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

HEAVY READING
WHITE PAPER

5G-Era Cloud Strategies for Network Operators

A Heavy Reading white paper produced for Huawei

AUTHOR: GABRIEL BROWN, PRINCIPAL ANALYST, HEAVY READING

DEFINING 5G-ERA SERVICES

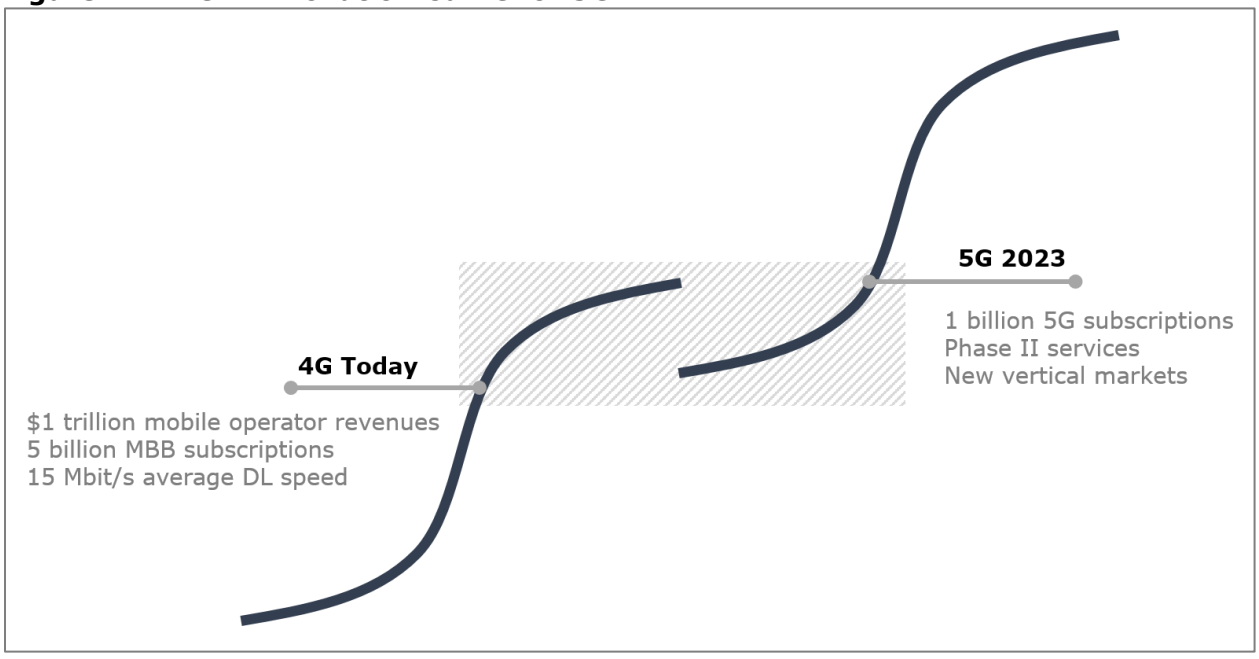
The 5G era is underway and will have a deep and lasting impact on operator networks and cloud strategies. Advanced 4G (LTE-A Pro) already offers services ranging from gigabit mobile broadband to the large-scale Internet of Things (IoT), and with the introduction of 5G from 2019 onward, the service opportunities available to operators will multiply. This will, in turn, introduce new requirements and challenges to their "network cloud" deployments.

This white paper investigates how 5G will impact operator cloud strategies, focusing on how advanced new services drive deployment of an "edge cloud" network architecture to ensure the performance, availability and reliability required by the most demanding 4G and 5G use cases. The paper emphasizes how the unique access and metro network assets owned by operators can serve as a physical footprint for distributed cloud and, in combination with inter-data-center connectivity, can create a unified "network cloud" architecture.

A New Innovation Curve

From one perspective, 5G can been seen as simply the next generation of network technology, which is an appealing prospect given the tremendous success of mobile. However, 5G is also an opportunity for the networking industry to participate in – and to enable – a new set of consumer, industrial and enterprise use cases. **Figure 1** shows that 5G is about a new innovation curve: going beyond smartphones – not abandoning them – to support many more diverse use cases. This is what makes 5G exciting from a business perspective and is why the industry is backing the technology with R&D and capital investment.

Figure 1: A New Innovation Curve for 5G



Source: Heavy Reading

Much of the broader technology industry is also looking for the next big platform transition after smartphones, and so technologies such as augmented reality (AR), virtual reality (VR), artificial intelligence (AI), IoT, connected car, etc., are in vogue. The phrase "software is

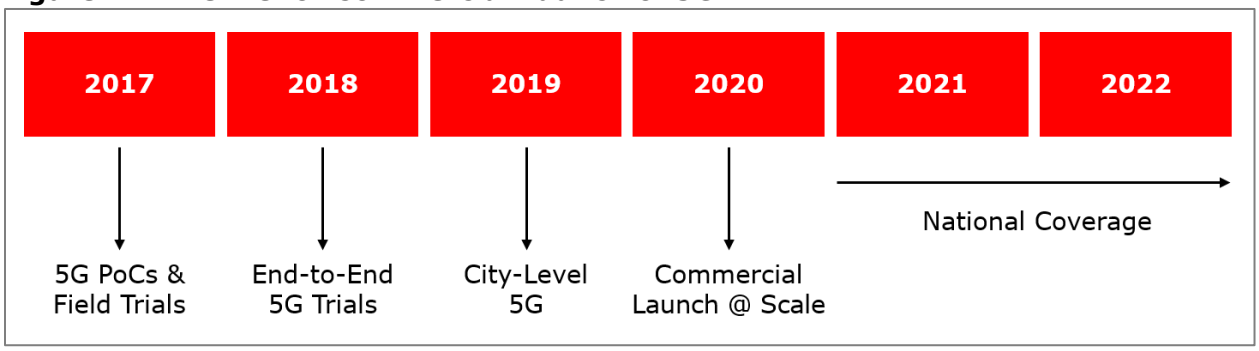
eating the world" is a good way to understand the transformation underway, in virtually all industries, to software-driven operating processes. The opportunity for 5G mobile is to participate in those processes and, in some cases, be the force that enables them. If progress continues on its current trajectory, then the industry could be at the start of a steep new innovation curve five years from now. In this sense, 5G is a foundational technology for the broader economy.

At the same time, mobile operator revenues – a huge $1 trillion in 2017 – are dominated by classic mobile services. These revenues will be far more significant than the revenue from new services, even under the most optimistic of scenarios, for at least the early years of 5G. It is imperative to the 5G business case, therefore, that mobile broadband and IoT services are also supported and improved. As this paper will argue, the edge cloud can also improve delivery of a wide range of 4G and 5G mobile broadband services.

5G Launch Timeline

Operators in all the major global regions have aggressive 5G deployment plans. **Figure 2** shows a timeline for the commercial introduction of 5G. The schedule is driven by standardization, chipset and device availability, spectrum assets and investment in new "cloud-native" networks. The very first launches of 5G services are expected in 2018 for fixed wireless and specialist use cases on a limited scale, with smartphones services expected from mid-2019 onward.

Figure 2: Timeline for Commercial Launch of 5G



Source: Heavy Reading

The timeline and buildout varies by country, but roughly speaking 2019 will see the first city-wide deployments, with mass-market commercial services available from 2020 onward. National coverage will materialize over time and will be dependent on frequency bands, customer demand, etc. The capabilities of 5G will also evolve: The first Phase I services will focus on broadband access; more advanced capability will be delivered in Phase II, around 2022.

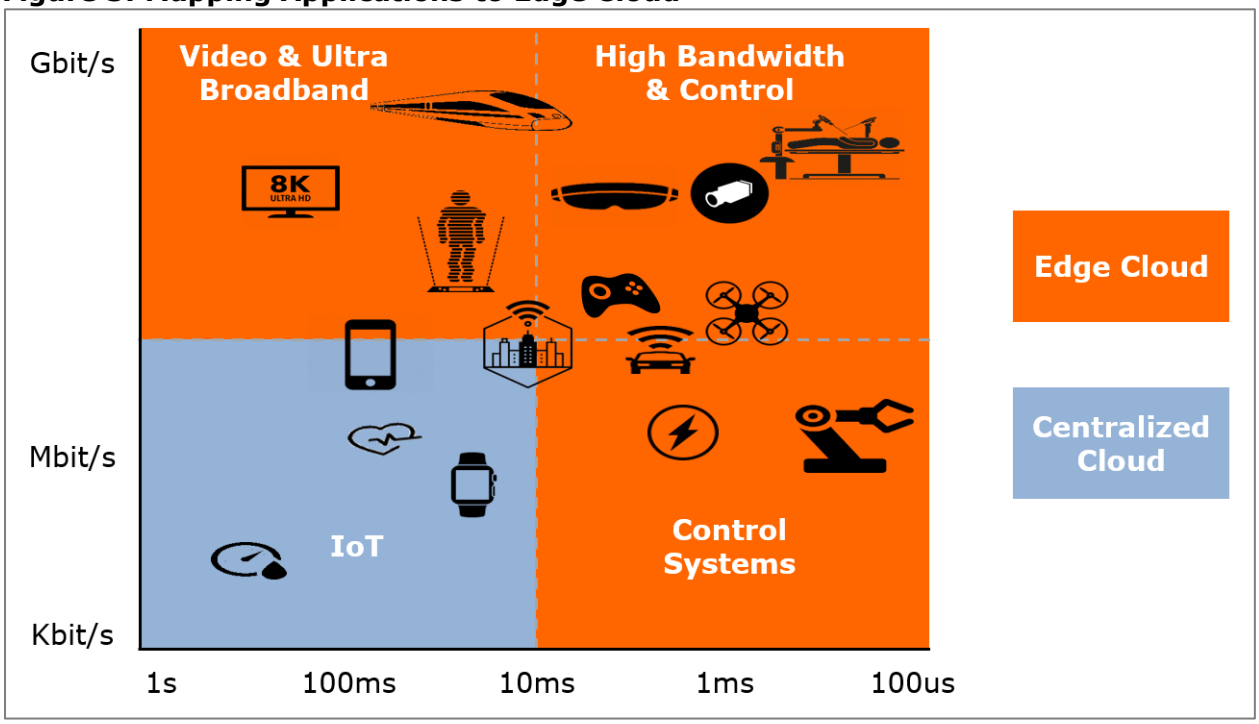
5G Service Types

With smartphone penetration above 80 percent in advanced markets, new service types are the major new growth opportunity. Some of these new services will be consumed via smartphones, others will use specialist devices, using 5G radio modules integrated in a wide range of end-user equipment from heavy machinery, to healthcare, vehicles and more. The mobile industry has developed a wide range of new service types in various fora, including the ITU,

the 3GPP, trade associations, industry alliances and by private enterprise. Huawei X-Labs, for example, has created a paper to identify and analyze the Top 10 5G Use Cases.

The performance requirements of these new service types are a critical influence on the design of telco cloud infrastructure. Where very low latency is required, edge cloud becomes critical. **Figure 3** maps some of the services proposed to run over 5G according to their latency and throughput requirements (packet loss and jitter is another axis that would be useful, but is not shown here). To the top left of the chart are services such as 8K video delivered using 5G wireless to the home. More demanding high-throughput services might include high-speed rail (the target 5G performance is 15 Gbit/s downlink and 7.5 Gbit/s uplink per train, at up to 500 km per hour), or even, in time, holographic communication. To the bottom left are services that require very low latency and low packet loss, but perhaps not as much throughput. High-voltage electrical monitoring, for example, requires only around 10 Mbit/s throughput, but needs less than 5 ms latency, four-nines availability and 1 ms jitter.

Figure 3: Mapping Applications to Edge Cloud



Source: Heavy Reading

To the top right are services that need both ultra-high capacity and ultra-low latency. Given how demanding they are in terms of performance requirements, some of these services are likely to come later in the cycle – for example, tactile Internet applications, such as remote surgery. Other advanced services, such as VR, drone control, intelligent transport services and machine vision, are more forgiving and are likely to come sooner.

Note that not all services need edge cloud and that operators (and application and content providers) need flexibility in where they host services – in many cases it will continue to be more efficiency to host in large, centralized data centers. Longtail content, for example, would not need to be distributed to the edge. A unified cloud platform with common management tools across edge, regional and centralized locations is, therefore, desirable.

THE 5G EDGE CLOUD ARCHITECTURE

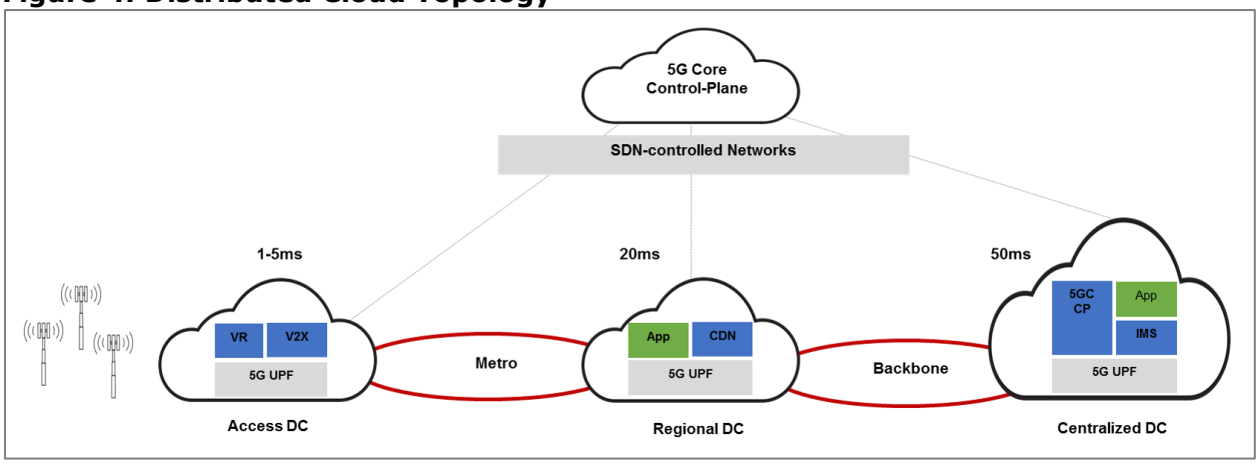
The service types the operator intends to offer determine how the network is designed and deployed. There are three major categories of applications that will determine the distributed cloud architecture. These are:

- **Network Functions.** This includes physical and virtual network functions (VNFs) and is probably the key determinant of how many edge locations are needed and where they should be deployed. An operator deploying virtual radio access network (vRAN), for example, will need more locations closer to the edge. There is also an opportunity to distribute 5G core functions at the edge to improve scalability and latency.
- **Video Content.** This can include ultra-high-definition 4K and 8K video to reduce transport costs and to reduce time to start and buffering. Video also includes VR and AR, which require low latency and high throughput to provide an immersive experience.
- **Application Logic.** Edge deployment is needed for time-sensitive applications. An assisted-driving application for example, would likely need roadside infrastructure to process and relay data and meet target system performance of 10 ms latency, four-nines availability, in a 2 km roadside service area, at a density of 10 Gbit/s per km^2.

Multi-Level Data Centers

Services and applications can be deployed at different locations in the edge or metro network. Operators have unique assets than can serve as a physical footprint for distributed cloud and can provide inter-data-center connectivity services to create a unified "network cloud" architecture. This network footprint is not easily replicated and is a source of competitive advantage relative to hyper-scale, centralized cloud players. **Figure 4** shows a distributed cloud topology with services deployed according to their performance/latency requirements.

Figure 4: Distributed Cloud Topology



Source: Heavy Reading

The most demanding services – in terms of latency, packet loss and jitter – are deployed in the access data center. This may be a refurbished central office, transport aggregation point, or even a cell site shelter or major cell site hub. The extent to which operators deploy vRAN – and the extent to which it uses a higher-layer or lower-layer functional split – is probably a

determinant of how many access data centers are needed and where they should be located. A vRAN deployment using a lower functional split and eCPRI fronthaul transport has latency requirements of just a few hundred microseconds, making it extremely demanding and driving the need for more access data centers. End-user services that might be deployed in the access data center include VR and vehicle-to-X.

The extent to which operators require a three-tier architecture, incorporating centralized, regional and access data centers, is debatable. One view is that the regional facility delivers an order of magnitude greater latency than the access data center and, therefore, the operator might as well bypass the regional facility and go directly to the centralized cloud. In this case, a two-tier architecture containing more access and more centralized data centers, but no regional locations, is preferable.

This decision is likely to be very dependent on the geographic size of the operator's market – three data center tiers would make more sense in, say, China or the U.S., than in Belgium or Switzerland. It is also influenced by the footprint and capabilities of the transport network – for example, where the operator has transport aggregation points close to Internet points of presence (PoPs) or peering facilities, these may make good locations for regional data centers. In general terms, ex-incumbent operators with wireline and wireless assets have a greater number and variety of facilities to select data center locations from.

Distributed Data Center Conversion

Alongside the service strategy, there are many other factors that impact how operators should think about distributed cloud. Some of the most important are outlined in **Figure 5**. The first point – demand forecasting and knowing when and how much to invest – is probably the most critical to highly distributed clouds. Having a large number of lightly used locations is obviously uneconomic. On the other hand, distributed cloud is a transformation project that needs strategic long-term investment; operators must develop enough distributed data centers to change the network architecture and means of production, but not so much as to generate runaway cost with low utilization.

Figure 5: Distributed Data Center Deployment Issues

Issue	Key Factors & Decision Points
Demand Forecasting	Clients pay for more resources (e.g., CPU, storage) as they need it, making it difficult to predict future demand based on past behavior. This means the operator has to deploy infrastructure in advance of customer demand.
Operations & People	Edge locations are lightly supported (not many people on site). May require additional staffing costs. Underlines need for very low maintenance equipment. Access to facilities can be an issue.
Cloud Platform	Applications and services should be portable between edge data centers and centralized data centers. Needs common deployment and management environment across centralized and edge cloud infrastructure.
Network Fabric	Networking underlay and overlays using an SDN-controlled fabric are needed to connect edge cloud locations. Near-real-time migration to enable software-defined resiliency is a critical feature.
Converting Legacy Facilities	Central offices often need major upgrades (power supply, admission control, fire suppression, etc.) to host modern cloud infrastructure. Space, cooling, access and regulation can make this challenging.

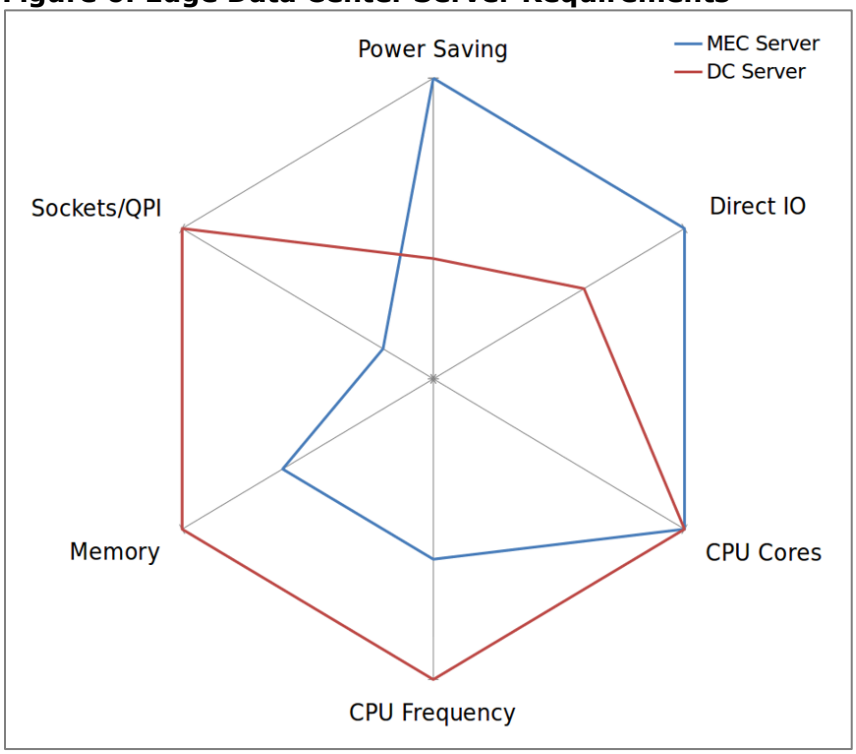
Edge Cloud Infrastructure

The facilities at edge locations are sometimes remnants from legacy networks and often need upgrades to host modern cloud infrastructure. This includes power supply, admission control, fire suppression, cooling, access, etc. There may also be challenges related to regulation (NEBS, etc.), physical space and floor weight.

Distributed data center conversion does not mean replicating centralized data facilities, but rather optimizing the infrastructure for the applications and functions deployed at the edge. In particular, operators need to consider high-throughput data-plane functions, such as switching and routing, and in 4G and 5G networks, distributed core network gateways with multi-gigabit throughput requirements. This drives different server requirements and infrastructure design than the standard, rack-based data center model.

Figure 6 contrasts the requirements for edge servers with those for centralized data center servers, and shows that power efficiency (which also reduces cooling) and input/output (IO) performance are relatively more important than memory, CPU performance or multiple CPU sockets per server. As a result, there are now specialist server platforms available for these edge environments.

Figure 6: Edge Data Center Server Requirements



Source: Verizon, Kontron (Light Reading Webinar, December 2017)

Operators, however, need to maintain common execution environment across edge and centralized data centers so that applications (VNFs, etc.) can run anywhere without adaptation. As a result, vendors are developing power- and throughput-optimized x86 blade servers and rack-mount server platforms that support edge runtime requirements. In the future, ARM servers are also expected to be suitable for edge locations. It is important at the cloud software level that VNFs do not need to be rewritten to run at the edge.

Cloud Management

Part of the economic rationale for cloud and network functions virtualization (NFV) deployment is to reduce labor costs. This is important at the edge because there are many facilities, and they are typically lightly staffed environments. This makes automated operations critical. Operators are generally open to using multiple toolsets to manage different types of cloud applications – for example, IT toolsets for IT applications and network orchestration tools for VNF lifecycle management. However, to achieve economies of scale, and to benefit from the flexibility to place VNFs where they want, operators need common toolsets for deploying across edge and centralized data centers. This enables them to optimize workload placement and resource use.

In addition to application workload management, operators also need to automate the management of the edge cloud infrastructure itself. This incorporates a wide range of tasks, including the physical, such as managing the facility, physical servers and network hardware; the physical/virtual boundary, such as hardware accelerators, NIC cards and mapping virtual network overlays to the wide-area SDN transport; and virtual resource management tasks not covered by management and orchestration (MANO) and VNF manager (VNF-M).

This is a multifaceted challenge, requiring solutions that can scale from just a few servers in a small facility to a massive edge node – it is plausible that an edge cloud node could contain more than 10,000 servers in certain cases. Operators can create their own cloud infrastructure management toolsets using a mixture of open source and commercial components or they can integrate vendors' solutions into their networks.

In terms of workload placement, especially for user-plane VNFs, a complicating factor is how to manage applications across virtual machine (VM) and container environments. There is a view among some operators that containerized user-plane VNFs can offer higher I/O performance, better hardware efficiency and lower latency than in VM-based environments and, therefore, are better suited to the edge. However, to achieve the most benefit from containers, VNFs need to be rewritten as cloud-native applications, which typically means decomposing them into microservices and developing new failover/redundancy schemes.

This view has wide support; however, like much of the rest of the IT market, telecom is in a transitional period, as vendors refactor VNFs for these new cloud environments. This is leading to the pursuit of interim schemes such as "containers in VMs" or "machine containers." There is an urgency to address this issue, because the 5G core is expected to be cloud-native.

One promising development in cloud workload placement is the use of machine learning (ML) to optimize resource use (e.g., CPU, storage, power and networking) and to ensure service continuity (e.g., using predictive analytics). This field is underdeveloped but has great promise as a way to interpret the vast data sets generated by the cloud platform and the service assurance software used to monitor and maintain customer service-level agreements (SLAs). These datasets overwhelm human pattern recognition and are well suited to optimization by ML-based tools. Similar ML and AI concepts can be applied to route optimization and service assurance in an SDN network, or even to customer sentiment analysis.

PaaS at the Edge

Operators, in general, prefer to deploy their own cloud platforms to run NFV and to run end-user services (VoLTE, unified communications, etc.). This gives them control of the

environment and, if they have sufficient scale and operating expertise, should generate superior economics. They will also seek to extend the platform to applications and content from external providers (over-the-top video, gaming, etc.).

In some cases, however, there is a role for deploying VNFs using a platform-as-a-service (PaaS) model. PaaS hides infrastructure details – typically infrastructure-as-a-service (IaaS) details – from application design. In this way, it enables VNF deployment across a range of infrastructure options (for example, incorporating FPGA/ARM/x86 and private/public cloud) without creating excessive platform dependencies for the application developer.

This is why the capability to operate using PaaS was one of the critical design criteria for NFV. In practice, PaaS for NFV can take several forms, including:

- The operator's cloud infrastructure division may offer PaaS internally to the network team – for example, the core network team can request resources/performance but doesn't run the cloud infrastructure itself.
- In large operator groups with multiple national operating companies a group cloud platform can be offered on a PaaS basis to each national operator. This generates economies of scale, subject to regulatory issues such as data protection.
- In roaming scenarios. This can be a standard roaming hub where an operator deploys a virtual gateway in a peer network, or in a federated network slicing scenario where international coverage is needed, in the future.
- Offer PaaS to other service providers. For example, an operator could build the edge platform and offer PaaS to a third-party partner wanting to host, say, an IoT platform or a connect car service.

Multi-Cloud Strategies for Network Operators

There are other areas where external clouds are useful, particularly when used in a hybrid cloud manner. Multi-cloud strategies are commonplace across large enterprises, and many of the same conditions and reasoning also apply to telecom operators. Some example multi-cloud scenarios include:

- Use public cloud to burst for extra capacity. There may be occasions where additional capacity is needed – for example, a special event, or during a maintenance or upgrade window. Public cloud can also be used for resiliency.
- For out-of-region deployments. This is particularly useful for operators with international enterprise business that may need to deploy services outside of their own geographic network footprint.
- Hybrid telco and public cloud for IoT. Aspects of an IoT service are often suited to processing in the edge cloud (e.g., to manage high-connection density or reduce power consumption), while other parts of the services (e.g., data processing, storage) are better suited to centralized public cloud.

It is also the case that some operator workloads may simply be better suited to the public cloud. This is most likely for back-office functions (billing, customer care, etc.), which are similar to standard enterprise IT applications and can be more efficiently hosted and operated externally and even consumed as software as a service (SaaS).

5G-ERA CLOUD SERVICE EXAMPLES

There are many 5G-era cloud services that require a unified distributed and centralized cloud infrastructure. Some examples are discussed in this section.

Network Slicing With "Open" APIs

The ability to use 5G to support diverse services on one network platform is a powerful idea, with compelling commercial opportunities. Network slicing enables operators to configure virtual network instances optimized to the customer type or application. If a network slice is defined as a processing path containing all the networking functions needed to deliver a service, it is clear that it extends across network domains, across operators and across industries. To deliver this capability requires industry collaboration and open application programming interfaces (APIs).

Common implementations are critical for operators, and for CIOs in industrial verticals. Automotive companies, device makers, media companies, cloud providers, manufacturers, distributors and industrials all seek to operate nationally, regionally and globally. A common format for 5G slicing and open APIs should make the decision to use 5G easier and quicker, as it will mean operators in different geographies have, to a greater or lesser degree, compatible platforms.

Inter-operator slicing is also important for roaming cases. To offer multinational customers a compelling service slice will often require inter-operator agreements, as demonstrated by Deutsche Telekom (Germany) and SK Telecom (South Korea) in 2017. In such cases, the host operator offers the guest operator PaaS on which it can run its own VNFs – for example, a guest operator could deploy a virtual gateway or roaming hub into a partner's host network to avoid long-distance "tromboning" of home-routed traffic.

Although 5G is associated with the radio access and core network, as defined by 3GPP, in practice, an end-to-end network slice will involve transport, cloud and service platforms, and the ability to isolate and assure traffic across these domains, while meeting the performance requirements of the slice. Cross-domain collaboration, therefore, is also critical.

IoT & PaaS

The IoT is a service category that can be deployed at the edge by operators themselves, and that operators can support by offering PaaS services to other providers. An example of the latter model is Amazon's AWS Greengrass, which operators can deploy in MEC data centers to, in effect, extend the AWS cloud to the edge of their mobile network. This changes their role to active players in cloud computing and positions the mobile network as a low-latency platform for critical IoT applications.

Greengrass is an IoT software platform that can be deployed on third-party hardware – in this case, the operator's edge data center – and comprises two components: a device software development kit (SDK) and the AWS Lambda cloud environment known as "Greengrass Core" (but deployed at the edge). Using these components, many types of devices, from heavy plant machinery to sensors, can communicate with the cloud application running on "Greengrass Core." In a smart city deployment, for example, a high-definition camera running the SDK may be deployed to monitor traffic conditions and would connect to a locally

deployed application, running at the edge on Greengrass Core, to analyze video and send the correct data back to the centralized control system.

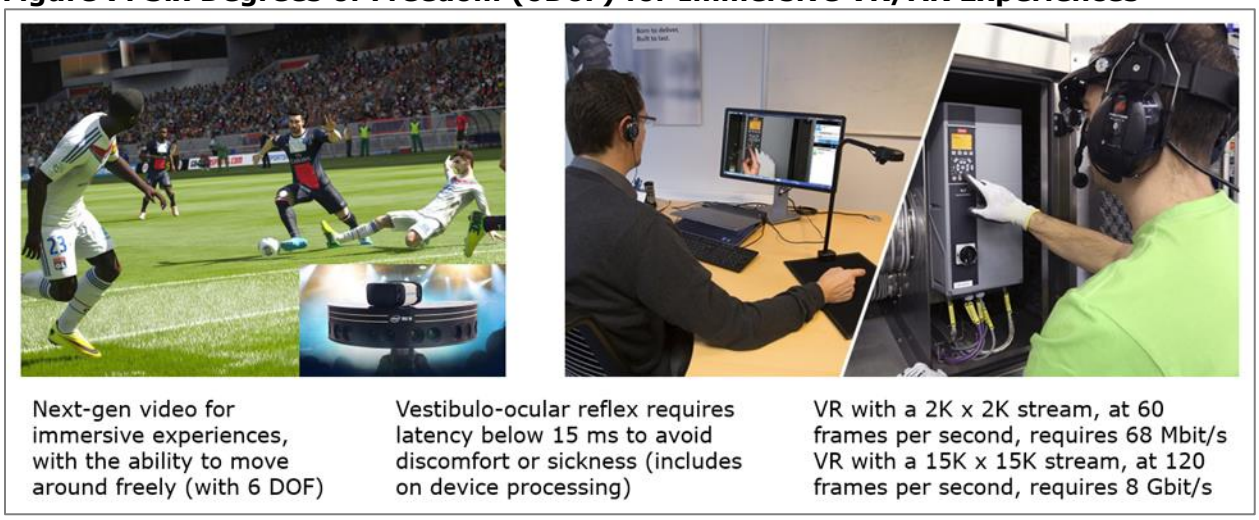
Operators have several business model issues to address in IoT. Their role in the value chain and how they interact with partners and customers will vary between markets and the service type. In some cases, operators will want to run their own IoT platforms in their own cloud environment; in other cases, forming partnerships with vendors such as AWS makes sense; and in others, the operator may choose to use IoT platforms hosted by vendors to run its own services. The build vs. partner decision for operator IoT platforms is a difficult one; it is likely that only the larger, more successful Tier 1 operators will be able to invest in and maintain leading IoT platforms over time, and that smaller and mid-sized operators will lean toward vendor solutions.

Virtual Reality; Mixed Reality; Augmented Reality

The performance requirements of VR in terms of bitrate and latency are very demanding and make it a good candidate application for edge computing. If operators can meet these performance targets from services hosted in the network edge, there is potential to increase the overall market for VR/AR considerably using 5G mobility to untether the technology from in-room use cases to a broader set of consumer and enterprise services.

Figure 7 shows how immersive mobile VR experiences require high-performance networks. To the left is a player-point-of-view cam, where the spectator can see the game from the perspective of the player on the pitch. This could also be the front row at a concert, the view from a racing car or similar. The examples to the right are remote expert use cases for industry verticals such as aerospace engineers, nuclear clean-up experts, agricultural plant operators, civil engineers, doctors, drone pilots and similar.

Figure 7: Six Degrees of Freedom (6DoF) for Immersive VR/AR Experiences



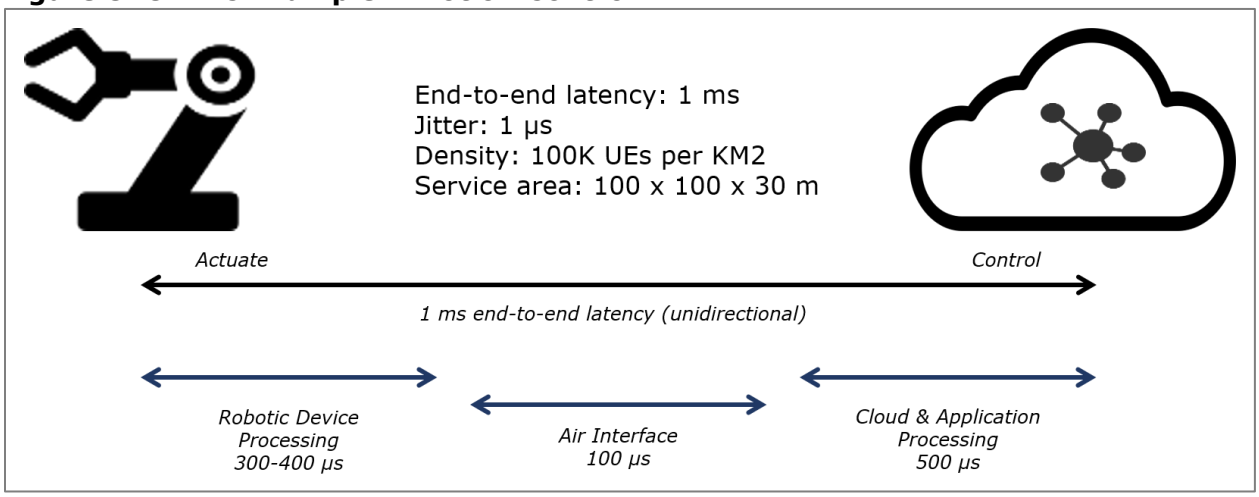
Source: Heavy Reading

In each case, an immersive experience requires the user to have six degrees of freedom (6DoF) to be able to look and move around freely. This requires very fast rendering with sub-15 ms two-way latency to avoid motion sickness. Resolution and frame rate are also critical: As a guide, a 2K stream would needs 68 Mbit/s and a (theoretical today) 15K stream would need 8 Gbit/s.

On-Premises Cloud for URLLC Services

Some advanced 5G services are so performance-critical that they are extremely difficult to serve from the wide-area 5G network, and instead need application logic to be deployed on-premises very close to the user. One example ultra-reliable low-latency communications (URLLC) service is robotic motion control – typically used for factory automation and "Industry 4.0" – which requires network latency on the order of a few hundred microseconds, as shown in **Figure 8**, to stay within the application's overall delay budget of 1 ms one-way latency. In recognition of this, the 3GPP performance targets for 5G state that this performance need only be delivered within a service area of 100 x 100 x 30 meters. For operators to serve this market, they will likely have to extend the edge cloud into the customer premises.

Figure 8: URLLC Example – Motion Control



Source: Heavy Reading