

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

**HEAVY  
READING**  
**WHITE  
PAPER**

# 5G Service Velocity & the Distributed Cloud

*A Heavy Reading white paper produced for  
Affirmed Networks and Juniper Networks*



AUTHOR: GABRIEL BROWN, PRINCIPAL ANALYST, HEAVY READING

---

## A SERVICES-DRIVEN NETWORK ARCHITECTURE

Service provider networks are on the verge of significant and lasting change. Leading operators have started to re-architect their networks for the services that will underpin their commercial success in the 5G era. Driven by performance (especially latency) and scalability demands, these operators are distributing cloud infrastructure to network edge locations to enable rapid deployment of advanced network services. Around the world, fast-followers are racing to understand how these interrelated themes of distributed cloud and "service velocity" can apply to their networks.

With specific reference to evolved 4G and emerging 5G mobile networks, this paper discusses the role of edge computing, the IP-based "network fabric" that connects edge data centers, and the critical question of "service velocity." It makes the case for deploying edge infrastructure in combination with network slicing mechanisms to enable operators to create and deploy granular services, optimized to the use case and customer, using cloud-native network functions and management tools.

The performance requirements of advanced 4G and 5G services are a key influence on the design, deployment and operation of this new network architecture. By colocating network functions with end-user content and applications at the edge, there is an opportunity to enable service types that would be expensive, or even impossible, in the classic centralized mobile network. Success, of course, is dependent on the operator's ability to deliver and manage these services in a fast and agile manner.

### New Architectures; New Services







Emerging service requirements – such as low latency, extreme connection density, high reliability/availability and gigabit speeds – are driving the design of the new architecture. There are three major categories of application that are likely to be deployed at the network edge:

- **Network Functions:** Changes to RAN and core architectures will result in functions being deployed at edge locations. This is likely to include RAN control, and in some cases baseband, functions colocated with distributed 5G core user plane and critical ancillary functions, such as security and traffic management. Virtual network functions are very demanding in terms of performance.
- **High-Performance Applications:** For ultra-low-latency services, it may be necessary to host applications at the edge. This could include Intelligent Transport Services, Industry 4.0 applications, and in the longer term, tactile Internet services. At the other end of the scale, high-density Internet of Things (IoT) applications can also be aggregated and processed at the edge. Computational capability is often sacrificed to lower the cost of IoT devices; edge infrastructure, with low processing time and low network latency, allows for processing to be offloaded from the device.
- **Video Content & VR/AR:** Video caching is the original mobile edge use case. Operators use CDNs to reduce transport costs and improve the subscriber experience through lower "time to start" for video streams and reduced video stalls and rebuffering events. Video is also becoming much more than the classic viewing experience: machine vision, augmented reality (AR) and virtual reality (VR) promise new experiences (in stadiums, in the workplace, etc.) and will need high-performance delivery mechanisms and computation offload. Edge computing, in combination with 5G access, can help unlock the mobile VR/AR market across consumer and professional sectors.

## From URLLC to Massive IoT

A range of 5G service types, and related performance requirements, are described in the 3GPP document "Service Requirements for the 5G System" (TS 22.261). Many require the low-latency networks and high connection densities enabled by distributed cloud. Some example service types, with promising commercial outlooks, are shown in **Figure 1**.

**Figure 1: 5G Service Examples for the Distributed Cloud**

<b>Dense Urban MBB</b>  User data rate: 300 Mbit/s DL Traffic density: 750 Gbit/s DL km <sup>2</sup> User density: 25,000 UEs km <sup>2</sup>	<b>Electricity Distribution</b>  End-to-end latency: 5 ms Jitter: 1 ms Survival time: 10 ms Availability: 99,9999% User data rate: 10 Mbit/s	<b>Mobile AR/VR</b>  Immersive video: 6 DoF Ocular reflex: <15 ms latency 2K stream: 68 Mbit/s 5K stream: 8 Gbit/s
<b>High-Speed Train</b>  Per train: 15 Gbit/s DL & 7.5 Gbit/s UL Users per train: 1,000 @ 30% activity rate Mobility: up to 500 km/h	<b>Intelligent Transport</b>  End-to-end latency: 10 ms Availability: 99,9999% Service area dimension: 2 km along road Traffic density: 10 Gbit/s per km <sup>2</sup>	<b>Robotic Motion Control</b>  End-to-end latency: 1 ms Jitter: 1 µs Density: 100K UEs per km <sup>2</sup> Service area: 100 x 100 x 30 m

Source: Heavy Reading

Many of these services can only be delivered from the edge cloud. Robotic motion control is an example of an ultra-reliable low-latency communications (URLLC) service type. Being sensitive to packet loss and jitter, and with a ~1 ms end-to-end one-way latency requirement (which means ~300 µs network latency), it can best be served from highly distributed locations, such as the factory floor – in fact, the design targets for this application are based on a service area of just 100 x 100 x 30 meters.

VR/AR applications are slightly more forgiving in terms of latency and but are nevertheless best served relatively locally (e.g., a stadium or network edge). To meet the "6 degrees of freedom" (DoF) needed to produce immersive experiences requires less than 15 ms round-trip latency to avoid motion sickness and 68 Mbit/s for a 2K stream at 60 frames second. A theoretical (today) 5K stream at 120 frames per second would need 8 Gbit/s.

The IoT is often associated with non-critical, small-data services, which can benefit from edge processing to manage high connection densities (up to 1 million connections per km<sup>2</sup> is the design target for 5G) and to support low-power devices. In other cases, IoT refers to mission-critical services such as high-voltage electrical grid monitoring, Intelligent Transport Service, and high-speed rail, each of which need specialist, distributed cloud infrastructure to be deployed in the grid, roadside or railway service area.

## Service Velocity & Cloud-Native Networking

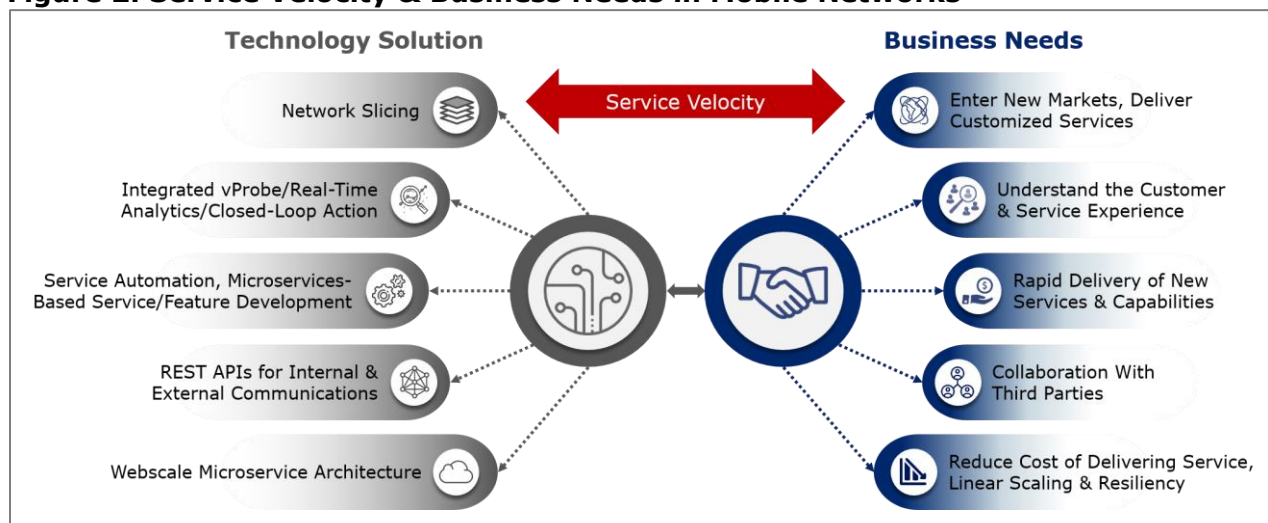
Services and architecture evolution go hand-in-hand. It is not sufficient to speculate on future service needs and then build supporting infrastructure; operators also need to think in terms of "service velocity" and how to go from ideation, to creation, to marketing of a new product in a short time period. Large cloud providers have shown how services can be

launched quickly, and then scaled rapidly if successful, or quickly replaced if market fit does not emerge, and how infrastructure doesn't have to change for each new service launch.

This performance of cloud providers, commercially and technically, is a source of inspiration for network operators. From a customer perspective, operators should take these lessons and apply them so as to make network services as easy to consume as cloud services: An IT department, for example, should be able to configure and manage a mobile enterprise service through a console in the same way it can buy and consume cloud services.

**Figure 2** shows how business and technology must combine to enable service velocity. The technology solutions derive from greater "softwarization" of the network and the ability to decouple service configuration from the underlying hardware resources. This capability enables granular and dynamic network slicing; however, to manage large numbers of slices within a limited operating budget requires automation. This in turn requires visibility into the service and infrastructure stack, through telemetry and analytics, which in combination with network programmability can, over time, enable closed-loop automation.

**Figure 2: Service Velocity & Business Needs in Mobile Networks**



Source: Affirmed Networks

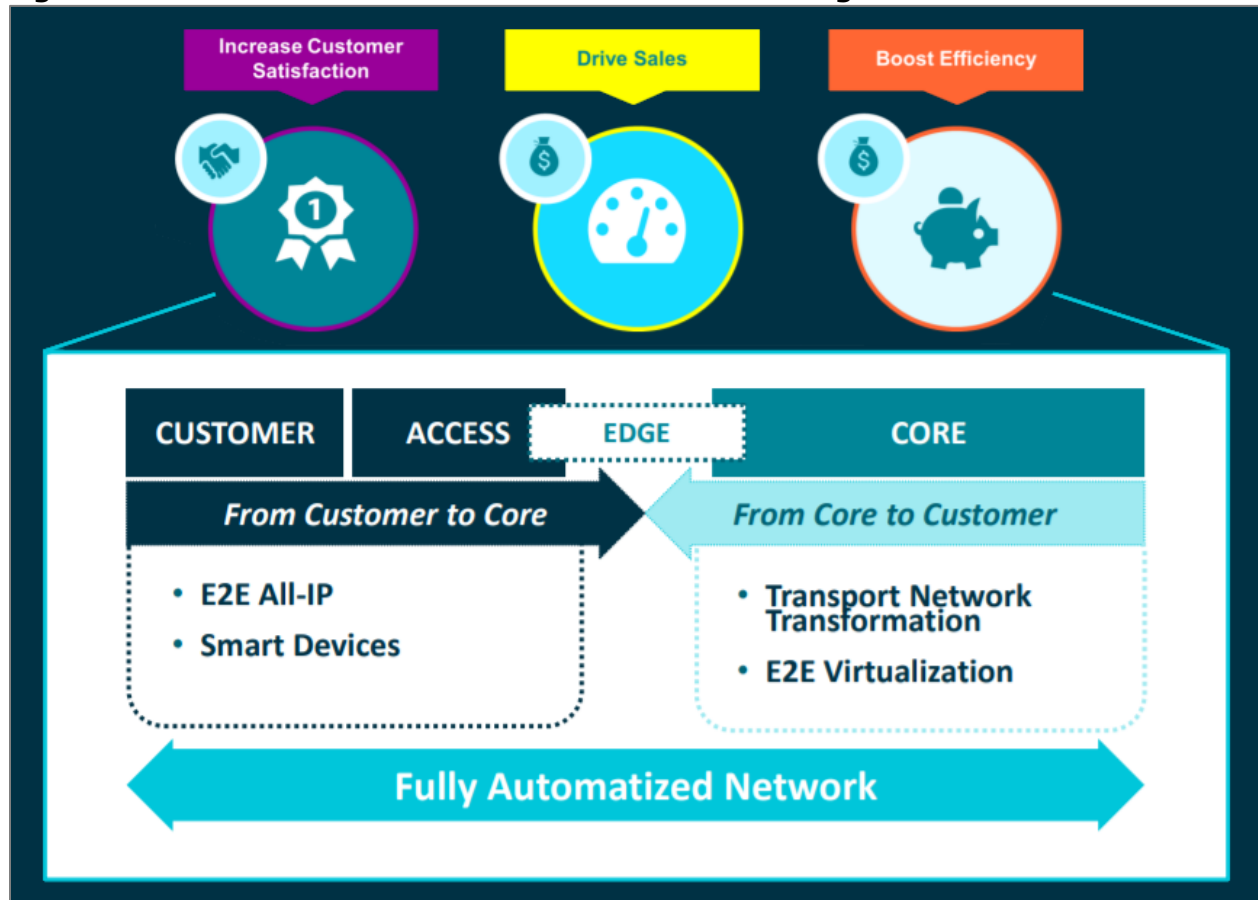
For operators, 5G also posits a change in service mix, with a swing to vertical industries. Consumer remains important, of course, but is less dominant and may also be delivered through an enterprise channel (e.g., connected car services via an insurance company or vehicle manufacturer). This means that many of the commercial opportunities in 5G are specific to particular industries and companies. Accordingly, there is substantial opportunity in the enterprise long tail. To serve these users, operators need to create "network slices" that offer unique services, optimized according the customer type and use case.

As discussed later in this paper, a major change in 5G is the adoption of service-based interfaces in the core network. These REST APIs, common in the Internet and cloud worlds, make it easier to change and configure network services in software. Over time it is possible that certain interfaces could be exposed externally – for example, such that an enterprise could control user policy within the operator's 5G network. There is a direct link between cloud-native 5G core design, network programmability and the ability to offer differentiated, customized services without escalating operational costs.

## Technology & Business Converge at the Edge

The business imperatives of 5G and associated technology enablers converge at the edge. As shown in **Figure 3**, customer requests for services with demanding performance and economics are best served from the edge, while what were previously core network functions are being distributed to meet these new demands.

**Figure 3: Network & Business Transformation at the Edge**



Source: Telefónica, MWC 2018

To achieve both efficiency gains and greater sales success requires automation of end-to-end network processes and in how customers interact with the operator to order, consume and pay for services. The chart and concept are, obviously, high-level, but this line of thinking is driving long-term change to network architectures and operating models.

## DISTRIBUTED CLOUD-NATIVE NETWORKS

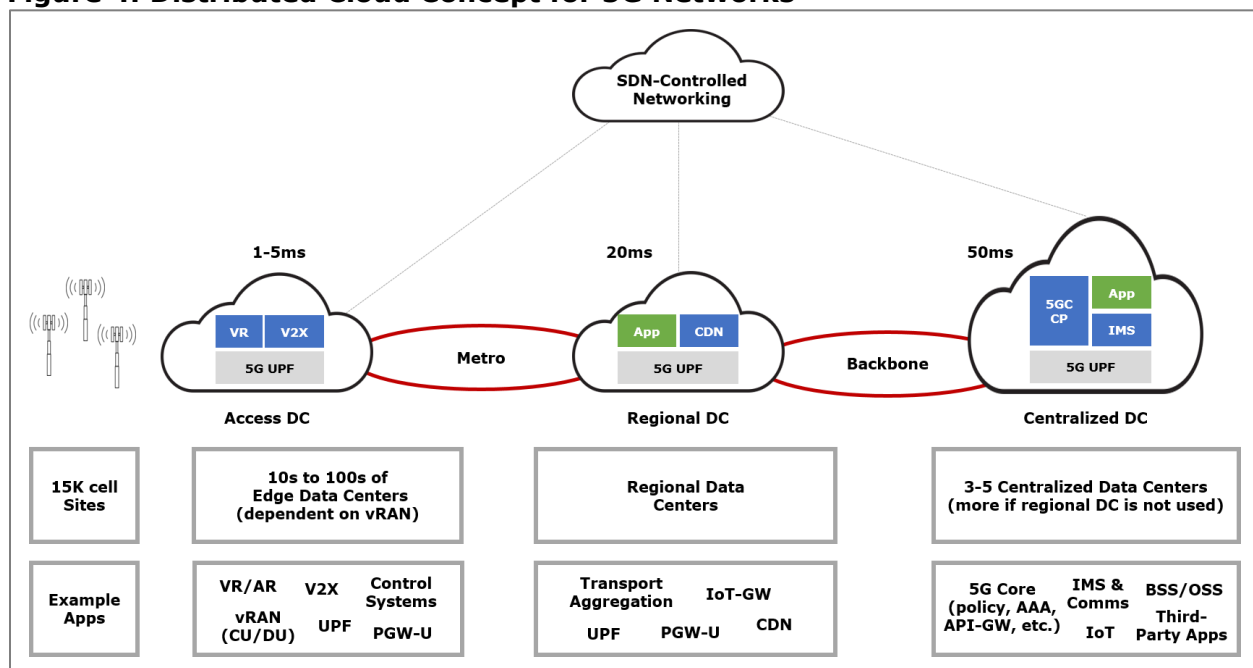
Ownership and control of access and metro networks is a unique source of competitive differentiation for network operators. The opportunity is to use these assets to host services closer to the user to improve performance, reliability and cost structure. Operators need to carefully evaluate how much, and where specifically, they invest in edge cloud based on the performance requirements of the proposed applications and the revenue potential.

## Distributed Service Provider Cloud

The reinvention of service provider networks and the adoption of cloud-native architectures is inspired by hyperscale cloud infrastructure. By applying cloud technologies and processes to the unique circumstances of network owners, operators can take advantage of having direct physical access to customers to build assets that generate long-term, defensible value.

The distributed cloud concept for 5G networks is shown in **Figure 4** (in this example referencing an operator with 15,000 cell sites). The operator deploys cloud infrastructure at multiple locations connected over SDN-controlled networks to create a unified "network cloud" in which services (workloads) can be hosted and moved according to performance needs, available capacity and other criteria determined by service and network policies.

**Figure 4: Distributed Cloud Concept for 5G Networks**



Source: Heavy Reading

An access data center, for example, might be used to host mobile VR and V2X services; the regional data center might host a CDN and third-party applications (shown in green); and the centralized data center might host IMS communications services, 5G core control plane and OSS/BSS. 4G/5G user-plane functions (UPFs) can be deployed in each data center, perhaps with multiple instantiations per location, to terminate the radio access traffic and apply IP services (routing, security, etc.). The wide-area SDN network is responsible for routing traffic from 5G access to the correct UPF location. More on this in the next section.

The major components of the distributed service provider cloud are as follows:

- Cloud-native network functions:** To be agile, and to scale in the cloud, 5G core network functions need to be (re)written as cloud-native applications. This means decomposing functions into smaller microservices that can be deployed across a multi-vendor, multi-site cloud infrastructure. This has major implications for resource efficiency, redundancy, scaling, etc.



- 
- **Distributed, yet unified, cloud infrastructure:** Deploying storage and compute at edge network locations is the foundation of the distributed service provider cloud. However, operators also need a common software environment (and "single pane of glass") to enable them to deploy services where and when they need to; it is less useful to have multiple cloud environments, with differing management tools. This is challenging because edge locations demand simpler technologies, with minimal operating overhead, that are nevertheless capable of high performance.
  - **Network fabric for distributed inter-data-center connectivity:** Connectivity between edge centers, the access network and the core is fundamental. The requirement is that this network is software-controlled, secure and dynamically configurable. These twin capabilities of inter-data-center connectivity and SDN-controlled networking are critical to the distributed cloud. For integrated wireline and wireless operators, this is also a multiservice network, and is not necessarily 5G-specific.

## Edge Cloud PoPs – How Many?

Network operators have unique physical assets that can be converted into micro data centers to create distributed, highly available clouds. Example locations include cell sites, street cabinets, transport aggregation sites, central offices/local exchanges, enterprise premises and public venues. Which locations, and how many, to upgrade, is dependent on geography, the existing transport infrastructure and demand density.

Generally speaking, ex-incumbent converged wireline/wireless operators have more sites that are candidate locations for distributed cloud – their challenge is to select which sites are most useful from a wealth of options. Pure-play mobile operators, especially challenger operators, with less owned transport network, may need to acquire/lease edge facilities

For the access data center, latency is the major determinant of location choice and how many edge facilities an operator needs. As an approximate guide, to consistently deliver 5 ms or less of one-way latency, distances up to 30 km or so from the end user to the edge data center is the sweet spot.

To achieve low latency nationally across an owned transport network, a typical mid- to larger-sized European operator may need only a dozen or so distributed sites. However, in practical terms, at a minimum, deployments in major cities are likely to be needed. There will also be cases where the operator needs to deploy on-premises at the enterprise, stadium, airport, factory or other venue.

The number of edge data centers the operator needs is largely determined by RAN architecture. If the operator is pursuing a virtual RAN/centralized RAN architecture, it typically needs more locations than an operator with a classic RAN deployment. And even where virtual RAN/centralized RAN is deployed, the nature of the functional split between distributed unit (DU) and centralized unit (CU) impacts latency requirements and, therefore, the data center topology. A lower-layer split requires very low latency and therefore very high-quality transport and more access data centers, versus a higher-layer split, which is more tolerant of latency and therefore requires fewer access data centers.

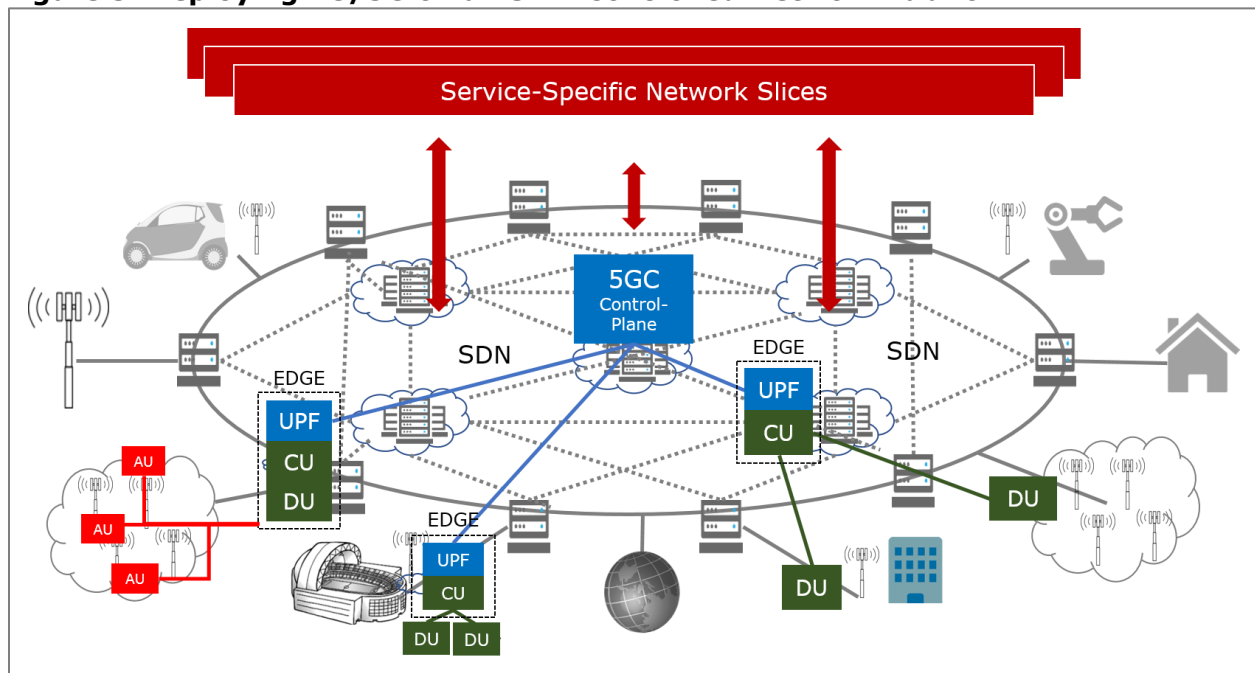
There is debate about the role of a regional data center, which would typically be an important transport aggregation site close to an Internet PoP, relative to access and national (centralized) data centers. One issue is that if the regional data center, shown in **Figure 4**, adds an order of magnitude of latency compared to the access data center, then traffic might

as well continue to the centralized data center, which means a two-tier data center architecture is likely to be sufficient for most operators.

## 5GC Deployment Architecture

The 3GPP specifies a logical 5G architecture, which includes RAN and core components, that must be implemented and then deployed on a physical network. **Figure 5** shows a high-level view of a network fabric used to interconnect edge cloud locations with access and core networks. 5G RAN and core functions are deployed on this "network cloud" platform.

**Figure 5: Deploying 4G/5G on an SDN-Controlled Network Fabric**



Source: Heavy Reading

Some important aspects of this distributed 5G deployment architecture are as follows:

- A new 5G core and/or 4G core upgraded to CUPS, as shown in blue. This means deploying distributed user-plane functions (UPFs) to terminate GTP-U tunnels from the access network at an edge location. Generally speaking, the 5G control-plane will be more centralized than the user plane, although in some cases both control and user plane will need to be locally deployed (e.g., for a robotic control system).
- RAN controllers, shown as CUs, are deployed at the same location. This is enabled by the new 5G RAN architecture split into distributed unit (DU) and centralized unit (CU), both shown in green. The idea is to deploy CUs at the same locations as the UPF (or in 4G terms, at the PGW-U location), creating an opportunity to map network packet processing to radio resource control to improve service quality and reliability for critical applications. This integration will be important to high-value network slices with demanding service-level agreements.
- In some cases, the operator may also deploy DUs (roughly equivalent to baseband units) at the same edge location and use eCPRI transport to connect to the antenna



---

units (AUs) shown in red. This is similar to a 4G centralized RAN using CPRI. In other cases, DUs will be deployed closer to the cell site in a shelter or cabinet – or even integrated in the antenna itself – in which case, carrier-grade IP/Ethernet transport between CU and DU is sufficient.

There are, naturally, many variations on this model. As indicated previously, in some cases GTP-U tunnels from the access network will terminate at a more centralized location, or where the operator has an Internet PoP. Moreover, distributed data center locations will vary based on operator circumstance, the capabilities and footprint of the underlying transport network, the nature of the edge facilities, and the service type itself. A machine vision application for factory automation, or, say, a facial recognition security application at airports, may need a core network deployed on-premises. Similarly, a mobile VR application at a sports stadium would likely see 5G network functions deployed locally at the stadium itself. A strength of the 5G architecture is to allow these forms of flexibility.

Once this physical deployment is achieved, the operator must then be able to overlay discrete, secure network slices across the entire network. Mapping 3GPP-defined slices into the SDN transport network is critical and requires multilayer integration. Today, the U.S. market leads investment in wide-area SDN for 5G.

## SERVICE VELOCITY IN THE 4G/5G CORE

To deliver high-value, high-performance services, operators need mechanisms and toolsets to deploy and manage 5G network functions across this distributed infrastructure. This includes where to place workloads, how to scale them and how to make services easily consumable by customers in the form of network slices.

### A Service-Based 5G Core Architecture

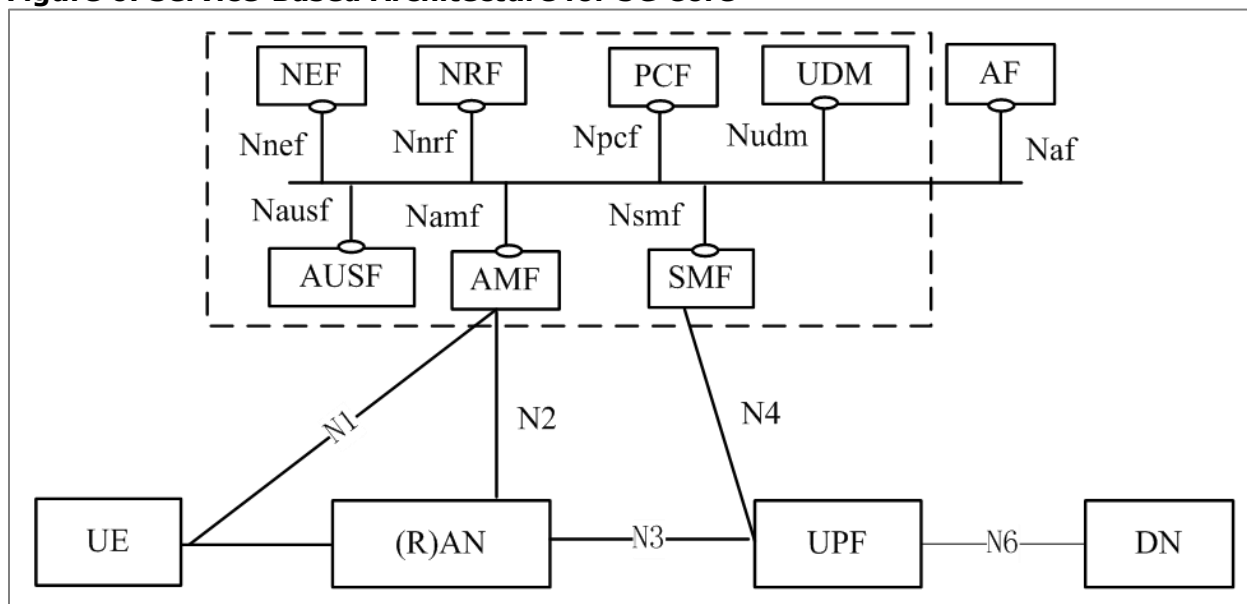
The 5G network architecture, including core network, is specified in TS 23.501 "System Architecture for the 5G System." This document defines the overall system and identifies some key design principles, including that the 5G core should be: 1) cloud-native; 2) distributed; 3) state-efficient; and 4) multi-access. These principles are incorporated in the service-based architecture (SBA), shown in **Figure 6**.

The SBA formally separates control plane and user plane to allow independent scaling and deployment. This is an extension of the CUPS concept in 4G core that, among other things, enables deployment of user-plane functions in the edge cloud. This decoupling will encourage innovation in virtual user plane and associated IP services. Transporting GTP-U over the S1 (4G) and N3 (5G) interfaces and then terminating it at the distributed cloud location allows operators to deploy additional services, such as security and traffic management, close to the user. These were often known as *Gi/SGi* services in 3G/4G networks.

The control-plane functions, shown within the dotted line in **Figure 6**, connect over "service-based interfaces," which are REST APIs over HTTP 2.0 transport. The use of REST APIs is a game-changer in the mobile core because it reduces the dependencies between functions that proliferated in the 4G core, making it faster to add and remove instances from a service path – in effect, core network software becomes more like a modern cloud application, in that functions can be added or removed from the service path rapidly, without impacting

adjacent functions. This is useful for standard maintenance and upgrade tasks and for more dynamic service management because it provides operators with the ability to adapt or optimize a network function and then quickly deploy the new version into the live network. This capability is useful, for example, to deploy and modify network slices customized for different enterprise users or service types.

**Figure 6: Service-Based Architecture for 5G Core**



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

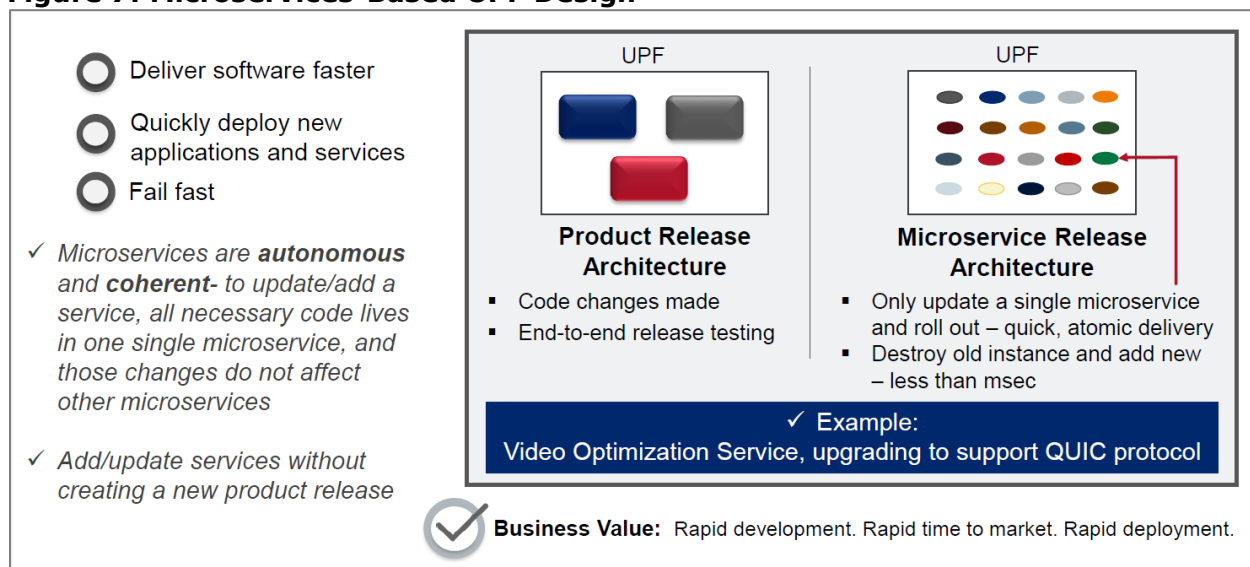
## Microservices for 4G/5G Core

By adopting cloud principles and processes, operators can increase service velocity. This starts with the use of "cloud-native" network functions using microservices. Mobile core functions and interfaces are standardized by 3GPP; however, vendors are free to implement them as they see fit. Some progressive vendors, supported by progressive operators, are now offering mobile core functions composed of multiple microservices. This has important, and very positive, implications for end-to-end service design and operation.

**Figure 7** shows an example of two virtual UPFs: to the left is a classic VNF design made up of relatively large "monolithic" components; to the right is a microservice design. The monolithic design would be deployed in VMs, typically centrally, and is characterized by relatively long release cycles. This means going from ideation of a new feature to commercial operation can be a lengthy process (typically six months or so), because to make even minor changes to the code base requires extensive regression testing before a new or updated function can be deployed commercially. The microservices-based UPF, by contrast, is made up of smaller functional modules that can be upgraded independently of each other, radically reducing the time from ideation, to software development, to commercial deployment.

Microservices-based design concepts are now commonplace in the cloud world; their adoption in telecom is a recognition that software change is inevitable in networking. With a cloud-native network design, operators are more able to adapt and seize the opportunities in the dynamic online services market. This is a form of DevOps for telecom.

**Figure 7: Microservices-Based UPF Design**

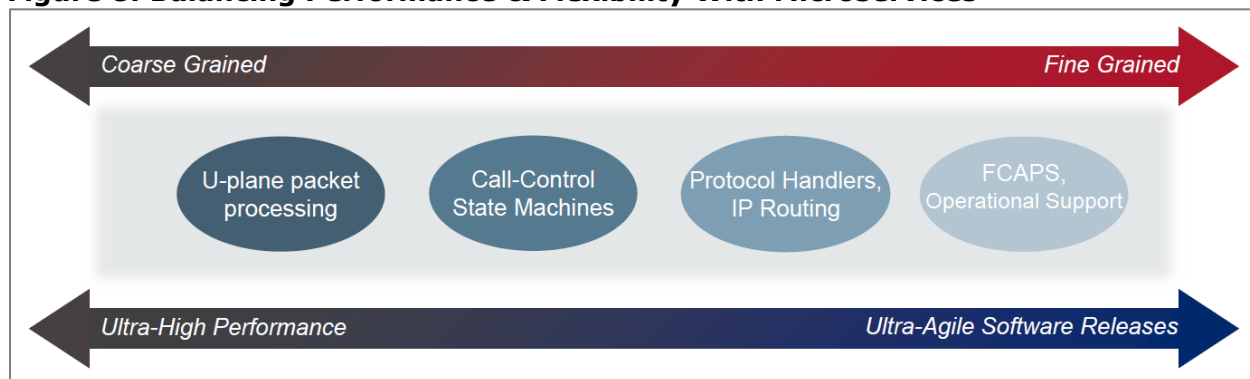


Source: Affirmed Networks

## Flexibility & Performance With Microservices

Not all network functions are created equal. **Figure 8** shows how different functions use microservices and how they scale, according to the task's requirements. To the left of the spectrum is a virtual UPF: This is a user-plane function that should be optimized for high-performance packet processing and is coarse-grained, in that it incorporates a relatively low number of microservices and would typically need relatively few customizations and changes. To the right are control-plane-centric functions: These are much more likely to require customization and be finer-grained – for example, each operator may need to adapt FCAPS modules to integrate with their specific OSS.

**Figure 8: Balancing Performance & Flexibility With Microservices**



Source: Affirmed Networks

Being able to optimize network functions using microservices is very powerful – e.g., being able to scale call-control state machines, or user-plane forwarding, according to the needs of, say, smartphone users vs. IoT devices allows for efficient operation of the end service. This means, in effect, that operators have an opportunity to increase performance of mobile core networks at a rate faster than Moore's Law. There is also a direct relation between the time

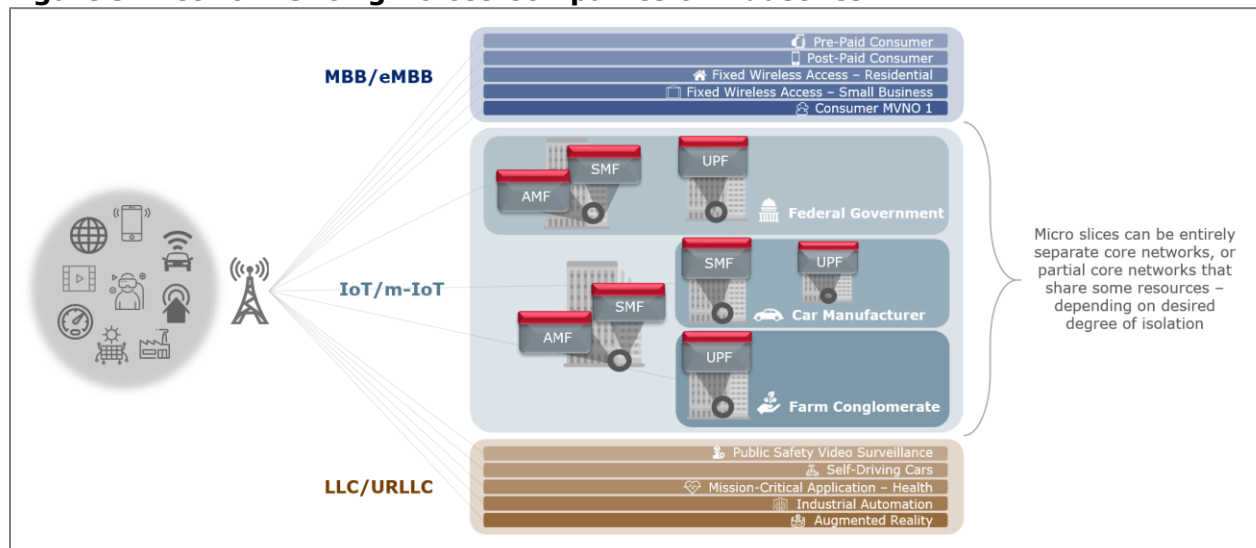
it takes to test updates to individual microservices relative to monolithic applications. Microservices have fewer lines of code, and hence shorter test cycles, and less need for regression testing of code not impacted by the update. This reduces test time at all levels of integration.

## Network Slices & Microservices

The target for operators is not to redesign their networks, but to better serve customers and seize market opportunities. Many 5G opportunities are in the enterprise and are often specific to companies and industries. Operators, therefore, need a way to address the "enterprise long tail" without escalating the costs of sales and operations. Similarly, consumer services will also be segmented, potentially at fine-grained level – for example, into home broadband, smartphones, IoT, etc. A connected car may need multiple network slices (assisted driving, infotainment, telemetry, etc.). And so again, operators need a way to serve smaller user groups with specific service requirements.

Network slicing and the ability to support diverse services on a common network platform is, arguably, the defining commercial concept driving 5G. Conceptually, it is quite simple: the provision of virtual networks adapted, in software, to the needs of the application on an end-to-end basis. In practice, it is technically challenging in several dimensions. In the first instance, operators need to determine how granular network slices should be. Typically, the industry talks in terms of coarse-grained slices optimized for the major service types (eMBB, MIIoT and URLLC); however, if the commercial opportunity is in fine-grained services optimized for the customer or user group, operators need to scale slices to the individual enterprise level – perhaps as far as just a few thousand devices per slice. This has cost and management implications.

**Figure 9: Network Slicing Across Companies & Industries**



Source: Affirmed Networks

As **Figure 9** shows, a network slice is made up of multiple network functions – a UPF, AMF, SMF, etc. – composed into a service. These are standard 3GPP functions connected over standard interfaces. However, because the needs of each slice may differ significantly, there is an opportunity to optimize network functions using microservices. For example, in a slice dedicated to high-speed rail services, the AMF may need to be dimensioned according to the needs of a high-mobility service; whereas for a small-data IoT slice, for example utility

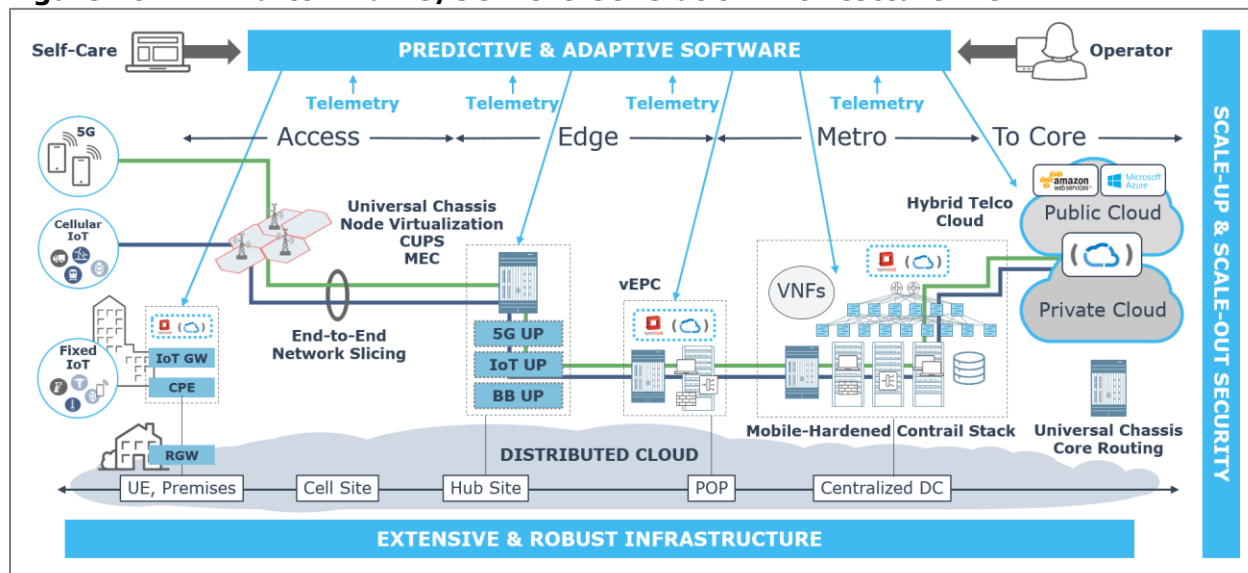
meters, where mobility is limited but the number of connections is high, the AMF may need to be optimized for a high number of sessions.

Resources can be dedicated to some slices, but shared in others. For example, some services may share a user-plane function, but use dedicated control-plane functions (it may be useful to have separate authentication and subscriber databases, for example). In other cases, the customer may require a slice made up of dedicated network functions – for example, a government department or an enterprise with important intellectual property or mission-critical tasks – to preserve performance or privacy. Being able to create slices and manage their lifecycle at this level of granularity, without a commensurate rise in operating complexity, is therefore very attractive to operators commercially. This underlines the need for automation, real-time analytics and programmability as inherent to the architecture.

## A NEXT-GENERATION REFERENCE ARCHITECTURE

Service agility and network slicing should apply end-to-end and across the network stack. A reference next-generation architecture is shown in **Figure 10**, including multi-access, core network, cloud (centralized and edge), telemetry and analytics, and management tools and the underlying transport network.

**Figure 10: An End-to-End 4G/5G Next-Generation Architecture View**



Source: Juniper Networks

Transport and SDN are critical to automated operations. The telecom industry's effort to automate with OSS has not been a great success historically. In combination with real-time, in-depth knowledge of the network state provided by telemetry and analytics, SDN allows operators to integrate at a much lower level with the network infrastructure to enable a closed-loop architecture where the operator can decide to implement a service or route, monitor it, make changes and start over, across a range of transport types. This includes FlexE, MPLS and segment routing. Finally, in converged networks, slicing concepts can (and should) be extended to wireline services. This delivers greater service capability to the operator and, crucially, generates economies of scale in the transport infrastructure.