ACCEPTED FROM OPEN CALL

Network Function Virtualization in 5G

Sherif Abdelwahab, Bechir Hamdaoui, Mohsen Guizani, and Taieb Znati

5G wireless technology is paving the way to revolutionize future ubiquitous and pervasive networking, wireless applications, and user quality of experience. To realize its potential, 5G must provide considerably higher network capacity, enable massive device connectivity, with reduced latency and cost, and achieve considerable energy savings compared to existing wireless technologies.

ABSTRACT

5G wireless technology is paving the way to revolutionize future ubiquitous and pervasive networking, wireless applications, and user quality of experience. To realize its potential, 5G must provide considerably higher network capacity, enable massive device connectivity with reduced latency and cost, and achieve considerable energy savings compared to existing wireless technologies. The main objective of this article is to explore the potential of NFV in enhancing 5G radio access networks' functional, architectural, and commercial viability, including increased automation, operational agility, and reduced capital expenditure. The ETSI NFV Industry Specification Group has recently published drafts focused on standardization and implementation of NFV. Harnessing the potential of 5G and network functions virtualization, we discuss how NFV can address critical 5G design challenges through service abstraction and virtualized computing, storage, and network resources. We describe NFV implementation with network overlay and SDN technologies. In our discussion, we cover the first steps in understanding the role of NFV in implementing CoMP, D2D communication, and ultra densified networks.

INTRODUCTION

In the last decade, wireless technology has emerged as one of the most significant trends in networking. Recent statistics show that mobile wireless broadband penetration has exceeded that of fixed wireline broadband networks. In addition to general broadband access, recent advances in wireless communications and node processing capabilities have made it possible for communication networks to provide support for a wide variety of new multimedia applications and compelling wireless services, which are rapidly and steadily becoming national priorities. This trend is expected to continue in the future at much faster growth rates. By 2018, the global mobile traffic will increase from 2.6 to 15.8 exabytes. Addressing the expected exponential growth of rich media underscores the need to evolve cellular networks. To this end, the fifth generation (5G) will support 1000 times the current aggregate data rate and 100 times the user data rate, while enabling a 100 times increase in the number of currently connected devices, 5 times decrease of end-to-end latency, and 10 times increase of battery lifetime [1].

To meet the expected three-orders-of-magnitude capacity improvement and massive device connectivity, 5G centers its design objectives around efficiency, scalability, and versatility. To sustain its commercial viability, 5G networks must be significantly efficient in terms of energy, resource management, and cost per bit. Connecting a massive number of terminals and battery operated devices necessitates the development of scalable and versatile network functions that cope with a wider range of service requirements including: low power, low-data-rate machine-type communication, high data rate multimedia, and delay-sensitive applications, among many other services. The efficiency, scalability, and versatility objectives of 5G direct the 5G community toward finding innovative but simple implementations of 5G network functions.

5G network functions face critical functional and architectural challenges in spite of their performance superiority. Coordinated multipoint (CoMP), for instance, can improve the cell edge user experience by using coordinated and combined transmission of signals from multiple antennas, cells, terminals, and sites to improve the downlink (DL) and uplink (UL) performance (e.g., by coordinated scheduling, coordinated beamforming, and interference alignment). However, CoMP achieves this gain with increased computation, increased signaling overhead, and increased backhauling and equipment cost. Moreover, the massive number of devices requires ultra densified networks, specialized hardware, and device-centric architecture that are not well defined yet. Finally, 5G must coexist with legacy technologies like 2G, 3G, and 4G. This requirement alone increases cost and complexity indefinitely. These challenges can be effectively addressed by implementing the 5G network functions as software components using the network functions virtualization (NFV) paradigm.

A growing group of companies and standardization bodies are pushing research and development of the NFV paradigm to improve cost efficiency, flexibility, and performance guarantees of cellular networks in general.[1] In NFV, vendors implement network functions in software components called virtual network functions (VNFs). VNFs are deployed on high-volume servers or cloud infrastructure instead of specialized hardware. For example, NFV pools the signal processing resources in cloud infrastructure rather than using dedicated baseband processing units (BBUs) at every site. Such resource pooling

[1] https://portal.etsi.org/TBSiteMap/NFV/NFVMembership.aspx

Sherif Abdelwahab and Bechir Hamdaoui are with Oregon State University; Taieb Znati is with the University of Pittsburth/ Mohsen Guizani is the University of Idaho.

0163-6804/16/$25.00 © 2016 IEEE

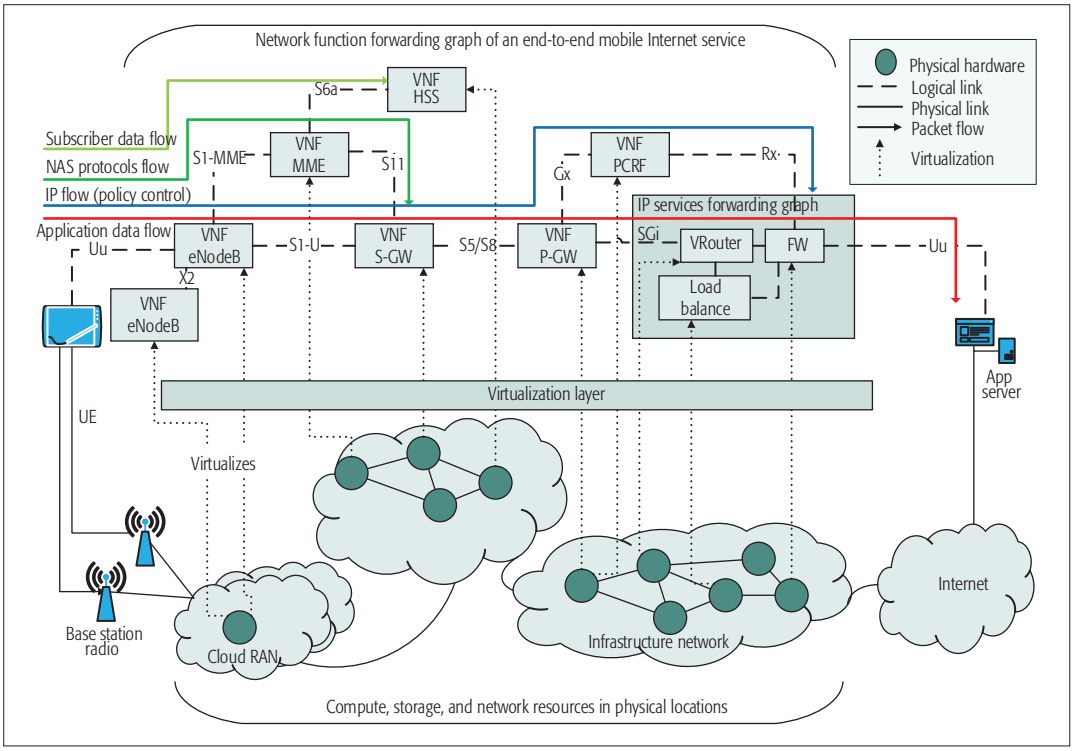

Figure 1. Virtualization of a forwarding graph implementing mobile Internet service.

Mobility management and NAS protocols flow through different network functions for mobility management, authentication, and policy enforcement. Unlike the current cellular networks where a particular feature is activated network wide, forwarding graphs enable 5G operators to activate features per service.

reduces computational and signaling overhead, optimizes cost, and improves flexibility so that a service provider activates a particular signal processing resource for only specific terminals in the whole network instead of activating all processing resources unnecessarily at each site.

Generally, NFV can overcome some challenges of 5G by:

- Optimizing resource provisioning of the VNFs for cost and energy efficiency
- Mobilizing and scaling VNFs from one hardware resource to the other
- Ensuring performance guarantees of VNFs operations, including maximum failure rate, maximum latency, and tolerable unplanned packet loss
- Ensuring coexistence of VNFs with non-virtualized network functions [2]

Unlike other work on application of NFV and software defined networking (SDN) technologies in generic 5G networking, virtualized Long Term Evolution (LTE) evolved packet core, and software defined radio (SDR)-based sites [3–6], this work focuses on the implementation of an NFV framework that meets 5G radio access network (RAN) technology requirements and enables several complex 5G functions while smoothing its coexistence with other technologies. We also demonstrate the effectiveness of NFV in reducing the capital expenditures (CAPEX) and operational expenditures (OPEX) of the 5G RAN.

In this article, we first survey service abstraction, architecture of NFV, and network virtualization via the network overlay model. As NFV enabling technologies, we describe how to use SDN and OpenFlow to virtualize and interconnect VNFs. Second, we focus on 5G virtualizable radio functions and describe CoMP, inter-cell device-to-device (D2D), and ultra densified network implementation using NFV. Finally, we discuss open research problems specific to NFV in 5G RAN.

NFV and Network Overlay

With NFV, services are described as a forwarding graph of connected network functions. A forwarding graph defines the sequence of network functions that process different end-to-end flows in the network. For example, Fig. 1 shows a simplified forwarding graph of a mobile Internet service where data flows traverse network functions from the evolved NodeB (eNodeb) to the service gateway (seGW) to the IP backbone until it reaches the application server. Mobility management and non-access stratum (NAS) protocols flow through different network functions for mobility management, authentication, and policy enforcement. Unlike current cellular networks, where a particular feature is activated network-wide, forwarding graphs enable 5G operators to activate features per service (e.g., CoMP becomes active only for predefined service classes). The network functions are virtualized using a separate virtualization layer that decouples service design from service implementation while improving efficiency, resiliency, agility, and flexibility. Network functions that can be virtualized in general include:

- Evolved packet core functions such as the mobility management entity, serving gateway, and packet data network gateway
- Baseband processing units functions, including medium access control (MAC), radio link control (RLC), and radio resource control (RRC) procedures [7]
- Switching function
- Traffic load balancing
- Operation service centers

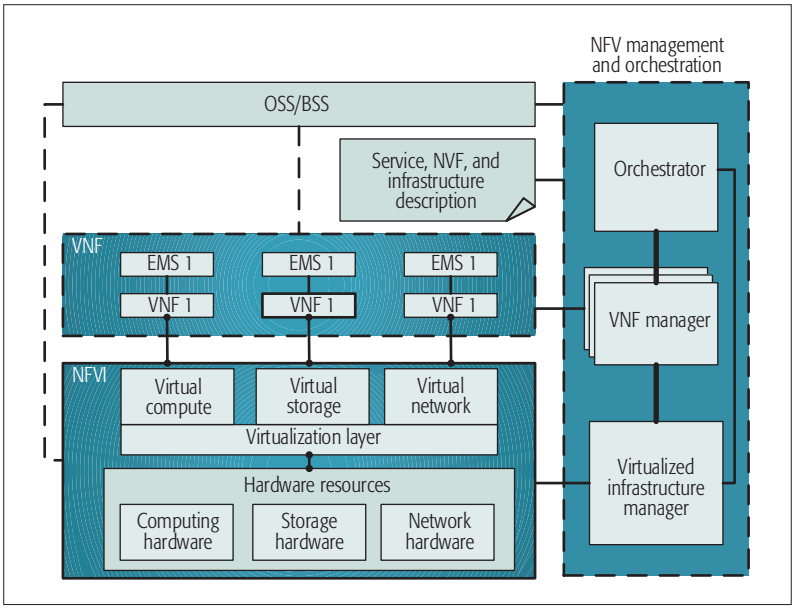

Figure 2. The network function virtualization reference architecture.

The NFV reference architecture (Fig. 2) supports a wide range of services described as forwarding graphs by orchestrating the VNF deployment and operation across diverse computing, storage, and networking resources [2]. As shown in Fig. 2, the computing and storage hardware resources are commonly pooled and interconnected by networking resources. Other network resources interconnect the VNFs with external networks and non-virtualized functions, enabling the integration of existing technologies with virtualized 5G network functions. NFV management and orchestration comprises resource provisioning modules that achieve the promised benefits of NFV.

The VNF managers(s) (Fig. 2) perform two main functions: operation and resource provisioning. VNF operation consists of infrastructure management, fault management, performance management, and capacity planning and optimization. Resource provisioning ensures optimal resource allocation (e.g., allocate virtual machines, VMs, to servers), optimal connectivity between VNFs, energy conservation, and resource reclamation. Moreover, resource managers discover computing, storage, and network resources in the infrastructure. Efficient design of a VNF manager leverages the peak benefits of NFV to reduce CAPEX and OPEX in 5G by means of dynamic resource allocation, traffic load balancing, and easier operation and maintenance [8].

In the rest of this section, we detail the NFV design trade-offs and the main networking problems associated with them. Then we introduce the network overlay concept as a solution to these problems.

Networking Problems in NFV

NFV faces several networking problems; some are inherited from multi-tenant data center networking, while others are specific to NFV. Designing NFV platforms for carrier-grade availability that exceeds five nines requires fail-over times between redundant 5G VNFs of less than a second. Also, almost all cellular services are dynamic in nature, and the physical resources must expand and shrink as service demand changes (elasticity). Cellular traffic has regular daily and weekly patterns, but also changes spatially in case of special events (e.g., football matches), so resources must be assigned optimally to cope with these changes. VM mobility is one technology that can support these rapid traffic changes, but it comes with *networking* design challenges. First, migrating VMs from one server to another must retain VMs' network states, including at least physical location, and IP and MAC addresses. Second, as a VM implements 5G radio functions, it must have access to devices' data, radio states, and channel information, and it becomes critical that VM migration solutions provide real-time capabilities of distributed state management through localized caching and acceleration agents. Third, from an operational efficiency viewpoint, resource utilization must be kept as high as possible to ensure profitability. An optimal NFV system design incorporates efficient and flexible allocation of resources and optimal forwarding of traffic by which an operator can realize and mobilize virtual networks of VNFs on any hardware across the infrastructure.

The flexibility of NFV is also associated with overhead. If we place multiple VNFs on the same physical server, the server will not have a single address but many. The switching network will have to learn addresses of individual VMs, and we can witness an uncontrolled increase in forwarding table sizes. Additionally, if an infrastructure is shared between multiple service providers, VNFs address separation becomes a must as we need to perceive the address use flexibility of a single provider, while the address space may overlap between providers. Specifically, as traffic from different providers share the same networking resources, not only security becomes challenging, but also flexibility and optimal forwarding of traffic from one virtual network (network of VNFs) to the other without compromising security and address separation. Additionally, NFV shall maintain the scalability characteristics of the current highly distributed cellular networks while exploiting the discussed benefits of NFV; hence, features such as load balancing and VM placement in the cloud environment shall become real-time aware and support thousands of back-end cellular virtual functions. We discuss the network overlay concept as a typical solution to the networking problems in such a virtualized environment.

Network Overlay

Network overlay is an approach to address NFV networking problems by implementing virtual networks of VNFs as overlays. The first-hop network device connected to a VNF, called the network virtualization edge (NVE), encapsulates the original packets from the VNF and identifies the destination NVE that will decapsulate the packet before delivering it to the next VNF. The network forwards the packet based on the encapsulation header oblivious of the packet payload. The NVE is basically a physical switch, router, or a virtual switch in a network hypervisor.

Network overlay enjoys several appealing

characteristics. A key feature of network overlay is the decoupling of the VNF addresses from the physical network addresses, and isolation of traffic from multiple virtual networks. The traffic isolation is achieved by the fact that forwarding traffic between virtual networks requires a gateway entity to forward such traffic. If this gateway is missing, forwarding traffic between virtual networks is not possible. With such a feature, the overlay provides both traffic isolation and flexibility to forward traffic between virtual networks (with adequate gateways).

Moreover, overlay works well in environments that are highly distributed, which involves thousands of VNFs. The expected number of NVEs required to implement a virtual network is generally low, which is important for scalability, while these NVEs provide the needed flexibility to mobilize VNFs with highly dynamic traffic. In principle, migrating a VNF implies quick reconfiguration of a single NVE to maintain routing flows from that VNF.

Looking at its drawbacks, network overlay generally requires changes, possibly using existing encapsulation or tunneling protocols in order to support packet (en/de)capsulation. For example, the Generic Routing Encapsulation protocol (RFC 2784) can be used to encapsulate — in principle — any arbitrary protocol over IP and to create any virtual layer 2 network on top of a physical layer 3 network.

SDN is another approach that simplifies network overlay implementation. The idea is to program switches at the NVE to modify packet headers from different NFV flows according to a global mapping of virtual network addresses (e.g. MAC and IP addresses) to physical network addresses. This can be done without changes to the data plane protocols. A central SDN controller maintains global mapping of virtual/physical network addresses and install rules in switches to implement this mapping. We overview SDN via OpenFlow first and give more details on network virtualization using SDN in the next section. After that, we provide specific use cases of SDN in virtualization of 5G RAN functions.

Virtual Network Functions Overlay via SDN

SDN adopts two main ideas: *logically* centralized control of the data plane, and network state management across distributed controllers. Separating the control and data planes accommodates increasing traffic volumes and improves network reliability, predictability, and performance. Such separation allows a controller to deploy forwarding table entries in data plane programmable switches (or routers) and frees switches from performing control functions.

The controlling function does not need to be centralized in principle, but logically centralized. How distributed controllers manage their states to improve performance, reliability, and scalability is a challenging problem. Support from an underlying SDN platform is required from one side to achieve distributed state management. This platform incorporates sophisticated algorithmic and protocol solutions for optimized network control and state management [9].

OpenFlow [10] is a standardized protocol for programming the data plane using control plane application programming interfaces (APIs). Openflow programs the forwarding behavior of the traffic flows in switches based on different packet header fields, which are specified in flow table rows, matching. An OpenFlow switch matches protocol header fields (e.g., ports, MAC, and IP) in an incoming packet, and performs actions against matched packets. A router matches the specified header fields and either floods, forwards the packet on a predefined port, or drops the packet. The router is also capable of rewriting header fields before forwarding the packet.

OpenFlow made the idea of a network operating system possible. A network operating system is software that controls the behavior and state of the network through:

- Data plane forwarding rules programming
- Network state management
- Network behavior control

Network state management is challenging in distributed SDN controllers to maintain network state at different controllers. The open network operating system (ONOS) is an example of a distributed controller [11] that maintains consistent shared network state information across all controllers represented by a graph database. For fast read/write of network states, it maintains the network data in low-latency, distributed key-value storage along with in-memory topology information cache. The question now is why SDN and OpenFlow are particularly important for NFV.

OpenFlow and NFV

NFV does not necessarily require SDN and OpenFlow. However, NFV and SDN are related in many ways. First, SDN is an enabling technology to NFV, where it can simplify the implementation of the network overlay model. Second, virtualizing network functions like routers and switches is complicated with conventional networking technologies, while SDN provides a natural solution. Imagine the complexity of a router that is running several virtual routers, each implementing its own control plane. Third, SDN flexibly allocates pooled computing resources to a particular VNF, elastically manages these resource allocations according to traffic demands, and easily mobilizes VNFs with quick modification to NVE rules. In this subsection, we discuss the first two possibilities and leave the third one to the next section.

Unlike adding an encapsulation layer to implement network overlay, an SDN controller just rewrites packets' addresses to implement overlays.[2] This idea does not require changing the data plane at all and still leverages the same benefits of separating virtual networks' address spaces. A controller maintains mapping between virtual networks and physical networks including routes through which traffic of a virtual network traverse. The controller installs a flow in the OpenFlow switch's (NVE switch at the edge) flow table with an action to rewrite a matched source and destination IP/MAC address of a packet from a VNF to addresses in the physical network. The controller also installs rules in the OpenFlow switches in the network to implement

Unlike adding an encapsulation layer to implement network overlay, an SDN controller just rewrites packets' addresses to implement overlays. This idea does not require changing the data plane at all and still leverages the same benefits of separating virtual networks' address spaces.

[2] http://ovx.onlab.us/

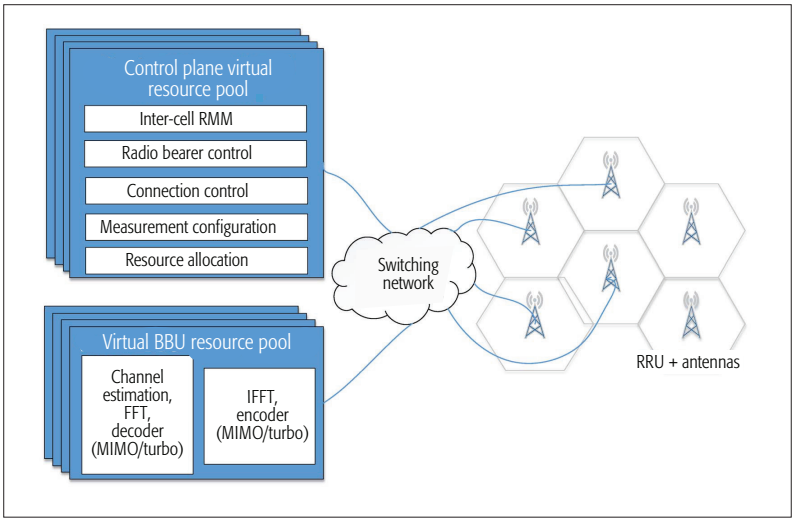

Figure 3. Common RAN network functions in 3GPP control and user planes.

a particular route between two chained VNFs. In this process, the controller is not aware of every single packet rewriting event, but just installs the flows in the switches that optimally implement a particular network overlay.

A rigorous method of traffic isolation between virtual networks with SDN-based virtualization is to define multiple physical IP addresses ranges for the same physical network. Packet addresses from one virtual network are translated to a particular physical IP addresses range, while packet addresses from another virtual network are translated to another physical IP address range. This separation allows flexible isolation of traffic between virtual networks as flows from one virtual network can be controlled to follow a disjoint route from another virtual network's flows. The main drawback of this approach is the increased IP address space that is needed in the physical network, which is not necessarily required in the encapsulation approach. Nevertheless, rigid traffic separation is of paramount importance when the infrastructure is shared between multiple service providers.

The second flexibility of the SDN approach is the independent networking behavior design of different virtual networks of VNFs. Even if network virtualization is not implemented via SDN, a separate SDN controller can control each virtual network behavior independent from other virtual networks. The network behavior not only includes how traffic flows are routed, but also how individual VNFs process traffic (control plane) flows (e.g., firewall, load balancing, deep packet inspection). This discussion reveals that SDN is a natural choice for implementing 5G VNFs. Using OpenFlow for SDN or not is another arguable choice due to some limitations in the OpenFlow standard that we discuss later.

NFV and its implementation using SDN can be applied to legacy cellular network functions, virtualization of data centers networks, infrastructure as a service in cloud computing, and so on. What are the network functions that shall be virtualized in 5G RAN?, How does the third advantage of SDN mentioned at the beginning of this section benefit 5G related technologies? How do NFV and SDN meet 5G architectural and functional challenges? We try to give an answer to these questions by discussing current and forthcoming research activities that leverage the benefits of NFV and SDN towards an advanced but yet simpler 5G network.

Virtualization of 5G RAN

Several control and user plane network functions in 3GPP RANs are candidates for virtualization. Figure 3 shows typical 3GPP network functions, which will also be in 5G, that are virtualizable in principle. Virtualizing these functions lowers footprint and energy consumption through dynamic infrastructure resource allocation and traffic balancing. It also eases network management and operations, and enables innovative service offerings. We study potential CAPEX and OPEX savings to be incurred from virtualizing BBUs in a typical cellular network.

CAPEX and OPEX in NFV

Consider a scenario in which a VNF implements baseband processing in virtual BBUs, as illustrated in Fig. 3. This scenario is known as Cloud-RAN [7], where NFV provides the needed orchestration layer for Cloud-RAN to virtualize layers 2 and 3 of the radio interface. and the necessary framework to incorporate specialized hardware and accelerators for baseband processing. The virtualized infrastructure manager deploys a pool of virtual BBUs near the network edge infrastructure. The cell site in this scenario simplifies to antennas, remote radio units (RRUs), and switching functions. The switching functions interconnect the virtual BBU pool to the RRUs via optical links and a high-speed OpenFlow switch to meet strict latency requirements [7, 12]. Every virtual BBU has exactly the same processing capability as the non-virtual BBUs being deployed in every site. According to traffic demand, the VNF Manager allocates particular slices of BBUs' VNFs to active cell sites. For this allocation, the VNF Manager programs an overlay virtual network to switch physical layer flows to/from the RRUs connected to the site and from/to the RRUs to the allocated VM hosting the BBU VNF for processing. We study the impact of VNF on CAPEX by comparing the total number of needed BBUs in virtualized and non-virtualized deployments given the same maximum traffic. We also study the impact of NFV on OPEX by showing the average number of active BBUs in both cases.

We consider the real traffic mixture of a cellular network.[3] The network consists of 85 cells, and the traffic traces were collected for a period of six hours. A speech call in these traces requires one processing unit per second, and a packet session requires two processing units per second. This assumption is quite realistic and follows dimensioning rules of major hardware vendors. A single BBU capacity, whether virtualized or not, ranges from 64 to 256 processing units. We assume that a BBU is active if at least one processing unit is active, and when the BBU is idle it consumes no energy.

Figure 4 shows the total number of required BBUs in virtualized and non-virtualized scenarios. As the maximum capacity of a single BBU increases, the total number of the required BBUs

[3] The data source is anonymized as per the providing operator request.

decreases significantly with VNFs to reach 25 percent if a single BBU supports 256 processing units (typically found in major vendors). The saving is attributed to two facts. First, with NFV a single virtual BBU can serve traffic from multiple cell sites by ideal traffic allocation to pooled virtual BBUs instead of a specific BBU. Second, the total number of required virtual BBUs depends on the maximum of the aggregate traffic of the network, unlike the non-virtualized case where it depends on the maximum traffic of each individual cell. Since the maximum traffic of each cell occurs at a time interval that varies from one cell to the other, the maximum aggregate traffic of the network becomes significantly less than the sum of maximum traffic of all cells. The savings in total number of required BBUs translates directly to CAPEX savings.

OPEX saving in this study can be observed from the average number of active BBUs shown in Fig. 5. The fewer the active BBUs, the lower the aggregate energy consumption of the whole system (contributed only by BBUs). In the proposed NFV architecture, we allocate traffic from any cell site to an already active virtual BBU first with sufficient utilization before activating another virtual BBU. At any point in time, a virtual BBU becomes active only if the current aggregate network traffic cannot be served by the already active BBUs. By this approach, we can observe around 30 percent savings comparing current non-virtualized architecture and VNF. The saving reaches up to 55 percent with increasing the maximum BBU capacity to 256. The saving in CAPEX and OPEX is clear from this study on a small-sized network. We can anticipate more significant impact on networks with thousands of cells and heavier traffic. But the benefit of NFV is not only expenditures savings, but also flexibility in implementing 5G functions.

NFV for CoMP and D2D

NFV and SDN can be viewed as enabling implementations of advanced 5G technologies such as CoMP and D2D communication. Figure 6 illustrates this architecture. The VNF Manager, embodying the OpenFlow controller, easily and effectively realizes DL CoMP, UL CoMP, and high-speed inter-cell D2D connectivity by installing the flows shown in the flow table in Fig. 6 in the switch.

DL CoMP requires all BBUs from multiple 5G cell sites to communicate while delivering parallel terminal data from one to all involved cell sites. Similar communication is required in UL CoMP in the reverse direction from multiple cell sites to a single BBU. Additionally, two terminals communicating in inter-cell D2D require BBUs of the cells to communicate directly and to handle high-speed low-latency traffic. That type of D2D communication required exploiting the mobile backhaul network in legacy architectures to route traffic through the core network.

The NFV/SDN approach in Fig. 6 instantiates DL CoMP in which terminal data from BBU-1 are forwarded to two different sites. A flow modification message installs an OpenFlow flow that matches traffic from input port 1, and takes two parallel actions to output flow packets to output ports a and c. This realizes both DL CoMP from two cell sites to a single terminal at aggregate rate and forwards the same aggregate message to multiple terminals at user data rate. A two-match single-action flow entry realizes UL CoMP similarly. Input flow matched on ports b and d are forwarded in a single action to output port 4.

The OpenFlow controller implements D2D communication in the inter-cell scenario by establishing high-speed low-latency connection of different BBUs. At the same time, another high-speed low-latency connection is established between the correspondent cells. This is illustrated by the two multiple-match multiple-action flows in Fig. 6. Multiple matches and multiple

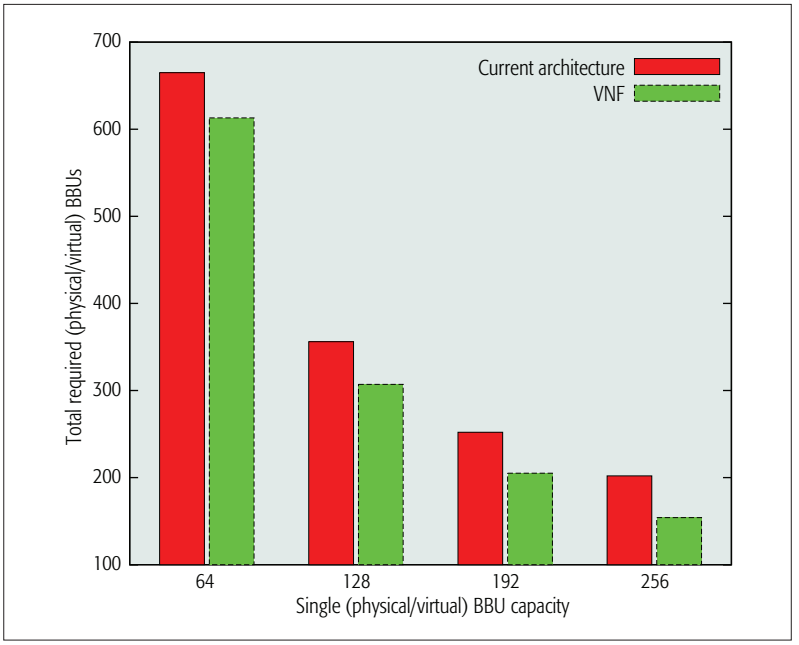

Figure 4. Up to 25 percent saving in total required BBUs, comparing current (non-virtualized) architecture and VNF.

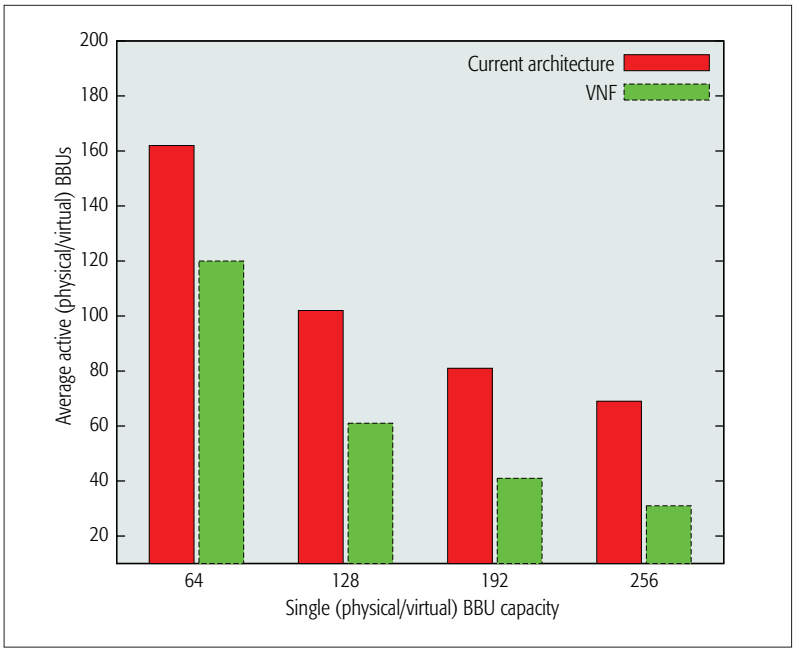

Figure 5. Up to 55 percent saving in active BBUs, comparing current (non-virtualized) architecture and VNF.

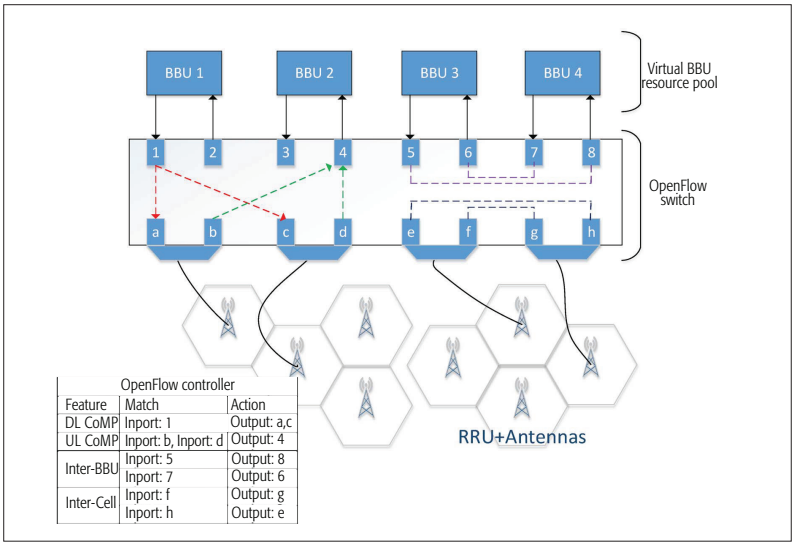

Figure 6. NFV/SDN enabling implementation of DL/UL CoMP and inter-cell D2D communication [7].

actions are needed in this case as both UL and DL traffic are involved in the connection. We could also use four parallel single-match single-action rules in a less optimized flow table size. In all these scenarios, the NFV manager keeps track of active flows' rules and BBU allocation.

Evolving Densification with NFV

Another 5G technology where NFV and SDN are of great benefit is ultra densified networks. 4G network design was based on the assumption of sparse deployments where cell sites make nearly autonomous radio resource management decisions. This is not the case in ultra densified networks. The terminal connects to the network through a cluster of closest cells, which cooperatively minimize the impact of interference from neighbor clusters to which the terminal is not connected [13]. The terminal will also exhibit rapid handover decisions, adding and removing cells from its cluster. The solution to this is to logically centralize the radio resource management decision like legacy 3G and 2G networks. However, unlike 2G and 3G, we are challenged by scalability problems, which prevents providing a commercially viable centralized controller that manages resources in chaotically deployed massive numbers of cell sites.

NFV can provide a solution to scalability issues by deploying all control decisions that mainly require cooperation of a large number of cells in VNFs near the network core and rapid decisions in NFVs near the network edge. Handovers, transmit power allocation, and cluster selection are control decisions that must be made cooperatively as they impact inter-cell interference. Alternatively, control decisions as radio resource allocation are done near the network edge as a decision must be available as frequent as every transmission time interval (TTI) [14].

In addition to the optimized deployment of VNFs, the logical centralization enables advanced algorithms to have access to an accurate and updated view of network status, interference maps, flow parameters, and operator preferences. Mobility management functions can base their decisions on network statuses beyond local radio quality at the cell site (e.g., energy, traffic, and interference awareness), while still providing minimal service interruptions during handovers. For example, operators can implement efficient VNFs that offload user traffic at the network edge, and load balance traffic at the core. And small cells clustering can be done more efficiently with network-supported decisions rather than terminal-based decisions.

Open Problems

The previous discussion envisioned several research problems to efficiently employ NFV in 5G RANs. RANs rely heavily on digital signal processors in the base station hardware to meet strict real time requirements. Virtualized SDR technology can virtualize BBUs, and generally requires support of real-time constraint processing in both VMs and the interconnecting networks. The CoMP example presented earlier [12] uses fiber communication to ensure meeting time constraints of the BBUs. However, OpenFlow does not provide native support of time-critical packet switching and leaves this task to controllers. Performance of virtualized SDR-based BBUs interconnected to RRUs through OpenFlow switches is unexplored.

OpenFlow is currently limited by the lack of programmable data plane support across different network stacks, by which packet payload can be inspected, modified, or reassembled. The work of Bansal *et al.* in [15] is an example approach that addresses data plane programmability across the wireless stack by decomposing the data plane into two main components, processing and decision. The processing plane includes data stream processing operation (e.g., signal processing), and the decision plane includes rules that define the sequence of processing operations required to process the data stream.

Moreover, the programmable control plane is currently limited in available solutions (e.g., OpenFlow) as it supports limited protocol spectrum to suit all needs of 5G protocols. Non-access stratum protocols, RRC protocols, and packet data conversion protocols are examples of protocols above layer 3 that require OpenFlow modifications to match their header fields and specify relevant actions to interconnect VNFs in RANs.

Computing resource allocation is also challenging with strict real-time requirements and dynamic allocation according to network traffic demands, service descriptions, and operator cost constraints. One particular challenge previously discussed is where to place the VNF pool initially; that is, near the edge or near the core of the network. Although this split is somewhat intuitive — deploy VNFs with real-time constraints near the edge and those with coordination requirements near the core — the deployment scenario where both requirements are present is still unstudied.

Support of deployability and interoperability with legacy and non-virtualized network functions is not investigated yet as the NFV is far from maturity. Possible solutions include integration of special-purpose hardware in data centers such as digital signal processing and graphics processing

units, optimized placement of VNFs in proximity to non-virtualized functions to avoid performance degradation during interworking procedures, and extension of I/O virtualization beyond Ethernet network interfaces to include other legacy interfaces such as time-division multiplexing transport interfaces, specialized acceleration units (e.g., crypto hardware accelerators), and SoCs. Performance evaluation of early proof-of-concept deployments along with legacy technologies shall enforce policy and research directions in developing open and standardized protocols, programming interfaces, infrastructure federation, and orchestration algorithms. The orchestration algorithms in particular shall not orchestrate virtualized resources only but also manage dependencies and information flows between virtualized and non-virtualized functions.

Conclusions

As mobile computing continues to evolve and access to computing clouds becomes ubiquitous, mobile users expect highly reliable, anywhere and anytime wireless connectivity and services. The need to evolve future wireless networks toward supporting, reliably and efficiently, a wider range of networking and multimedia services and applications becomes a critical design requirement of next-generation wireless networks. Cognizant of emerging trends in wireless services and applications, the article focuses on exploring the potential of NFV to address the daunting challenges and design requirements of 5G RANs. The article underscores that NFV approaches to enable advanced, cooperative, rapidly changing baseband processing and radio resource management in 5G must be flexible, cost effective, and elastic. NFV naturally inherits these benefits from virtualization, cloud computing, and SDN paradigms. New challenges related to carrier-grade network functions must be addressed. To this end, the article discusses critical open problems, including the need to adhere to strict real-time processing, support a programmable data plane, achieve efficient local and global resource management and orchestration, and explore NFV placement trade-offs.

Acknowledgment

This work was supported by National Science Foundation (NSF): grant CNS-1162296.

References

[1] J. G. Andrews *et al.*, "What Will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1065–82.

[2] ETSI GS NFV, "Network Functions Virtualisation (NFV): Architectural Framework," 2:V1.1.1, 2013.

[3] I. Giannoulakis *et al.*, "On the Applications of Efficient NFV Management Towards 5G Networking," *Proc. 1014 1st IEEE Int'l. Conf. 5G for Ubiquitous Connectivity*, 2014, pp. 1–5.

[4] R. Guerzoni, R. Trivisonno, and D. Soldani, "SDN-Based Architecture and Procedures for 5G Networks," *Proc. 1st IEEE Int'l. Conf. 5G for Ubiquitous Connectivity*, 2014, 2014, pp. 209–14.

[5] H.-H. Cho *et al.*, "Integration of SDR and SDN for 5G," *IEEE Access*, vol. 2, 2014, pp. 1196–1204.

[6] A. Basta *et al.*, "Applying NFV and SDN to LTE Mobile Core Gateways, the Functions Placement Problem, *Proc. 4th Wksp. All Things Cellular: Operations, Applications, & Challenges*, ACM, 2014, pp. 33–38.

[7] A. Checko *et al.*, "Cloud RAN for Mobile Networks, A Technology Overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 2014, pp. 405–26.

[8] E. Hernandez-Valencia, S. Izzo, and B. Polonsky, "How Will NFV/SDN Transform Service Provider OPEX?" *IEEE Network*, vol. 29, no. 3, 2015, pp. 60–67.

[9] N. Feamster, J. Rexford, and E. Zegura, "The Road to SDN: An Intellectual History of Programmable Networks," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 44, no. 2, 2014, pp. 87–98.

[10] N. McKeown *et al.*, "Openflow: Enabling Innovation in Campus Networks." *ACM SIGCOMM Comp. Commun. Rev.*, vol. 38, no. 2, 2008, pp. 69–74.

[11] P. Berde *et al.*, "Onos: Towards an Open, Distributed SDN OS, *Proc. Third Wksp. Hot Topics in Software Defined Networking*, ACM, 2014, pp. 1–6.

[12] N. Cvijetic *et al.*, "SDN-Controlled Topology-Reconfigurable Optical Mobile Fronthaul Architecture for Bidirectional Comp and Low Latency Inter-Cell D2D in the 5G Mobile Era," *Optics Express*, vol. 22, no. 17, 2014, pp. 20809–15.

[13] N. Lee *et al.*, "Base Station Cooperation with Dynamic ClusTering in Super-Dense Cloud-Ran," *Proc. 2013 IEEE GLOBECOM Wksps.*, 2013, pp. 784–88.

[14] A. Gudipati *et al.*, "Softran: Software Defined Radio Access Network," *Proc. 2nd ACM SIGCOMM Wksp. Hot Topics in Software Defined Networking*, ACM, 2013, pp. 25–30.

[15] M. Bansal *et al.*, "Openradio: A Programmable Wireless Dataplane," *Proc. 1st Wksp. Hot Topics in Software Defined Networks*, ACM, 2012, pp. 109–14.

Biographies

Sherif Abdelwahab [S'07, M'11, S'14] received his B.S. and M.S. degrees in electrical and communications engineering from Cairo University in 2004 and 2010, respectively, and is currently working toward a Ph.D. degree at the School of Electrical Engineering and Computer Science (EECS), Oregon State University, Corvallis. Before pursuing his Ph.D. degree, he was in the mobile networks industry with Alcatel-Lucent from 2004 to 2007 and with Etisalat from 2008 to 2013. His research interests include distributed networking, networked systems and services, and mobile and wireless networks.

Bechir Hamdaoui [S'02, M'05, SM'12] is an associate professor in the School of EECS at Oregon State University. He received his Ph.D. in electrical and computer engineering from University of Wisconsin at Madison (2005). His research interests span various topics in computer networking and communication. He won an NSF CAREER Award (2009), and currently serves on the Editorial Boards of several journals. He has served as Program Chair/Co-Chair and TPC member for many conferences..

Mohsen Guizani [S'85, M'89, SM'99, F'09] is currently a professor and chair of the Electrical and Communications Engineering Department at the University of Idaho. He received his Ph.D. in computer engineering in1990 from Syracuse University. His research interests include computer networks and wireless communications. He currently serves on the Editorial Boards of six journals. He is a Senior Member of ACM.

Taieb Znati is a computer science professor at the University of Pittsburgh, Pennsylvania. He served as senior program director of Advanced Networking Research at NSF (1999–2005) and later as the director of the NSF-CNS Division. He served as General Chair of INFOCOM 2005 and SECON 2004. He is a member of the Editorial Board of *IEEE Security and Privacy*, *Wireless Sensor Networks*, and the *International Journal of Sensor Networks*.

NFV can provide a solution to scalability issues by deploying all control decisions that mainly require cooperation of a large number of cells in VNFs near the network core and rapid decisions in NFVs near the network edge.