Independent market research and
competitive analysis of next-generation
business and technology solutions for
service providers and vendors

HEAVY READING
WHITE PAPER

Service-Based Architecture for 5G Core Networks

A Heavy Reading white paper produced for Huawei Technologies Co. Ltd.

AUTHOR: GABRIEL BROWN, PRINCIPAL ANALYST, HEAVY READING

A SERVICE-CENTRIC 5G CORE

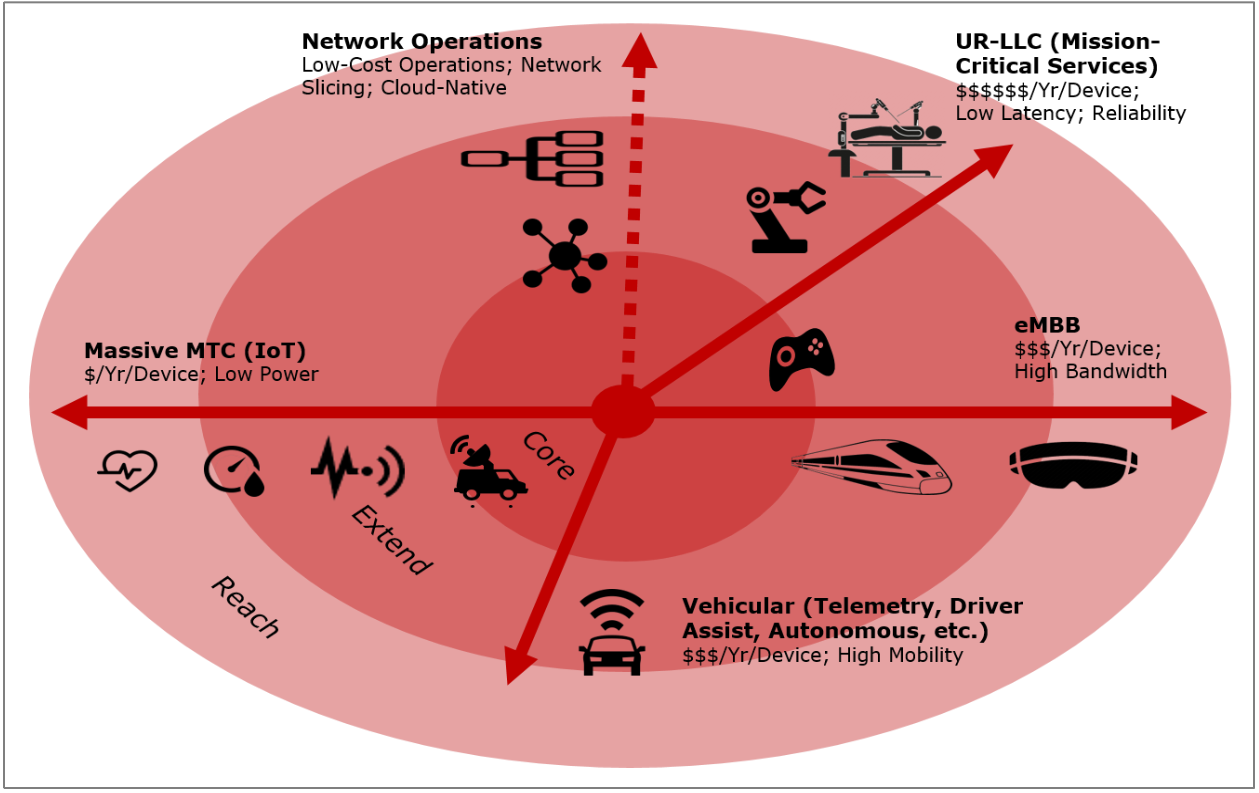
To meet the needs of new services, with diverse and demanding performance requirements, across a wide variety of industries, the 3GPP standards development organization is developing a new 5G system architecture (5GS), including 5G New Radio (NR) access and a new 5G core network (5GC). This new core is fundamental to the commercial success of 5G because it will enable new service types and benefits from cloud economics.

This white paper focuses the service-based architecture (SBA) under development for 5GC and discusses why it is suitable for deployment on cloud infrastructure and can meet future service needs. The paper compares the SBA with the more familiar point-to-point (P2P) architectures defined for mobile core networks. It argues that the SBA is better aligned with the new cloud-native networking models being pursued by leading operators and makes the case for operators to adopt SBA as the preferred option for 5G core network deployment.

Services Drive 5G Architectures

Heavy Reading's firm and established view is that services will drive 5G architecture development and deployment. The opportunity is to support a very wide range of service types, each with associated economic models and technical requirements. By supporting this range of services – examples are shown in **Figure 1** – operators can build a diversified revenue base with massive growth potential. By making 5G inherent to a multitude of consumer and business processes, the industry can transition to a new s-curve of innovation. The new 5G system architecture and core network must facilitate this service diversity at a reasonable cost.

Figure 1: An Expansive Vision for 5G Services



Network Operations
Low-Cost Operations; Network Slicing; Cloud-Native
UR-LLC (Mission-Critical Services)
$$$$$$/Yr/Device;
Low Latency; Reliability
eMBB
$$$/Yr/Device;
High Bandwidth
Massive MTC (IoT)
$/Yr/Device; Low Power
Vehicular (Telemetry, Driver Assist, Autonomous, etc.)
$$$/Yr/Device; High Mobility
Core
Extend
Reach

Source: Heavy Reading

Functional & Service Agility With the Service-Based Architecture

The mobile core network is responsible for functions such as session management, mobility, authentication, and security. These are critical to delivering a service. The 5G system architecture under development in the 3GPP's Technical Specification (TS) 23.501 working group – entitled System Architecture for the 5G System – identifies two representations of the 5GC architecture: one services-based, and one point-to-point based. In the first instance (Release 15), service-based interfaces apply to control-plane functions, while user-plane functions connect over point-to-point links.

The point-to-point architecture has been used in 2G, 3G, 4G and now 5G. In this model, different network functions are connected over standardized interfaces that allow for multi-vendor networks. This is well understood conceptually and operationally and has served mobile operators for decades.

Now, however, with the transition to cloud infrastructure, and the need for greater "service agility," the point-to-point model is no longer the best option. For operators that view 5G as an opportunity for transformative change – both in terms of functionality and cost-per-bit – the SBA looks more attractive.

The challenge with the P2P architecture is that is contains a large number of unique, or quasi-unique, interfaces between functional elements, each connected to multiple adjacent elements. This "tangle" of connections creates dependencies between functions and makes it difficult to change a deployed architecture. If a new function is introduced, or an existing function expanded or upgraded, the operator needs to reconfigure multiple adjacent functions, and test the new configuration, before going live. This raises the business case threshold to experiment with, and deploy, new services.

This is sometimes referred to as "network ossification." In effect, the end-user service is tied to the network and the operators' addressable markets are artificially limited. This is acceptable where the service-set is simple (voice, broadband, etc.), but in a 5G world, where operators expect to offer diverse services and must be able to adapt to fast-changing demand or industry-specific requirements, a more dynamic and agile architecture is needed.

The SBA decouples the end-user service from the underlying network and platform infrastructure and, in doing so, enables both functional and service agility. By virtue of SBA being designed to operate using the cloud model, in which different functions can be composed into an end-to-end service over standardized application programming interfaces (APIs), it is simpler for an operator to add, remove or modify VNFs from a network processing path (functional agility) and create new service-specific service paths on-demand (service agility).

5G SERVICE-BASED ARCHITECTURE

The 5G core is now under development across the industry. 3GPP standards work is currently ongoing as part of Release 15, scheduled to freeze by mid-2018. At the same time, operators and vendors are co-developing systems and using live trials to feed back into the standards process. This iterative development process is a notable difference between 5G and previous generations of mobile technology. China Mobile, for example, has already tested SBA for 5G core and will move on to multi-vendor testing in 2018.

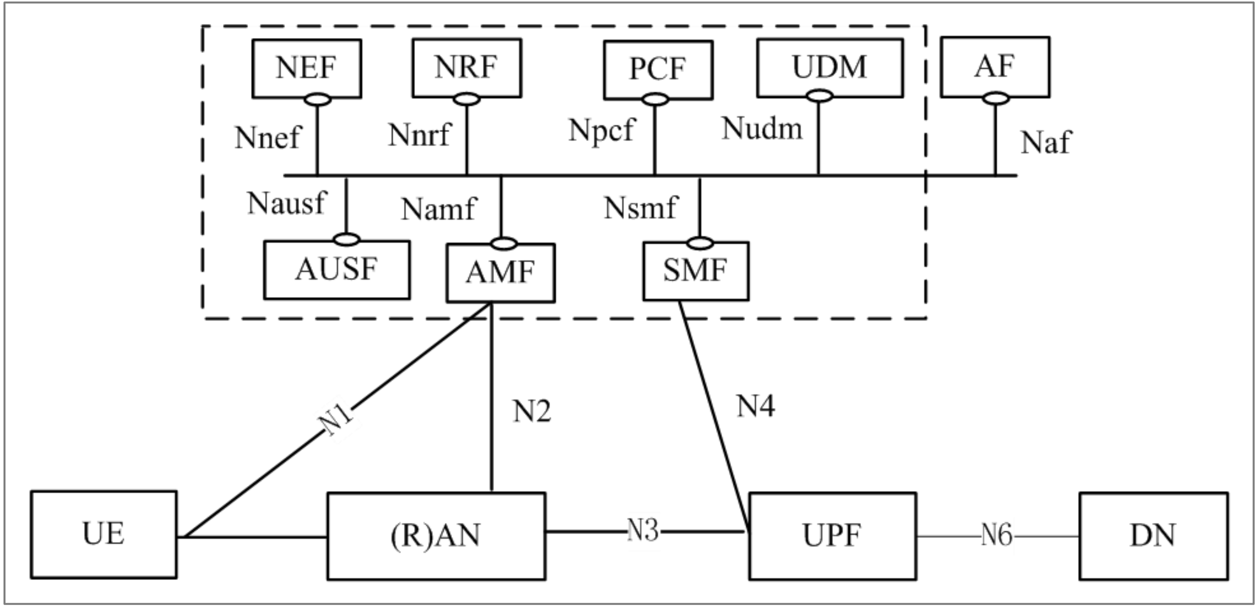
The objective is for operators to be able to deploy 5G in "standalone" mode, using a 5G core and without dependencies on the LTE network, by 2020. Over time, Heavy Reading expects standalone 5GC to become the common core for 5G NR, evolved LTE and fixed access.

Service-Based Architecture for 5GC

The reference document for the 5GC is Technical Specification 23.501, the System Architecture for the 5G System. This document defines the core architecture, the functional elements, and the high-level interfaces between them. Work to define protocols and the detailed interfaces is now also underway.

The non-roaming architecture is shown in **Figure 2**. In this representation of the architecture, control-plane functions, shown within the dotted line, connect to each other over service-based interfaces. The Access Management Function and Session Management Function connect to the user-plane nodes over N1, N2 and N4 to manage subscriber attachment, sessions, and mobility. The N2, and N3 interfaces are determined by how the 5G radio presents itself to the core and, therefore, are dependent on the 5G RAN architecture.

Figure 2: Next-Generation Service-Based Architecture



Source: 3GPP TR 23.501, July 2017, Figure 4.2.3-1

The major components of the 5G core are listed below:

- **Access and Mobility Management Function (AMF):** Manages access control and mobility. The AMF also includes the Network Slice Selection Function (NSSF).
- **Session Management Function (SMF):** This sets up and manages sessions, according to network policy.
- **User Plane Function (UPF):** UPFs can be deployed in various configurations and locations, according to the service type. These are equivalent of GWs in 4G.
- **Policy Control Function (PCF):** This provides a policy framework incorporating network slicing, roaming and mobility management. Equivalent to a PCRF in 4G.

- **Unified Data Management (UDM):** Stores subscriber data and profiles. Similar to an HSS in 4G, but will be used for both fixed and mobile access.
- **NF Repository Function (NRF):** This is a new functionality that provides registration and discovery functionality so that Network Functions (NFs) can discover each other and communicate via APIs.
- **Network Exposure Function (NEF):** an API gateway that allows external users, such as enterprises or partner operators, the ability to monitor, provision and enforce application policy, for users inside the operator network.
- **Authentication Server Function (AUSF):** as the name implies, this is an authentication server (being specified by SA WG3).

5G Core Design Principles

The 5G core architecture is designed to be "cloud-native," in the sense that it should make use of network functions virtualization (NFV) and software-defined networking (SDN) techniques, and use service-based interactions between control-plane functions. As discussed below, SBA aligns well with a microservices view of network function composition.

The expectation, in other words, is that 5GC will be deployed on a shared, orchestrated cloud infrastructure and should be designed accordingly. Some of the key 5GC design principles are:

- Control- and user-plane separation (CUPS) to enable independent scalability and decoupled technical evolution. This will also support flexible deployments, such as at centralized and edge locations. CUPS can also be applied to the EPC in 4G.
- Modular function design. This is a form of functional disaggregation such that a function composed of multiple modules can be created according to the use case's requirements – for example, network slices A and B may have different requirements.
- Minimize dependencies between the Access Network (AN) and the Core Network (CN). This will enable operators to build a multi-access, converged core network, with common AN-CN interfaces, which integrate different 3GPP and non-3GPP access types.
- A unified authentication framework – this is useful in multi-access core, for efficiency and to enable operators to offer "follow-the-user" services, independently of access method.
- Support "stateless" network functions, where the "compute" resource is decoupled from the "storage" resource. This concept is derived from cloud applications. It enables much more efficient creation and consumption of network processing paths.
- Network capability exposure. Exposing information about the network's capabilities to internal and external applications is expected to be more important in 5G. This is especially the case where operators want to integrate 5G with vertical industry processes. Standardizing this makes life simpler for vertical customers, especially those with international operations and multi-operator relationships.
- Concurrent access to local and centralized services. This is to support access to low-latency services hosted in local data centers. Typically, user-plane functions might be deployed remotely, while the control plane is centralized. In very low-latency, mission-critical applications, the control plane may also be distributed.

The actual deployment and operation of the 5GC is independent of specification development and is not prescribed by the 3GPP. Operators and vendors have considerable freedom to implement the architecture in ways that are suited to the use cases or customer.

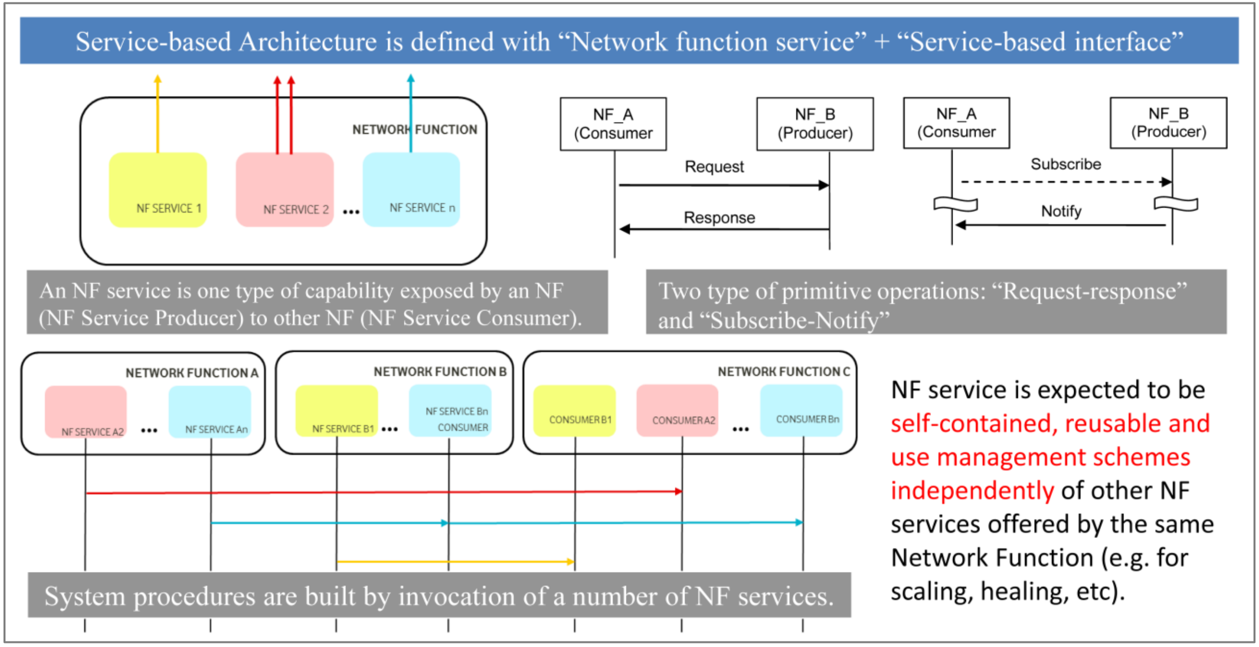
Service-Based Interfaces, Management & Maintenance (Using NRF)

In the SBA, the NF Repository Function (NRF) provides service discovery between individual network functions. It maintains profiles of network function instances and their supported services (for example, function ID, function type, network slice identifiers, capacity information, supported services, and endpoint information such as IP addresses). In this sense, it is an important "pivot" in the SBA.

An example of how the NRF is used is for establishment of a new session. In this case, the SMF discovery and selection request is initiated by the AMF when a request to establish a data session is received from the UE. The NRF is used to assist the discovery and selection of the correct SMF. In a network slice context, the same process occurs: the AMF queries the NRF to select an SMF that is part of a Network Slice instance based on S-NSSAI, UE subscription profile and operator policy, when the UE requests a session to be set up.

Control-plane functions communicate with one another, via the NRF, over service-based interfaces (using HTTP 2.0 transport). **Figure 3** shows how that a network function is itself made up of "network function services" (top left the diagram). These are self-contained software modules that are reusable independently of each other and could be thought of as microservices. The network function is either a producer or consumer of services (top right of the diagram), with two modes of interaction: either a consumer NF can request a response from a producer NF – for example, to request subscriber policy information; or it can subscribe to a producer and be notified if needed – for example, if a subscriber's state changes to inactive mode.

Figure 3: Network Function Services & Service-Based Interfaces



Service-based Architecture is defined with "Network function service" + "Service-based interface"

An NF service is one type of capability exposed by an NF (NF Service Producer) to other NF (NF Service Consumer).

Two type of primitive operations: "Request-response" and "Subscribe-Notify"

System procedures are built by invocation of a number of NF services.

NF service is expected to be self-contained, reusable and use management schemes independently of other NF services offered by the same Network Function (e.g. for scaling, healing, etc).

Source: China Mobile

The end-to-end network service, made up of Network Function A, B and C, contains many NF services that connect to each other over HTPP. Interaction between NFs and NF services is currently under development in 3GPP. Simplistically, a good analogy is to think about how microservices interact to create a cloud applications.

Network Slices & the Service-Based Architecture

Network slicing enables operators to offer logical virtual networks on shared 5G infrastructure. This ability to support multiple customer and service types, each with individual performance requirements, is perhaps the most important commercial driver for 5G. Taken to its logical conclusion, 5G network slices can be thought of as the network adapting itself, in software, to the needs of the application.

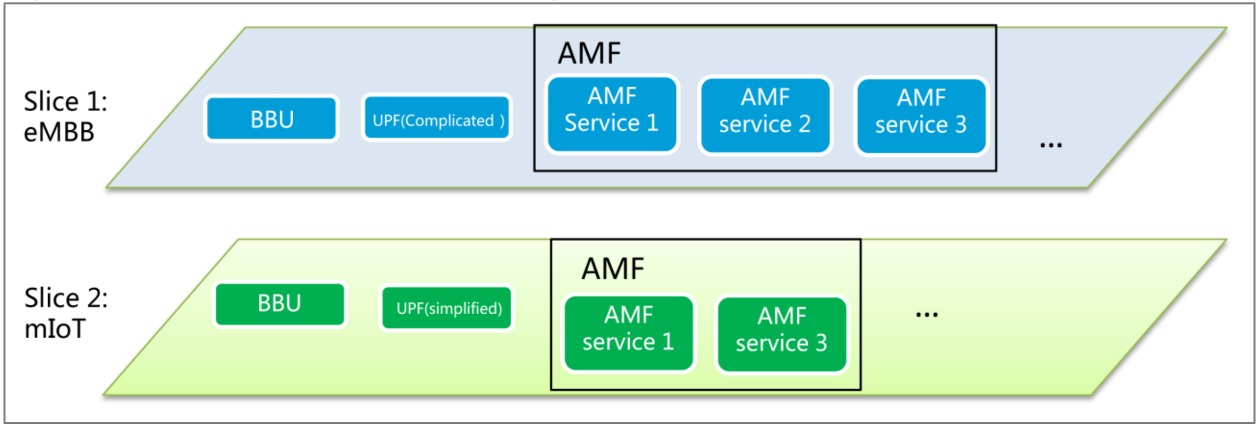
Slices can be fine-grained, at the per-user or per-service level, or can be more coarse-grained, at a company or industry level (e.g., a connected car slice, or meter-reading slice). Examples include a connected car service, an MVNO, an enterprise, a utility, a financial trading slice, and so on. Ideally, the slice should run end-to-end across the RAN, transport and core domains, and extend as far as the UE on one side or SGi services on the other.

A useful way to define a network slice is using an adapted version of the Next Generation Mobile Network Initiative (NGMN) definition, as follows: "A set of network functions instantiated to form a complete logical network that meet the performance requirements of a service type(s)." A network slice consists of three layers: 1) Service Instance Layer, 2) Network Slice Instance Layer, and 3) Resource layer.

Network functions combine to a create an end-to-end service – such as Network Function A, B and C, as in **Figure 3**. The SBA is very suitable for network slicing because it allows for simple reuse of network function services and customization across slices.

In **Figure 4**, for example, the AMF is deployed into different slice types. The AMF in the eMBB slice is incorporates functionality of three different NF services (AMF service 1, 2, and 3) needed to meet the demands of this high-mobility slice. In contrast, the AMF in the mIoT slice incorporates only AMF service 1 and 3 because it is relatively simpler from a mobility perspective. In this way, slices can be created using only the functional components needed to support the specific requirements of the service or customer.

Figure 4: Customized Networks Using Service-Based Architecture



Source: China Mobile

4G CORE TO 5G CORE MIGRATION

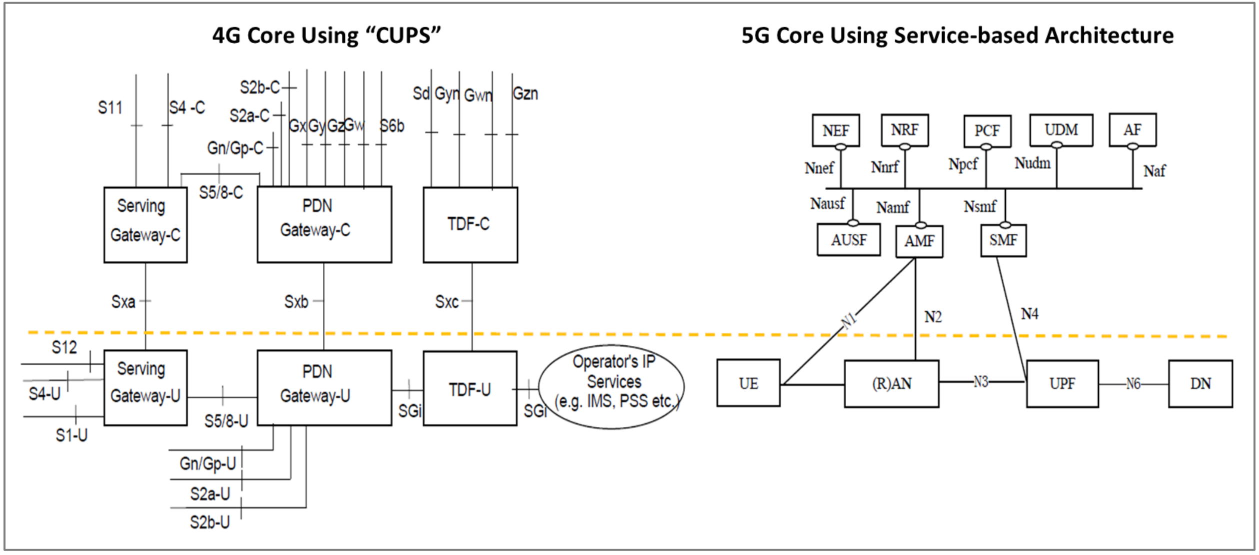
A critical aspect of a 5GC strategy for most operators is to effectively manage the migration from the 4G core, and to optimize investment, timing and technology.

Control- & User-Plane Separation

An important feature of the new architecture is the separation of control- and user-plane processing. This principle of centralized control and distributed processing is used in large cloud platforms to scale traffic and transactions efficiently. It is part of what makes the 5G core a "cloud-native" design. In fact, this trend toward abstracting control functions has been underway in mobile networks since at least Release 4 well over a decade ago and is now, in Release 15, inherent to the underlying 5G system design.

In the packet core, control- and user-plane separation (CUPS) is currently being introduced to 4G core networks ahead of 5G. This upgrade enables the EPC to meet increasing traffic demands at lower cost-per-bit, and to serve low-latency applications hosted in edge locations. It also provides an important migration path from 4G to 5G. **Figure 5** shows the conceptual similarities between 4G and 5G core.

Figure 5: 4G & 5G Core Architecture Comparison

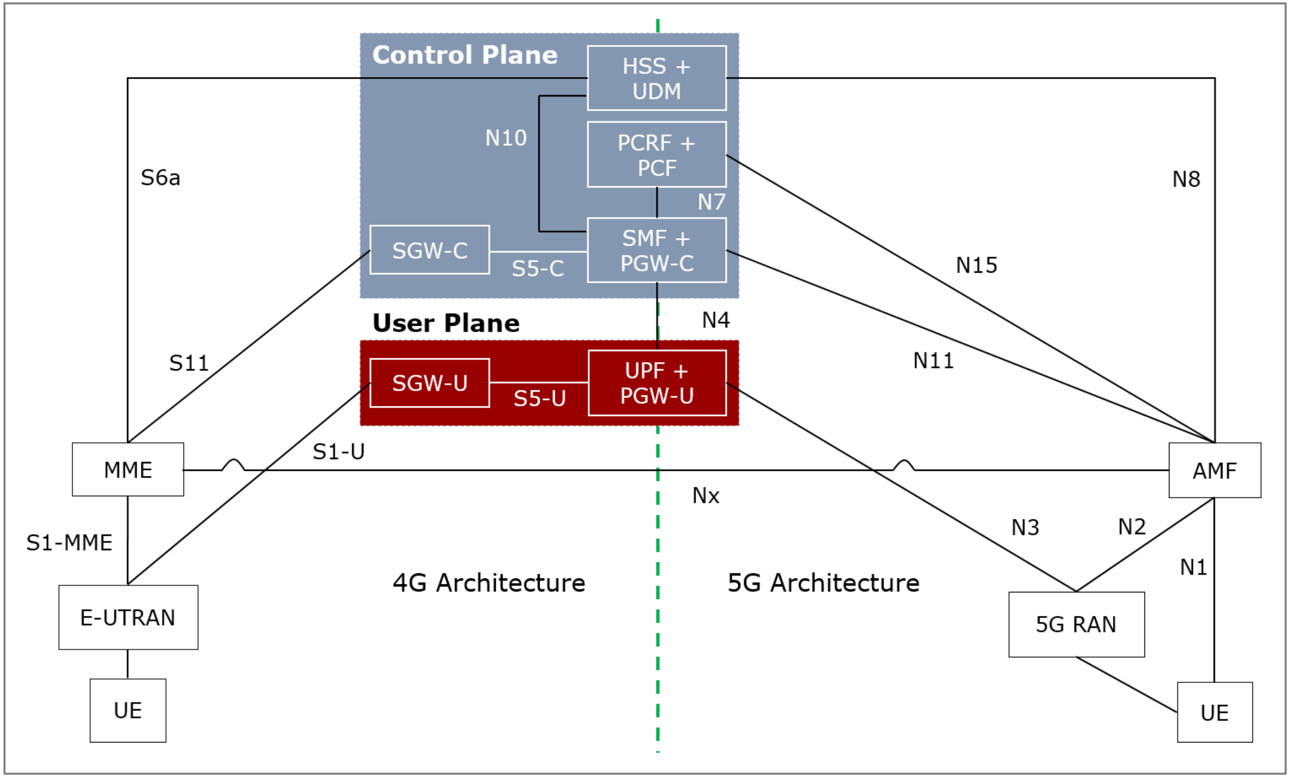


Source: 3GPP, Sprint

This logical mapping between the 4G and 5G architectures can also be applied, in practice, to core network migration. **Figure 6** shows how the network can be configured to serve both network types, using internal or external interfaces according to the deployment. In the user plane, converged "gateways" supporting S/PGW-U and UPF can be used for subscriber termination and traffic forwarding. These may be physical appliances or virtual instances. In the control plane, PGW-C can be combined with the 5G session management function, the PCRF with the 5G policy control function, and the HSS with UDM function. Because there is no MME in 5G (the AMF and SMF now serve these functions), this remains a 4G-specific function.

This hybrid 4G/5G core is a way for operators to migrate investment in the EPC to the new core network as 5G subscribers and traffic grow. It allows them to continue to invest in advanced EPC in the near term, with a roadmap to a full standalone 5G core over time. This is important because operators with LTE-Advanced Pro networks are launching Gigabit LTE, NB-IoT and other advanced services which will drive demand for capacity and performance. This model also supports common traffic processing environment on the SGi and N6 interfaces, such that common firewalls, Web proxies and so on, can be used in both networks. This is sometimes described as a 5G-ready core.

Figure 6: CUPS Evolution to 5G Architecture



Source: 3GPP, Huawei

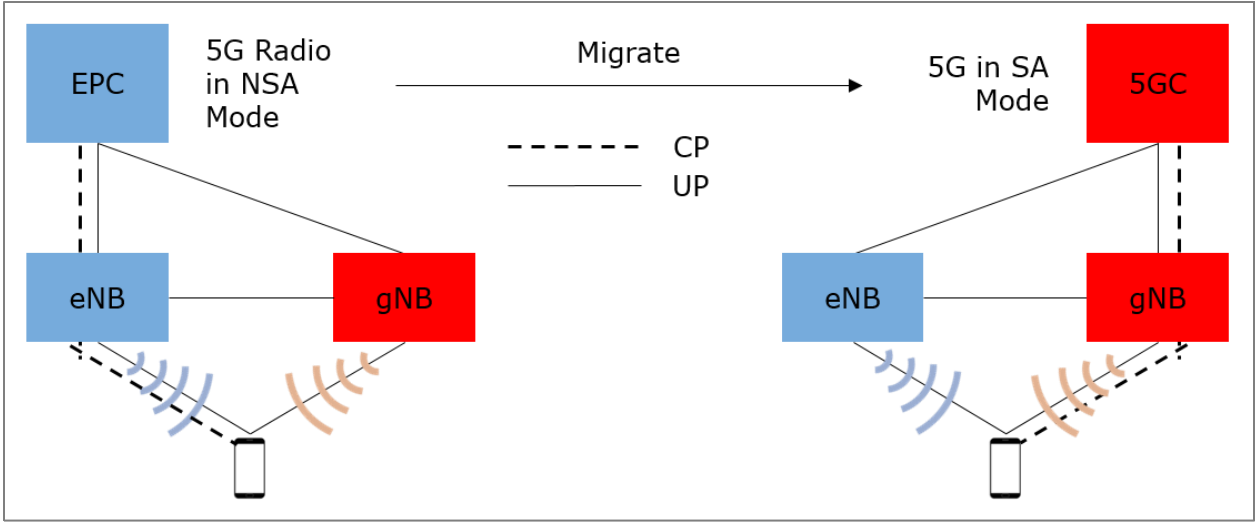
Non-Standalone & Standalone 5G

Many operators expect to introduce 5G in non-standalone mode. This means 5G RAN connected to an LTE network and a 4G core. In this model, the LTE network provides control-plane services over-the-air and in the core network, while the 5G radio is deployed to increase capacity and peak rate throughput. This is shown in **Figure 7**.

This will be simpler and faster to deploy than standalone 5G, and the first commercial services using commercial devices are expected to launch using this architecture in 2019. Over time, however, operators will need to migrate to a new 5G core using the SBA, in order to meet the performance requirements of advanced services and reduce their overall system cost.

To support the data traffic generated by 5G radio, non-standalone mode will require investment in the host 4G core, perhaps linked to the introduction of a CUPS architecture and edge cloud deployment. Operators do not want this investment to be stranded when they move to 5GC and, in many cases, will also want to move their evolved LTE access networks to this new 5G core. This migration strategy, therefore, has both functional and economic benefits.

Figure 7: Non-Standalone & Standalone Modes for 5G Deployment



Source: Heavy Reading

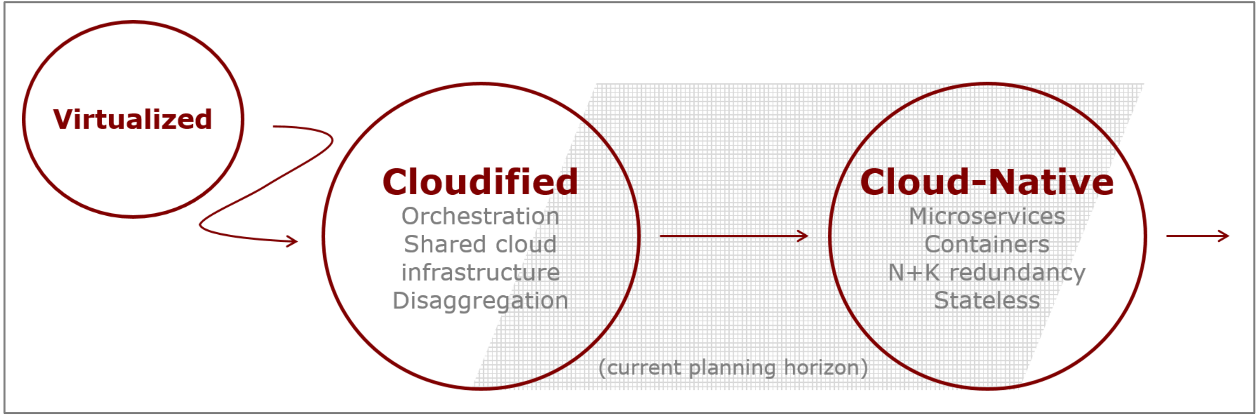
5G CORE DEPLOYMENT

The timing of 5G aligns with the adoption of cloud and SDN technologies in telecom networks. Therefore, to ensure that 5G is future-proof and evolvable, it should be designed for deployment on this next-generation infrastructure. Progressive operators are now deploying the distributed data centers, cloud platforms, SDN connectivity and control and management software on which they wish to deploy 5G. In this context, the 5G core can be thought of as an application on an SDN infrastructure.

Software-Defined 5G Core

This new environment is forcing companies to design "cloud-native" 5GC functions. As shown in **Figure 8**, cloud-native VNFs are designed using disaggregated software components (a.k.a. microservices), and deployed on cloud infrastructure as workloads that can be scheduled on demand by an orchestrator, and can scale in/out on demand.

Figure 8: Toward Cloud-Native VNFs



Source: Heavy Reading

The Cloud Native Computing Foundation defines cloud-native as container-packaged, dynamically managed by a central orchestrator, and microservices-oriented. These apply to 5GC in a number of ways; similarly, there a number of ways in which the SBA is better suited to this infrastructure model, including:

- **Microservices:** By deconstructing VNFs into smaller processes, and scheduling these processing to run on the optimal available infrastructure, operators are able to drive efficiency and performance gains – possibly at a rate greater than Moore's Law improves processing efficiency at the silicon level, according to operators such as AT&T. For example, transactional components of a network function, or network service, would be scheduled to run on cloud infrastructure optimized for compute, while user-plane services could run on servers optimized for packet processing.
- **Stateless Operation:** Most modern cloud applications are stateless, whereas most mobile core network functions are stateful. Retaining state information, such as the status of a user's session, in databases integrated within an appliance or VNF creates interdependencies between VNFs that make it riskier to add or remove functions from the service path. In a services-based architecture the idea is decouple state from the network functions by storing it in a separate high-availability, stateful back-end. This is especially useful in the user plane, making it easier to add and remove virtual gateways (i.e., UPFs) from the network processing path. For control-plane VNFs, a microservice for state store and processing may make more sense.
- **Automated Operations:** To place and schedule workloads in the cloud infrastructure is the task of the cloud orchestration platform. 5G core VNFs should therefore be compatible with these platforms and orchestration tools (Kubernetes, for example, where containers are used as the host; OpenStack where VMs are used, etc.). Features such as shared configuration stores, auto configuration, and automated discovery (such as provided by the NRF) make this easier. With service assurance interworking with domain orchestrators, operators will be able to use closed-loop automation as tools mature and confidence in them increases.
- **N+K Redundancy:** Mobile core network elements are often deployed in N+1 redundant pairs. This is valuable for failover, but is expensive. In the cloud model, the intent is that there is less need for active-active standby, because VNFs can be geographically distributed and rapidly instantiated in the event of failure. Potentially, new predictive analytics and machine learning techniques can be used to identify possible problems and instantiate new instances ahead of a failure. Note, however, operators remain cautious of cloud-level redundancy for mobile core functions.

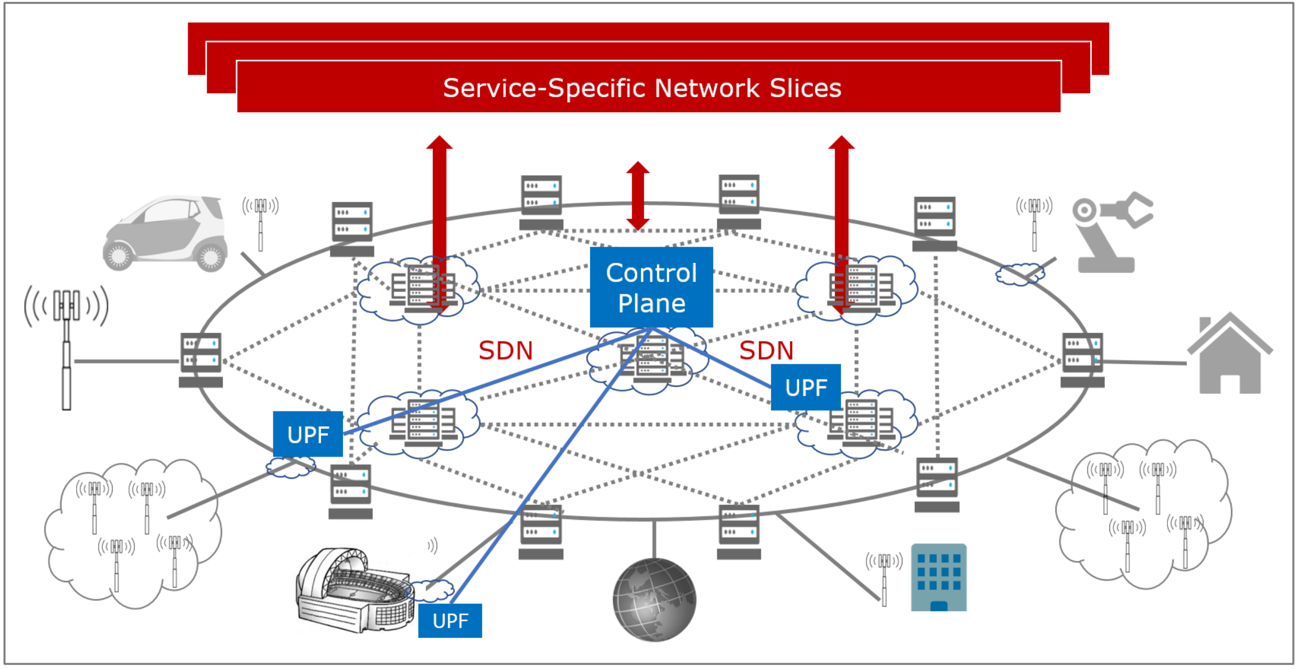
5G Core & Edge Computing

Scalability, resiliency and performance are essential to 5G. Operators are evaluating, and in some cases deploying, edge cloud platforms. These edge locations are suited to 5G because of the extreme low-latency requirements of some service types (such as URLLC applications) and to scale to meet the growing traffic demands.

Operators have a physical footprint in the form of base station controller sites, and transport aggregation sites, that they can convert into micro, distributed data centers. Depending on the service requirements, these data centers can be used to terminate access connections from the 5G RAN and become the obvious place to deploy 5G core functions, especially user-plane functions, and to host latency-sensitive content and applications. Converged fixed and mobile operators have an even greater range of edge locations, such as central offices and local exchanges where they could host 5G network functions.

This in turn drives a need for a high-performance wide-area IP services fabric. This fabric connects distributed data centers in a meshed architecture that provides resiliency and scalability, and should be programmable such that it can support dynamic, service-specific network slices. In effect, the IP services fabric makes distributed centers act in a unified manner – i.e., behave as one integrated data center. The concept is shown in **Figure 9**, with 5GC functions overlaid in blue.

Figure 9: Distributed Data Center Fabric for 5G



Source: Heavy Reading

SUMMARY & CONCLUSION

The 5G core network (5GC) is fundamental to the commercial success of 5G. In addition to the basic functions needed to set up manage user sessions, it enables new and diverse services across many different vertical industries. The service-based architecture for 5GC under development in the 3GPP and across the industry offers many attributes that make it suitable for progressive operators seeking to accelerate deployment of cloud-native 5G networks and services.

This paper has discussed how the architecture is well aligned with key operator requirements and how the ability to easily update the production network by quickly adding and removing functions from a service path enables both functional agility and service agility. As operators deploy 5G in standalone mode from 2020 onward, we expect the services architecture to be adopted in the 5GC control plane.