, we envision that both efficient harmonized slicing of RAN and 5GC will enable meeting 5G performance characteristics. However, there are still concerns about any distributions to active networks in such sliced network models when service and user assignment may not be well implemented or maintained. From a 5GC perspective, the main challenge will always be the hosting platform and efficient orchestration of various network operational status.

## 1.2 Cloud Computational Platforms and Networking

The 5GC deployment over cloud offers magnificent opportunities for mobile operators to restructure their networks by employing distributed clouds. Before exploring the impacts of 5GC, it is important to understand what cloud platforms are and the technologies that support cloud-based networks. A cloud is typically a set of data centers that are connected to each other and exposed to the external world through the Internet. A simple example is the Google Cloud data centers that inspire ongoing transitions to cloud-dependent networks considering technology and connectivity [23]. Any consumer can procure Google virtual machines over the Internet and allocate standard or customized numbers of computational resources to run applications or store data. Similarly, operators are looking to deploy their own clouds that will also reserve the virtual machines of their core network entities and data. However, a considerable number of operators will be using third-party data centers for their cloud operations to reduce the staffing cost. Each data center has a certain number of servers that are cooled and interconnected to serve as one pool of resources. Some of their data centers might be small and have tens of servers that provide the backbone for some network clusters/domains, while others may have hundreds of servers that provide the backbone for large network domains or national clouds. In large data centers, tens of servers are deployed in containers (also known as *pods*) to reduce the real-time impacts of small component defects. Generally, data centers may employ various vendor equipment, with different specifications and capabilities. Nevertheless, these data centers need to provide coherent platforms that

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

meet strict performance characteristics to help operators' standard service requirements. Cloud providers will be responsible for deploying the basic operating systems (OSs), hypervisors, and probably orchestrators in the form of infrastructure as a service (IaaS).

The 5GC is not contained within one single cloud as conventional proprietary hardware for 4G EPC. In fact, it will distributed over multiple clouds and instantiated as VNFs whenever needed. To clarify, the 5GC has many SBA cores that can be launched at different data centers to process their own incoming traffic. From a physical point of view, the 5GC looks segregated, but all of the cores are logically and virtually interconnected. The same concept applies to internal components of any core that can also be distrbuted over multiple servers and data centers to utilize local physical computational resources. Moreover, any 5G core could be instantiated within any data center to calls or data downloads that do not require interaction outside that network cluster. In the last case, the user profiles can be downloaded from the operator's national cloud and the call detail record (CDR) would be updated and loaded back to central network storage. Therefore, the 5GC will operate as a core as a service that can be created and removed at various network sites subject to use cases.

Considering virtualization platforms, the hypervisor is basic cloud software that supports abstracting virtual computational resources. Although the majority of hypervisors were created as OSS, some of them are provided only through vendors. From a technical perspective, all types of hypervisors have the same basic features for generating virtual machines or virtual containers but with different configuration options and commands. In virtual machines (VMs), a virtualized entity software is installed on top of a Linux OS. Container type supports two types: Linux basic containers, where virtual application is installed on top of an OS, and process containers that include stateless functions in smaller containers without any OS. The latter have their own driving software engines that consume less computational resources. Although these software options are not directly exposed to network developers, they have significant impact on systemic performance, which reflects easily on 5GC due to the very intensive processing delays. For example, process containers are more resilient to accident failure than other virtualization options. However, they increase the complexity of platform and require a significant amount of work during deploying and updating. Overall, containers allow almost eight times the concentration compared with virtual machines, but the business model for container deployment is more complicated and requires detailed work to create networking in between. This confirms our previous statement in this chapter that operators in 5G era are getting more involved in software development than any previous generation.

The virtual networks connect the VMs internally and provides the flat public network to access the external world through hosting servers. Since hypervisors are installed on top of a Linux OS, there is a degradation in the capacity of the traffic transferred from VM to physical ports because of multilayers of software installations. This is a major challenge for any operator network that runs its supporting core functionalities in cloud. The Open vSwitch (OVS) is a multilayer virtual switch that connects virtual networks in hypervisor platforms to the underlying Linux layer. Since 5GC will run in VMs, it is necessary to understand what the different networking options to connect VMs to physical ports with higher-capacity transfers. At this point, we refer to different

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

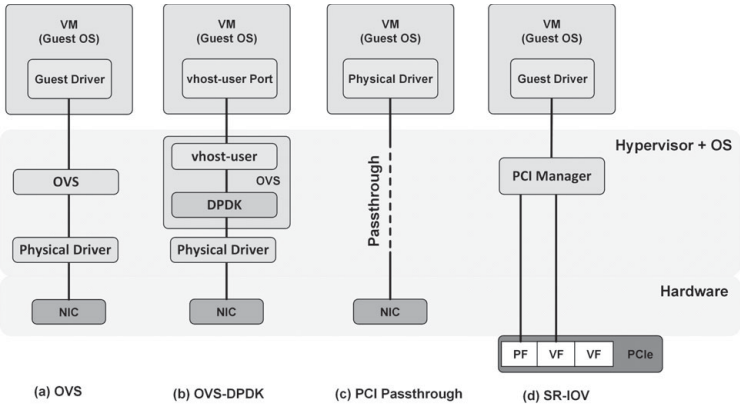


**Figure 1.5** Hypervisor and hardware input/output virtualization

types of connectivity between virtual networks and physical Peripheral Component Interconnects (PCIs), including network interface cards (NICs). The different options for virtual to physical connectivity are shown in Figure 1.5 and can be listed as follows:

- **OVS.** An independent project that provides open-source virtual switches in the form of bridges to connect hypervisor networks to OS layer, as shown in Figure 1.5(a). However, the OVS can easily become the bottleneck in high-traffic communications due to the design of flow lookup tables, which sequentially search each subtable until a match is found. Therefore, operators may choose alternative options to bypass OVS and connect their core VMs directly to NICs. Nevertheless, this does not eliminate OVS from consideration as it is still a key component for SDN controllers, as explained in later sections.
- **OVS-DPDK.** The Data Plane Development Kit (DPDK) is an open-source package that boosts packet processing performance and throughput through a set of data plane libraries and network interface controller drivers. An OVS-enabled DPDK creates a DPDK driver that is loaded to the NIC card to the bypass the OS and allows VMs to directly access NICs. The other side of the OVS has a vhost-user interface the connects to a VM installed vhost-user port, as shown in Figure 1.5(b). An OVS-DPDK allows higher capacity and better performance than a conventional OVS. However, it is necessary to highlight that not all drivers are supported by NICs, and procedures for configurations are quite different from one platform to the other. This raises many challenges when dealing with different vendor servers, which requires a lot of effort for developers.
- **PCI pass-through.** This feature allows full access and direct control of any PCI device from virtual machines provided that the correct driver was installed in VM, as shown in Figure 1.5(c). Pass-through methodology is simply bypassing the hypervisor and OVS for direct access to NIC from VM.
- **Pass-through with SR-IOV.** The single-root I/O virtualization (SR-IOV) requires a certain type of PCI express card (PCIe) that supports virtualization

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

feature. The PCIe can initiate a number of physical functions (PFs) with a similar number of virtual functions (VFs) depending on the number of actual ports on PCIe card, as shown in Figure 1.5(d). The hypervisor can then map virtual functions to VMs for direct device access with high default card performance and complete isolation from server environment.

It is important to understand that each one of the above connectivity options has its own configurations that differ according to hypervisor and OS types. For SBA, any additional time delays incorporated due to platform techniques will significantly impact the overall time latency for that slice service. Therefore, it is necessary to benchmark the performance of different connectivity options in the form of a separate cloud that runs 5GC prior to real deployment to life networks. This provides network developers the necessary information about the underlying platforms and the capacity obtained with each type of networking configurations. Moreover, any testing needs to emulate UDNs that employ a large number of users and application downloading to define the upper thresholds for individual servers. This will help to evaluate the overall number of computer nodes required for that data center or cloud.

## 1.3 Orchestrators and 5GC

The NFV orchestrator provides the necessary mechanism for onboarding new network services (NSs) and VNF packages. It also defines the various NS life cycle transitions through predefined policies for each application type. The orchestrator validates and authorizes various automated installations and resource requests, instructs the VNF manager (VNFM) to perform an installation, and stores VNFs in catalog. The NFV management and organization (MANO) is the basic model for an orchestrator. The MANO structure was defined by the European Telecommunications Standards Institute (ETSI) NFV model [24]. This standard defines the interfaces used to create VNFs and also to allocate resources using the virtualized infrastructure manager (VIM). However, orchestrators continue to evolve beyond the ETSI NFV model to automate large-sized clouds and support complete life cycle management. Recently, the Open Network Automation Platform (ONAP) emerged as a new initiative created by multiple vendors and operators to handle the full life cycle management of VNFs through SDN controllers and other active monitoring functionalities [25].

Operators can use orchestrators to create a suite of service by defining the business model that needs to run over the cloud. For example, orchestrating the VNFs for 5GC in one data center requires defining the overall architecture for the solution and the connectivity. The business model descriptors are represented by Topology and Orchestration Specification for Cloud Applications (TOSCA), which defines the basic building blocks: nodes and relationships. A node can be defined as an infrastructure or software component, such as a subnet, server, pod, or service or a run-time environment.The relationship describes how those nodes are connected with each other and the external world. The service providers will also need to provide policy files for life cycle

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

management and also a package of HEAT templates that orchestrate VNF creation through APIs. Both TOSCA and HEAT templates are written using YAML language to ease the development process. Moreover, the different 5GC components are provided into suitable software packages along with the descriptors to automate the installation and operation of 5GC VNFs. To this end, it appears that the main challenge here is not maintaining the actual 5GC functions but rather providing a suitable cloud environment that enables the 5GC VNFs to dynamically adjust to platforms and incoming traffic.

### 1.3.1 Boarding Virtual Network Functions

In UDNs, the dynamic changes in user numbers and downloaded applications require allocating sufficient computational resources to the core network. Virtualized 5GC in the form of VMs can be scaled in response to dynamic traffic changes at the RAN side. Since the traffic volumes are changing dynamically, 5GC VNFs are also instantiated, switched off, shelved, or terminated subsequent to any traffic changes. The orchestrator is the facilitator for all in-cloud dynamic operations. Specifically, the orchestrator provides the mechanism for enforcing operations, while the definition of those operations should be predefined by service providers. In reality, operators will be procuring 5GC VNFs from vendors who provide the necessary orchestration templates that comply with each operator platform. Therefore, there is a need to validate the VNF compatibility with the operator platform prior to instantiating the service. This process is called the *VNF validation and assurance*, as shown in Figure 1.6. This process happens first when the VNF is received from the vendor and is repeated when there is an upgrade or change in the software packages. Once the VNF is verified, it will be stored in the orchestrator catalog for subsequent installations. The operators may have their own additional validation procedures to secure their platforms and prevent cyberattacks or faulty situations where service may crash down during network life operations. The alarm for any crash is provided by service monitoring, which is discussed in the next section.

The various network life cycle management operations require certain time intervals for implementation. For example, instantiating new VNFs requirs time for launching the service to reconfigure networking to incorporate the newly created VNF with the relevant ports to handle the targeted traffic. Moreover, life ycle management operations are also requiring time to perform any state change transitions. There are also other time delays that may be incorporated during resource allocations and service assignment. All these operations are still software-tuned processes that requires a response from the underlying hardware servers. There is no current measurement of the response time required by virtualization layers to perform any of the aforementioned operations. This raises concerns that the platforms hosting 5G VNFs may impact the overall network E2E latency, causing SBAs to fail in meeting the performance requirements of connected services. Therefore, maintaining a suitable cloud environment is the real enabler for successful deployment of 5GC VNFs. Due to the absence of a reference cloud install that has been benchmarked for compatibility with 5G performance requirements, each operator continues to develop their own cloud setups, demanding VNF vendors to certify compatibility using that operator labs. The challenge is amplified when clouds

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

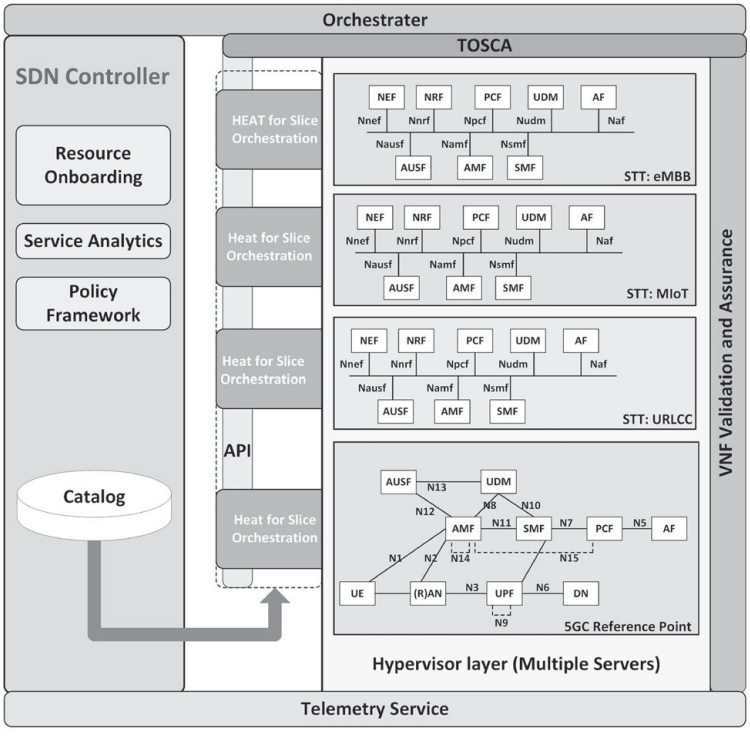


**Figure 1.6** Orchestrating 5GC VNFs using TOSCA and HEAT descriptors

are provided by third-party providers who may not be willing to permanently change their physical platforms to meet the need of temporary users who are leasing their services. This also raises concerns if 5G performance requirements can be met without comprehensive studies for cloud status and performance analysis.

A VNF resource should be optimally utilizing all allocated computational resources (e.g., cores, RAMs). However, VNF vendors need to be aware that overloaded VMs may also require additional resources for processing and to maintain stable operations. The API access from orchestrators and telemetry services may require a fractional millisecond when assessing VNF/VMN services. However, even a microsecond time interval process adds to the application processing delays that are accumulated as a flow passing through a platform E2E service. It is also understood that the plug-ins added to support third-party software may also incorporate some processing power, which is reflected in additional time delays. Therefore, developers need to examine every single scenario for software performance to make sure that 5GC VNFs will have the supporting resources for their operations. This also means that additional computational resources may be kept as a backup for the system during earlier real operations of 5G.

### 1.3.2 Orchestrator Layers

The software developers are always thinking about one orchestrator that manages a platform or cloud. Typically, this orchestrator will be managing the virtual domain

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

that resides on that cloud center. However, this concept does not align with 5GC deployments or distributed clouds. Considering 5GC, there are multiple logic networks that instantiate VNFs and networking subject to connected users. Each one of those logic networks has different performance requirements that need to be enforced through relevant policies. Therefore, each logic network will have its own orchestrator that manages the resource allocations and service delivery for that SBA. This means that we are looking to have at least four orchestrators to manage each service type for 5GC. However, all those logic networks are contained within the virtual layers, and there should be another orchestrator that manages their operations. This type – namely the "service orchestrator" – may also need to be incorporated within operator's cloudification framework. This higher-level orchestrator can be called the physical network orchestrator. In this way, we are looking at having multilayered orchestrators that can share the domain, but each handles different category of services.

From a technical perspective, multilayers of orchestrators require developing new interfaces that allow those orchestrators to connect to each other and enforce policies and authentications in a hierarchical order. Since each orchestrator handles a different type of service, it is not mandatory for multilayered orchestrator platforms to have the same service catalog or even VNF library. Therefore, vendor VMFs will be made available to the associated orchestration service and not necessarily to all orchestrators. To reduce complexity, the validation process for new VNFs can still be performed by the physical orchestrator, but certified VNFs may be forwarded to relevant orchestrator catalogs. Obviously, the number of deployed orchestrators may continue to increase with newly services, and this will also raise many challenges for the developers to maintain consistent and reliable services. Therefore, orchestrators will continue to evolve, considering technology, interfaces, and orchestration templates to provide more agile and large-scale services.

The concept of employing multiorchestrator layers with hierarchical order is necessary to manage distributed cloud environments. In Section 1.2, we clarified that different 5G cores may be instantiated and operated at various clouds over network sites. However, those clouds need to comply with a central management model to allow sharing computational resources and authenticating services. This adds an additional type of orchestrator identification to previously mentioned types of physical and logic. It is more likely that orchestrators may be defined by IDs that include their data center initials, type of network orchestrated, and initials to define whether the ID is to be permanent or temporary. To conclude, the current understanding of orchestrator functionality is still limited to one service type that operates in one cloud, and this type of thinking will definitely change over the course of 5G deployments.

## 1.4 SDN and Overlaid Networks

The SDN improves network adaptability and automates the management of large networks. It leverages OpenFlow as device-level abstraction to a programming model that edits traffic flows across the entire cloud [26]. Typically, the SDN decouples the control plane from the data plane, keeping the data plane on the switch while

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

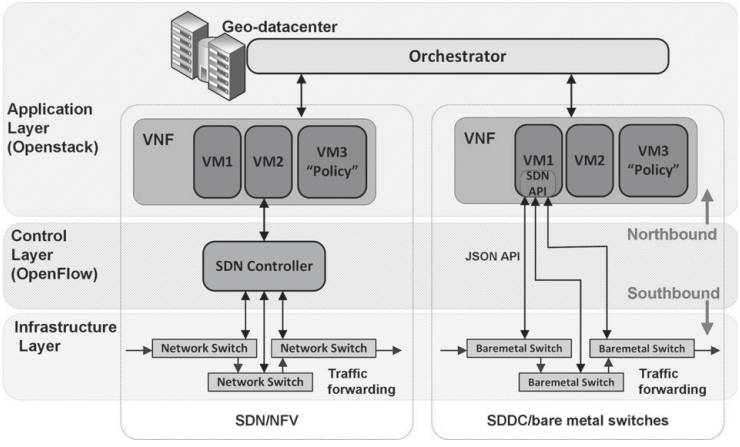


**Figure 1.7** Software-defined networking in cloud traffic managements

maintaining the control plane on the controller. This allows for traffic engineering, traffic steering, and enforcing security enforcement across platforms. In the cloud, it is easy to instantiate multiple SDN controllers that run in a hierarchical order to avoid overheads, potential bottlenecks, and potential single points of failure and to promote scalability and interoperability. The SDN controller edits the flow table on network switches according to the policies received from the orchestrator for new flow-forwarding procedures. The controller uses the network information to overlay a new virtual network on top of physical topology that reduces congestion or overloading at certain network sites. This distribution of traffic at the cloud is essential to improving performance for virtual 5GC.

Considering orchestrated hypervisor platforms, the SDN controller is using RESTful API to connect to OVS switches. This allows the SDN to set the flow entries for all packets at any OVS switch and define the networking to other OVSs on cloud. It creates multiple logical networks layered over the same physical plane by defining an L2/L3 network with no dependency on the physical topology. Therefore, the SDN detects the physical topology and maps the logical representations by installing the necessary flows on network elements (OpenFlow switches [OFSs]) [27].The SDN creates multiple virtual network abstractions that interconnect various network sites to edit traffic direction. This can be seen in Figure 1.7(a) where SDN connects OpenStack hypervisor OVS switches. The ONAP is relying on the OpenDayLight (ODL) project to provide the necessary SDN functionalities for orchestrating the necessary overlaid virtual networks [28].

The SDN controller can also be deployed in physical form to divert traffic on physical networks and forward the traffic to the VMs on the northbound, as shown in Figure 1.7(b). The physical SDN can also be chained with a virtual SDN for cascaded traffic forwarding across the cloud. This scenario is crucial for certain SBA applications that require extreme capacity, such as virtual reality (VR) in vehicle-to-everything

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

(V2X) communications. The enabler for such an approach is the NFV domain, which offers new features that were not seen before. However, achieving successful deployment requires creating large API databases at the hypervisors to allow universal calls of chosen API all across the cloud. It also requires SDN to utilize carefully the IP infrastructure for efficient packet forwarding over logic networks.

## 1.5 Monitoring Service and Platform

The monitoring of cloud platforms requires the development of alarm sensors that track the platform key performance indicators (KPIs) or even the incoming traffic flows. In the following subsections, we discuss the telemetry services and also application deployment options relative to 5GC.

### 1.5.1 Telemetry Services

The telemetry provides the networking infrastructure that connects the KPI alarming sensors on a platform with the hypervisor dashboard. However, the current telemetry services may not be enough considering operator needs. For example, there is a strong need to integrate the alarming KPIs on the hypervisor with the server's hardware alarming modules. Such integration allows real-time status monitoring of the cloud as a service and not just the upper virtual layer of it. Since telemetry tools are API and plug-in, the telemetry services need to span from the cloud center throughout access to RAN. This can be achieved easily with the current technologies, and it will be only a developer task to crate the necessary software modules to be installed on top of existing technologies. These plug-in interfaces can also help to monitor the multi-RAT interfaces that perform multigigabyte transmissions to instantiate the necessary 5GC VNFs that support service type upon RAT link creation. This procedure can be adopted as one of the requirements for service upgrades of current orchestrators to make them more operator related rather than IT. The virtualization technologies will continue to evolve and lean toward operators' needs at the 5G dawn.

### 1.5.2 Application Deployment Options

The 5GC applications may be deployed over the cloud in different formats subject to the platform configurations. The most popular option is to use VMs for instantiating an application over hypervisors. However, hypervisors are evolving, and new options of deployment are used from legacy software. The main motivation is to improve the resource allocations in clouds considering the number available (e.g., cores, RAMs). Most recently, hypervisors started to use containers for deploying applications in the cloud rather than VMs. This achieve almost eight-times the concentration of applications on top of the same server compared with VM installs. Also, there is a lot of attention nowadays to microservices that allow breaking entities/applications into small functions that are orchestrated as individual small stateless containers for higher

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

resiliency against failures and cyberattacks. These deployment options may not have a direct impact on an application running in virtual cores networks but will definitely impact the performance of supporting infrastructure and any cases of system crash. The main challenge is that those deployment options are for developers and will not necessarily be a matter of interest for the network designers. This breakdown between what a network engineer sees from an E2E level and cloud technologies provided by IT developers can only make the network more complicated and noncompliant with standard 5G performance requirements. It is more responsible to use test beds that are connected to large-sized networks as well as clusters of active networks to verify appropriate platform options prior to any 5GC deployments.

## 1.6 Conclusions

In this chapter, we reviewed the 5GC and various SBA slices in 5G networks. We have investigated the burden imposed by UDN on a core network and the strict requirements for different service types in 5G. We dived into the cloud configurations and various technologies for networking to review the traffic distribution in virtual core networks. The challenges of software and hardware were explained from real deployment perspective to show that E2E performance may be significantly impacted by different settings of underlying platforms and not just by the heavy loads of users and applications in 5G. The focus was to show different research aspects that need to be addressed in NFV/SDN to achieve a highly efficient cloud-based network.

## References

[1] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G backhaul challenges and emerging research directions: A survey," *IEEE Access.*, vol. 4, pp. 1743–1766, 2016.

[2] M. Koivisto, M. Costa, J. Werner, K. Heiska, J. Talvitie, K. Leppänen, V. Koivunen, and M. Valkama, "Joint device positioning and clock synchronization in 5G ultradense networks," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 5, pp. 2866–2881, May 2017.

[3] P. Kela, J. Turkka, and M. Costa, "Borderless mobility in 5G outdoor ultra-dense networks," *IEEE Access*, vol. 3, pp. 1462–1476, 2015.

[4] G. P. Koudouridis and P. Soldati, "Spectrum and network density management in 5G ultra-dense networks," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 30–37, Oct. 2017.

[5] N. Al-Falahy and O. Y. Alani, "Technologies for 5G networks: Challenges and opportunities," *IT Professional*, vol. 19, no. 1, pp. 12–20, Jan. 2017.

[6] H. Zhang, L. Song, Y. Li, and G. Y. Li, "Hypergraph theory: Applications in 5G heterogeneous ultra-dense networks," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 70–76, Dec. 2017.

[7] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving sustainable ultradense heterogeneous networks for 5G," *IEEE Comm. Mag.*, vol. 55, no. 12, pp. 84–90, Dec. 2017.

[8] H. Zhang, Y. Dong, J. Cheng, M. J. Hossain, and V. C. M. Leung, "Fronthauling for 5G lte-u ultra dense cloud small cell networks," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 48–53, Dec. 2016.

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

[9] C. Niu, Y. Li, R. Q. Hu, and F. Ye, "Fast and efficient radio resource allocation in dynamic ultra-dense heterogeneous networks," *IEEE Access*, vol. 5, pp. 1911–1924, 2017.

[10] Y. Wang, Z. Miao, and L. Jiao, "Safeguarding the ultra-dense networks with the aid of physical layer security: A review and a case study," *IEEE Access*, vol. 4, pp. 9082–9092, 2016.

[11] G. Yu, R. Liu, Q. Chen, and Z. Tang, "A hierarchical SDN architecture for ultra-dense millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 79–85, June 2018.

[12] K. S. Munasinghe, I. Elgendi, A. Jamalipour, and D. Sharma, "Traffic offloading 3-tiered SDN architecture for densenets," *IEEE Netw.*, vol. 31, no. 3, pp. 56–62, May 2017.

[13] M. Fiorani, A. Rostami, L. Wosinska, and P. Monti, "Abstraction models for optical 5G transport networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 8, no. 9, pp. 656–665, Sept. 2016.

[14] X. Zhang, M. Yang, Y. Zhao, J. Zhang, and J. Ge, "An SDN-based video multicast orchestration scheme for 5G ultra-dense networks," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 77–83, Dec. 2017.

[15] L. Zhou, J. Zhang, B. Seet, L. Zuo, X. Hu, J. Li, S. Wang, and J. Wei, "Software defined small cell networking under dynamic traffic patterns," in *2016 IEEE 14th Intl. Conf. on Dependable, Autonomic and Secure Computing, 14th Intl. Conf. on Pervasive Intelligence and Computing, 2nd Intl. Conf. on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2016, pp. 10–17.

[16] S. Vural, N. Wang, P. Bucknell, G. Foster, R. Tafazolli, and J. Muller, "Dynamic preamble subset allocation for ran slicing in 5g networks," *IEEE Access*, vol. 6, pp. 13 015–13032, 2018.

[17] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking* (MobiCom '17). New York, NY: ACM, 2017, pp. 127–140. [Online]. Available: http://doi.acm.org/10.1145/3117811.3117831

[18] 3GPP, "5G: System architecture for the 5G system," 3rd Generation Partnership Project (3GPP), Technical Specification (TS)123 501, 06 2018, version 15.2.0 Release 15. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144

[19] C. Zhang, X. Wen, L. Wang, Z. Lu, and L. Ma, "Performance evaluation of candidate protocol stack for service-based interfaces in 5g core network," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[20] IEEE1932.1 WG, "Standard for licensed/unlicensed spectrum interoperability in wireless mobile networks," 2018). [Online]. Available: http://sites.ieee.org/sagroups1932-1

[21] S. Al-Rubaye, A. Al-Dulaimi, J. Cosmas, and A. Anpalagan, "Call admission control for non-standalone 5g ultra-dense networks," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 1058–1061, May 2018.

[22] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "RFC 2616, Hypertext Transfer Protocol–HTTP/1.1," 1999. [Online]. Available: http://www.rfc.net/rfc2616.html

[23] Google, "Data center innovation: Google's data centers and global network provide low latency, low cost, and high availability," (2018). [Online]. Available: https://cloud.google.com/about/data-centers/

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press

[24] 3GPP, "ETSI GS NFV," 3rd Generation Partnership Project (3GPP), MAN 001,12 2014, version 1.1.1. [Online]. Available: www.etsi.org/technologiesclusters/technologies/nfv

[25] F. Slim, F. Guillemin, A. Gravey, and Y. Hadjadj-Aoul, "Towards a dynamic adaptive placement of virtual network functions under ONAP," in 2017 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), Nov. 2017, pp. 210–215.

[26] T. Dargahi, A. Caponi, M. Ambrosin, G. Bianchi, and M. Conti, "A survey on the security of stateful SDN data planes," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1701–1725, third quarter 2017.

[27] ONF, "OpenFlow switch specification," Open Networking Foundation, ONFTS-015, 09 2013, version 1.3.3 (Protocol version 0x04). [Online]. Available: www.opennetworking.org/wp-content/uploads/2014/10/openflowspec-v1.3.3.pdf

[28] OpenDayLight, "Platform overview," (2018). [Online]. Available: www.opendaylight.org/what-we-do/odl-platform-overview

https://doi.org/10.1017/9781108671323.001 Published online by Cambridge University Press