White Paper

Mobile Cloud Service Core for 4G & 5G Networks

Prepared by

Gabriel Brown
Senior Analyst, Heavy Reading
www.heavyreading.com

on behalf of

www.cisco.com

February 2016

A New Era for Mobile Core

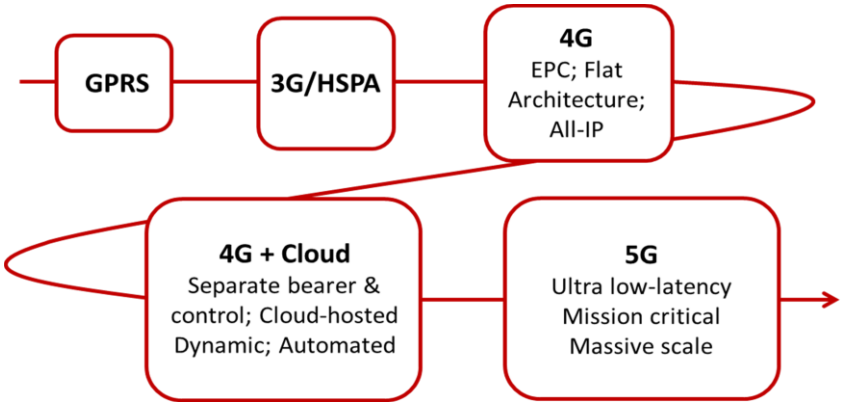
Mobile networking is at an important nexus: High-speed 4G access is now widely deployed and, from the perspective of many applications, connectivity is considered ubiquitous. At the same time, cloud technologies are evolving rapidly and changing the nature of online services. Used in combination, these technologies offer mobile operators an opportunity to create an efficient, automated service delivery infrastructure that can scale to support the diverse applications needed in modern societies.

This white paper argues that, to achieve this goal, operators need to evolve to a new service-oriented core network – the Mobile Cloud Service Core – to create an infrastructure with the flexibility to support new service models and a cost-of-production driven by "cloud economics," rather than by specialized hardware. This new core network comprises distributed user plane components deployed in smaller data centers close to the radio and a "mobility controller" hosted more centrally in the cloud to create an architecture that can radically improve performance and prepare the network for 5G services.

A New Era in Mobile Packet Core

The current mobile core is based on the evolved packet core (EPC) introduced with the deployment of 4G-Long Term Evolution (LTE) around 2010. This very capable, all-IP core can support high capacities and multiple use cases. Over the past few years the industry has sought to virtualize the EPC to run on commercial-off-the-shelf (COTS) server hardware. Virtualization has had some success; however, rather than a direct porting of a hardware-oriented software architecture to virtual machines (VMs), the next-generation mobile core should be designed, from the ground up, to run in the cloud. And, as always, the new core should interwork with existing core technologies and radio access networks (RANs) to enable service continuity.

Figure 1: A New Era in Mobile Core for Cloud & 5G



Source: Heavy Reading

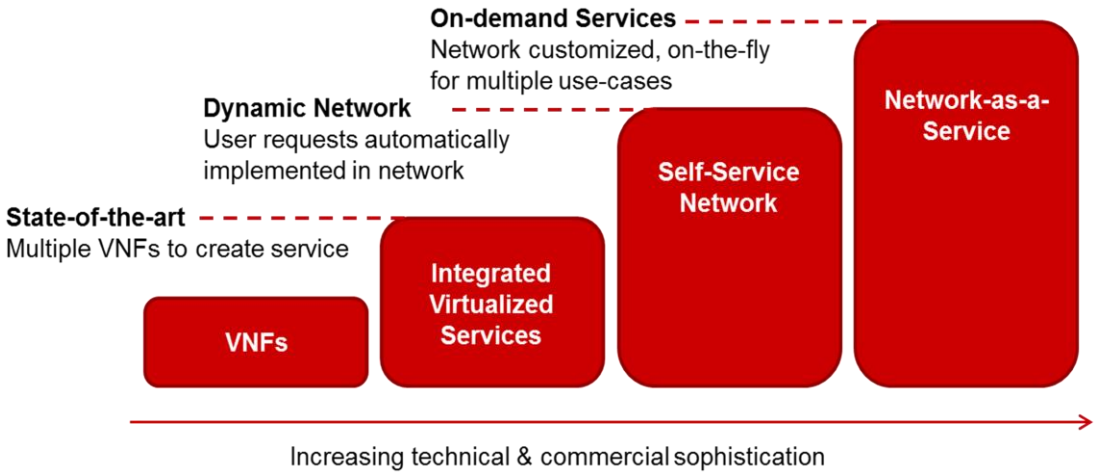
Figure 1 shows that a new generation of mobile core technology is typically introduced every five years (the overall lifecycle for each technology is considerably longer as generations overlap one another). The market is currently entering the 4G + Cloud era and will soon need to support 5G services if operators launch in, or before, 2020, as is expected. The new Mobile Cloud Service Core must bridge these two

phases to optimize monetization of 4G through the most productive part of its lifecycle and to support the introduction of advanced and diverse 5G services.

New Business Models & Network Slicing

A cloud-native mobile core should enable new service models. Using software-based networks with automated resource and service orchestration, operators should be able to dynamically create network services optimized for the needs of an application or user group. This is sometimes referred to as "network slicing" and is expected to be introduced in advanced 4G networks and to be inherent to the 5G system architecture. **Figure 2** shows how the technical and commercial capabilities of the mobile core increase in step-wise fashion.

Figure 2: A Service-Centric Mobile Core



Source: Heavy Reading

- **Step 1:** Most vendors and operators are currently in this phase. Some functions have been virtualized and are now deployable as virtual network functions (VNFs) in a hybrid physical/virtual environment.
- **Step 2:** Operators are able to use service orchestration tools to deploy integrated, end-to-end virtualized services. This is today's state-of-the-art and is now being deployed by a small number of progressive operators.
- **Step 3:** Make network services configurable by the customer – particularly enterprises – through a portal that can automatically implement service requests and changes in the network without the need for manual intervention on the part of the operator. This requires a dynamically configurable network infrastructure.
- **Step 4:** This is the "network as a service" model, where different users and applications can request connectivity configured according to their needs, on demand. For example, a service may require high transaction capability, or specific quality-of-service (QoS) attributes and can demand this configuration from the network. In the first instance, we envisage operators will use this capability to offer private mobile networks as a service to enterprises that seek greater control and customization of their environment.

Cloud Principles & the New Core Architecture

Operator networking is emphatically different from the provision of cloud services; both obviously include networking, but the challenges of cloud providers are not the same as those faced by network service providers. Operators have a different physical footprint, with many more points of presence, and are focused on connectivity in the first instance. And yet there are important lessons operators can learn from the cloud related to automation, control and scalability. Bringing these cloud principles to operator networks can drive efficiency and innovation. Ultimately, the aim is to make networking as easy to buy and consume as cloud services.

A Web-Scale Mobile Packet Core

One reason cloud providers are able to operate at "Web scale" with relatively little manpower is because they have developed infrastructure that combines centralized control with distributed processing. This enables cloud providers to very efficiently process data and make rapid changes to how and where it occurs, according to resource availability, policy, location, service type, and so on. These capabilities are now increasingly required by mobile operators that are dealing with very rapid traffic growth and need to support a wide diversity of applications.

Similarly, mobile operators would do well to emulate user portals that allow customers to purchase and configure network services on a self-service basis. Using software-defined networking (SDN) and virtualization technologies, customers' requests should be automatically implemented in the network, allowing operators to "close the loop" between services and infrastructure automation. Implementation of this model will greatly increase the speed of service change in operator networks and equip operators to better respond to and create opportunities in the dynamic online services market.

Distributed Processing, Centralized Control

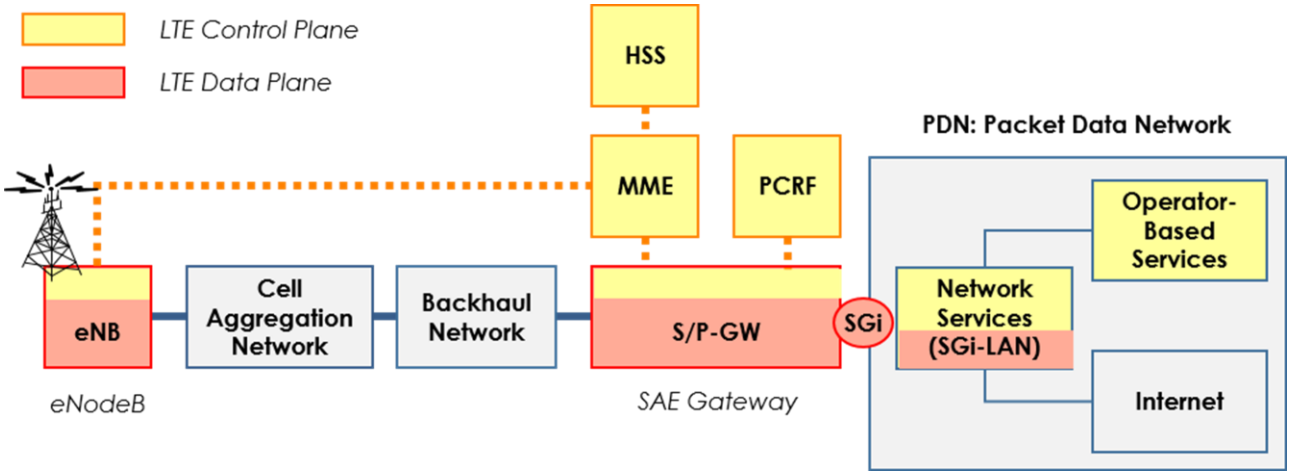
The 4G-LTE mobile network architecture comprises control plane and data plane elements that are, to a large degree, distinct and optimized. This delivers well-understood benefits. However, there remain important functions that specify both data plane and control plane functions within the node. This includes the packet core network gateways (S-GW, P-GW, a.k.a. SAE gateway) shown in in **Figure 3**. SAE gateways carry out critical functions, such as GTP encapsulation/termination, session management, mobility anchoring, policy enforcement, metering and routing. They are typically deployed in a centralized data center. The hardware nature of the nodes, and the associated software architecture, means both the control plane and data plane must scale simultaneously.

The SAE gateways interface with a large number of other "pure" control plane functions, such as online charging, policy control, AAA and the MME (which manages local mobility in the RAN and gateway selection). Overtime, the number of control plane interfaces the SAE gateway needs to support increases complexity and limits scalability and deployment flexibility. Lots of complex integrations with surrounding nodes means the operator's ability to makes changes is limited, the cost of change is too high and, as a result, the service portfolio remains static.

One proposed solution is to split the gateway node into two components: (1) a user plane node, which terminates the data plane, anchors sessions, forwards traffic according to policy and is deployed in smaller, distributed data centers; and (2) a

management and control plane node hosted in the cloud location, which acts as "mobility controller." This control and user plane separation (CUPS) approach effectively creates a new architecture that can scale cost-effectively and is closer in concept to mainstream SDN and cloud networking, but also supports interworking with the "classic" EPC, existing radio assets and existing operator services, such as IP Multimedia Subsystem (IMS) or SGi-LAN services.

Figure 3: Mobile Network Architecture Is Control-Plane Intensive



Source: Heavy Reading

The separation of control and bearer planes means each component can scale independently and be deployed where it makes sense for the service in question. For example, distributed SAE gateway elements – known as "user plane nodes" – can be deployed close to the user to reduce latency, or to optimize delivery of content from a content delivery network (CDN) or applications from a distributed cloud node. Heavy Reading research has shown there is a strong appetite among operators to make use of their distributed infrastructure (central offices, aggregation points, old base station controller sites, and so on) to create software-driven, cloud-optimized networks. Several large, progressive operators have commented publicly that this is their strategic direction for next-generation networks.

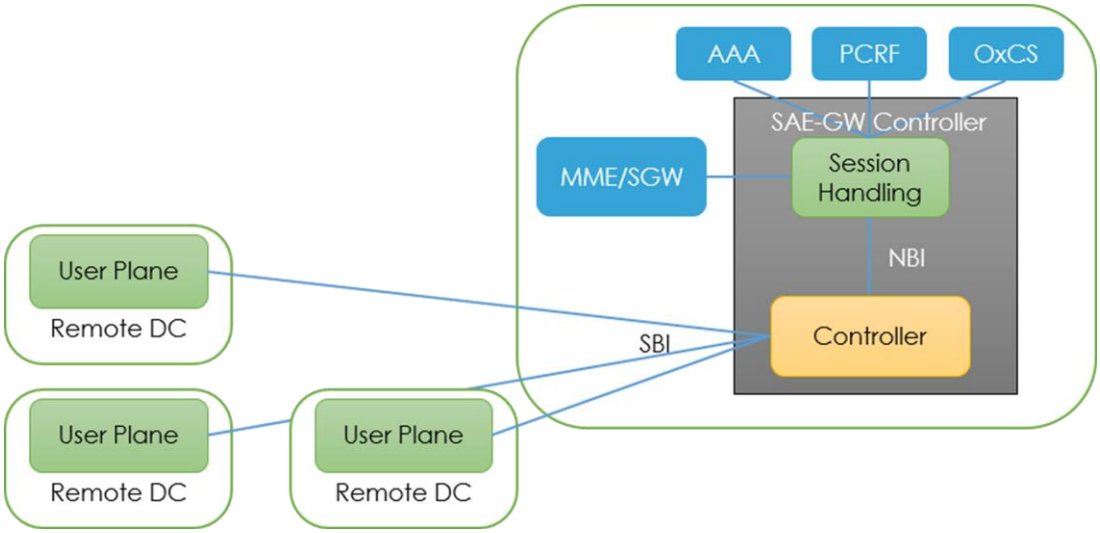
A New Architecture – the Split Gateway

This control and user plane separation – the so-called "split gateway" approach to EPC evolution – begets a new architecture that can better serve novel 4G services and prepares the network for 5G. In this model, "user plane nodes" are deployed in distributed data centers to process traffic according to policy and act as IP anchors. The SAE-GW management and control plane component (which we'll refer to as a "controller") is deployed in a centralized data center and is responsible for setting up and managing user sessions in the user plane nodes. This controller also interfaces with the surrounding control plane functions (PCRF, OxCS, etc.) deployed centrally and, as such, offloads the overhead of control plane integration from the user plane node, which can now be optimized for low-cost packet processing. This is shown in **Figure 4**.

The distributed user plane node can be colocated with distributed VNFs, content or other applications running on distributed cloud infrastructure – such as envisioned by

the Mobile Edge Computing (MEC) initiative, for example. Over time, a logical step could be to deploy user plane nodes and associated IP services at a cloud RAN "hub" location, in an enterprise small cell network, or even at the cell site. By placing GTP encapsulation/de-encapsulation at the same location as distributed cloud infrastructure, there is an opportunity to service customers locally and avoid having to backhaul traffic over long distances to a central P-GW. This is particularly important for low-latency applications and, clearly, the ultra-reliable, ultra-low-latency 5G use cases will need application logic to be hosted close to the radio access.

Figure 4: Separation of Control & Bearer in SAE Gateway



Source: Cisco, Heavy Reading

On the control plane side, this model offers other advantages. It can simplify the integration and management of control plane interfaces in the mobile core, making it faster and easier to modify deployed services or add new services to the portfolio. There is also potential to combine multiple control plane functions, including the MME, to create a larger "mobility controller" node.

The southbound interface (SBI) between the controller and user plane component (see **Figure 4**) is interesting and potentially challenging. Operators need to better understand how "chatty" that link is, what the latency requirements are, and which protocol to use. These issues are being investigated in the current wave of operator trials and by industry forums and standardization bodies. To pass messages between the controller and user plane node, some vendors may continue with a protocol like GTP to stay consistent with 3GPP standardization; others are investigating Open Flow, PCEP or BGP for the SBI. Currently, we believe it is most important to establish the principle of this architecture on the understanding that multiple different SBI protocols could be used in different vendor and operator implementations.

Another issue under active consideration is which features are ported over from the legacy gateway model to the new user plane node. Typically the mobile industry, through 3GPP, seeks extensive backward-compatibility between technology generations to enable a smooth, gradual migration. However, for a forward-looking architecture that targets LTE-Advanced Pro and 5G, this may not be a practical

approach – for example, support for 2G handover my very well be redundant for these operators and, therefore, would not justify the extra complexity. This discussion is ongoing in 3GPP and is being addressed through trials and hands-on experience.

Distributed User Plane Nodes

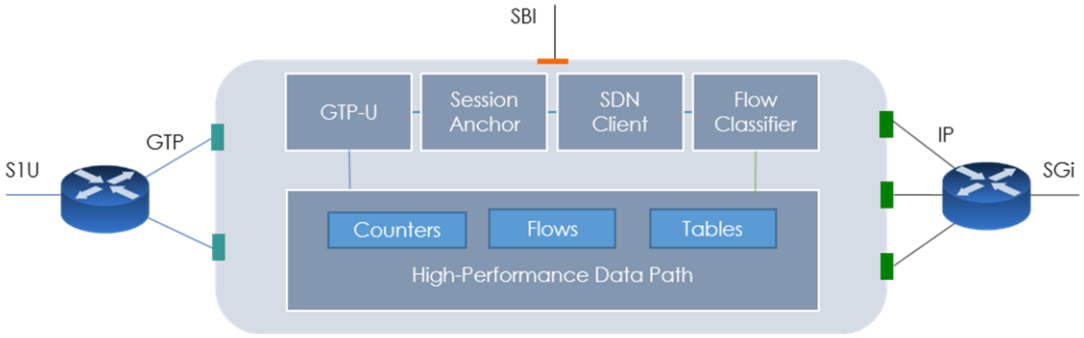
The user plane nodes must be simple, low-cost and scalable. In contrast to the traditional model in which subscriber control-plane functions and throughput scale in tandem, in this architecture it is possible to deliver user plane scalability (i.e., data throughput) without capex investment spiraling out of control. As discussed below, there is also an opportunity to integrate IP services as VNFs at the same location.

The user plane nodes can be hardware based – for example, based on white box Ethernet switches – or run as VNFs on x86 COTS servers. Both options have appeal: the hardware model is better from a performance footprint point of view today; however, over time, we believe the performance of virtual user plane nodes will improve through innovation in virtual switching and that the greater flexibility over the hardware model will win out. Many vendors are now heavily invested in vSwitch development and so the choice may be primarily a question of timing.

Fundamentally, the user plane node terminates subscriber access (GTP tunnels), anchors subscriber sessions, applies traffic management according to policy and forward traffic as IP. Like the P-GW it is deployed at the border of the 3GPP mobile environment and conventional IP networks.

The use of GTP has been beneficial to operators to enable delivery of policy-based services, but encapsulating traffic in GTP tunnels at the radio base station limits opportunities to process traffic until it reaches the core network gateway. In an SDN world, this now looks old fashioned and less appropriate. In the new architecture, the distributed user plane node strips off GTP at the distributed site and the operator is now able to process mobile traffic in the same way as any other IP access. The components of the user plane node are shown in **Figure 5**.

Figure 5: Distributed Forwarder Terminates GTP



Source: Cisco, Heavy Reading

Service Chaining & NSH for Distributed SGi-LAN

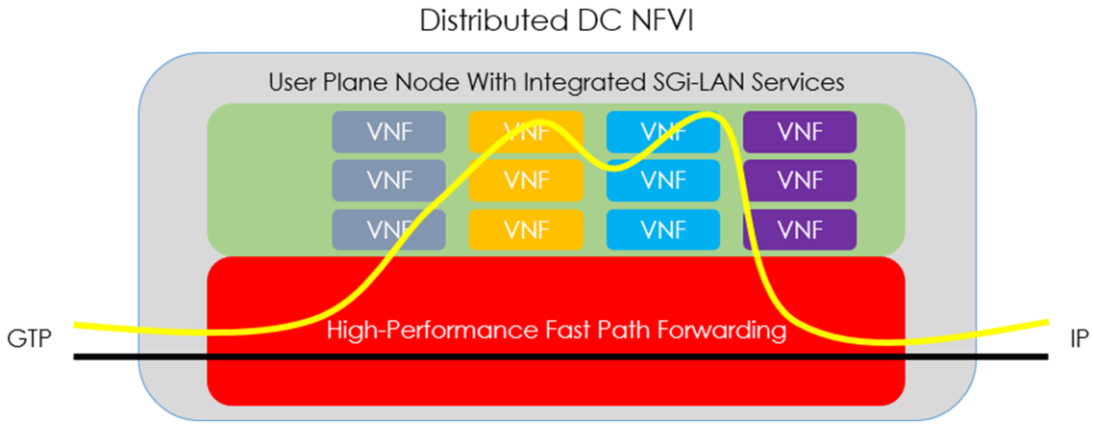
Because the user plane node terminates GTP it can act as an ingress (and egress) into a service chain comprised of in-line services on the SGi-LAN, which itself can

now also be distributed. These in-line services can also be deployed locally at the remote data center, and the operator can execute the majority of SGi-LAN processing (parental control, security, optimization, etc.) at this location and then forward IP traffic onward to the Internet. In essence, the service chain control and management are hosted centrally, but implemented locally.

A strength of this model is deployment flexibility. Even though there is less need for a large, centralized SGi-LAN service complex, operators can choose which services to centralize or distribute. For example, when launching a new service in a "fast fail" model, the operator can start with a centralized deployment in order to get to market rapidly and then scale the service by distributing it when it becomes successful. In this model, using transport-independent service chaining technology means in-line services can be located anywhere in the network.

This is a powerful idea that can help operators segment the network into virtual "slices." Using Service Function Chaining (SFC) technology, the user plane node acts as a "classifier" that adds metadata to the egress SGi packets using the network services header (NSH), making it possible to tag and segment traffic such that the operator can offer per-subscriber service processing within the node, as shown in the yellow line in **Figure 6**.

Figure 6: Distributed User Plane With Local IP SGi-LAN Service Chaining



Source: Cisco, Heavy Reading

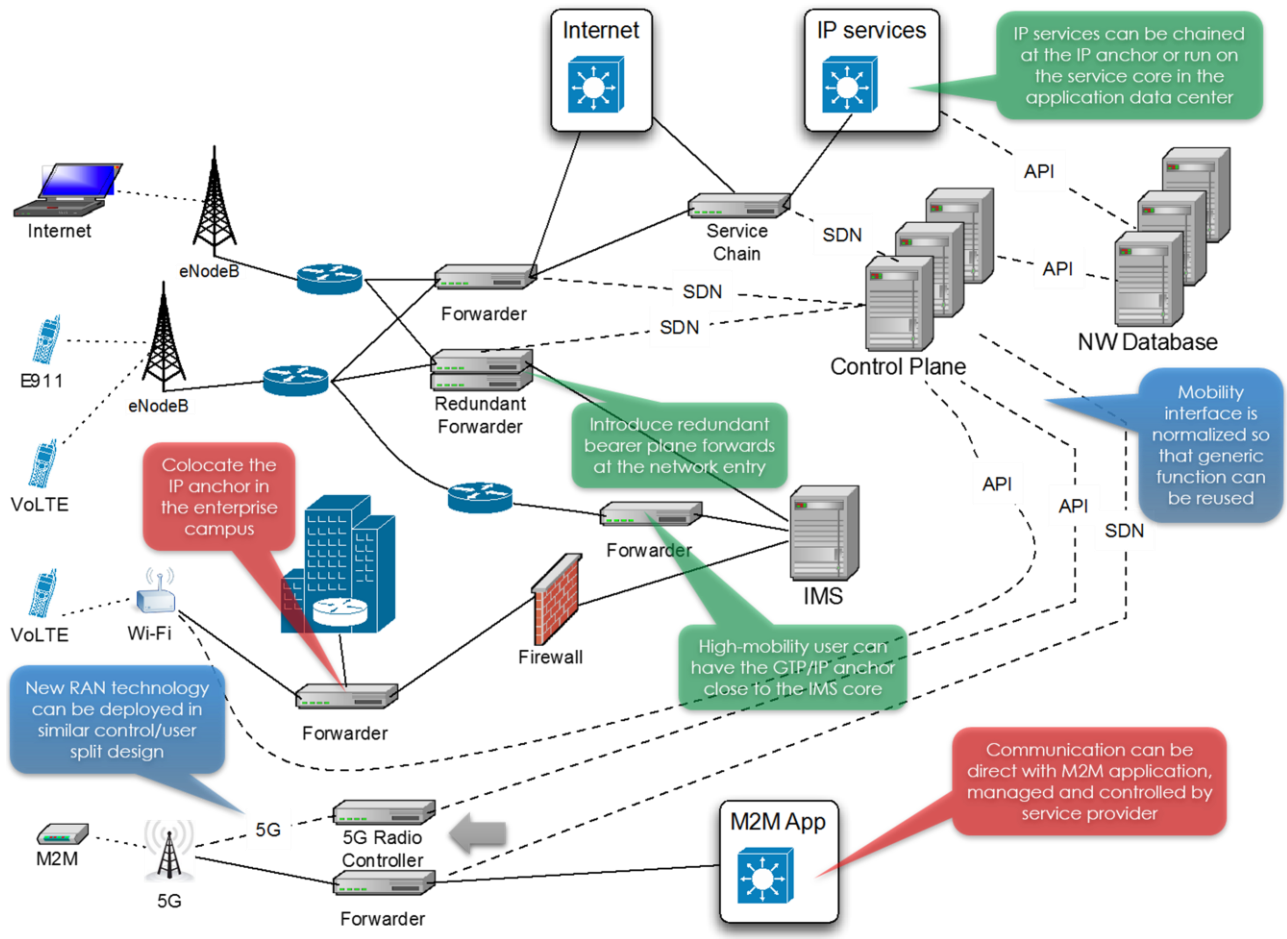
For operational reasons there is a need to keep the distributed data center simple and standardized. In-line services running in the distributed data center alongside the user plane node must not require special capabilities in the local NFV Infrastructure (NFVI) or any additional configuration that would impact on operational efficiency. In cases where the required processing is not available in the local NFVI, sessions can be tagged using the NSH header and forwarded to an appropriate, centralized cloud location.

Conversely, the majority of traffic on an operator network is destined for the Internet and probably needs very little, if any, additional processing. In this case, it can take the fast path across the user plane node (shown in black in **Figure 6**) for onward routing to its destination. In this way, the operator avoids the expense of processing traffic in which it has very little economic interest, in its own data centers. This form of local break-out has proven difficult to implement in the classic EPC architecture.

Deployment of Distributed User Plane Nodes

The new architecture offers deployment flexibility. Notably, the user plane node and associated SGi-LAN IP services can be deployed on a relatively simple NFVI – perhaps just one 2RU server – at many different locations. As shown in **Figure 7**, this could enable, for example, a user plane node with relevant SGi-LAN functions to be deployed on premises for an enterprise customer or as a dedicated deployment for an Internet of Things (IoT) service or public safety network. In some cases it may not be important to distribute the node, but to place it close to the services the customer is using – for example, a user plane node could be deployed alongside an a IMS infrastructure for voice and rich communications services.

Figure 7: Distributed User Plane Enables New Service Models



Source: Cisco, Heavy Reading

In principle, in an orchestrated NFV environment, the user plane node and the various SGi-LAN functions, can be instantiated on demand at distributed data centers equipped with an NFVI. This "programmability" of the infrastructure will be primarily useful to change or add services to already deployed locations, but on occasion could be useful to meet unexpected or temporary demand (e.g., a music festival).

Selection of User Plane Nodes

When a user connects to the network, it must associate with one or more user plane node, according to security policy, service type, and so on. Selection of the correct node (or pool of user plane nodes) is important for efficiency and load balancing, and to enable different service types – for example, a corporate user may have to connect to a user plane node associated to a specific virtual private mobile network. In this way, node selection can be important to the creation of an end-to-end "network slice" and to private, secure networking. (Although note that multiple slices can also be supported on one user plane node if needed).

Classically, a gateway selection is made by the MME when a device first attaches to the network. The new distributed "CUPS" architecture follows the same, standard 3GPP mechanisms to select a controller (which may now include the MME). The controller then selects the user plane node the subscriber should connect to, based on availability and service policy. In this way, the node selection process is "service-aware" and maps to the monetization opportunity.

Classic SAE gateways are typically multitenant and users are routed to the correct node using a type of virtual network known as an access point name (APN) – for example, operators often use an Internet APN and a voice over LTE (VoLTE) APN, depending on the service, and in some cases offer enterprise APNs. Another option is to use the multi-operator core network standards (MoCN) to route user traffic to the desired gateway. This is used effectively for RAN sharing and for mobile virtual network operator (MVNO) services, but it requires a specific network ID to be broadcast by the RAN and the user device to have the correct SIM card.

These methods work reasonably well, but are not very granular, have device dependencies, and a generally high configuration and operations overhead, which means they do not scale especially well. To address this, operators are seeking new methods for gateway – or user plane node – selection. One option is to use an emerging specification known as Dedicated Core Networks (DECOR), as this enables operators to direct users to the correct node or gateway without changing either the RAN equipment or the handset configuration (as would be needed using APNs). This initiative appears to have reasonable support and will likely be part of the Release 14 specifications. In the meantime, vendors and operators are working on practical schemes that will enable node selection and load balancing in the near term and can inform longer-term standards development.

Operational Impact

Adopting a new architecture that requires deploying user plane nodes in distributed locations, such as the RAN or on enterprise premises, inevitably impacts the operational model. Whereas previously operators have had a clear demarcation between core and RAN – with different teams, vendors, etc. for each specialty – in this new model, responsibilities will also shift in line with how equipment is deployed and managed. To fully capture the benefits of distributed user plane, operators will need an operations model that reflects the shift in network architecture.

This is, to an extent, already occurring. Many operators are realigning organizational structures to support their adoption of NFV and some are already conceptually committed to a distributed cloud model, with SDN connectivity between locations, and are creating teams with lines of responsibility and expertise accordingly. The Mobile Cloud Service Core is very much encompassed by this broader organizational change.

Toward a 5G Core Network

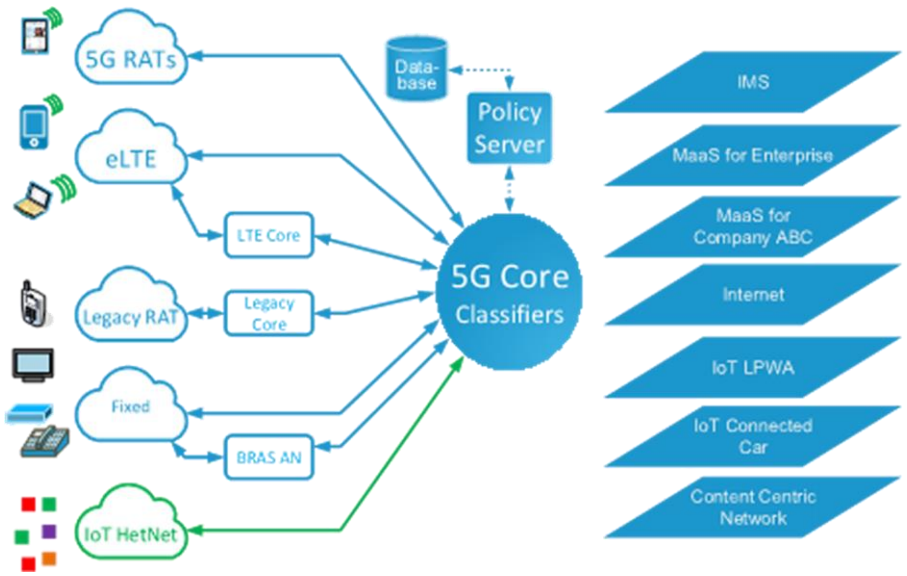
With 4G-LTE deployed for over five years, progressive mobile operators are now considering the next-generation mobile technology. Guided by the ITU's IMT 2020 process, the industry is now developing technologies and use cases for 5G with a view to standardization and deployment within five years. Several operators plan to launch trial services before 2020. SK Telekom (South Korea), Docomo (Japan), Verizon (U.S.), MTS (Russia) and TeliaSonera (Europe) are good examples.

It is likely that the first 5G radio access standards (scheduled for Release 15 at the end of 2018) will be focused on the radio interface, and specifically on the user plane, and will not include a control plane to set up and manage user sessions. The expectation, therefore, is that the first 5G networks will use the existing LTE network to manage connectivity – known as LTE Assisted Access for 5G – and that the 4G core network will, therefore, need to evolve to support 5G access in the near term before a new 5G system architecture and core is developed. This will probably apply to a limited number of operators, for a limited amount of time. Over time, a common 4G/5G core will be important across the wider market.

5G System Architecture

There is an expectation that to meet the performance requirements specified by IMT 2020, a new 5G system architecture, and related 5G core network, will be needed. This is expected to be specified in Release 16, due for completion toward the end of 2019, to enable commercial launch from 2020 onward. Formal development on the 5G system architecture in standards is at an early stage, but the industry already has ideas about what this new 5G core should look like. **Figure 8** outlines some of the key principles of the new 5G core.

Figure 8: Toward a 5G Core



Source: Cisco

On the access side (left of the diagram) the new core should support multiple RATs, and on the network side should be able to steer user traffic into an appropriate "network slice" for processing (shown to the right).

Varied use cases, with associated performance requirements, mean each service, or customer type, consumes a specific type of network configuration. In 5G a "network slice" is used to describe the end-to-end networking requirements needed for a particular service. Creating and operating slices, through orchestration of network and radio resources, is therefore critical to commercial success.

Features of the 5G Core

Although formal work on 5G system architecture is only recently underway, the major characteristics of the new 5G core are reasonably well known, as follows:

- **Formal separation of control plane and data plane:** This is expected to be formalized in 5G standards and is a logical extension of the "split-gateway" model discussed in this paper. The overall 5G core architecture is expected to be control plane heavy because user plane peculiarities, such as GTP, will be de-emphasized or eliminated altogether.
- **Software-driven:** As a new generation of technology, 5G will be the first major mobile network buildout that will be software based and cloud native from inception. It is widely expected that 5G will leverage expertise and capability from the cloud, SDN and virtualization worlds.
- **Distributed deployment models:** A classic core network, hosted centrally hundreds of miles from the radio access, will not be able to support the performance needed for some higher-value 5G use cases. In these cases it will be necessary to place content and application logic close to the user and, therefore, a distributed user plane node, such as discussed in this paper, would be appropriate. A 5G radio controller, or mobility controller, is also needed, and would be distinct from the user plane.
- **Access-agnostic:** There will be some 5G-specific aspects to the 5G core network; however, the new core will be largely access agonistic in that it should also support evolved LTE access, fixed broadband access, and should interwork with legacy 3G/4G core networks. Many subscriber management functions are similar across accesses types, so this is technically relatively straightforward. More significant is that end-user services are themselves agnostic, and operators will need to offer "follow-the-user" network services across fixed and mobile.
- **Automation:** Perhaps the key lesson from Web-scale cloud providers is the value of automation and the use of software tools to make decisions on how and where to process data without the need for manual intervention. These concepts are infiltrating telecom – for example, via self-organizing networks – but the need for automation in 5G is acute if operators are to meet the cost-of-production targets that will enable them to profitably support new use cases.

About Cisco

Cisco (Nasdaq: CSCO) is the worldwide leader in IT that helps companies seize the opportunities of tomorrow by proving that amazing things can happen when you connect the previously unconnected. For ongoing news, please go to newsroom.cisco.com. Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. A listing of Cisco's trademarks can be found at www.cisco.com/go/trademarks.

For more information on the Cisco Packet Core or other information on Cisco's Mobile Solutions, please go to www.cisco.com/go/lte.