Research Brief on 5G Networks

Software-Defined and Cloud-Native Foundations for 5G Networks

Sponsored by: InterDigital

Introduction

The fifth generation of mobile networks, commonly known as 5G, holds a lot of promise. Historically, 2G brought us mobile voice, while 3G introduced us to mobile data. 4G and LTE enabled usable mobile broadband services, and now 5G is supposed to unlock further value from our mobile networks with faster broadband data speeds and lower latency. 5G is expected to bring about a sea of change that will come as a result of new use cases and business models for consumers, businesses, and the industry. And it will lay the foundation for massive adoption of Internet of Things (IoT) everywhere.

5G and the applications it enables require a significant upgrade of carrier infrastructure beyond the radio access networks (RAN). Putting in 5G radios is just part of the overall effort. To support the diversity of applications, an explosion in device types, capacity and scale of traffic, communications service providers (CSPs) are turning to the adoption of cloud data center innovation both adapted and re-developed for carrier networks. Virtualization, software-defined infrastructure, cloud-native computing technologies in the form of network functions virtualization (NFV), software-defined networking (SDN), multi-access edge computing (MEC), containers and micro-services architectures will have significant roles to play in 5G infrastructure upgrades.

Communications service providers (CSPs) are turning to the adoption of cloud data center innovation both adapted and re-developed for carrier networks.

5G - An Unstoppable Wave

Some might ask why cloud-native and software-defined technologies are needed within 5G networks. The reason lies in the impending scale of the mobile traffic that the 5G core (5GC) and the edge will have to process shortly. For instance, Ericsson's mobility research report estimates that mobile traffic is growing at a rate of around 40%, and Cisco's Visual Networking Index estimates that wireless and mobile devices will account for more than 63% of total IP traffic by 2021.

The wave has already started. AT&T recently rolled out 5G in 12 cities at the end of 2018, and will deploy service in 19 more markets in early 2019. Verizon has already rolled-out 5G fixed wireless service across four US cities and demonstrated live 5G at CES 2019. It plans to expand more aggressively in 2019. All carriers are laying the groundwork for 5G by increasing existing LTE capacities with solutions like LTE Advanced and LTE-LAA (Licensed Assisted Access), which provide 400Mbps to 1Gbps peak speeds. Moreover, with

5G data-only devices starting to become available, operational 5G networks are just around the corner.

In addition, we expect that module-based 5G devices for IoT will come online around 2020. We already have significant IoT traffic on today's mobile networks, from surveillance cameras to agricultural sensors, to building HVAC monitors. Research firm Statista predicts that 12.86 billion IoT sensors and devices will be in use in the consumer segment by 2020, growing at 35% CAGR from 2017. Whether LTE or 5G, IoT devices will make up a good number of mobile end-points.

What all this means at the end of the day is increased traffic, an increased number of devices and higher demands on the mobile edge, transport and core infrastructure. This all leads to more challenges for global telcos.

What all this means at the end of the day is increased traffic, an increased number of devices and higher demands on the mobile edge, transport and core infrastructure.

New 5G-Enabled Use Cases

So what can 5G enable that 4G can't handle? The 3GPP has defined a large number of different 5G use cases, categorized into a few different groups including massive IoT, critical communications, enhanced mobile broadband (eMBB), and network operation. Also, the Next Generation Mobile Networks (NGMN) Alliance, in its February 2015 whitepaper, described 25 use cases grouped into eight families: broadband access in dense areas, broadband access everywhere, higher user mobility, massive IoT, extreme real-time communications, lifeline communication, ultra-reliable communications, and broadcast-like services.

All these new use cases depend on the lower latency, higher-bandwidth and increased flexibility that 5G promises. The mobile consumer has been bombarded with images of new applications like augmented reality/virtual reality (AR/VR), connected and autonomous vehicles, telemedicine, and smart cities. So consumers can't wait to get their hands on these new applications, but there's work on the backend to make all this happen. For instance, to support realistic AR/VR, the network needs to provide sub-10ms latency and high bandwidth. Moreover, remote surgery in telemedicine requires extremely low latency to be successful. Meanwhile, premium HD, 360° and 4K video require both high-bandwidth and the support for edge caches.

5G and Network Slicing

To support these different use cases, 5G infrastructure provides for a capability known as network slicing. This is the ability to enable multiple separate networks on top of a common shared infrastructure. From a carrier perspective, a network slice is a virtualized end-to-end logical network on top of a physical infrastructure that provides a specific quality of service that is negotiated during the provisioning of the network. This network slice might use dedicated physical resources, or perhaps shared resources, from the base station of the radio access network all the way through the transport layer and into the core.

One of the use cases for network slicing is to support mobile virtual network operators (MVNOs). With virtual network slices, the extended networks of multiple virtual carriers or even enterprises can stretch beyond enterprise boundaries into the mobile sphere as well, running on top of the shared physical infrastructure.

Other examples of network slices include a high bandwidth slice for movie streaming or an ultra-low latency slice for telemedicine. Another example would be an ultra-reliable slice for autonomous driving or assisted driving.

Network slicing can facilitate new business models for carriers to sell or share parts of the network in a secure and isolated manner to different companies. This is similar to AWS or Microsoft Azure selling cloud services as shared resource slices of computing, networking, and storage to various enterprises.

Network slicing can facilitate new business models for carriers to sell or share parts of the network in a secure and isolated manner to different companies.

Key Technologies in the Cloud-Native World

Speaking of cloud technologies, to bring innovative new 5G applications and network slicing to market, carriers have realized that they need to turn to new technologies. Carriers and industry organizations like 3GPP and NGMN have recognized that their best bet lies in adopting technologies pioneered within cloud data centers. Only through adopting these technologies can carriers provide the scale and performance required by 5G applications.

SDN

SDN is a movement that started in the early 2010s. It involves the separation of the control plane of network devices from the data plane, allowing a centralized approach for networking control that provides simplification and global optimization for the routing and switching of network packets. SDN also advocates open APIs and a programmatic approach to networking.

SDN is an essential element of 5G that enables fast service provisioning (and de-provisioning) as well as the optimal use of the underlying transport. Through its support for a programmable network, SDN is critical in virtualization of the underlying network.

SDN architectures can also be implemented to ensure that end-to-end paths are provisioned efficiently to maximize the transport from the mobile user equipment (UE) to the data center or edge services that they are likely to consume.

NFV

SDN itself will facilitate the coordination and virtualization of the underlying network, but a 5G network also requires network functions that support critical elements in each network slice. Specifically, network slicing will require flexible network services per slice that perform the kind of network functions that are found today in the SGi/Gi-LAN, such as firewalling and other security functions; caching and acceleration; and any metering or policy elements.

NFV is a complementary movement to SDN, leveraging virtualization to take proprietary physical hardware that is non-portable and hard to manage, and convert the network functions into virtualized software-only versions. These software virtual network functions (VNFs) can be quickly moved around as needed, and scaled up or down dynamically.

NFV architectures support the ability to provision network slicing services flexibly, and this will likely take the form of data centers spread across the network from core to edge to reduce the latency of these services and for load distribution.

> The combination of NFV and SDN is critical in achieving flexible network topology and the realization of 5G targets.

Even in core 5G networks that don't need network slicing, it is still essential to have the flexibility to create network services. Virtualized mobile cores (vEPC) enable flexible network management while maximizing its availability based on virtualization aspects.

The combination of NFV and SDN is critical in achieving flexible network topology and the realization of 5G targets such as 1,000-times higher system capacity; 100-times increase in data rates (10-Gb/s speeds); connectivity enablement for 100-times more devices; latency reduced to 1 millisecond from 5 ms; and energy savings.

Containers

SDN and NFV are well-understood and already deployed in today's 4G LTE networks. However, to enable the next level of flexibility in mobile networks, operators might have to turn to more recent innovations in the data center world — containers and cloud-native capabilities.

One of the most fundamental changes to occur in cloud platform the last few years has been the ascendency of containers and their associated microservices. While not necessarily a new concept, containers and container solutions like Docker, have made it much simpler for developers to package their applications in a way that makes it possible for them to

be deployed, both on-premises and in the cloud, at will. By providing a higher level of abstraction, containers eliminate the need for developers to navigate multiple types and classes of virtual machines.

Containers allow applications to be separated from the underlying infrastructure, making them more portable. They also provide speed and agility to allow applications to be tested and deployed quickly. Further, containers provide lower overhead than virtual machines while still providing isolation from infrastructure and other tenants on the same hardware and operating system.

Microservice Architecture

Historically, applications were written as a monolithic entity. The more complex the application, the larger the deployable binary. Microservices architecture is an approach that breaks a monolithic application down into a collaborating collection of components, called microservices. For instance, in an application that involves a database, there might be a microservice that is responsible for search and another responsible for backup. The goal is to simplify the application into many components, usually running on containers, each of which can be independently deployed, upgraded, and patched. The thought behind this architecture is that each component tier can be scaled accordingly to accommodate scale, and each component can be upgraded independently to facilitate agility.

To orchestrate the bring up of multiple components running on containers, there are a few popular orchestration systems, with Kubernetes from Google being the dominant solution in the cloud community today. Kubernetes can orchestrate the instantiation of hundreds or more separate components necessary to instantiate an application.

Each microservice component tier can be scaled accordingly to accommodate scale, and each component can be upgraded independently to facilitate agility.

Applications of Cloud Frameworks in 5GC

To tie all these elements together, we need to examine the 5G system architecture as described by the 3GPP's Technical Specification (TS) 23.501 working group. Historically, in 2G, 3G and 4G networks, the different network functions that come together (e.g., HSS, MME, etc) are integrated as point-to-point links. This point-to-point model has served the mobile industry well in the past.

However, going forward, for 5G, the 3GPP has recognized that there is a need for increased service agility. As such the 3GPP working group has designated a services-based architecture (SBA) that mirrors cloud-native architectures and is believed to scale better.

Just like containers and micro-services architectures did for cloud computing, SBA will facilitate 5G network functionality becoming more granular and decoupled.

SBA Enables Scalable Network Services for 5G

Just like containers and micro-services architectures did for cloud computing, SBA will facilitate 5G network functionality becoming more granular and decoupled. This should allow for increased automation and an agile operational process (not unlike the continuous integration/continuous deployment or CI/CD movement in the cloud computing world). The result should be an overall reduction in deployment times, better operational efficiencies and improved resilience.

Within the context of SBA, a service is an atomic entity (like a microservice) that serves a specific function. A service can be updated independently and deployed as needed. In release 15, 3GPP introduced the concept of Network Functions Services (NF Services) as part of SBA. In this first release, SBA only applies to control-plane functions but will evolve beyond in release 16.

Under SBA, the 3GPP has defined common supporting infrastructure to make service-based architecture a reality. These capabilities include service registration, which provides a list of available services, their status and how to reach them. They also include service authorization and authentication which controls whether services can contact each other and service discovery, which provides for appropriate selection of the right services to use, and can facilitate load-balancing.

Also, 3GPP has picked common protocols (in release 15) reminiscent of cloud architectures for SBA. These currently include an overall RESTful framework utilizing HTTP/2 for application layer communication, TCP-based transport and JSON for serialization.

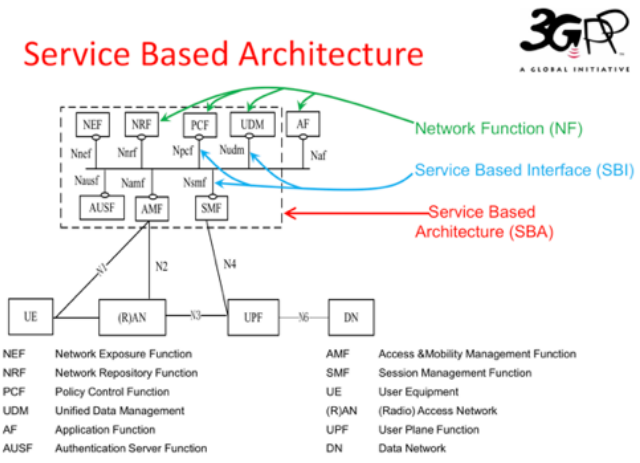

Figure: 5G Service Based Architecture (source: 3GPP)

As depicted in the SBA diagram from the 3GPP above, the major components of the 5G core are listed as follows:

- Access and Mobility Management Function (AMF): Manages all UE related functions, especially access control and mobility.
- Session Management Function (SMF): Session establishment, modification, and release, as well as other control plane functions.
- User Plane Function (UPF): Provides functions specific to U-plane processing, act like the S and P Gateway in 4G. UPFs can be deployed in different locations to perform different functions.
- Policy Control Function (PCF): Equivalent to PCRF in 4G. Dictates the policy that governs overall network behavior.
- Unified Data Management (UDM): Stores subscriber data and profiles.
- NF Repository Function (NRF): Provides a service registry and service discovery for NFs.
- Network Exposure Function (NEF): API gateway that allows external services or users to integrate with the mobile network and provision or deprovision services as well as push application policy. Works like the Service Capability Exposure Function (SCEF) in 4G.
- Authentication Server Function (AUSF): Supports authentication services.

These and other services identified above are meant to recreate under the SBA framework the components that were available within the pre-5G mobile core.

Network Slicing, SDN, NFV, and SBA - Tying it Together

Today, outside the mobile core, SDN, and NFV are used in conjunction to provide network functions within multi-tenant frameworks. By using SDN to provide network virtualization and service function chaining, a specific tenant can be served by a subset of available VNFs in an NFV deployment by merely associating a specific chain of VNFs for flows belonging to that tenant. This framework works regardless of whether these NFs are VNFs or CNFs (cloud-native network functions packaged as containers).

> The good news is that numerous technologies can be borrowed from the cloud world, including Kubernetes, service meshes, message bus frameworks, monitoring, and telemetry, and distributed key-value stores to help make SBA a reality.

Taking the same approach within the SBA framework, 5G vendors have set up an early demonstration of a similar setup on the 5G control plane. For instance, a mobile

operator can request a network slice for surveillance video feeds that consist of a next-generation firewall to protect these IoT devices from malware, followed by some video optimization function. At the same time, this network slice can be allocated resources to ensure that it has sufficient bandwidth to deliver the videos while not necessarily needing the best latency.

Just as AWS, Microsoft Azure and Google Cloud Platform use containers to power some of the world's most successful web services, the vision is to use the same technologies to enable even more massive scale at incredible efficiencies on 5G networks.

Under the SBA framework, and in concert with SDN and NFV, we can envision an example with the instantiation of two different NF services, such as AMF service-firewall and AMF-service-video-opt solely for this network slice through the use of VNFs or CNFs under NFV. We would then orchestrate the underlying network via SDN to ensure the traffic from this slice, as it comes through the RAN via the transport network, is routed to these AMF instances running within our data center, for appropriate processing.

While this is going on it is possible to leverage the same technologies to simultaneously service a telemedicine network slice that requires different NFs and extremely low latency using the same shared physical resources. Also, when one of the slices is longer needed, the services would be de-provisioned dynamically, and the resources returned to the shared pool.

Indeed, there is a lot more work needed before all this becomes a reality. Resource management, admission control on whether a network slice can be instantiated to guarantee the appropriate SLAs, what to do when a node in the SBA fails, and SLAs are put at risk. How to best orchestrate the AMFs and enable efficient service discovery etc. are all questions that must be resolved. The good news is that numerous technologies can be borrowed from the cloud world, including Kubernetes, service meshes, message bus frameworks, monitoring, and telemetry, and distributed key-value stores to help make SBA a reality. We've only seen the tip of the cloud-native iceberg within the 5G ecosystem.

Conclusion: Looking Towards Release 16 and Beyond

While we are early in these demonstrations, the hope is that when Release 16 is frozen in March 2020 (delayed from December 2019), the full concept of SBA has been fleshed out and that we are using a microservices-based container infrastructure (perhaps orchestrated by Kubernetes) to power multiple 5G network slice at scale. Just as AWS, Microsoft Azure and Google Cloud Platform use containers to power some of the world's most successful web services, the vision is to use the same technologies to enable even more massive scale at incredible efficiencies on 5G networks.