ETSI WORKSHOP ON FUTURE RADIO TECHNOLOGIES – AIR INTERFACES
27-28 January 2016, Sophia Antipolis, France

End-to-end Network Slicing in 5G Wireless Communication Systems

Qian (Clara) Li, Geng Wu, Apostolos (Tolis) Papathanassiou, Lili Wei
Intel Corporation, USA
Emails: {clara.q.li, geng.wu, apostolos.papathanassiou, lili.wei}@intel.com

***Abstract*— Network slicing is considered as one of the key technologies to fulfill the diverse requirements and the diverse services and applications expected to be supported in 5G communication systems. The technology development on vertical slicing that already started in late 4G and early 5G is mostly focused on slicing the core network. We envision this trend to continue with the development of horizontal slicing and with the implementation of air interface slicing and radio access network slicing. In this paper, we explain the concept of network slicing, propose an end-to-end (E2E) 5G system framework built on vertical and horizontal slicing, and illustrate the techniques in the air interface, the radio access network (RAN) and the core network (CN) for enabling 5G system with E2E network slicing. This paper aims to initiate discussions and spur development on the long-range technology roadmap and solutions for E2E network slicing in 5G and beyond.**

Keywords— 5G; network slicing; vertical networking slicing; horizontal network slicing; system architecture; air interface; radio access network; core network

I. INTRODUCTION

The 5G wireless communication system is expected to serve diverse applications with various traffic types and requirements, various network and user equipment with diverse communication and computation capabilities, and diverse market segments. To meet these requirements and to develop a network with future-proof scalability and flexibility, network slicing is considered as one of the key technologies [1].

With network slicing, one network can be sliced into multiple networks, each architected and optimized for a specific vertical application or service. Such network slicing has already started to occur in the wireless industry, usually in software defined core networks through virtualization technology and sometimes in the radio air interface through resource partition and the application of tailored radio access schemes. We call this "vertical network slicing", since it turns one big network into multiple parallel vertical end-to-end networks.

We envision this trend will continue, with virtualization technology applied across radio access network and portable devices, as well as across portable devices to wearable devices. The computing resources in the base station and the portable device will be horizontally sliced, and these slices together with the wearable devices will be integrated to form a virtual computing platform though a new 5G air interface design to significantly augment the computing capability of future portable and wearable devices.

Vertical slicing enables resource sharing among services and applications. Horizontal slicing, as a step forward, enables resource sharing among network nodes and devices. That is, high capable network nodes/devices share their resources (e.g., computation, communication, storage) to help the less capable network nodes/devices. Horizontal slicing forms network hierarchy and underlay networks.

Network slicing implementation is end-to-end. It includes slicing in the core network (CN) and in the radio access network (RAN). In the CN, network function virtualization (NFV) and software defined network (SDN) can be the technical enablers for network slicing: NFV and SDN virtualize the network elements and functions so that easily configured/reused network elements and functions in each slice meeting specific requirements can be enabled. In the RAN, slicing can be built on physical radio resources (e.g., transmission point, spectrum, time, etc.) or on logical resources abstracted from physical radio resources.

Each slice can have its own network architecture, engineering mechanism and network provision. As a result, splice-specific operation is needed. In the CN, each slice can have its own core network signaling and functionalities. In the RAN, the following could be necessary: Slice-specific on/off operation, slice-based access control and slice-based load-balancing.

In this paper, we first introduce the concept of network slicing and describe the system architecture with E2E vertical and horizontal network slicing, and then discuss the technologies in the air interface, the RAN and the CN to enable 5G system with E2E network slicing. This paper aims to initiate discussions and spur development on the long-range technology roadmap and solutions for E2E network slicing in 5G and beyond.

II. NETWORK SLICING CONCEPT AND SYSTEM ARCHITECTURE

Vertical network slicing slices one network into multiple network slices, each architected and optimized for dedicated services or applications. Different from physical networks dedicated to the particular services or applications, network slicing allows for sharing of the physical resources among the network slices. Based on the demand, each of the network

slices requires a certain amount of physical resource. As the resource allocated to a network slice can be dynamic (e.g., subject to the overall system load and the load of the network slice), it is desirable to minimize the correlation between the operation of the network slice and the physical resource. To this end, a virtualization layer can be added to map the layer of physical resource into a layer of logical resource for the network slices to operate on, as illustrated in **Error! Reference source not found.**.

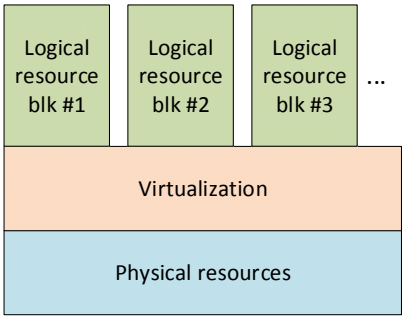

Fig. 1 Illustration of resource virtualization

Each of the network slices can be self-contained from end to end, i.e., vertical E2E slicing. Devices in one network slice may not be aware of the other, physically co-existing slices. Based on the service requirements, each of the network slices configures its RAN and CN functions. These functions can be specifically built for the network slice or can be common functions used by multiple slices.

Vertical slicing enables resource sharing among services and applications. Horizontal slicing, as a step forward, enables resource sharing among network nodes and devices, that is, high capable network nodes/devices share their resources (e.g., computation, communication, storage) to help the less capable network nodes/devices. Examples of horizontal network slicing include: 1) Portable devices take some of the computation load of the wearable devices so that wearable devices can meet user expectations (which are usually beyond the capability of the wearable devices) with small form factor and low cost; 2) relay nodes share their radio resources (e.g., time, frequency, spatial) to help deliver traffic to and from end terminals; 3) macro base stations (BSs) with wide coverage and/or reliable connections share their radio and computation resources to help small cell BSs provide coverage (e.g., control-plane anchor at macro BS and user-plane fallback to macro BS when the small cell link fails, especially for small cells running at millimeter wave bands). Since this type of network slicing forms network hierarchy and underlay networks, we refer to it as horizontal slicing to highlight the difference from vertical slicing.

Horizontal slicing can be applied in one vertical network slice or operate across multiple vertical network slices. When operating across multiple vertical slices, the resources in the horizontal slice will be virtualized and sliced for each of the member vertical network slices. To each of the member vertical network slices, the horizontal slice is dedicated to the vertical network slice and is a component of the vertical network slice architecture.

Fig. 2 illustrates the concept of vertical and horizontal network slicing. In the following section, we will further illustrate the techniques enabling network slicing in the air interface, the RAN and the CN. Our illustration will be focused on the communication techniques; however, we would like to note that implementation of network slicing also relies heavily on computation and information techniques.

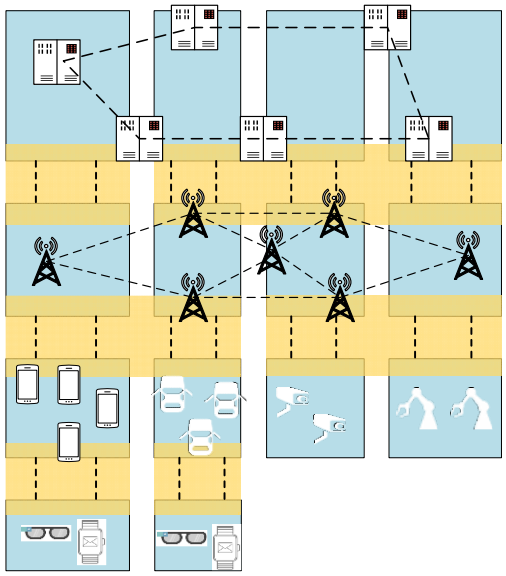

Fig. 2 Conceptual illustration of vertical and horizontal network slicing

III. Enabling Network Slicing in the Air Interface

Besides meeting the 5G requirements (e.g., data rate, latency, number of connections, etc.), the desirable features of the air interface to enable network slicing and in general 5G include:

- Flexibility: Support flexible radio resource allocation among slices
- Scalability: Easily scale up with the addition of new slices
- Efficiency: Efficiently use the radio and energy resources

One way to achieve the desired features is setting out from the physical (PHY) and medium access (MAC) layer architecture, as illustrated in Fig. 3. The basic components are a flexible partition of the PHY resources, a map of the physical PHY resources into logical resources, and a build of the operations of MAC and higher layers based on the logical PHY resources. The MAC layer can be partitioned into two layers: Layer-1 MAC performs intra-slice scheduling; Layer-2

MAC performs inter-slice scheduling. The two-layer MAC partition avoids the complexity of jointly scheduling multiple slices and allows better scalability.

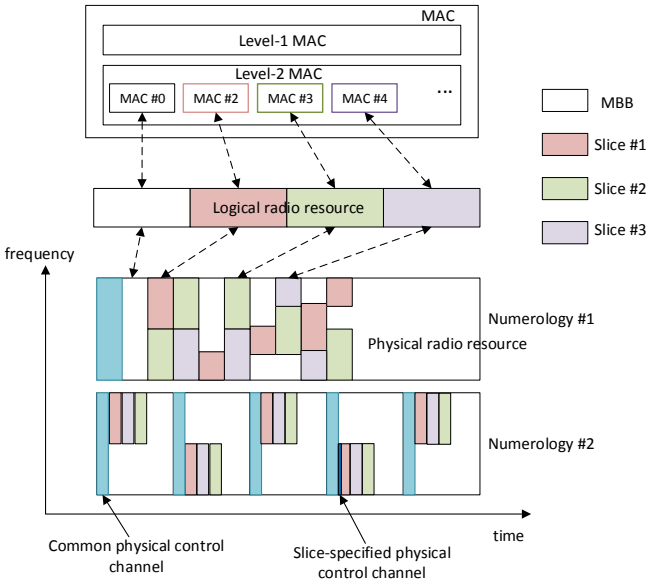

Fig. 3 Illustration on PHY and MAC architecture for air interface slicing

To identify a network slice, a network slice ID (sNetID) is assigned to the network slice. The sNetID is known by devices accessing the network slice and can be used to address all the devices in the network slice. The sNetID can be broadcast in the system information to indicate whether the slice is active in the BS.

For the key physical channels in the air interface, such as physical downlink (DL) and uplink (UL) control channels, physical random access channel and physical shared channel, we can have slice-specific physical channels, as well as common physical channels. The common physical channels can be used by all slices, and the dedicated physical channels are dedicated to the respective slices.

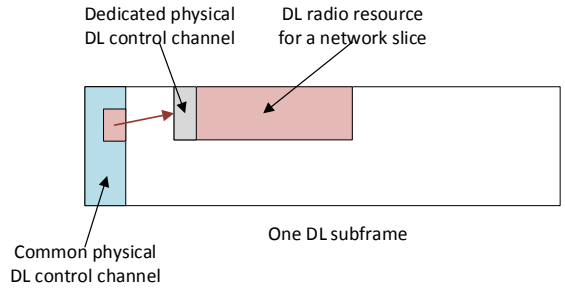

Fig. 4 Illustration of the physical downlink control channel

Fig. 4 shows an example of the physical downlink control channel in one DL subframe. The common physical DL control channel carries resource allocation information for the network slices. The sNetID is used to address the scheduled network slices. All the devices accessing a scheduled network slice can detect the common physical control information addressed to the corresponding sNetID. The dedicated physical downlink control channel for a network slice is located in the radio resources assigned to the network slice. The dedicated physical downlink control channel carriers scheduling information for the devices in the network slice. Similarly, in the uplink, common and dedicated physical uplink control channels can be designed. Devices accessing multiple network slices can aggregate the uplink control information and transmit it using the common physical control channel.

For random access, dedicated random access channels can be used to differentiate the contention resolution and admission control of the network slices, so that a crowded network slice with high random access collision probability should not affect devices accessing another network slice. The resource for the dedicated random access channel of a slice can be indicated in downlink system broadcasting, which requires the slice to be active in the BS or access point. For slices that have not been activated in the BS or access point, or for slices that do not require a dedicated random access channel, the common random access channel can be used. In this case, the common random access channel can be used as a way of activating a slice in the BS or access point.

IV Enabling Network Slicing in the RAN

As each slice can have its own RAN architecture, RAN operations such as mobile association, access control and load balancing schemes would be slice-specific instead of cell-specific as it is currently the case in mobile networks. Slice on/off operation would be enabled at each BS or access point. The control-plane (C-plane) and user-plane (U-plane) configuration could be tailored considering the slice-specific operation. In a sense, the slice-specific operation blurs the concept of physical cell site and makes the network operation more service/traffic/user oriented instead of physical cell oriented.

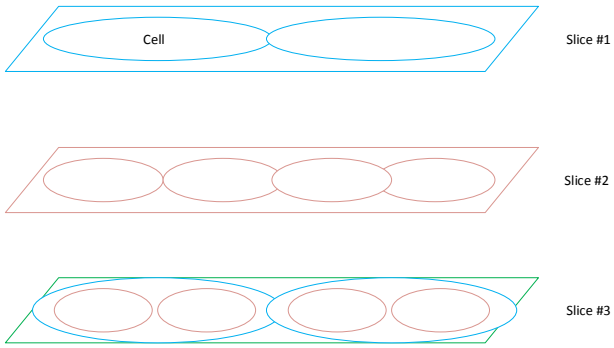

Fig. 5 Illustration on slice-specific RAN architecture

Fig. 5 shows an example of slice-specific RAN architecture. Depending on factors such as traffic type, traffic load and QoS requirement, the RAN architecture of each of the slices can be dynamically configured: In one instance, slice #1 can only operate on a macro cell, slice #2 can only operate on small cells, and slice #3 can operate on both macro and small cells. In another situation, slice #1 could expand its operation to small cells, while slice #3 can terminate operation on some of the small cells. The slice-specific RAN architecture would require slice-specific control-plane/user-plane operation, slice on/off operation and slice-based treatment on access control and load balancing.

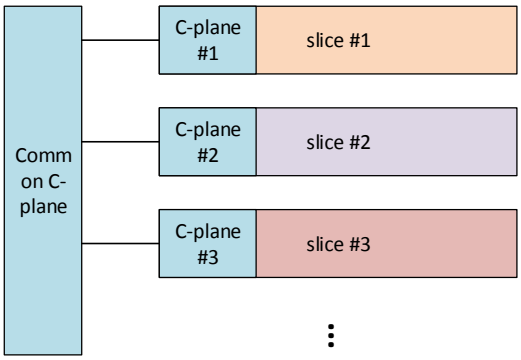

Fig. 6 Example on C/U-plane configuration

Fig. 6 shows one option of the slice-specific control/user-plane configuration. Some of the control-plane functions such as the functions in idle mode (e.g., paging, cell reselection, tracking area update) can be categorized into common C-plane slice functions, while the functions in connected mode (e.g., handover, dedicated bearer setup) can be categorized into slice-specific control plane functions.

The slice-specific RAN architecture inherently supports slice on/off, slice-based admission control and slice-specific load balancing. The triggers for turning on a slice at an access point could include: 1) Traffic load of that slice goes beyond a certain threshold; 2) The number of active devices operating on that slice goes beyond a certain threshold; 3) Need to keep service continuity; 4) Need to meet certain QoS requirements, such as low latency, ultra-reliability, etc. Slice-on at an access point can be triggered by a device or by the network. When triggered by a device, the network can decline the device slice-on request if, for example, the network sees that the cost/overhead of serving the slice outweighs the service benefit.

Likewise, admission control and load balancing will be based on the availability of the requested slice at the BS or access point as well as the load conditions and the overall system performance. The system information of a BS or access point carries information on the active slices in the BS or access point. Based on the system information, a device can decide whether to access the BS or access point. If the intended slice is active in the BS or access point and the channel condition is sufficiently good, the UE would generally choose to access the BS or access point. If the intended slice is not supported by the BS, the UE may still decide to request access. In this case, factors affecting the decision can be the link condition, QoS requirements, traffic load of the neighboring cells, etc. If the device makes the access request but the slice is not currently active in the BS or access point, the BS or access point would have to decide whether to accept the request. Signaling exchanges among the BSs/access points or between the BS/access point and the central controller may be needed to facilitate the decision. Once the BS/access point decides to accept the access request, the BS/access point will need to turn on the slice using the slice on/off procedure.

V Enabling Network Slicing in the CN

NFV and SDN are technical enablers of network slicing in the CN. The SDN technology was mainly developed by the IT industry for efficient operation and management of datacenters and large IP and Ethernet networks. The goal of SDN is to separate the control plane from the data plane, and to make the control plane programmable through APIs in order to bring flexibility in how networks are deployed, operated and managed. Network function virtualization is primarily driven by network service providers. The goal of NFV is to virtualize network functions into software applications that can be run on industry off-the-shelf standard servers or as virtual machines running on those servers. NFV and SDN virtualize the network elements and functions to easily enable configured/reused network elements and functions in each slice to meet its own requirement.

VI Conclusion and Future Research

In this paper, we introduce the concept of horizontal network slicing and 5G system architecture with E2E vertical and horizontal slicing. Techniques in the air interface, the RAN and the CN in enabling 5G system with E2E network slicing are discussed. Many questions need to be addressed to enable E2E vertical and horizontal network slicing such as network slicing criteria and granularity, air-interface and protocols, slice-specific RAN operation, and coordination and co-existence of the slices.

As the industry moves towards vertical slicing as the phase-1 implementation, provision for enabling horizontal slicing as the phase-2 implementation is needed. Without a long-range vision, the industry may end up doing a version of Phase 1 network slicing, and then realize the limitations when Phase 2 arrives. We hope this paper serves the purpose of triggering the discussions and technology development on E2E network slicing for 5G and beyond.

References

[1] NGMN 5G white paper, version 1.0, Feb. 2015