AvidThink®

Sponsored by Western Digital.

Impact of 5G on the Future of Storage

RESEARCH BRIEF

Table of Contents

Impact of 5G on the Future of Storage

Executive Summary

5G, the next generation of mobile connectivity, promises dramatic speed increases, ultra-low latencies, and the capacity to handle 1 million Internet of Things (IoT) devices per square kilometer. 5G encompasses multiple technologies, from new cellular radios with innovations that increase network performance in a power-efficient manner to a revamping of the network support software infrastructure to improve flexibility and scalability. It also brings new capabilities, such as network slicing which allows the creation of multiple virtual networks on the same infrastructure, each securely isolated from each other and each with different performance characteristics.

We've analyzed and integrated our viewpoints with those of over 40 experts from diverse backgrounds: top-tier network operators, independent software vendors, semiconductor companies, automotive companies, manufacturers, hyperscale cloud providers, and enterprises, as well as researchers in academia. From the data gathered, we were able to probe how 5G might impact data and storage across the top use cases in 5G, including content delivery networks, connected car and autonomous vehicles, Industry 4.0, video surveillance, cloud-based gaming, and telemedicine.

Across our research, it is clear that the main drivers of storage use in the 5G era have less to do with 5G connections and more to do with the digitization of everyday processes. The increased use of video data and the improved resolution of image sensors will have a significant impact on storage capacity. So too will the underlying virtualization of the infrastructure and the move towards cloud-native architectures. And our penchant for logging and capturing every available piece of data for troubleshooting and machine learning is yet another driver for storage capacity growth.

The consensus from experts is that existing storage technologies, such as flash and SSD, as well as current storage access technologies, like NVMe, NVMe-oF, and UFS, appear to be sufficient to meet the performance needs of 5G networks, with only a few caveats.

It seems, therefore, that 5G will be primarily an enabler of digital transformation and have a significant impact on where data is stored and processed. As 5G networks roll out and become more reliable, we expect a migration of storage from end-user devices to 5G edge locations and the public cloud, where storage and processing are cheaper. Ultimately, 5G is somewhat orthogonal to the progress of storage technologies and capacities, providing mostly second-order effects as 5G enables new and innovative use cases. The impact on storage will be seen primarily in the overall compute and software architecture needed to serve these use cases, and it appears that today's storage technologies will stand up well to the challenges ahead.

Introduction and Methodology

5G represents the next generation of mobile connectivity, promising dramatic increases in speeds, eye-popping reduction in latencies, and a host of new applications. At the same time, there's plenty of hype and poorly set expectations around 5G and the associated topics of edge computing and the Internet of Things (IoT). While much has been discussed around 5G and networking and 5G and computing — particularly edge computing — less has been said about 5G's impact on storage.

Western Digital, a leader in the storage industry, commissioned AvidThink to analyze the impact of 5G on storage. The goal of this research project was to understand the relationship between 5G and data and both the near-term and longer-term impact on storage.

To create this report, AvidThink collaborated with Western Digital to collect insights from over 40 interviews globally with a range of experts at top-tier network operators, independent software vendors, semiconductor companies, automotive companies, manufacturers, hyperscale cloud providers, and enterprises as well as researchers in academia. We analyzed the results of the interviews conducted in Q2 2020 (during the height of the COVID-19 pandemic) and compiled the findings in this research brief.

For readers less familiar with 5G, this report also includes a primer on 5G technologies. That is followed by the top use cases surfaced in our interviews. As we analyze the 5G use cases, we share how 5G and edge computing will impact data flows and storage. Given the recency of these interviews, we have also captured a multifaceted view on the pandemic's impact on 5G and edge computing.

Readers already familiar with 5G can skip ahead to the section titled "5G Use Cases and the Impact on Data" on page 11.

We hope you find the information in this research brief helpful and welcome your feedback. You're welcome to contact us at **research@avidthink.com** should you have follow-on questions. We would like to extend our thanks to the many interviewees for this research report as well as Western Digital for their collaboration and support!

Primer on 5G

The fifth generation of mobile networks, 5G, is a little harder to define than previous generations. 2G systems as GSM and CDMA technologies were primarily about mobile voice services, while 3G first introduced the world to cellular data. Long Term Evolution (LTE) and 4G opened the floodgates by enabling true mobile broadband services. To differentiate 5G from 4G, we have to look beyond treating 5G as just faster broadband — being able to download movies faster doesn't justify that "G" jump. The expectation for 5G is that it will drive new use cases and business models for consumers, businesses, and industry — not to mention the Internet of Things (IoT).

Business Drivers for 5G

Many in the industry view 5G as a business-driven initiative. Whereas many applications of today's 4G networks are a mix of consumer and business use cases, the bulk of revenue and profit generation on 5G networks is expected to come from the business-to-business (B2B) market. From a mobile network operator (MNO) perspective, the insatiable demand for increased bandwidth requires a new generation of both hardware and software that can provide a high level of performance across the consumer and business segments. From a cost standpoint, 5G networks are expected to lower an MNO's operational and capital expense through improved radio frequency efficiency, better power management, use of virtualization, and scalable software architecture, in addition to leveraging commodity hardware.

5G networks are expected to lower an MNO's operational and capital expense through improved radio frequency efficiency, better power management, use of virtualization, and scalable software architecture.

At the same time, MNOs realize that a single mobile network with limited ability to differentiate traffic will not meet the needs of businesses. New business applications demand lower latencies that facilitate real-time interfaces and control systems at <10ms, <5ms, and eventually < 1ms, support for large scale IoT deployments (millions of devices per square mile), and the ability to create virtualized isolated networks that have different security profiles and support varying service-level agreements (SLAs). As an analog, what innovation in public clouds did for computing, 5G will do for mobile networks.

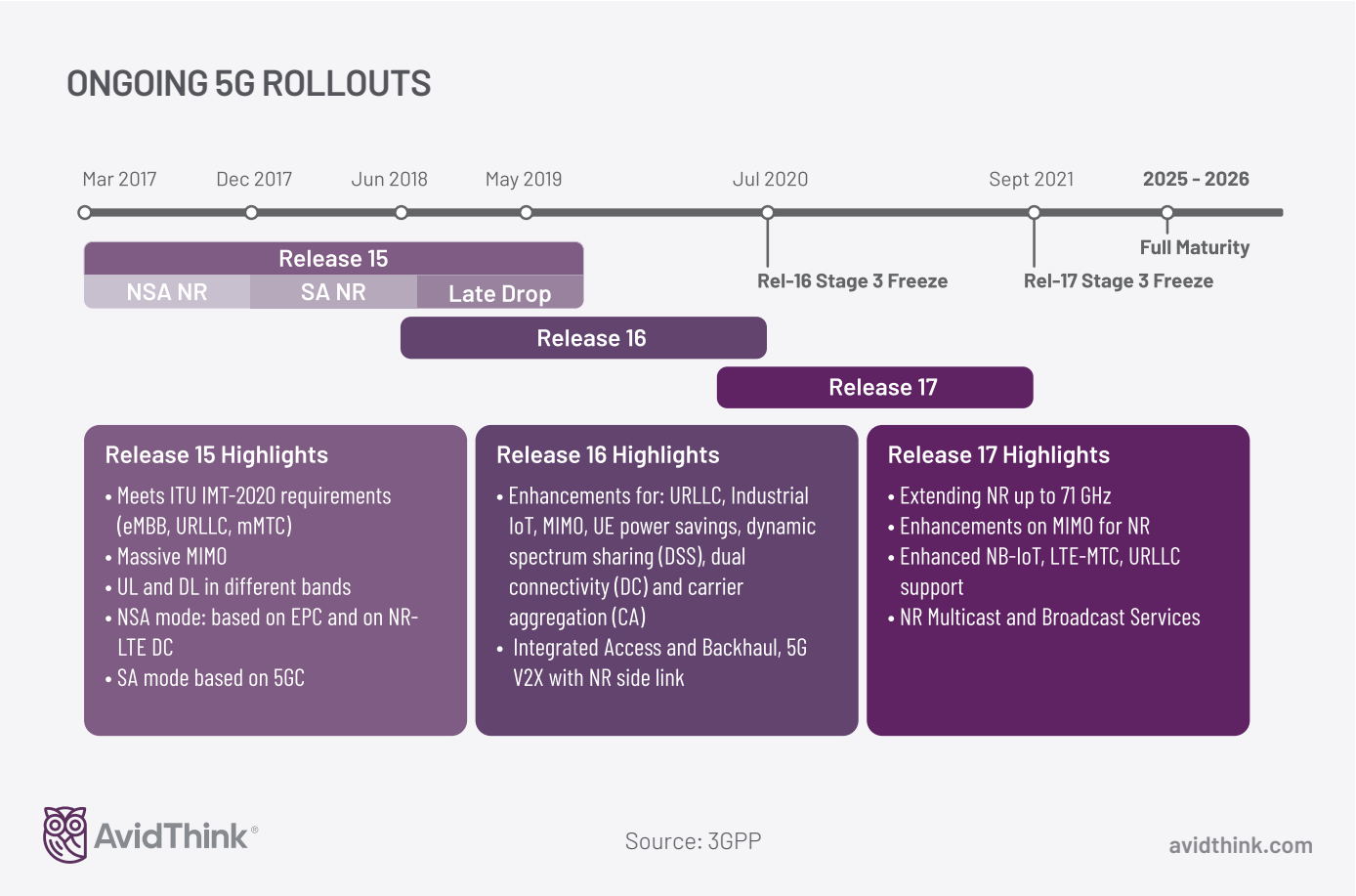
Benefits of 5G

When fully deployed, network operators and industry experts expect that 5G will provide almost ubiquitous gigabit speeds almost everywhere, reaching peaks of 20Gbps or higher, with millisecond latency (or sub-millisecond for specific cases). Not only will people be connected to each other, but also to machines, vehicles, city infrastructure, public safety, and more. As we will detail later in this paper, 5G also makes possible new applications that 4G networks cannot support, such as virtual reality (VR) or augmented reality (AR), real-time connected vehicles, and mobile robotics.

5G Standards

5G is not a single technology but a collection of multiple capabilities under a standards framework. Ultimately, the International Telecommunication Union (ITU) controls the 5G standard, which is also referred to as International Mobile Telecommunications-2020 (IMT-2020). The often-mentioned 3GPP is a related mobile industry standards body that submits proposed specifications to the ITU to be part of the IMT-2020 standard. Both mobile operators and multiple network equipment providers (NEP) participate in the 3GPP specification process.

5G is in the process of rolling out, even though the full set of standards is not complete. Those in the industry will refer to different release numbers, Release 15, 16, etc., to denote the ongoing process of adding more details to the standards supporting 5G. As of the publication of this report, the bulk of deployments are based on Release 15 (R15). R15 laid down the ITU IMT-2020 essential requirements for enhanced broadband (eMBB), which supports high-speed uploads and downloads, ultra-reliable low-latency communications (URLLC) and massive machine-type communications (mMTC) — the three main categories of 5G network capabilities.



Release 16 was finalized on July 3, 2020 by the 3GPP. R16 provides improvements for managing IoT devices, dynamic spectrum sharing for better coexistence with 5G, and carrier aggregation for increased bandwidth. It also includes improvements for vehicular support (V2X) and flexibility on the backhaul (uplink connections) to include wireless options in addition to fiber. Looking beyond, R17 and its improvements for increased capacity handling will not be ratified until late 2021. We'll discuss below the implications of standards release timeframe on actual rollouts. In general, vendors and MNOs can take up to a year or two after the standard is locked-in to implement the actual capabilities in production.

5G Radio and RF Spectrum

One of the key elements of 5G is the arrival of a new wireless radio standard. As mobile technology, 5G uses the spectrum more comprehensively than any other generation before it. 5G spectrum is broken up into three major categories: low band (sub 1 GHz), mid band (1-6 GHz), and high/ultra-high band (above 6 GHz and including millimeter-wave or mmWave spectrum in the 24/30 to 300 GHz range). The higher the band, the more carrying capacity: in other words, greater bandwidth for data. However, high bands travel shorter distances and are easily blocked by buildings, walls, or foliage, which means more radios must be deployed in denser configurations to achieve coverage. This means MNOs will roll out 5G radio access networks (RAN) using a combination of different spectrum bands to accommodate different terrain types, population density, and application needs.

Beyond the broader spectrum, there are additional enablers needed to achieve the carrying capacity and performance needed for 5G. For instance, advanced antenna techniques, such as massive multiple input, multiple output (MIMO), which utilize massive arrays

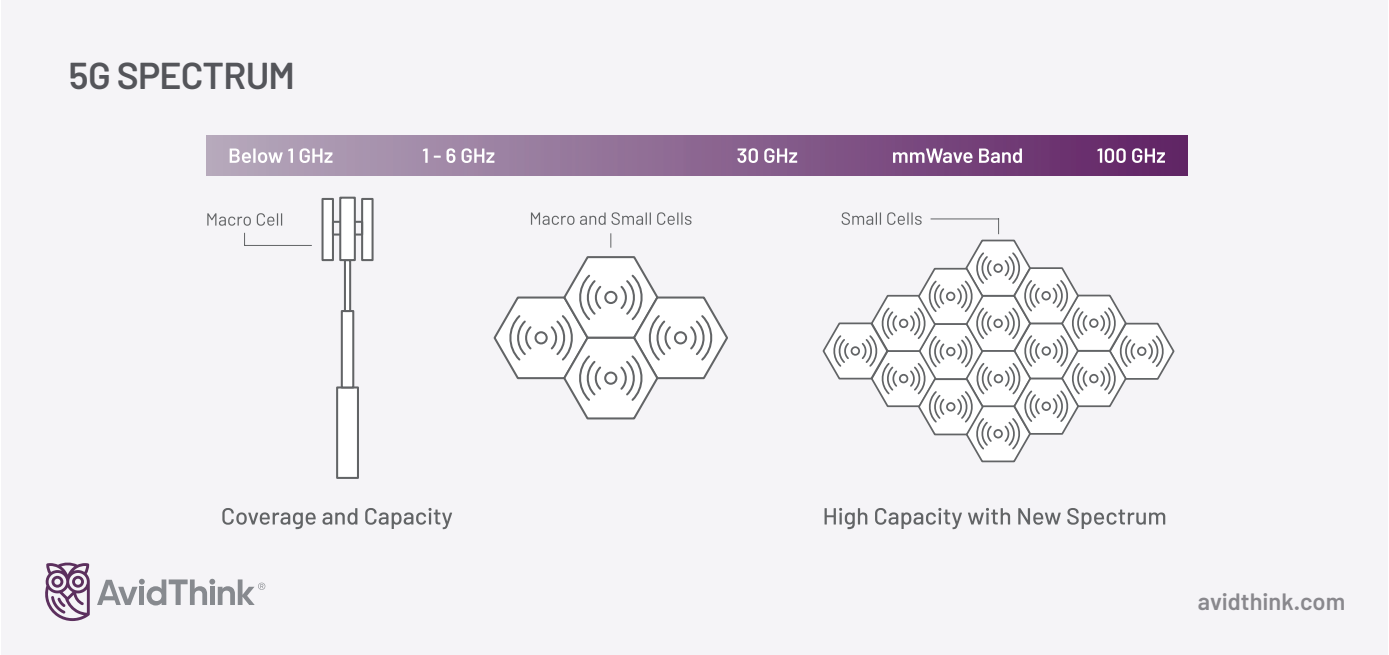
of antennas to deliver faster, more uniform data rates to users; and adaptive beam-forming and beam-tracking techniques to enable robust mobile broadband communications at millimeter-wave spectrum bands.

Regardless of the underlying technologies, different carriers in different regions are taking slightly different rollout strategies. Some carriers start with mmWave first in urban areas, while others blanket an entire nation with low band first before moving to mid band then high band. In Europe, most of the carriers are using mid band and some high end. In Asia-Pacific, most initial deployments are mid band combined with the high/ultra-high bands. The spectrum strategies they employ are usually tied to their business goals. Some carriers want to achieve the mantle of being the first nationwide 5G carrier, while others prefer to claim the fastest-network-in-the-region crown.

Some carriers want to achieve the mantle of being the first nationwide 5G carrier, while others prefer to claim the fastest-network-in-the-region crown.

Before we leave the topic of radios and spectrum, we would like to correct the critical misconception that 5G is only about the wireless spectrum. To achieve the network capacity that these new radios can enable, there needs to be an underlying network that can carry the traffic. These networks tend to be built on top of optical fiber technologies. Fiber powers the wireless backhaul infrastructure, ferrying packets from cell towers to aggregation points and to the core network. Sometimes copper can be used, but in general, 5G networks will primarily consist of wireless traffic riding on top of optical fiber in the fronthaul (the links that connect the radio units to their controllers) and backhaul (the links that connect the mobile network edge to the core network). Fiber has the necessary capacity to support large numbers of devices with high bandwidth, and fiber buildout is a crucial part of 5G network buildout.

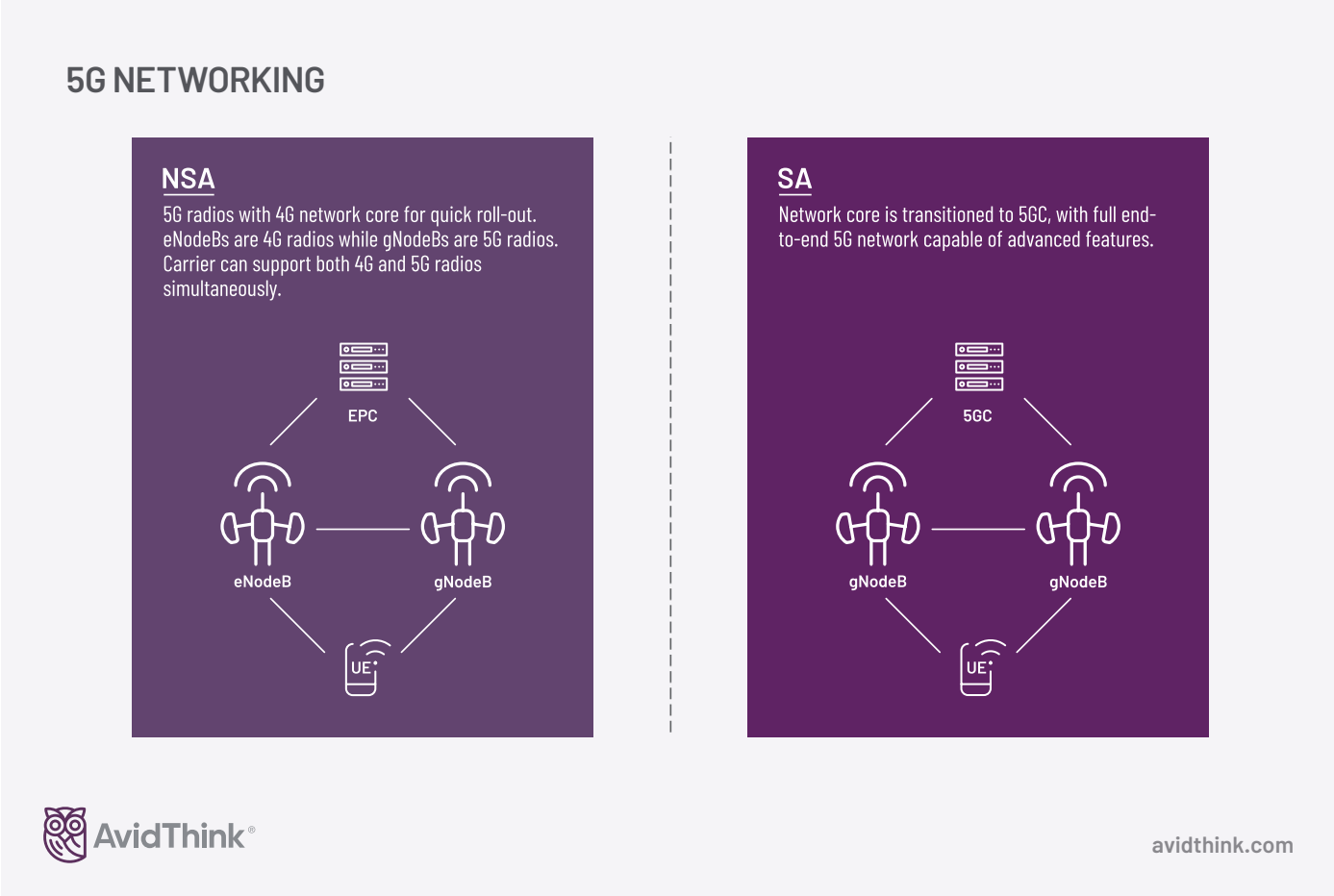
Likewise, to enable 5G and some of the more advanced features, there is a software infrastructure stack that manages the radios and handles all the wireless traffic. From the packet handling to orchestrating how all the software components of a 5G network are interconnected, this 5G core (5GC) stack is another critical element for 5G networks and no less important than the radios.



5G NSA vs 5G SA

To provide coverage across the spectrum bands as described above, MNOs must upgrade their radio access networks from existing standards to a 5G New Radio (5G NR) network. There are two main paths of getting from today's 4G LTE networks to a 5G network:

- **5G Non-Standalone (5G NSA):** Deploy new radios but tie those radios to an existing 4G LTE evolved packet core (EPC) software control stack.
- **5G Standalone (5G SA):** Deploy new radios along with a new 5GC software stack that both controls the radios and processes the packets that come from the radios.



Many of the initial rollouts to date have taken the 5G NSA path because the 5G core software stacks from the leading network vendors were not quite ready. Most carriers already have existing 4G EPC stacks available, and it is easy to integrate the new 5G NR radios into these stacks. Additionally, the new 5G core stacks are often built using a software architecture based on Linux container technology instead of the virtual machine (VM)-based or physical appliance-based approaches that MNOs are familiar with. We'll visit this topic in more detail in the next section.

The ultimate goal is to get to 5G SA. Without this, MNOs will not be able to achieve the ultra-low latencies that 5G promises nor the flexible network slicing that 5G provides.

5G, NFV, and SDN

Network functions virtualization (NFV) and software-defined networking (SDN) are technologies that are expected to play fundamental roles in 5G. NFV represents the use of commodity servers commonly found in enterprise and cloud data centers to replace proprietary and specialized telco equipment. NFV, which brings virtualization and a software-centric architecture, can provide the necessary scale to handle the 100-times increase in data rates (10-Gb/s speeds) and connectivity enablement for as many devices. By leveraging virtual network functions (VNFs) and, more recently, cloud-native network functions (CNFs), the cloud-scale software architecture that has been proven within hyperscale cloud environments can be used to power telco 5G networks.

SDN is an architecture for managing networks that provides scalability and dynamism. It allows a central orchestrator to control how different network elements process packets. It then combines central intelligence with distributed granular controls to ensure the network can move packets around efficiently. SDN is viewed as one of the critical factors in moving 5G networks from 10 ms to 5 ms to under 1 ms — one of the hardest challenges for MNOs to meet.

5G and Edge Computing

In addition to SDN and NFV, the other key to achieving low latency in 5G (sub-5 ms or even 1 ms) is the use of edge computing. Note that edge computing isn't 5G-specific. It already existed in the 4G LTE world under a variety of names, from mobile edge computing to the updated multi-access edge computing (MEC) to fog computing and distributed cloud. Some of these frameworks, like MEC, were mobile specific, while others were more general. Regardless, the ability to run application workloads closer to the RAN is critical to lowering latency. To get to a sub-5-ms end-to-end latency, the packet processing functions need to be within 50 km or 30 miles of the RAN[1] and for sub-1ms, possibly even less. Therefore, there's no likelihood of backhauling all the RAN traffic to a central cloud data center for processing. Most 5G deployments will have distributed data processing elements located at the edge. These packet handling functions are usually called User Plane Functions (UPF) in 5G nomenclature. These are separate from the control logic which sets up the actual flow of data. This separation of the control plane from the UPFs that process the data at the edge as packets travel from the source RAN to their destination is another fundamental principle of 5G called Control and User Plane Separation or CUPS for short.

The success of 5G is, therefore, dependent on having edge infrastructure available to run these UPFs and keep overall latency low.

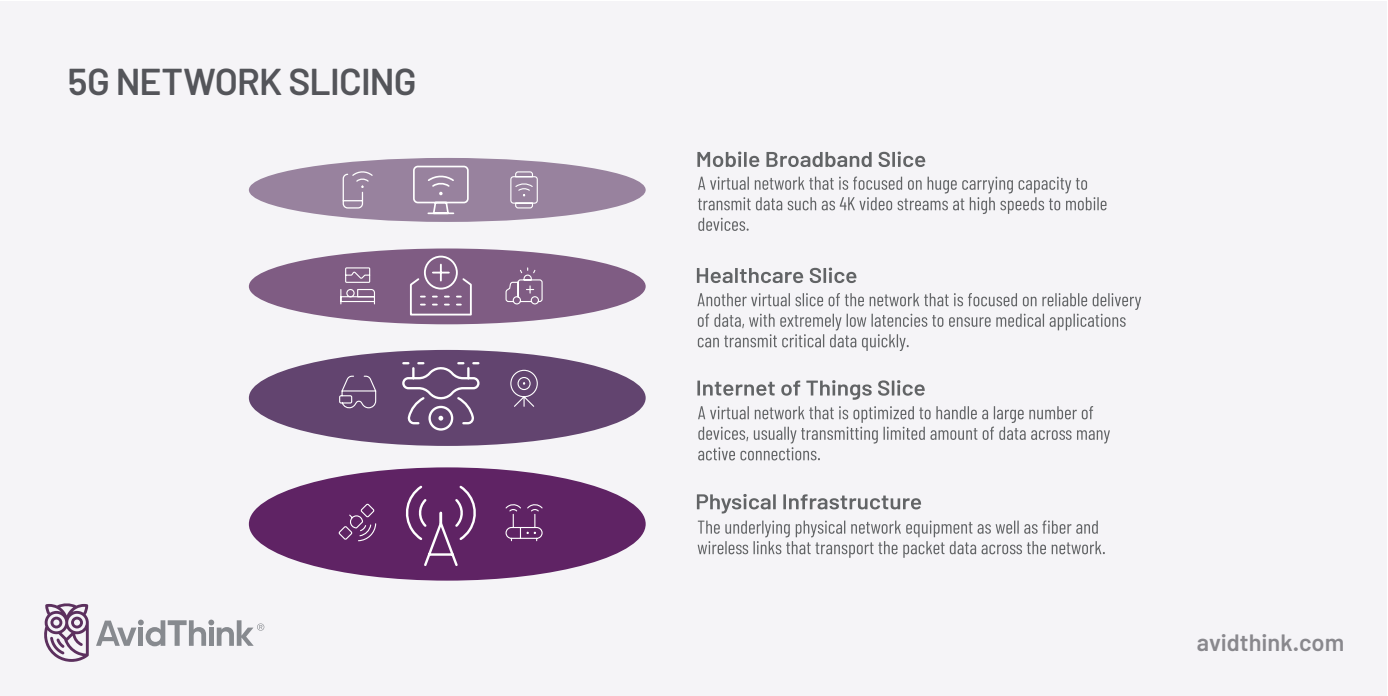
The success of 5G is, therefore, dependent on having edge infrastructure available to run these UPFs and keep overall latency low. Given that the edge infrastructure needs to be built out to support 5G, MNOs see the opportunity to run other functions at these edge locations too. We'll discuss the role of the edge shortly when we look in detail at the various 5G and edge use cases. While the rollout of edge services isn't necessarily dependent on 5G, as we shall see, 5G brings new use cases to the edge and the combination is highly synergistic.

5G Network Slicing

When all the different 5G technologies are put together: 5G NR, edge computing, NFV, SDN, and 5GC, new capabilities unique to 5G, like networking slicing, can be achieved. As has been alluded to, network slicing provides MNOs with the ability to create multiple network slices running on the same physical network, each with different characteristics and virtually isolated from each other. For instance, you could have a network slice that was super reliable and with ultra-low latency designed to carry messages from roads to the vehicle as part of a hazard warning system. Or you could have one that had massive bandwidth but wasn't necessarily

[1] Light travels about 200 km per millisecond in optical cables. However, there's delays in processing in the networking equipment (opto-electrical conversation, physical layer processing, etc.) as well as the packet-handling and application logic in the software stack, that adds to overall latency.

optimized for latency to enable fast downloading of movies in an airport prior to takeoff, so travelers can grab their last-minute television drama downloads.



Fundamentally, network slices can satisfy different service-level agreements (SLAs) that the MNOS would sell to enterprises. The ability to guarantee an end-to-end quality of service (QoS) is what carriers are hoping businesses will pay for. Further, enterprises might value the ability to create isolated slices that ensure only that enterprise's devices can reach each other. Automakers might pay money for slices for only their vehicles to use. Municipalities might ask for a civil defense slice that had a high priority for their use during natural disasters. Gaming companies could offer their cloud gaming patrons access to an ultra-low-latency slice to improve the gaming experience. We're still early in the 5G journey, but we expect lots of creative offerings and business models that will emerge as a result of network slicing.

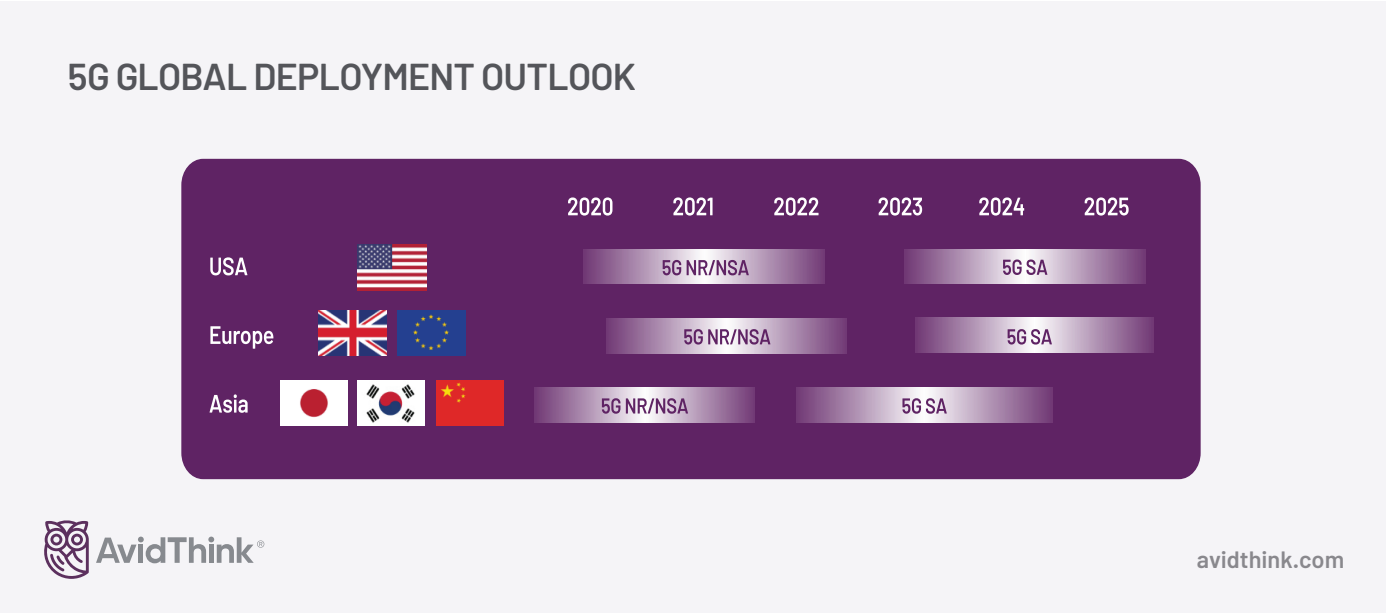
COVID-19 Impact on 5G Rollout

We discussed MNO strategies with regard to spectrum use; now, we'll look at expected deployment timing worldwide. MNOs have been rolling out 5G deployments as the standards were being defined. Most of the 5G deployments utilize features from Release 15. Beating the U.S., China and Japan, South Korea was the first to roll out a 5G commercial network on April 3, 2019[2]. A year later, in April 2020, South Korea boasted over 5 million 5G subscribers. The three MNOs in South Korea have rolled out mid-band 3.5 GHz 5G NSA networks, and they plan to deploy mmWave 5G SA later in 2020 despite COVID-19.

In general, most of the interviewees for this report believe that COVID-19's impact on the timing of 5G rollouts will be minimal. Due to the pandemic, some spectrum auctions have been slightly delayed, and local permits required for a cell tower or fiber buildouts are taking longer than usual. The prevailing view is that the pandemic will delay rollouts in some countries by at most six months but have little to no impact on deployment in the Asia-Pacific markets.

[2] ZDNet "South Korea's 5G goes live earlier than scheduled to claim 'world's first' title" https://www.zdnet.com/article/south-koreas-5g-goes-live-earlier-than-scheduled-to-claim-worlds-first-title/

As we look across the different regions in the diagram below, we note that most 5G deployments start with 5G NSA and are followed a few years later by 5G SA. The SA networks are usually associated with mmWave densification in areas with high populations. American and European deployment schedules are about the same, with the U.S. leading slightly. But Asia-Pacific leads, particularly in China and South Korea, who have more aggressive 5G deployment timelines.

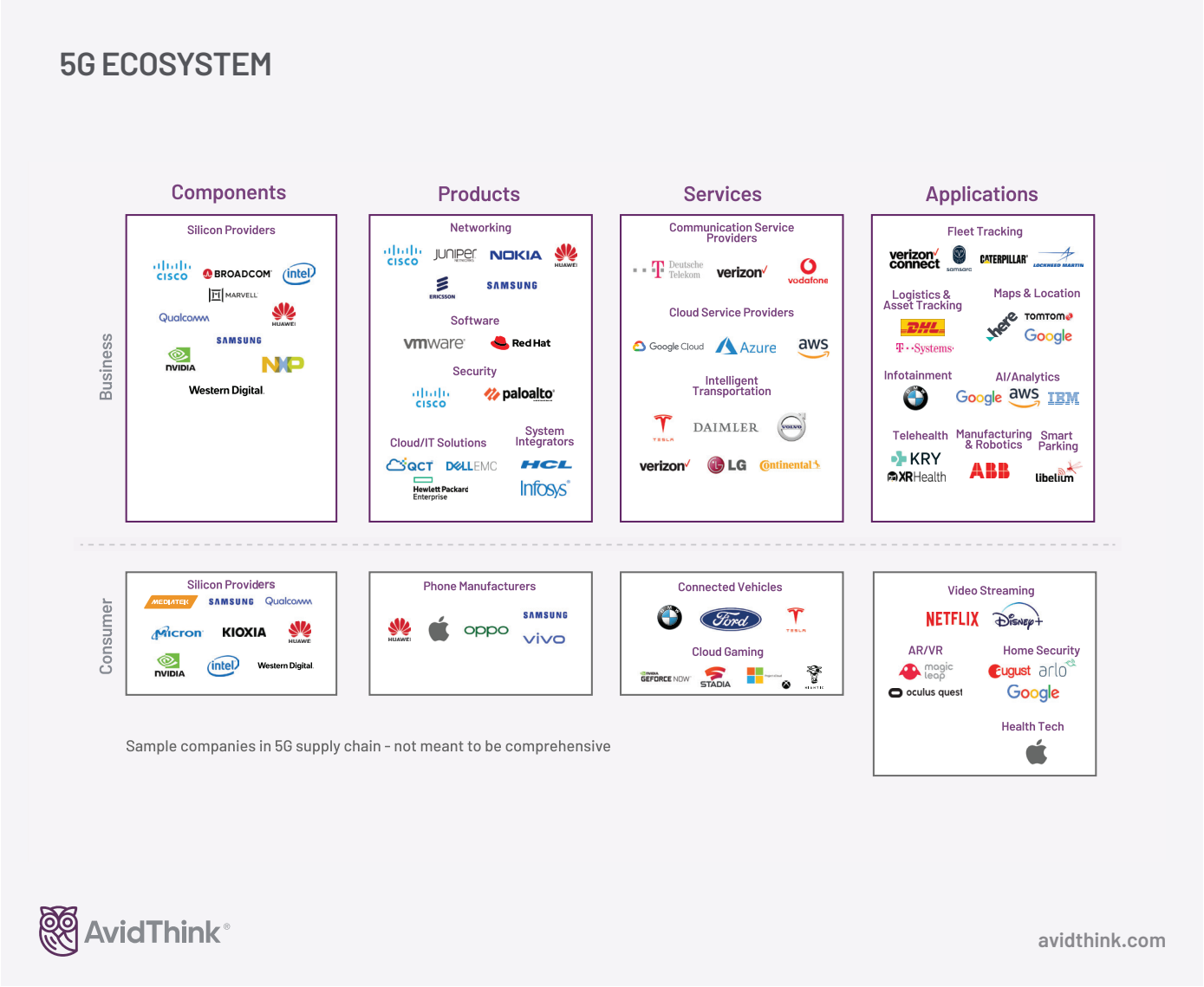


Other countries will lag by a few years, but 5G infrastructure rollouts will become more cost effective over time as best practices are learned and codified. The expectation is that by 2025, approximately 55% of the world's population will be covered by 5G with most of the subscribers coming from the U.S. and North-East Asia, followed by Western Europe. By 2030, the expectation is that all carriers will migrate to 5G as the primary network, though by that time, some interviewees believe that we'll already be started on 6G deployments (note: 6G is not well-defined yet).

The 5G Ecosystem

Historically, from 2G to 4G, the mobile ecosystem was quite stable. Network Equipment Providers (NEPs) would write their own unique software in-house, while sourcing hardware components from semiconductor companies and commissioning original design manufacturers (ODMs) to build hardware appliances to their specifications. In turn, the NEPs would sell the appliances and their software to the MNOs, who would then use them to build and operate mobile networks. Aside from standards being defined by the standards bodies at the interface level, there was limited sharing between NEPs, or MNOs for that matter. On the user equipment (UE) side of the equation, phone manufacturers would build their feature phones — later smartphones — out of a diverse set of components, including chips from semiconductor companies, and certify them for use with different RANs and carrier networks.

While that is still true today, the current ecosystem (depicted below) is a lot more diverse. In addition to the previous generation players, 5G brings about more openness on both the software and hardware front. The move towards disaggregating proprietary hardware network equipment and the rise of NFV has opened up the market to data center server manufacturers and independent software vendors. With 5G software stacks that can be run on general x86 or ARM CPUs and increased availability of open source for 5G RAN, many more can play in the 5G market. Likewise, through the adoption of cloud-type platforms by 5G software players, even hyperscale cloud providers can act as viable hosts for running the operational software that power MNOs.



MNOs also have to contend with the evolution of device diversity in user equipment, and they are scrambling to certify a host of new device types on their network. While an open ecosystem can help drive costs down for MNOs who now have more architectural choices and are less beholden to the NEPs, the same openness creates new competition for the MNOs and threatens their ability to profit from their 5G buildouts.

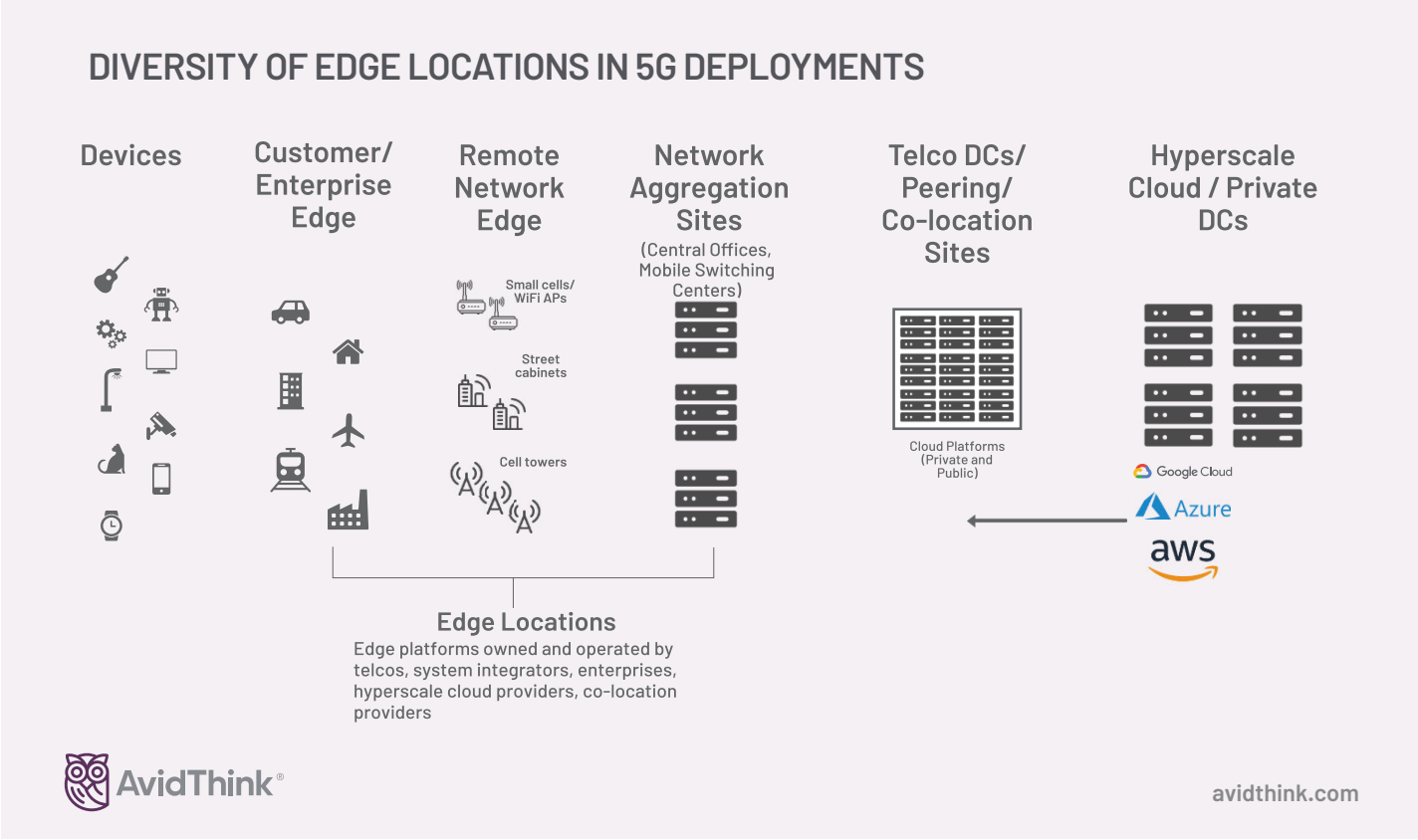
5G Use Cases and the Impact on Data

Now that we've covered the basic tenets of 5G and the importance of edge computing, we'll explore some of the top use cases as surfaced during our interviews. By examining the data flow for these use cases and understanding how they create, process, transmit, and store data, we can understand the impact of 5G on storage.

As established earlier, it's hard to treat 5G as a pure connectivity play. The edge is an integral component of 5G, and before we go further, let's drill down a little further into the various types of edges.

The Edge is a Continuum

Perhaps it's easiest to start with a diagram that shows the diversity of edge locations in 5G deployments.



Generally speaking, there's three classes of edge that we refer to in 5G use cases.

Cell Site or Cabinet Level (Remote Network Edge)

The first, which we can dispense with quickly, is the cell-site or street cabinet edge. This is the edge that is within 10 km of the RAN and the likely location to process sub-1 ms workloads in the future. While some carriers and new-generation edge data center providers are experimenting with deploying edge infrastructure in these locations, most of the MNOs we interviewed are not seriously pursuing buildout in such sites at this time. There are numerous reasons for this, including difficulty with power budgets, space constraints, temperature, environmental issues, physical security, and the immaturity of software orchestration to manage

so many locations. In addition, there are limited use cases today that require these remote deployments, even less that can generate the necessary revenue to justify the massive costs of building out and operating these highly remote edge locations.

However, as 5G increasingly densifies, and new use cases for low-latency arise, these locations could see increased buildout with unique storage needs in the form of higher endurance, better reliability, and very likely, a need to support encryption-at-rest due to the potential for theft of these remote systems.

Customer or Enterprise Edge

The next edge that we will discuss is the enterprise edge that lives on or near the enterprise campus. As discussed, the initial 5G revenue will be driven and funded by B2B use cases. MNOs hold the belief that if they can convince enterprises to let them manage edge infrastructure on their campuses, MNOs can use that infrastructure to power campus-wide 5G networks as well as host cloud services for enterprises. These applications would have very low latency for enterprise users since they would be running onsite, but on infrastructure managed and provided by the telcos. It could allow telcos to compete with hyperscale cloud providers. Unsurprisingly, the hyperscale cloud providers have the same ideas, and Amazon, Microsoft, and Google all have various offerings that provide a cloud edge on enterprise campuses.

Private Networks — LTE/CBRS/5G

One of the hottest opportunities in 5G is ironically not necessarily 5G-specific, nor does it exist on the mobile network. Instead, it's the use of mobile technology stacks — the 4G EPC and 5GC we've discussed in conjunction with 4G or 5G radio access points — within enterprise campuses, hospitals, factories and warehouses, public venues or in shipyards or airports. Mobile technology turns out to be more efficient, reliable, and secure than WiFi networks and can be an excellent alternative to replacing existing wireline networks. These private networks, in conjunction with a local edge computing platform, represent a potential opportunity for MNOs, cloud providers, and systems integrators. Most of the networks that are built today use LTE technologies, though our research indicates a willingness to switch to 5G mmWave systems when available. The expectation is that the industry will learn how to build out mmWave for private networks at a low enough cost to make them a feasible alternative in a few years. In the U.S., there are also high expectations for the use of the lightly licensed Citizens Broadband Radio Service (CBRS) spectrum. A new shared spectrum framework within private LTE networks, CBRS provides enterprises the use of the 3.5GHz spectrum in an unlicensed manner with the ability to bid for licenses to slots of local spectrum rights (Priority Access Licenses) as an add-on for more bandwidth.

Telco or Cloud Edge (including network aggregation sites)

We now come to the telco edge or cloud edge, if you're a hyperscale cloud provider. For MNOs, these are usually the mobile switching centers that aggregate traffic from multiple downstream cell sites. In the U.S., depending on the carrier, there may be anywhere from 50 to 200 of these mobile switching centers. They look like small data centers with multiple racks, often with raised floors, cooling systems, and significant power systems with backups. Hyperscale cloud providers are looking to install their edges in roughly the same locations, either in partnership with MNOs or with other colocation providers. The switch centers are ideal locations given their proximity to the RAN while providing a data center-quality environment for server hosting.

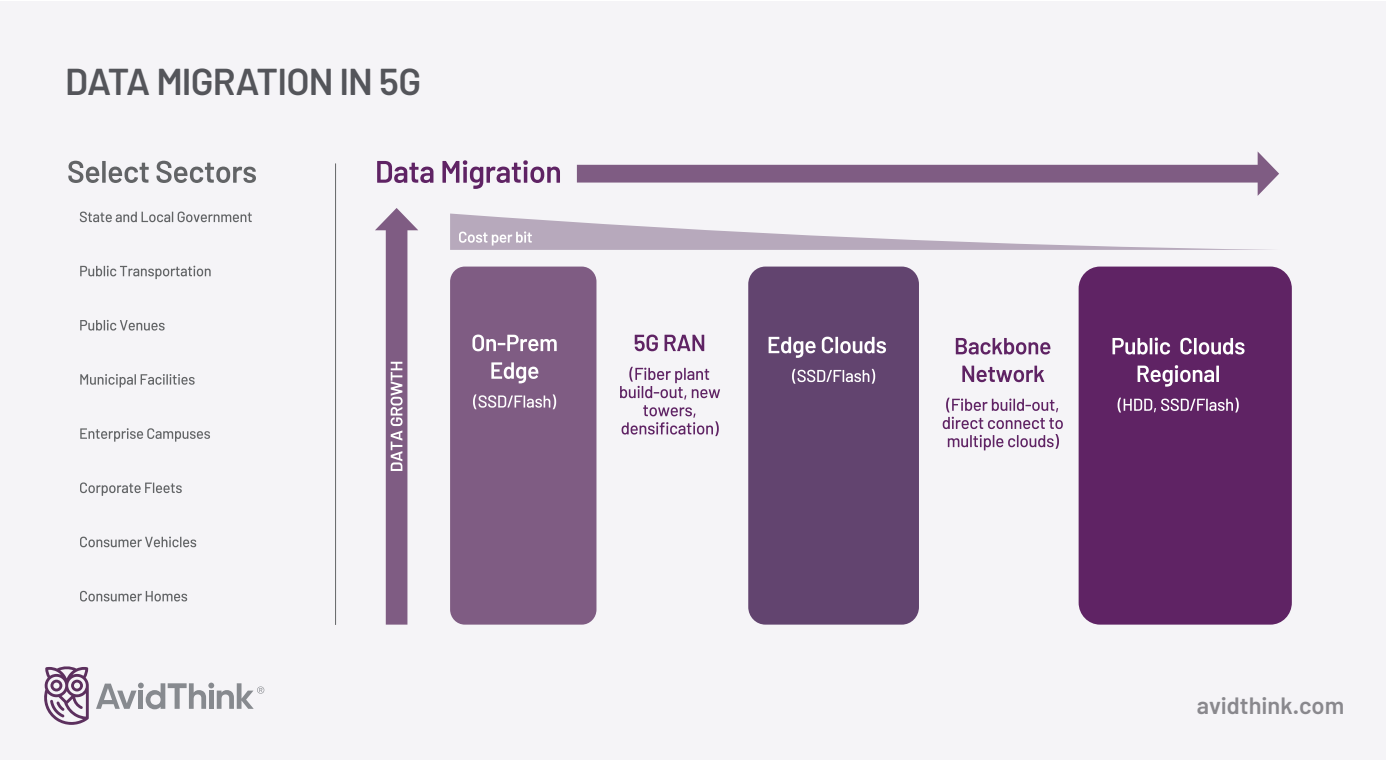
In the end, there is no single edge: from the telco edge to the enterprise edge and the cell-site edge, there are multiple options to host 5G applications.

Observations on 5G and Data

Now that we've laid the groundwork of 5G and the edge, we'll share observations and analysis based on our research. First and foremost, it turns out that 5G is mostly orthogonal to storage growth trends. Of more significant impact to storage is the increased resolution of cameras (still and video), growth in the use of sensors and IoT, and the ongoing digital transformation of our daily life and businesses (automation, robotics, telemedicine). Second, 5G does impact where data gets stored. A high-speed, low-latency network facilitates the movement of data from end-user devices to upstream locations where processing and storing could be more appropriate. Third, 5G impacts the viability and adoption rate of new use cases and applications that generate and consume more data, and so has second-order effects on storage.

Data Flow in 5G

What we expect to see with 5G is a set of data flows that we depict in the diagram below. For example, a data flow that goes from end-user or enterprise device to perhaps an enterprise campus edge before making its way to the telco edge cloud and then off to a centralized cloud (telco or public). For a consumer out and about, the data would go from the mobile phone to the telco edge cloud and then perhaps to a centralized cloud.



At each of these edge locations, data processing can be performed, and some of the data can be stored locally. As we proceed from left to right, the cost to process and store a single bit of data decreases due to economies of scale. Likewise, the compute and storage capacity increases as we move to the right, as does latency. Following this logic, absent any other factors, as the 5G network provides higher capacity at a lower cost, then data should migrate from the left to the right. Thus, end-user devices should see a decrease in storage capacities since they can quickly push data they've captured or generated to a more cost-effective storage location, where it can be retrieved rapidly again with low latencies.

User Device Storage Trends with 5G

Despite the natural tendency of data to migrate upstream into the cloud, our research and analysis indicate that there are mitigating factors which slow down this process:

- Image and video sensors will continue their dramatic increase in resolution, capturing 4K then 8K video and ever-higher-resolution images.
- The consumption of higher-resolution video content at 4K or more.
- Lack of complete coverage across all locations — 5G network coverage will not be 100%, and the highest speeds will likely be spotty for many years — will require a significant buffer on the mobile device to ensure ongoing operation when the network is slow or unavailable.
- An increase in the size of the working data set on the phone to support more sophisticated manipulation or consumption — e.g., local video processing and editing, local buffer for image browsing.
- Disconnected operations with air travel or deep indoor locations.

The expected result is that we'll continue to see growth in storage capacities on phones and other end-user devices — don't be surprised to see phones with an average of 1 TB of flash storage soon. Though as 5G becomes more pervasive, it will slow the rate of growth of capacities on these devices. Nevertheless, the most significant increase in storage capacity will be at the edge and public clouds, as the offloaded data continues to migrate upstream for processing and storage.

In terms of the speed of storage on user devices, our expectation is that local needs rather than network needs will again drive the speed of the storage interface. Experts believe that existing technologies like Universal Flash Storage (UFS) 3.0/3.1 and NVMe are more than adequate to handle the transfer from and to 5G networks. Devices capturing 4K-8K video will be dealing with speeds of 4-20Gbps or more, based on color depth, dominating any data transfer rates from 5G networks.

Edge Device Storage Trends with 5G

With the growth of data being upstreamed or processed at the edge, there will be a growth in demand for flash-based storage. For example, the edge is a harder place to perform a hardware refresh. As such, there's a desire for higher reliability and smarter storage controllers that can increase the endurance or lifespan of flash. Flash also provides the necessary low latency for roll-out of URLLC network applications. And for cabinet or cell-site installation, the environmental endurance of certain flash types can be a good fit, as they tolerate larger temperature and humidity ranges than would be experienced in a data center.

Outside of cell-sites or street cabinets, most of the interviewees believe that hardware architectures at these telco edges will be similar to that of data centers. As such, from a storage interface perspective, we expect NVMe on local servers and NVMe-oF (ROCEv2) or NVMe/TCP for centralized storage in a rack or half-rack.

In addition to usual data center needs, due to the remote nature of these installations and reduced physical security — whether at switching offices or enterprise edge or in cabinets and cell sites — there's more sensitivity towards theft and physical compromise. As a result, experts interviewed are citing encryption and security as even more critical in these locations when compared to central data centers. Platform capabilities like Silicon root-of-trust and trusted platform modules came up in multiple discussions. The consensus is that just as data-in-motion is encrypted today across the web to protect privacy, data-at-rest in edge locations will be encrypted in the event of location breach, loss, and theft.

Top 5G and Edge Computing Use Cases

Moving from the general to the specifics, we'll examine the impact of the most popular 5G and edge computing use cases on storage. AvidThink has been tracking multiple use cases as part of our research, and the viability of many of these top use cases was confirmed in our series of interviews. The principal use cases that came up in our study were:

- Telco network functions (not an external use case).
- Content delivery networks.
- Connected car and autonomous vehicles.
- Industry 4.0.
- Video surveillance.
- Cloud-based gaming.
- Telemedicine.

There were others, including drone-based surveillance and property/site inspection, smart buildings, smart cities (which we reference under video surveillance), sports and entertainment venues, AR/VR, fixed wireless access (primarily a straightforward replacement for wireline connectivity), mining, smart ports, disaster, and first-responder support, plus numerous more. To keep the report length manageable, we've picked the critical use cases that were consistently highlighted by our interviewees.

Storage Impact

Increase in the use of storage at core and edge data centers due to use of NFV, virtualized RANs, collection of real-time analytics and use of AI/ML.

Telco Network Functions

Unsurprisingly, network functions (NF) as part of the 5G core will be one of the vital edge purposes. As discussed, one of the premises of 5G networks is that NFV and cloud technologies are needed to achieve the flexibility and scale efficiencies required to bring about the improved capacity, reduced latency and reliability.

Whether NFs are part of a virtualized RAN deployment or in support of other network services (such as security), we'll see many variants running at the edge. The use of virtualized or containerized NFs will require storage for keeping copies of the virtual images of software functions, as well as additional space for snapshots (to assist in recovery). Compared to the fixed-function proprietary hardware equipment of the past (4G and earlier networks), an NFV software-centric platform with multiple disaggregated software components will prove more agile and scalable but will also consume more storage.

Many of these network functions also capture a large amount of telemetry data and perform ongoing real-time analytics, all of which use storage. Finally, with the increasing complexity of managing network functions, telcos have started to employ AI and machine learning (ML) to augment human operators and eventually run these systems autonomously. AI/ML training will require high-performance storage for training, likely in the core data centers, but also high-performance data stores for real-time inference at the edge.

Content Delivery Networks (CDNs)

CDNs are a natural use case for the edge, and with the increase of video resolution from HD to 4K to 8K, the storage requirements for CDN will continue to balloon. As AR and VR use grows with 5G, CDN systems will store a considerable amount of data to augment the AR/VR systems. For instance, in AR-enhanced tourism, a CDN might be called upon to provide fast access to short overlay movie clips superimposed on a specific view.

Depending on the evolution and maturation

Storage Impact

Increased use of storage for CDN at the edge will be primarily driven by increased content resolution (4K, 8K), and, potentially, CDNs serving AR content as overlays now and VR content in the future. 5G is not a direct driver but facilitates and promotes delivery and consumption of high-resolution content across mobile use cases, in stadiums, in enterprise locations and at home, and therefore acts as a secondary driver of storage growth. Note though that the more capable CDNs become, the need for storage on devices or in homes will go down.

of video transcoding capabilities, edge locations may turn into real-time transcoders of video streams, as mobile devices pull down media at the most appropriate resolution for the device type and network speed. In terms of the impact of 5G on CDNs, the net increment will be due to new use cases, like AR, but predominantly due to the consumption of higher-resolution content — 4K or 8K video, at up to 15 times the file size of 1080p HD video, simply requires more storage.

Until we get to mmWave and ultra-low latency networks, it is unlikely that VR content will be stored at the cloud or telco edge. Requiring high-speed flash and SSD storage onsite, it is more likely to be hosted at the enterprise on-premises edge. However, this is not directly related to 5G as a driver.

Connected Car and Autonomous Vehicles

The connected car is one of the most popular and significant use cases for both 5G and the edge. With the increase in compute, storage, and networking capability, many vehicles will look like a mini data center on wheels. They might even qualify as a nano-edge location. Regardless of in-vehicle applications, the very digitization and conversion of the vehicle into a computing platform will already increase storage use. The car, like a computer, will need a place to store the operating system, application modules, and ongoing log data.

There has been significant hype in this area, which we'll try to tamp down. Let's start by distinguishing two main classes of automobiles:

- **Autonomous vehicles or self-driving cars:** these are vehicles with the ability to drive with little human intervention. They may be trucks or lorries that can ferry goods from one point to another or self-driving taxi cabs that pick up passengers and ferry them to their destination.
- **Connected cars:** these automobiles are highly connected and automated but still piloted by humans. Like autonomous vehicles, they may use AI/ML onboard to optimize the driving experience and connect to the roadside infrastructure to obtain data or stream infotainment from the network. Still, a human is in charge at all times.

Based on our research, fully autonomous vehicles will take a while before they are on the road in any significant numbers — five or more years. Most likely, commercial fleets traversing a set of common routes between warehouses and factories will go fully autonomous first. They may be followed by public transportation and eventually consumer vehicles. However, this may vary by region.

Regardless, the concept of using edge compute and 5G to handle the autonomous driving for vehicles, touting the use of sub-1 ms latency is mostly hype and conjecture today. All interviewees we spoke with indicate that the intelligence for autonomous cars will be on the vehicle. However, the vehicle could utilize localized information feeds on road conditions from edge servers. Therefore, most of the compute and storage systems will stay on the vehicle. This approach also protects against situations where the 5G coverage is inadequate or when the network goes down.

Nevertheless, 5G will enable vehicles to be infotainment hubs, allowing fast streaming or

Storage Impact

The primary storage impact in vehicles comes from the use of more sophisticated control systems and computing intelligence, as well as gathering of sensor, video and telemetry data. These are independent of 5G, especially since most experts today don't envision offloading the data in real time over 5G networks. And while 5G will enable real-time updates to maps, road conditions, and eventually, autonomous systems aided by edge infrastructure compute, these too have limited impact on the type or amount of storage on vehicle. However, 5G will enable streaming of high-resolution infotainment into vehicles, likely with mixed impact on storage — on one hand, driving up consumption of high-resolution content and potentially requiring larger in-vehicle buffers to maintain continuity during network glitches, but on the other, also improving download speeds and enabling on-demand consumption, reducing in-vehicle storage.

downloading of videos to the car. 5G will enable more interactive browsing or game playing from cars. With the rise of 4K and 8K videos, and because 5G coverage will still be spotty over the next three to five years, most vehicles will continue to see increases in onboard storage to accommodate downloaded content. Likewise, personalized consumption, where each passenger watches their own stream, will add a multiplier effect on storage needs.

Independent of 5G, the connected car will drive up storage use. With more sensors and video cameras — a vehicle could easily have four or more ongoing video feeds — the need to store these streams while processing them or archiving till they can be offloaded will push up storage capacities on vehicles significantly. In both connected and autonomous cars, storing ongoing telemetry data and video data in a black box will also create new storage requirements for fast writes, increased durability and high reliability. Especially for commercial fleets, these types of data will be used for debugging and post-trip optimization — mainly ongoing AI/ML training. And in the early days of autonomous driving, fleets will choose to capture more information until systems prove reliable, thus driving up the use of storage. This is the one area where interviewees indicated a need for innovation to solve the black box problem. To achieve the constant high write speed, durability and reliability while keeping costs down appears to be a current challenge with today's flash and SSD technologies.

Ironically, even with the rise of 5G network capacity, most involved in connected and autonomous vehicles do not believe the bulk of vehicle data will be offloaded in real-time. They cite costs as being the primary issue —transferring that large a data set over the air will be expensive — and suggest instead that data will be offloaded during charging or refuel or at depots at night. At the same time, 5G will enable over-the-air (OTA) updates of vehicular applications, including software, mapping information, AI/ML model updates, and AR feeds.

5G will also play a role in safety. With Release 16 and beyond, there are updates for vehicle-to-everything (V2X) communications between vehicles, to pedestrians and to the roadside infrastructure. The communication will serve as an augmentation of on-vehicle safety systems helping to avoid collisions and improve the driving experience. If vehicles detect (through driver intervention or otherwise) that there are hazards on the road, that information can be transmitted in real time via 5G to the edge and the cloud where it can be processed and pushed as OTA to all other local vehicles, pedestrians, and infrastructure. Similarly, if maps are discovered to be outdated, or new hazards or conditions are detected, 5G serves as a fast way to transmit that data. Storage will need to be allocated for all this data and logging information.

Industry 4.0

Another promising area for 5G and edge is Industry 4.0, which will use 5G and edge locations to host factory and industrial automation systems. The premise is that with 5G, enterprises could run their IT applications on edge systems managed by carriers or others. 5G provides the bandwidth and low latency that allows these systems, which might be tens of miles/km away from the factories or warehouses, to control robots, forklifts, and other automated systems effectively. The advantage, of course, is that the IT systems are now managed and hosted by a third-party, reducing upfront capital investment and possibly providing improved SLAs. This will drive more storage growth at the edge but reduce it onsite.

Storage Impact

Over time, 5G will probably cause migration of storage from enterprise campus premises to public cloud or telco cloud edge locations, enabling Industry 4.0 workloads to run on cost-effective cloud services while maintaining low-latency access. Initially though, most Industry 4.0 workloads will stay on premises as the early use cases have proven. As for the capacity and growth of storage for Industry 4.0, the key drivers are less around 5G and more around digitization and industrial automation.

Another potential advantage of using the edge is data jurisdiction compliance. In some regions, centralized clouds may be located off-country, which makes using the compute and storage resources difficult when government mandates prevent specific sensitive data from leaving the country. An edge infrastructure could allow companies to benefit from cloud services

without violating regulatory compliance.

However, our research indicates that most of the initial Industry 4.0 initiatives will probably use private networks on campus — either private 5G networks or, more likely, private LTE-based networks. In these deployments, the edge infrastructure will sit on campus, though the racks could be managed by either telcos, system integrators, or even hyperscale cloud providers. In those scenarios, storage will continue to grow significantly on location. For large campus areas, like airports or ports run by municipalities or private firms, the current prevailing wisdom is that edge compute and storage will stay onsite. 5G radios could be leveraged for the campus network, with all the attendant benefits, and could be used as a connection to external edge and cloud systems as a transmission mechanism.

Perhaps the situation could change in two to three years with companies moving to external edge clouds. Still, the current consensus is that 5G, public and private, will not cause a migration of storage off campus.

Video Surveillance

Another hot use case is using 5G and edge to transmit, process and store video surveillance data. Whether deployed in homes, at enterprises, in stadiums, at roadside intersections, or public locations in smart cities, video cameras have become a pervasive part of our lives. The use of video analytics for safety and security has risen. Despite the issues with privacy, it doesn't look like global video surveillance will slow its march any time soon. 5G provides the ability to stream higher-resolution feeds (9 megapixels and above) into video analytics systems powered by AI/ML, all of which can now live at the edge.

In video surveillance, there are classes of intelligent cameras that run real-time algorithms for face detection, person detection, or event triggers (e.g., safety zones violations) and only transmit recognized events to save on bandwidth. However, our research indicates that many deployments on enterprise campuses and in smart cities will use centralized storage and processing.

New 5G-enabled cameras are showing up in smart cities. These cameras simplify the deployment of surveillance — high-quality images in real time with no need to run wired connections. Many of these cameras will have local storage buffers at minimum to handle retransmission or delayed transmission, but for public use cases, they will likely have limited storage. Experts feel that the potential theft or destruction of these cameras will be an issue and prefer to have little storage on the devices.

Storage Impact

5G will allow video processing and storage to move from on-premises to edge locations and eventually to cloud. It facilitates a migration of storage capacity while the growth in capacity is driven primarily by the image capture resolution.

Private 5G for enterprises and public 5G (perhaps a network slice reserved for municipalities) will facilitate the real-time streaming of videos to either edge or cloud data centers for ongoing processing and archival. By performing centralized processing in a secure facility, the cost of the cameras can be kept low, and footage will not be lost should the cameras be damaged or stolen. Algorithms can be easily upgraded, as well as any hardware acceleration needed to improve the accuracy or processing speed, without having to touch the cameras.

For this use case, storage capacity will be driven by the trends in video resolutions. Moving from 1080p HD to 4K translates to a four-fold increase in storage per stream, multiplied by the proliferation of the cameras. 5G's role is to facilitate the offloading of streams to either a local edge site, where the videos are stored on flash devices or even HDD (with a much lower cost per bit), or the cloud for longer-term storage.

Cloud-based Gaming

5G and the edge hold promise on the gaming front as well. With lower latency and the ability to stream large amounts of data from edge servers, the gaming industry is closely looking at creating new multi-player (or even single player) games by leveraging 5G. 5G cloud gaming is one of the flashiest demonstrations of 5G and the edge and is often spotlighted at tradeshows.

Storage Impact

While touted as a poster child for 5G and edge deployments, 5G is unlikely to enable streaming cloud gaming in the immediate future. However, once 5G networks mature, the expectation is that storage will migrate from end-user devices to the edge.

Early experiences from the carriers and hyperscalers, however, indicate that cloud-based games aren't quite ready for general consumption yet. Pro-gamers have avoided new systems like Google Stadia and its cloud console concept. This likely is a situation of chicken-and-egg where 5G coverage needs to be sufficiently ubiquitous and fast before cloud gaming takes off. Perhaps game application developers need to learn lessons in improving the user experience while accommodating tiny fluctuations in network latencies or speeds.

For the next three to five years, the expectation is that gaming consoles, with 4K or even 8K scene renders, will continue to consume an increasing amount of storage. This will remain true until 5G gets to the point where some of this storage can be offloaded and aggregated at edge sites. At that point, 5G may trigger the migration of storage from the user devices and homes into edge clouds.

Telemedicine

One of the areas that's risen in prominence due to COVID-19 is telemedicine. Not the dramatic demonstration of remote surgery enabled by 5G that's been the subject of overhyped presentations, but remote diagnosis during consults and wellness doctor visits. While experts believe that eventually remote surgery or remote-assisted surgery might happen, the regulatory environment, complexity, and liability will slow down that use case for many years, independent of 5G network capability.

However, 5G enables access to high-quality two-way video conferencing and also the transmission of diagnostic and medical sensor data. It provides remote consult in situations where the patient is hard to reach, has limited mobility, or there are needs to save time and travel. The COVID-19 pandemic has demonstrated that for many ailments, remote diagnosis and triage can eliminate the hassle of trudging to urgent care, thereby reducing cross infections and improving the quality of medical care by shortening the time to get to a nurse advisor. It's possible that, in time, some of these consults may have to be archived for insurance and regulatory reasons (though subject to strict confidentiality), thus driving up storage needs.

For wellness, there are also applications where elder care is augmented by systems that can detect issues like a fall or lack of movement. These video camera systems will often upstream the data but will also have a local storage component to buffer hours or a day of data in the event of a network disconnect. These systems, augmented by the use of video analytics with AI and ML, can detect and alert caregivers when potentially life-threatening events happen.

Ultra-low latency and high-reliability 5G networks can also expand the use of home care, where medical devices at home can transmit a rich set of information reliably and quickly to health monitoring systems in the cloud or edge, enabling fast detection and response to home-care patients, dispatching ambulances or paramedics if something untoward is detected, or at the very least, a same-day or next-morning urgent care follow-up.

Storage Impact

The impact of 5G on storage in telemedicine will be mixed. 5G enables medical care to be performed in a more distributed manner, resulting in more data being gathered at home and driving up some level of interim storage. At the same time, that data will be upstreamed to edge and cloud locations, capping the need for large storage capacity in-home. As with other use cases, the primary drivers around capacity will come from the digitization of medicine — from the use of more imaging and sensor data to gathering data for analysis with AI/ML to spot trends and assist with diagnosis.

While expansive, the data impact of 5G in telemedicine is limited. Experts believe that the primary impact is upstreaming sensor and medical data and reducing the in-home compute and storage footprint.

Wrap-Up and Conclusion

Across our research, it is clear that the main drivers of storage use in the 5G era have less to do with 5G connections and more the digitization of everyday processes — Industry 4.0, entertainment, healthcare, public safety, transportation. One of the most significant drivers is the increased use of video data (more cameras everywhere) and the improved resolution of image sensors from 1080p HD to 4K to 8K. Likewise, the move towards disaggregated cloud architectures based on general off-the-shelf hardware coupled with virtualization requires more storage — to support the virtualized software images, store snapshots for reliability, and record logs for troubleshooting. Finally, the pervasive use of AI/ML across the board necessitates large volumes of storage for sensor, log, and image data in AI training and subsequently AI inferencing to drive automated decisions.

5G itself is more of an enabler of the digitization and its impact is felt in the location of the storage and data processing. As 5G becomes more pervasive and reliable, we'll see a migration of storage from end-devices, like cameras or phones, to edge locations and core cloud locations, where they can be processed and stored more cost-effectively. While we believe storage capacities on end-user devices will continue to grow for the foreseeable future, the rate of growth will be capped by 5G-enabled offloading.

In terms of storage technologies, interviewees cited primarily the same type of technologies currently in use with flash and SSD playing a more prominent role over time, especially at the edge. 5G network speeds will not present a problem for flash and SSD devices, which will easily fulfill their bandwidth requirements — just as they already do in high-speed 25Gbps, 100Gbps data center networks today. HDDs will continue to provide practical and cost-effective large-scale storage, especially with the amount of video and image data we will be generating and uploading through 5G connections.

In summary, it appears that 5G is somewhat orthogonal to the progress of storage technologies and capacities. Storage will be primarily impacted by the compute and software architecture needed to service new 5G use cases. In any case, today's storage technologies will stand up well to the challenges ahead, especially as storage continues to evolve to support additional performance and scale.

AvidThink, LLC
1900 Camden Ave
San Jose, California 95124 USA
avidthink.com

Rev A