Gartner

Hype Cycle for Edge Computing, 2020

Published: 31 July 2020 **ID:** G00450508

Analyst(s): Bob Gill, Thomas Bittman

Edge computing has moved from IoT-focused to a broadly considered complement to the more centralized hyperscale cloud. The edge role in distributed cloud and digital business ensures that despite the nascent market, technologies and architectures, edge computing is here to stay.

Table of Contents

List of Tables

List of Figures

Analysis

What You Need to Know

Edge computing is defined as "a part of a distributed computing topology in which information processing is located close to the edge — where things and people produce or consume that information." Where is this "edge"? Well, the *absolute* edge is that place where the physical and digital world interact, such as where sensors measure something happening in the physical world. They then translate those observations or measurements into data, which can be used to make a decision to take action, aggregated to look for patterns, or simply passed back to a storage or analytics applications for further analysis. The complement to the edge is the core, that is, the upstream system or systems that guide what is passed down to the systems at the edge, often serving as the centralized storage, processing and archiving "back end." The core can be a private data center but, increasingly, it is located in the public cloud. We look at the space between the edge and core to represent a continuum with multiple "edges" serving different roles, often in a hierarchy. Location-sensitive distributed computing is not a new concept; what is new is the explosive growth in the number of highly intelligent edge devices, coupled with the broad array of advanced services and massive scale of the cloud, serving as both orchestrator and core.

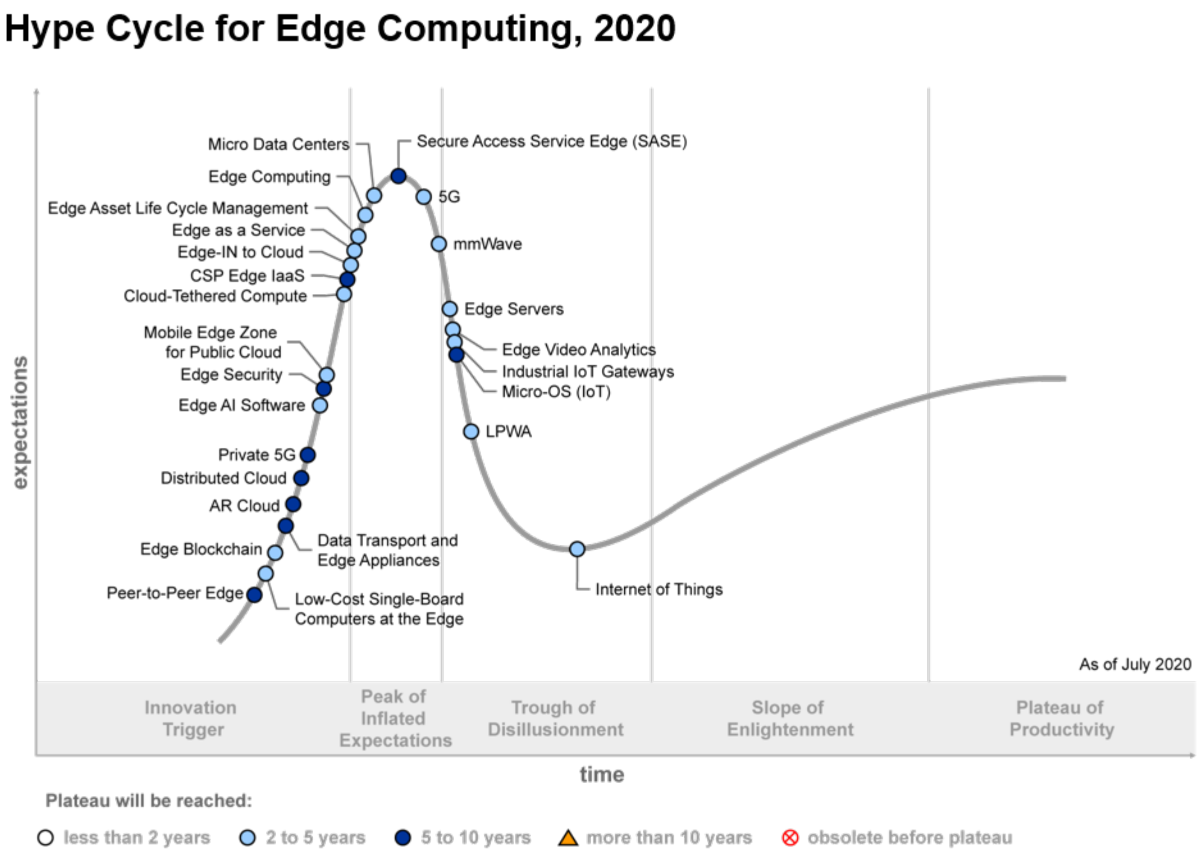
The Hype Cycle

Edge computing technologies and services are being driven by both newly emerging and integral complementary technologies, such as 5G multiaccess edge computing (MEC), as well as the application of adjacent horizontal technologies, repurposed and optimized for the edge, such as edge video analytics and edge AI software. The edge computing arena is filled with great promise; but, at this early stage, it also features significant hype. As system providers flock to fill the opportunity vacuum created by hyperscale cloud computing's dominance of the past several years, we are seeing "edge washed" or hyped offerings, from servers to communications services, marketed alongside true innovations. While the state of edge computing has evolved a great deal since last year's Hype Cycle, so has the hype, affecting the overall category's placement on the Hype Cycle curve. One of the goals of the Hype Cycle is to clarify what is hype versus what is valuable, including new roles and positioning for existing technologies, such as in the innovation profile on edge servers.

While edge computing is nascent, its position as the decentralized complement to rapidly maturing hyperscale cloud offerings, as well as the symbiotic relationship between edge and the Internet of Things (IoT), will compress the timeline to productivity. (See the profiles for IoT and cloud computing, as well as "Hype Cycle for the Internet of Things, 2020" and "Hype Cycle for Cloud Computing, 2020," for further context.) For example, IoT vendors have been building endpoint solutions for years now, and the hyperscale cloud providers, most notably Amazon Web Services (AWS), Microsoft and Google, are investing significant resources to bring edge-enabling software to market on an unusually aggressive timeline.

This Hype Cycle reflects the contrast between existing technologies adapted for edge on the plateau or the right side of the peak, while newer emerging technologies centered on edge and its use cases are shown on the rising, left side of the curve.

Figure 1. Hype Cycle for Edge Computing, 2020



Source: Gartner
ID: 450508

The Priority Matrix

Edge computing is composed of established technologies and architectures adapted for a new distributed computing model, as well as innovations that extend, or “fill in the blanks” around, such models to bring advanced application capabilities to adjacent spaces, such as cloud computing and IoT. This allows early adopters to prototype and implement solutions today, while standardized and accepted technologies to operate at the edge efficiently and securely at scale will emerge over the next two to five years. Over the five- to 10-year horizon, many such innovations are likely to be subsumed into mainstream application stacks, operating systems and management tools, such that edge computing becomes just an accepted and expected part of “computing,” much like protocol stacks and directory services are today.

Technologies and models expected to mature over the next two to five years include:

- Edge AI software
- Edge video analytics

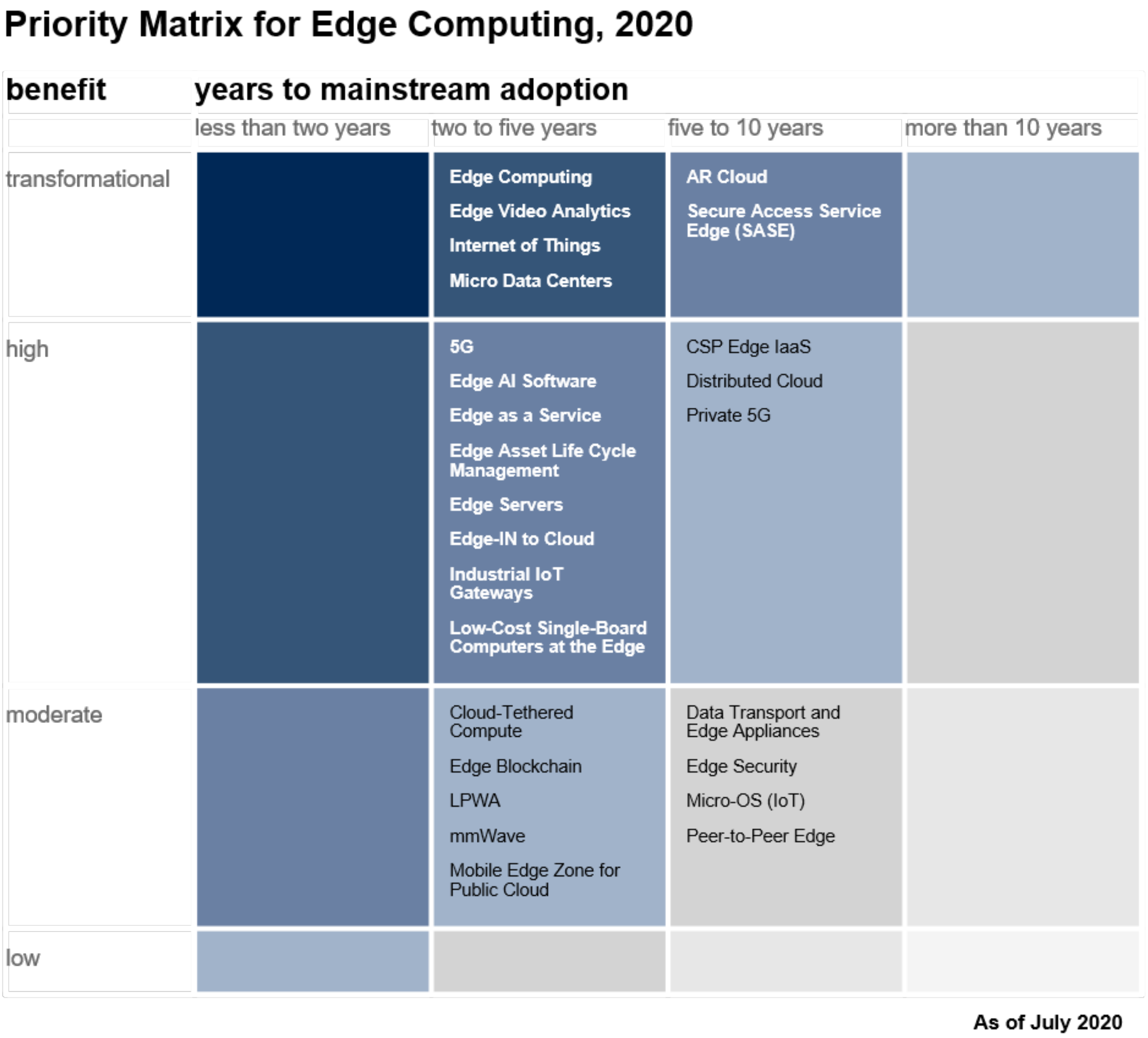
- Cloud-tethered compute
- 5G
- Edge as a service
- Edge-IN to cloud

Over the five- to 10-year horizon, technologies expected to reach mainstream status include:

- Augmented reality (AR) cloud
- Secure access service edge
- Peer-to-peer edge

Some of these technologies may be accelerated through the Hype Cycle if championed and invested in by either consortia effort (e.g., Linux Foundation's LF Edge) or a hyperscale cloud provider with extensive market following, such as Microsoft, AWS or Google.

Figure 2. . Priority Matrix for Edge Computing, 2020



Priority Matrix for Edge Computing, 2020

benefit	years to mainstream adoption			
	less than two years	two to five years	five to 10 years	more than 10 years
transformational		Edge Computing Edge Video Analytics Internet of Things Micro Data Centers	AR Cloud Secure Access Service Edge (SASE)	
high		5G Edge AI Software Edge as a Service Edge Asset Life Cycle Management Edge Servers Edge-IN to Cloud Industrial IoT Gateways Low-Cost Single-Board Computers at the Edge	CSP Edge IaaS Distributed Cloud Private 5G	
moderate		Cloud-Tethered Compute Edge Blockchain LPWA mmWave Mobile Edge Zone for Public Cloud	Data Transport and Edge Appliances Edge Security Micro-OS (IoT) Peer-to-Peer Edge	
low				

As of July 2020

Source: Gartner
ID: 450508

Off the Hype Cycle

- IT/OT hybrid servers, edge supercomputing, and HCI at edge have been consolidated into an enhanced edge servers innovation profile
- Cloud computing is considered mainstream.
- Edge networking has been split into several subsets for greater detail, including LPWA and peer-to-peer edge

- SD edge has been replaced with several profiles with greater specificity, such as mobile edge zone for public cloud
- Name data networking was dropped due to limited potential applicability to edge at this time.
- CSP multicloud edge providers was replaced with CSP edge IaaS
- Cloudlets has been subsumed into the distributed cloud profile
- NarrowBand IoT (NB-IoT) and Sigfox have been consolidated into the low-power wide-area (LPWA) profile
- IoT edge architecture has been subsumed into Internet of Things
- Telemetry-based routing has been relocated to the networking Hype Cycle
- Edge data fabric has been subsumed into the peer-to-peer edge profile

On the Rise

Peer-to-Peer Edge

Analysis By: Thomas Bittman; Bob Gill

Definition: Peer-to-peer edge computing enables distributed computing across an edge environment for resilience, workload orchestration, horizontal scaling, and interaction and cooperation between edge computing nodes using local or mesh networking as an enabling technology.

Position and Adoption Speed Justification: Much of the early focus on edge computing has been extending intelligent things to the cloud, and extending cloud capabilities closer to users and things at the edge. Edge computing is defined as a form of distributed computing, but that distributed computing is not only between the edge and the cloud, but also across edge processing nodes. The edge is very interactive — things need to interact with other things (e.g., sensors with actuators), systems of things need to interact with other systems of things (e.g., self-driving cars and smart cites), people need to interact with things (e.g., operators and maintenance workers in a plant), and people need to interact with other people (e.g., virtual and mixed reality collaboration between engineers). Edge computing will take place in layers, and in systems — embedded in cameras, aggregated to compute nodes, interacting with personal compute devices, working across complex systems. Peer-to-peer edge will evolve to enable:

- Resilience (when one edge node goes down), orchestration (balancing workload across compute capability based on latency and compute requirements)
- Horizontal scaling (leveraging compute power inherent in intelligent nodes)
- Cooperation (e.g., hand-offs between edge nodes as things move, providing services across nodes)

Diversity and lack of standards will limit peer-to-peer edge computing to specific use cases, but will gradually generalize to broader use cases as standards evolve.

User Advice: Peer-to-peer networking is relatively mature, but distributed processing at the edge is embryonic and the early evolution will evolve based on specific use cases rather than broad standards. Until broad solutions and standards appear, enterprises should be careful and only deploy custom solutions for peer-to-peer edge when the business case is very strong, and a complex and lengthy development and rollout process is worth the investment. Enterprises can expect solutions to enable basic connectivity and interaction at the edge to mature first. Resilience and workload orchestration (placement) solutions will take more time. Horizontal scaling and true distributed processing will take the longest, and will likely focus only on specific use cases (such as distributed analytics). As interactions at the edge between people and intelligent things continue to grow, peer-to-peer edge capability will grow in importance.

Business Impact: Peer-to-peer edge processing will enable more independence from centralized data centers, leverage compute and storage that will proliferate at the edge, reduce latency and response time for complex local computations, and enable collaborative immersive experiences. Peer-to-peer edge processing will also complement the power of centralized computing to unlock new use cases that cannot be achieved in a single edge node or in a data center far away. The power of machine learning and training at the edge will grow due to peer-to-peer edge capability.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: AlefEdge; Equinix; mimik; MobiledgeX; Storj

Recommended Reading: "Why and How I&O Should Lead Edge Computing"

"4 Steps to Successful Edge Computing Deployments"

"Exploring the Edge: 12 Frontiers of Edge Computing"

Low-Cost Single-Board Computers at the Edge

Analysis By: Tony Harvey

Definition: Low-cost single-board servers are small low-cost general-purpose systems that perform functions such as filtering data like anomaly detection or AI inferencing like image recognition at the edge. Based on a system-on-chip (SoC) solution, single-board servers are designed with the minimum capability to perform the tasks required. I/O interfaces will vary, but at a minimum will include a wired or wireless network. The operating environment will be based around a micro OS, VMs and containers to enable rapid delivery of updates.

Position and Adoption Speed Justification: Single-board servers at the edge are a relatively new development for processing data and delivering AI inferencing at the edge. Initially, the market was driven by the introduction of very low-cost single-board general purpose computers such as the Raspberry Pi. Now, the market has expanded with open-source microcontroller-based system like Arduino, and AI inferencing systems such as the Texas Instruments BeagleBone AI, and the NVIDIA Jetson Nano.

Unlike larger edge servers, which are generally repackaged x86 servers, single-board edge servers are fixed configuration single-board systems based on ARM CPUs. Opportunities exist for other CPU architectures such as x86 and RISC-V but the dominance of ARM in the SoC space will make it difficult for other chip architectures to meet the power and performance requirements to succeed in this space.

While the profusion of vendors and low costs associated with the hardware make prototyping and development very easy, the lack of standards and as importantly the lack of security features will make enterprise usage less likely. As the opportunities grow so will demand, for a more secure solution and a more standardized software environment for developers.

User Advice: Evaluate the use of single-board edge servers for edge projects where a large number of low-cost devices will be required to provide data processing, image recognition, voice recognition or AI inferencing capabilities. Expect this market to evolve rapidly over the next few years with improved performance and new capabilities being rolled out at a rapid cadence. Choose single-board edge servers that can be rolled out rapidly, without skilled staff on-site, and that can easily be managed and updated in the field. Security should be built into the system and potential vendors should be evaluated for security across all areas including, physical, data storage, communications, management, and updates. Integration with existing Internet of Things (IoT) and artificial intelligence (AI) frameworks should also be considered when selecting a single-board edge server.

Business Impact: Single-board edge servers can help enterprises realize the potential of the large pool of data that is generated at the edge. The ability to use this data has significant potential to generate cost savings, for example, by allowing real-time image processing to recognize faulty or damaged items on manufacturing lines. It also helps develop new areas of business that will be enabled through real-time data processing at the edge.

Enterprises that do not adopt single-board edge servers may find themselves left behind as enterprises that successfully integrate these systems into their digital transformation strategy will lower their costs and deliver new services to market faster.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Coral; NVIDIA; Raspberry Pi Foundation (Raspberry Pi); Texas Instruments (TI)

Recommended Reading: “Top 10 Strategic Technology Trends for 2020: Empowered Edge”

“How to Overcome Four Major Challenges in Edge Computing”

“Why and How I&O Should Lead Edge Computing”

Edge Blockchain

Analysis By: Martin Reynolds

Definition: Edge blockchain is a funding mechanism that enables edge computing devices to carry a financial balance in a blockchain wallet and perform independent transactions. The balance can be increased through transaction receipts, or funding by a managing entity. The mechanism allows for small transactions at very low friction. An edge blockchain system has one or few controlling entities.

Position and Adoption Speed Justification: In an edge blockchain implementation, edge devices carry their own funding, which can pay for services, or collect value. This funding enables new business models, where devices can be funded to provide a service. Initially, the services will be simple. For example, heating or air conditioning services, where remote billing is not practical. The device can replace an electricity meter, potentially contracting with multiple suppliers. The system also allows devices to be mobile, as transactions are associated with the device, not the location. The key advantage is the very low cost of edge blockchain, as it eliminates billing systems and reconciliations. Therefore, very small transactions on a large installed base become possible. Edge blockchain may also enable devices to purchase connectivity and other services.

Today, the technologies that edge blockchain depends on are not ready for robust deployment. Smart contracts and distributed oracles are critical components of the solution, enabling edge devices to transact shared tokens. As the next wave of business-oriented blockchain developments emerges supporting these technologies, edge blockchain will find traction.

User Advice: Although immature today, edge blockchain should be part of any distributed, intelligent device plan where autonomous transactions add value. Use cases encompass any system where a monthly rent or metering system is important, but the cost of maintaining and scaling such a system may be prohibitive in terms of the value. Use cases also cover shared assets and resources including, for example, wireless connectivity and location tracking. Edge blockchain systems should be planned to support partners or successors.

When preparing for an edge blockchain project:

- Use edge blockchain solutions when remote billing solutions are impractical or not cost-effective
- Plan edge blockchain as lightweight implementations, balancing the cost of security against returns and risks
- Design the blockchain control mechanisms for longevity through potential ownership transfers

Business Impact: Edge blockchain allows small systems to reliably bill for services rendered, without the cost of a central billing system. As such, the technology will enable business cases previously inaccessible because of billing overhead.

A typical use case would cover shared assets or resources, where edge blockchain can help allocate costs to otherwise competing partners. The alternative — flat-rate payments — is unattractive since users have different value propositions (for example, in electricity generation and consumption). Smart contracts will also emerge in this environment, where limited participants mitigate risks.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Embryonic

Sample Vendors: Helium; LO3 Energy; NetObjex; Power Ledger

Recommended Reading: "Top 10 Strategic Technology Trends for 2020: Empowered Edge"

"Get Ready for Blockchain to Reshape Society"

"Blockchain in Utilities: Promise and Reality"

Data Transport and Edge Appliances

Analysis By: Raj Bala; Julia Palmer; Santhosh Rao

Definition: Data transport and edge appliances are physical devices capable of transporting bulk data to cloud infrastructure and platform service (CIPS) providers via package carriers rather than solely relying on a network to transfer data. Such appliances are often designed with ruggedized cases to be self-contained shipping units. The appliances optionally can be equipped with compute capacity in order to preprocess data before being transported to the cloud.

Position and Adoption Speed Justification: Data transfer and edge appliances are emerging as an efficient means of transporting large quantities of data when no network or less-than-ideal network conditions exist. Data transfer and edge appliances are playing an important role in enabling data transfer from the data centers or edge location to CIPS environments for processing and analytics. Gartner clients are evaluating ways to enable a continuous collection of data centralized in the cloud at significant scale for which network-based transfer is simply too limited.

User Advice: Enterprises are increasingly interested in using CIPS for an ever-expanding set of workloads, but find migrating such workloads and the data they require to be a challenge. As a general guide, moving 10TB of data in a 24-hour period requires a 10 Gbps network link. Such large network links may not be feasible depending on the amount of new data being generated per day.

The physical movement of data is merely part of the challenge. Planning the procedure and preparing the data take more effort and often more time than the shipment itself. Start planning these data shipments well in advance of the date on which you need to ship the appliance.

Business Impact: Getting data to the public cloud can be challenging due to network bottlenecks. There are distinct advantages to shipping data using transfer appliances when the data is unwieldy and the network bandwidth is constrained. Enterprise backups, for example, can be seeded at the public cloud target such that only incremental backups need to be sent to the cloud. Data can be collected in low- or no-network conditions to then be processed using public cloud services.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Amazon Web Services (AWS); Backblaze; Google; IBM; Microsoft; Oracle; Wasabi

Recommended Reading: "Market Guide for Cloud IaaS Data Transport and Edge Appliances"

AR Cloud

Analysis By: Tuong Nguyen

Definition: AR Cloud enables the unification of physical and digital worlds by delivering persistent, collaborative and contextual digital content overlaid on people, objects and locations to provide people with information and services directly tied to every aspect of their physical surroundings.

Position and Adoption Speed Justification: Similar to AR, aspects of the AR Cloud have existed for decades, but the emergence of the AR Cloud as a concept has reemerged as recently as 2017. Many startups (see the Sample Vendors below) are developing revolutionary platforms and systems to bring the AR Cloud to fruition.

A common misconception of AR is that it is a database that will "feed" AR and mixed reality (MR) experiences. The AR Cloud is much more rich and complex than a simple database. Numerous, underlying elements will need to be created. For example, edge networking, high bandwidth and low-latency communications, standardized tools and content types for publishing into the AR Cloud, management and delivery of content, and interoperability to ensure seamless and ubiquitous experiences. These elements will enable a shift in how we organize and interact with digital content. While this is the long-term goal, early implementations of the AR Cloud are expected to be siloed (rather than ubiquitous) before eventually converging into an interoperable, seamless system. Moreover, we expect vendors to fiercely compete to establish dominating, potentially proprietary standards.

Traditionally, leading tech vendors and communications services providers have invested in distributed network infrastructure, but many of them (such as Amazon, Google, Facebook, Microsoft) are adapting to a new paradigm to support localized, persistent, collaborative, shared, multiuser interactions. Some of this infrastructure and requirements will be ushered in by the arrival of low-latency, wireless networking (mmWave 5G will serve as an enabling tech), while others are still being developed.

User Advice: The AR Cloud is in its infancy. The best way to prepare is to evaluate potential areas of impact of business outcomes — areas that can be exploited or affected by the AR Cloud. There are broad possibilities here. For example, in city management, collaborative, dynamic and contextualized maps of cities to highlight details such as public restrooms, public transit locations, traffic issues, wayfinding as well as public utility maintenance records, log fix-it requests, and government office locations.

- Create a roadmap to address potential obstacles caused by the AR Cloud.
- Define privacy parameters. For example, what should be captured by sensors (image, mapping, location, etc.) and how it will be stored.
- Create guidelines to manage content (such as segregating data into public and private realms).
- Establishing hierarchies for data capture and protection.
- Evaluate the impact of massive physical data collection vis-a-vis compliance for regulations such as GDPR.

Business Impact: In the next decade, the AR Cloud will form the multilayer digital doppelganger of people, objects and locations in the physical world. This will enable new interactions and in turn new business models and ways to monetize the physical world. The AR Cloud will change the way that enterprises think of physical and digital assets, how they seamlessly interact with each other their customers and the associated risks.

Benefit Rating: Transformational

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Amazon; Apple; Facebook; Google; Microsoft; Niantic; Ubiquity6; Visualix; Xperiel

Recommended Reading: “Venture Capital Growth Insights: Immersive Technologies”

“Top 10 Strategic Technology Trends for 2020: Multiexperience”

“Gartner’s 2020 Strategic Technology Trends for Product Leaders”

“Market Guide for Augmented Reality”

Distributed Cloud

Analysis By: David Smith; Daryl Plummer; Milind Govekar

Definition: “Distributed cloud” refers to the distribution of public cloud services to different physical locations, while operation, governance, updates and evolution of the services are the responsibility of the originating public cloud provider.

Position and Adoption Speed Justification: Distributed cloud computing is a style of cloud computing where the location of the cloud services is a critical component of the model.

Historically, location has not been relevant to cloud computing definitions. In fact, the variations on cloud (e.g., public, private, hybrid) exist because location can vary. While many people may claim that private cloud or hybrid cloud requires on-premises computing, this is a misconception. Private and hybrid cloud do not require that the private components are in any specific location. With the advent of distributed cloud, location formally enters the definition of a style of cloud services.

Distributed cloud supports tethered and untethered operation of like-for-like cloud services from the public cloud "distributed" out to specific and varied physical locations. This enables an important characteristic of distributed cloud operation — low-latency compute where the compute operations for the cloud services are closer to those who need the capabilities. This can deliver major improvements in performance as well as reduce the risk of global network-related outages.

User Advice: Begin identifying scenarios where a distributed cloud model will effectively obviate the need for a hybrid cloud model, and where hybrid cloud models, and connectivity and latency matter and will continue to be needed for years to come.

Business Impact: A major notion of the distributed cloud concept is that the provider is responsible for all aspects of the delivery. This restores cloud value propositions that are broken when customers are responsible for a part of the delivery as is true in some hybrid cloud scenarios. It should be noted that while the cloud provider does not need to own the hardware on which the distributed cloud substation is installed, it must take responsibility for how the system is managed and maintained. Otherwise, the value proposition of distributed cloud is compromised.

In hyperscale public cloud implementations, the public cloud is the center of the universe. There has been distribution of cloud services through worldwide regions in public cloud practically since its inception. The major hyperscale cloud providers have different geographic regions around the world, all are centrally controlled and managed and provided by the public cloud provider.

Now, with distributed cloud, we are extending that distributed concept out to the edge and into next-generation hybrid environments such as Microsoft's Azure Stack Hub, Oracle's Cloud at Customer, AWS Outposts, Google Anthos and IBM's forthcoming Satellite offering.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Amazon; Google; IBM; Microsoft; Oracle

Recommended Reading: "Top 10 Strategic Technology Trends for 2020: Distributed Cloud"

"'Distributed Cloud' Fixes What 'Hybrid Cloud' Breaks"

"The Cloud Strategy Cookbook, 2019"

Private 5G

Analysis By: Sylvain Fabre; Joe Skorupa

Definition: Private 5G is defined as a private mobile network (PMN) based on 3GPP 5G to interconnect people/things in an enterprise. CSPs and TSPs have the potential to offer 5G PMN to various verticals, such as Industry 4.0, mining, oil, utility and rail road companies; IoT service providers, university campuses, stadiums/arenas etc. Private 5G offerings provide a separate network from the public network. It can include voice, video, messaging, and broadband data and IoT/M2M use cases with higher performance requirements.

Position and Adoption Speed Justification: Most enterprises use cases that justify a need to deploy a Private Mobile Network (PMN) that requires cellular connectivity for mobility, will be adequately served with a 4G network. Where 5G is justified, a PMN can provide the required functionality earlier than what the local CSPs may have deployed on their public infrastructure. There is another class of use cases however, not focused on mobility but requiring a high performance backbone where wiring is complex and costly — as in a factory deployment. In that case, support of autonomous delivery vehicles on the shop floor, could apply, but as a complementary application. Some verticals may adopt 5G sooner as a response to the current COVID-19 pandemic, driven by cost optimization, resiliency and automation concerns.

Volkswagen is reported to plan 5G private deployments in 122 German factories in 2020.

An early implementations example is: BMW Brilliance Automotive (BBA, BMW's JV in China) claims complete 5G coverage in all of its factories.

User Advice: Enterprises looking to deploy 5G PMN should:

- Seek out quotations not only from CSPs but also other possible providers such as large equipment vendors and smaller specialist
- Consider SIs and consultancies for design, deployment and managed services.
- Consider licensed and also unlicensed/shared spectrum options where available (for example 5G was approved for use in CBRS in February 2020 such as CBRS; German regulator Bundesnetzagentur [BNetzA] has allocated spectrum in the 3.7 GHz to 3.8GHz band for industrial local 5G).
- CSPs offering Private 5G to industrial buyers need to work with IT service providers that have the required industry skills (and knowledge of other technologies) to provide a value proposition of how 5G supports a specific use case or business KPI to justify the additional investments required to make existing infrastructure 5G ready. For example, in manufacturing 5G is seen as a platform that combines connectivity with security and AI capabilities.

Business Impact: While 5G standard are defined by 3GPP, other bodies are contributing in order to improve applicability to connected industrial applications, for example 5G-ACIA (Alliance for Connected Industries and Automation). 5G PMN can offer enterprises improved security and independence and enable efficiency gains in several manufacturing and industrial processes.

An early implementations example is: BMW Brilliance Automotive (BBA, BMW's JV in China) claims complete 5G coverage in all of its factories.

Benefit Rating: High

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: Athonet; Ericsson; Huawei; Nokia

Recommended Reading: "4 Hype Cycle Innovations That Should Be on the Private Mobile Networks Roadmap for 5G Security, CSP Edge and Slicing"

"Market Impact: How CSPs Can Rebuild Resilience and Business Continuity During Disruption"

"Cool Vendors in Communications Service Provider Network Operations"

Edge AI Software

Analysis By: Chirag Dekate; Carlton Sapp

Definition: Edge AI software refers to the use of systems for orchestrating, integrating, deploying and monitoring ML models across edge and IoT environments. While predominantly focused on inference, advanced systems may enable training at the edge (or IoT) based on local context data.

Position and Adoption Speed Justification: IT leaders are increasingly challenged with curating a distributed architecture spanning data center (on-premises/cloud), edge and IoT, especially in the context of AI. Unlike traditional enterprise stacks, managing deployment of AI requires meticulous model versioning, A/B (champion/challenger) testing and finally optimized delivery for compute accelerators in data center, edge or IoT environments. Edge AI software technologies and toolkits, such as edge containers (e.g., Docker swarm) and high-performance programming frameworks (e.g., TensorFlow Lite, Caffe2), enable enterprises to accelerate the orchestration, integration and deployment of AI in IoT, edge and cyber-physical systems.

Edge AI software needs to support three common edge AI architectural patterns found in CSPs and IoT environments:

- **Swarm patterns:** A deployment configuration that focuses on decentralized edge environments or collective actions of IoT endpoints. For example, logistics fleet optimization and management to optimize routes and extend life of service vehicles.
- **Control patterns:** Supervisory deployment configuration that focuses on executing highly regulated actions of IoT endpoints. For example, smart city deployments and intelligent industry 4.0 applications.
- **Graph patterns:** Highly interconnected and relational IoT endpoints in distributed edge environments. Graph patterns are commonly deployed in MEC applications to solve bandwidth optimization problems.

In each of the above contexts, trained AI models need to be:

- Optimized for delivery (tuned for specific architecture, optimized for hardware or operating environment constraints, etc.)
- Managed (with rollback, and champion/challenger testing)
- Monitored (for changes in state, need for redeployment or retraining, etc.)

Many of these applications are in trial phases, and broad adoption is a few years away. Some use cases with comparatively mature system stacks (for example, surveillance cameras or video recognition systems) are closer to productization.

User Advice: Infrastructure and operations leaders responsible for architecting AI infrastructures should:

- Deploy analytics closer to where the data is created and managed, by curating AI edge/IoT aligned with relevant streaming data ecosystems.
- Architect middleware that enables orchestration, integration and deployment of AI in IoT, edge and cyber-physical systems.
- Augment IoT analytics architectures with edge AI architecture patterns to take advantage of data and model parallelism across IoT endpoints.
- Select edge AI software technologies that support the broadest set of relevant edge AI architectural patterns (swarm, control, graph).

Business Impact: By incorporating edge AI software technologies, enterprises can:

- Improve orchestration, integration and deployment of AI in edge, IoT and cyber-physical systems, such as enhanced visual inspection systems in manufacturing assembly lines.
- Accelerate decision making, with the use of streaming AI analytics using event-based architecture platforms and a constantly evolving streaming data context.
- Manage versioning, rollout/rollback and monitoring of AI models deployed in edge and IoT environments.
- Reduce communication cost, with less data traffic between the edge and the cloud.
- Increase availability even when the edge is disconnected from the network.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Amazon; C3.ai; FogHorn; Google; IBM; Intel; Microsoft; NVIDIA

Recommended Reading: “Deploying IoT Analytics, From Edge to Enterprise”

"Deploy Leaner AI at the Edge: Comparing Three Architecture Patterns to Enable Edge AI"

"Predicts 2020: Artificial Intelligence Core Technologies"

"Top 10 Strategic Technology Trends for 2020: Empowered Edge"

"Emerging Technologies and Trends Impact Radar: Artificial Intelligence"

Edge Security

Analysis By: Neil MacDonald

Definition: Edge security offerings protect the integrity of an edge workload, its data and its access at distributed edge computing locations.

Position and Adoption Speed Justification: Like any IT system, edge computing locations will be targets for attack, including data theft, data poisoning, denial of service and placement of malware (for example, bitcoin mining). Edge computing security combines the requirements of data center computing security with the scale, location and variety of heterogeneous mobile and IoT computing security. Because of this, we expect more traditional security vendors and offerings will target the security of edge computing platforms along with mobile and IoT security offerings. However, there are some key differences from on-premises and cloud-based data centers. First, edge computing locations must be assumed to be uncontrolled and subject to physical tampering and theft. Second, constant network connectivity must not be assumed — intermittent network connectivity will require security controls to continue to provide protection even if disconnected from their management console, requiring autonomous protection capabilities. Third, in some cases compute capacity will be limited, so a strategy of low-overhead, minimum viable protection must be an option. These differences will require new approaches for securing edge computing. However, movement and adoption have been slow for multiple reasons.

One issue is the breadth of capabilities required to secure edge computing. When evaluating offerings, data-at-rest encryption must be considered mandatory, with hardware-based protection of keys. Boot time integrity checks will be mandatory, with strong controls on software updates. Each edge computing device must have an associated identity that is provisioned and managed (and able to be decommissioned in the event of failure or compromise). Zero trust network access (ZTNA), also referred to as software-defined perimeters, will be the most commonly used secure communications pattern, often provided by connecting edge computing systems to secure access service edge (SASE) fabrics. Another issue is the variety of edge computing platforms and approaches. This complexity has slowed the emergence of security pure plays targeting this opportunity. As such, we expect edge computing platform providers such as Amazon Web Services (AWS), Dell Technologies, Microsoft, and VMware to build in and/or partner for edge security protection capabilities. For example, in 2019, StorMagic acquired KeyNexus for a software-based key management system.

User Advice: When evaluating edge security offerings, enterprise security and risk professionals should prioritize the following:

- Protection capabilities should be software-based to provide the most flexibility in placement.

- Vendors should offer a platform-based set of security capabilities — including data encryption, network security, workload integrity, application control, memory protection, behavioral monitoring and intrusion detection/prevention — and let customers choose which controls make sense for different edge scenarios.
- The offering should be centrally managed, ideally cloud-based, and provide for tightly controlled and role-based administrative access.
- Assume the network is hostile and intermittent. The offering must be capable of providing protection even if network connectivity is intermittent and compromised.
- Assume the hardware will be attacked, tampered with or stolen, and make tamper resistance a part of the security control evaluation.
- The offering should use identity-based policy management using provisioned unique identities of edge equipment, typically certificate-based.
- Network security is a key requirement, restricting access to the edge platform using a zero trust network access offering.
- At a minimum, the edge protection strategy should verify integrity at boot, and verify/control which executables are allowed to run using application control.
- It should be capable of supporting Linux containers and container-based security offerings, which are expected to be widely used for distributed edge computing architectures.

Business Impact: For edge computing strategies to succeed, the integrity of edge computing workloads and data must be protected. Without protection, edge computing capabilities could be rendered inaccessible and the data could be stolen, copied or tampered with, leading to potentially catastrophic results that could threaten health and safety if monitoring and control signals are lost, tampered with or spoofed.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Amazon Web Services (AWS); AppGate; Dell Technologies; Hewlett Packard Enterprise (HPE); Microsoft; Red Hat; StorMagic; Thycotic; Venafi; VMware

Recommended Reading: "How to Overcome Four Major Challenges in Edge Computing"

"Exploring the Edge: 12 Frontiers of Edge Computing"

"How to Mitigate Firmware Security Risks in Data Centers, and Public and Private Clouds"

Mobile Edge Zone for Public Cloud

Analysis By: Sylvain Fabre; David Wright

Definition: Mobile edge zone for public cloud covers an application platform that is both carrier-grade and public-cloud-managed, to be deployed at the cloud provider's edge or within a telecom communications service provider's network. It is suitable to host a variety of telecom workloads stretching from the telecom back office out to the mobile edge.

Position and Adoption Speed Justification: There are multiple capabilities needed to be a mobile supplier or a communications service provider (CSP), and the cloud providers moving into this space only cover some of these. They have no radio access products, or large professional services arms for installing and operating live networks, but can partner with CSPs or telecommunications service providers (TSPs) to acquire that capability. They are however making incremental moves toward some core network aspects which are well within their current capabilities and augment them. These aspects are the mobile edge as well as packet core domain, or insights on the radio access network (RAN) domain — all these options do not require actually providing RAN products. This is a systematic approach by public cloud providers to position as providers of mobile infrastructure. Lack of radio access and spectrum may not be an issue in the longer term however, with multiple initiatives working toward reducing barriers to entry for a new breed of nontraditional telecom operators. For example, network sharing such as Multi-Operator Core Network (MOCN) or Multi Operator RAN (MORAN), and unlicensed spectrum approaches such as Citizens Broadband Radio Service (CBRS).

A public cloud provider partner however brings something the telco CSPs have never been able to develop: a truly horizontal application hosting and management environment. The cloud providers bring a robust IaaS and PaaS service platform that will make it easy for telecom VNF providers and other telecom ISVs to build and deploy 5G-related solutions.

Examples so far include various incremental moves that public cloud providers are making toward building mobile network capabilities: 5G Multiaccess Edge Computing (MEC) with AWS Wavelength, Microsoft Azure Edge Zones for telco (optimized x86 plus virtual Evolved Packet Core [EPC] plus certified partner virtual network functions [VNFs]), and Google Global Mobile Edge Cloud (GMEC). Amazon Web Services (AWS) will self-OEM the hardware, Microsoft will partner for integrated certified hardware, and Google will run on a variety of third-party container-compliant hardware. Microsoft and Google are explicitly including the use of their own edge network access points and capabilities in the offering, AWS has not yet specified this aspect. In terms of workloads, GMEC is currently demonstrating mostly back-office Business Support Systems (BSS) functionality (like Amdocs and Netcracker), whereas Microsoft has purchased virtual EPC (mobile packet core) provider Affirmed Networks, and agreed to acquire Metaswitch Networks (cloud native telecom networks). AWS has been silent about VNF partnerships. In all three cases, the goal is to "tether" the mobile domains to the public cloud and manage them from there, at least partially.

The desire of telecom CSPs to sell services that monetize their 5G network and license investments will continue to propel them toward the cloud providers, and likewise the telecom VNF/app independent software vendors (ISVs) will see a scalable route to market. This combination will accelerate the adoption of this approach in market.

Finally this trend could accelerate with political initiatives that seek to promote a supply chain for 5G network infrastructure that is refocused on North American providers. Leveraging North American public cloud providers for 5G networks could be seen as an alternative to Asian vendors at a time where there are no traditional end-to-end vendors of 5G public networks left in North America (since Nortel Networks, Motorola and Lucent have left the space).

User Advice: The benefit for CSPs is a wider choice of vendors and potentially a faster route to OS innovation. But, CSPs should be mindful that using public cloud providers rather than traditional infrastructure may not reduce vendor lock-in but possibly a shift to a different set of suppliers instead.

There is also a risk for CSPs when public cloud providers compete in providing private mobile networks to enterprises.

CSPs considering partnership with cloud providers for their mobile edge should:

- Increase differentiation by integrating the edge computing offering from the public cloud provider(s) into your private mobile network solution for industries.
- Retain long-term value by designing edge architecture with edge public cloud provider(s) so that the CSP can secure most value from enterprise networks and related IoT data.
- Preempt competing offers by cloud providers for mobile private networks by offering complementary solutions such as slice as a service to cloud providers and mobile virtual network operators (MVNOs) (using shared as well as licensed spectrum options).

All things considered, there seems to be a bigger benefit for public cloud providers, who are positioning themselves to capture value away from both CSPs and their vendors, by gaining increased access to IoT data.

Business Impact: As the public cloud providers consolidate more network capabilities, they will over time become more of a competitor to:

- Traditional network infrastructure vendors (especially as VNFs [virtualization] and open source gain further adoption in how mobile networks are implemented).
- Existing CSPs (as unlicensed/shared spectrum options gain wider adoption for enterprise and industry use).

In a related initiative (although not a public-cloud-managed solution), other nontraditional vendors of infrastructure also venture into the mobile networking space, for example VMware acquiring Uhana for RAN analytics, signaling an era of more vendor choices for CSPs.

Finally, some of the cloud providers appear to also be edging toward their own fully independent wide-area networks — Google already has most of it in fiber, Amazon is attempting a long-play through satellites. We also see signs of public cloud providers entering provision of mobile networking to enterprises, such as private LTE over CBRS supplied as connectivity as a service by AWS and Microsoft (Google is also active with the Spectrum Access System [SAS] sensing database for CBRS spectrum).

One open question is where the profit will go — with app, cloud and telecom vendors in every solution, who controls the customer, who negotiates the sale and who will capture the most margin.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Alibaba Cloud; AWS; Google; IBM Cloud; Microsoft Azure; VMware; XGVela

Recommended Reading: "Emerging Technologies: Venture Capital Growth Insights for mmWave 5G"

"Market Guide for 5G New Radio Infrastructure"

"Assessing 5G Mobile Technology for Organizations"

"Market Trends: Strategies Communications Service Providers Can Use to Address Key 5G Security Challenges"

"Reduce Privacy Risks When Using 5G Products and Services"

Cloud-Tethered Compute

Analysis By: Tony Harvey; David Wright

Definition: Cloud-tethered compute is a model where MaaS, IaaS, or PaaS are delivered in a customer-controlled environment but managed by the vendor via a network tether from a public or private cloud. The system may require the tether to be continuously connected, for billing, or be able to operate disconnected with periodic connectivity. Updates are managed by the vendor, removing the responsibility of maintenance of the platform from the I&O team.

Position and Adoption Speed Justification: Cloud-tethered compute solutions are starting to become more common. Initially, they were developed by public cloud vendors to provide public cloud services such as IaaS and PaaS in client-controlled environments, for example, Azure Stack Hub and AWS Outposts. Other vendors have started to develop products in this space. For example, companies like Hivecell deliver an edge-focused solution; Dell EMC delivers a VMware-based private cloud with VMware Cloud on Dell EMC; and other vendors like HPE and Lenovo deliver metal as a service and other capabilities through consumption-based models. Current adoption is relatively low, although the potential for growth due to data sovereignty, connectivity and latency issues at the edge is significant.

User Advice: Cloud-tethered compute systems are relatively new to the market and may be missing key capabilities; and some entrants do not have clear SLAs and the commercial terms are not yet fully developed. Be careful when evaluating tethered compute systems to ensure that they provide the features that the development teams require and that you clearly understand the SLAs provided, what happens at the end of the contract, and how upgrades and expansions midterm are handled.

Key areas to consider when evaluating cloud-tethered compute systems:

- Provided as a service: Some or all elements are provided as a service and remain the property of the vendor. You must have a clear understanding of who owns which element and what are the consequences of a service shutdown.
- Connectivity: Does the system require a permanent connection for the tether, or can it operate in disconnected mode? What are the limitations when operating in disconnected mode?
- Response times and hardware maintenance services: Users more used to a Tier 1 vendor service contract for on-premises maintenance may struggle to adapt to a service model that is based on next-business-day, whole-unit replacement or lower.
- Contract terms: Understand what will happen at the end of the term and how any midterm changes, upgrades or additions will affect the termination date and costs.
- Security: Cloud-tethered compute systems typically use a “shared security model,” where some responsibilities belong to the customer and some to the vendor. Although the security of the vendor may be excellent, the customer’s security team must be engaged to ensure all security risks are mutually addressed.
- Variable pricing: Similar to cloud services, monthly or other recurring charges for a cloud-tethered system could vary based on usage, making costs and budgeting less predictable than traditional infrastructure.
- Data sovereignty: Although the data is held in a location under control of the user, the system is not under user control. At a minimum, any data being used on these systems should be encrypted at rest. In addition, cloud-tethered compute vendors should be able deliver secure data deletion and even storage device retention services to ensure that sensitive data does not leave the control of the user.

Business Impact: Cloud-tethered compute represents a new form of computing in which the customer’s IT gives responsibility for and control over some of its data center assets to a cloud provider. It reduces the need to perform many of the routine tasks traditionally performed by the IT staff, especially as related to server maintenance. It will, however, create new needs for skills in contractual analysis, security and spend management for these systems. Application- and data-level security and backup will still remain an I&O responsibility. In many cases, the I&O function may welcome the removal of basic duties related to system maintenance, which will enable it to focus on delivering higher-level services.

Benefit Rating: Moderate

Market Penetration: Less than 1% of target audience

Maturity: Emerging

Sample Vendors: AWS Outposts; Google Anthos; Hivecell; Microsoft Azure Stack; Oracle; VMware

Recommended Reading: “‘Distributed Cloud’ Fixes What ‘Hybrid Cloud’ Breaks”

“Prepare for AWS Outposts to Disrupt Your Hybrid Cloud Strategy”

“How to Bring the Public Cloud On-Premises With AWS Outposts, Azure Stack and Google Anthos”

“Best Practices for Tech CEOs to Manage Edge-to-Cloud Products”

“Top Emerging Trends in Cloud-Native Infrastructure”

CSP Edge IaaS

Analysis By: Sylvain Fabre; Ted Chamberlin

Definition: The new category of CSP edge IaaS facilitates CSPs’ edge deployment beyond their own network perimeter; this allows CSPs’ connection to an edge provider, on their footprint but with the edge providers handling the edge-related infrastructure and its operations. A diverse feature set is provided by multiple providers, including stateful applications, dynamic caching, CDN, application acceleration, or games and industrial IoT applications optimized for latency. Individual CSP deployments can be federated, creating an edge cloud framework.

Position and Adoption Speed Justification: This model also allows companies to connect to, and consume edge cloud as a service for enterprise networks — including private mobile 4G and 5G networks. Initial adoption indicates that early CSP edge infrastructure as a service (IaaS) providers are operational but defining clear business models and showing their viability is the next step.

Early initiatives include:

- Ericsson Edge Gravity, started in 2016, focused on wireline operators; provides a cloud native solution on both the application and the infrastructure side. With over 85 communications service providers (CSPs) signed up, they initially focused on a few use cases to validate their business approach. This provided a preintegrated full stack solution, such as CDN (deal with Limelight), caching, SD-WAN and gaming, with an intent to support more moving forward. The goal for Edge Gravity is to mature current CDN edge nodes to intelligent points of presence (PoPs). As of publication however, Ericsson Edge Gravity appears to have ceased activities.
- Deutsche Telekom’s MobiledgeX provides an abstraction layer across multiple CSPs, exposing as runtime to application developers to support gaming and industrial applications. These include Holo-Light (German, AR/VR for industrial applications); 1000 realities (Polish, AR/VR applications); Edgemesh (U.S., enhance visitors’ browser); Niantic (U.S. software, started with Google); VRee (Netherlands, AR/VR suits).
- Macrometa provides a distributed database platform that can leverage CDN and data center providers to provide distributed edge capability, underpinned by a distributed database and ledger functionality. This enables workloads to run locally at the edge with low latency.
- AlefEdge provides a software stack that works with existing networks as a plug-and-play overlay, using proprietary APIs. Their cloud-native software at the edge enables multiple edge sites to form the “Edge Internet.”
- Gaming/e-sports use case: E-sports and gaming experience suffer from lag time and excessive latency which impede user experience. Providers like Edgegap and Swarmio Media provide AI-based algorithms that optimize edge hosting locations with proximity to gamers.

While nascent, these approaches seem to provide a faster avenue for enterprises, and service providers, to connect and deploy edge cloud topologies at scale, beyond what traditional local solutions can deliver. However, profitability and revenue need to be demonstrated for many of these offerings, which will require wider adoption. The fact that Ericsson's Edge Gravity appears to have run into difficulties is confirmation that this approach is still unproven.

User Advice: Enterprises, service providers and CSPs should consider cloud edge providers to support use cases that require one or several of the following:

- Fast deployment and cycle times.
- Deployment of development environments and toolchains for deployments closer to end users.
- Wider area of operation (e.g., an edge supporting vehicle to infrastructure over a road network).
- Access to preoptimized edge and application offering (such as low-latency gaming).

The trend of public cloud providers stepping in to offer edge and other telecommunications-related infrastructures, is likely to continue to put significant pressure on the vendors in this space.

Business Impact: Consider the following areas of impact:

- The CSP edge IaaS provider can be beneficial for enterprises wanting to quickly connect and leverage industrially relevant applications, typically latency dependent and of an immersive nature.
- CSPs can quickly access an edge environment and focus on a specific set of use cases while leaving operational concerns to a third-party provider.
- In a market where multiple edge approaches will coexist, (such as ETSI Multi-access Edge Computing [MEC], OpenFog, ONAP etc.) this new breed of providers can enable edge cloud operation at a wider scale, sooner than that if extensive interoperability testing and integration was to occur. Skipping testing could accelerate adoption, but may not be optimal.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: AlefEdge; CloudBackend; Macrometa; MobiledgeX

Recommended Reading: "4 Hype Cycle Innovations That Should Be on the Private Mobile Networks Roadmap for 5G Security, CSP Edge and Slicing"

"Market Insight: Telecom CSP Sector Industry Product Planning, Strategy and Execution"

"Market Trends: Strategies Communications Service Providers Can Use to Address Key 5G Security Challenges"

Edge-IN to Cloud

Analysis By: Bob Gill

Definition: Edge-IN to cloud describes an architecture where edge applications, servers, and gateways are designed and built independently from any hyperscale cloud they may connect to for application services, such as analytics. Rather than using the programming models, or IAM of a hyperscale cloud platform to define the edge platform or application, the application is centered around its role at the edge first, and its need for hyperscale cloud application services second. A known example would be the Linux Foundation's LF Edge reference model.

Position and Adoption Speed Justification: The rise of edge-IN to cloud architectures results from the need to pull cloud services to the edge and use them selectively in edge-specific ways, rather than push public cloud architecture to the edge as a complete platform solution. Today, many cloud-connected edge applications being proposed use the edge services provided by one or more of the hyperscale cloud providers, including but not limited to identity and access management, programming model, and an architecture defined in often proprietary fashion by the cloud provider. This may not provide the greatest flexibility when dealing with the massive installed base or brownfield considerations of mature sectors such as retail or industrial. Many ISVs and system integrators already active in the edge space create their applications either as an extension of their existing proprietary stack, or based on an open framework such as that of the Linux Foundation's LF Edge. Frameworks and technology commonly used for such edge-IN applications might include containers, Kubernetes, and control plane offered from a centralized core or SaaS implementation. We expect the popularity of implementations based on open source frameworks and tools to rise rapidly given the maturing feature sets and desire for cloud independence.

User Advice: When planning, proposing or trialing edge applications, I&O leaders should:

- Balance the focus of whether application is an edge extension of a cloud-based application vs. an edge application that may use some cloud resources on the back-end.
- Determine whether the edge and IoT portfolio will be used in conjunction with a single hyperscale cloud provider, or be multicloud.
- Determine the viability of replacing edge endpoints or accessing them through a gateway.
- Determine whether the ISV or integrator you plan to use for edge and IoT applications favors a "cloud out" vs. "edge-IN" design pattern.
- Be aware that an edge-IN model assumes a degree of assembly and integration that may be beyond a standard enterprise's in-house skill set.

Business Impact: While an edge-IN approach may require greater expertise and integration on the part of the implementor, by considering an edge-IN design pattern, enterprises can:

- Minimize any potential lock-in from using a hyperscale-sourced and optimized stack
- Maximize the potential flexibility for multicloud implementations
- Focus on what the edge delivery team wants to use for building the application

- Build a vendor neutral, vendor independent architecture for a broad portfolio of edge applications

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Emerging

Sample Vendors: Linux Foundation (LF Edge); Rancher Labs; ZEDEDA

Edge as a Service

Analysis By: Bob Gill

Definition: Edge as a service describes a model in which the edge platform, or even the edge applications themselves, are offered via provider-owned and operated assets, requiring little to no infrastructure on the part of the customer. Edge as a service can take the form of 1) "edge platform as a service" where a CSP offers development and runtime services such as within a CDNs interconnected POPs, or 2) completely managed application solutions, e.g., a hospitality package running on provider-owned hardware on the customers premise.

Position and Adoption Speed Justification: One of the greatest impediments to deploying edge computing in 2020 is the lack of broadly accepted infrastructure and operation models, leaving the enterprise to piece together solutions from still emerging technology stacks and operational models. For such critical and far reaching applications, the potential for selecting and implementing an eventual "dead end" are high, while few universal deployment models to simplify implementations exist. Edge as a service focuses on a delivery model for edge computing where a communications services provider, cloud provider, or even ISV or system integrator provides all or nearly all of the infrastructure required to deliver edge-based applications, shielding the customer from technical and market volatility. In essence, the customer is free to focus on business enabling applications rather than infrastructure building.

User Advice: While I&O must view edge infrastructure as an extension of the corporate data center or cloud, "edge as a service" allows implementers to look at edge deployments as they might a PaaS platform or hosted application model.

From the PaaS perspective, CDN and interconnection providers have begun offering "edge PaaS," whereby a customer can utilize bare metal, serverless infrastructure, or Kubernetes and containers on POP-located hardware, that is already extensively networked and interconnected on the providers' backbone networks. These platforms are particularly suitable for communications centric applications, such as optimized interactive content delivery (e.g., gaming), network path optimization (e.g., telemetry based routing), and security and management overlays.

In something akin to the traditional hosting models, "edge SaaS" packages the edge infrastructure, management, orchestration, and applications themselves into turnkey, function-specific, outcome-based application suites; examples include point of sale, restaurant or store-specific applications, and even IoT functionality, for a restaurant or QSM (Quick Service Mart) implementation. In such cases, the IT expertise of the store or franchise may be low, the number of locations high, and the application suite tightly vertically focused. Elasticity and the need for “boots on the ground” may be supplied by overnight shipment of preconfigured server/software bundles that can be simply unpacked and plugged in, allowing the systems to “phone home” and obtain configuration and update software. “Edge SaaS” provides the customer with the traditional benefits offered by a vertical-specific ISV or system integrator, for those in need of a focused and complete solution, while minimizing the technical expertise required.

Business Impact: As enterprises begin to implement edge computing, it stands to reason that long standing “build vs. buy” models will be recycled, much as they have for more traditional applications over the last four or five decades. Enterprise needs for customization and differentiation must be balanced against in-house technical expertise, their stance toward “opex vs. capex,” and the breadth of the solution (targeted specific application set versus a more general distributed infrastructure platform). Edge as a service may speed the time to market for many use cases, by lowering the technical hurdles, operational expertise required, and “platform risk” present in such a nascent market.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Akamai; AlefEdge; CDNetworks; Cloudflare; Reliant Solutions; Rigado; VMware

Edge Asset Life Cycle Management

Analysis By: Santhosh Rao

Definition: Edge asset life cycle management includes onboarding, monitoring, change management and decommissioning of edge computing systems. Edge computing systems can be traditional servers and workstations deployed in remote offices or sites, or IoT systems such as gateways, embedded devices or IoT servers.

Position and Adoption Speed Justification: Edge asset life cycle management can differ based on the type of edge environment. Traditional edge environments such as remote offices are often managed by today’s data center system management software, this capability is relatively mature. However, system management for IoT Edge is still nascent. In this case, IoT asset management software is usually provided by the hardware vendors as a SaaS offering. More recently, cloud vendors and emerging asset management vendors have started to offer IoT platforms that support remote management of IoT gateways and embedded devices.

The IoT platform can periodically update the assets with the latest firmware/OS, device drivers, security patches or configuration settings. The success of the update is often dependent on the network bandwidth and availability. IoT assets are connected to the platform via Ethernet or cellular networks. While it is relatively easy to update remote assets through wired networks, OTA updates via cellular networks are often challenging and may result in inconsistent results. The problem is compounded in case of low power constrained devices operating on LPWAN networks.

User Advice: I&O leaders must view edge infrastructure as an extension of the corporate data center or the cloud, and must apply similar policies to the systems at the edge. It is unlikely management of both traditional and IoT-edge environments can be carried out through a common platform in the near future. Therefore, it is recommended that both these environments are managed separately. For managing IoT assets, choose platforms that support easier packaging and delivery of software updates (such as containers) to the edge system. Aim to consolidate various IoT functionality — analytics, visualization, application and device life cycle management onto a common IoT platform. Ensure that updates are delivered in a secure manner, while ensuring that the base OS/firmware of the asset has a roll-back option. Focusing on the usability aspect of these platforms is equally important, particularly when the enterprises begin to scale from a small number of assets to thousands.

Business Impact: As enterprises progress toward distributed IT architectures, I&O leaders must ensure that they are able to centrally manage the health and life cycle of these systems, irrespective of location. The role of system management software is therefore crucial. Similarly, when organizations begin to scale their IoT projects and deploy a large number of endpoints, gateways and edge servers, centralized management of these systems is imperative. This ensures better visibility and control of the environment.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Emerging

Sample Vendors: Amazon Web Services; ioTium; Litmus Automation; Microsoft; Pixeom; ZEDEDA

Recommended Reading: “Market Guide for Edge Computing Solutions for Industrial IoT”

“Market Guide for Industrial IoT Gateways”

At the Peak

Edge Computing

Analysis By: Bob Gill; Philip Dawson

Definition: Edge computing describes a distributed computing topology in which data storage and processing are placed close to the things or people that produce and/or consume that information. Drawing from the concepts of mesh networking and distributed computing, edge computing strives to keep traffic and processing local and off the center of the network. Edge balances latency requirement and the bandwidth required for an application, allows for autonomous operation, enables the placement of workloads and data that satisfies regulatory/security demands.

Position and Adoption Speed Justification: Most of the technology for creating the physical infrastructure of edge data centers is readily available. However, widespread application of the topology and explicit application and networking architectures are not yet common outside of vertical applications, such as retail and manufacturing. As IoT demand and use cases proliferate, the acceptance of edge computing as the topological design pattern (namely, the "where" a "thing" is placed in an overall architecture) has dramatically increased interest in edge technologies and architectures. However, the still-nascent state of non-IoT edge applications has prevented more rapid movement along the Hype Cycle since 2018.

User Advice: We recommend the following:

- We urge enterprises to begin considering edge design patterns in their medium- to longer-term infrastructure architectures.
- Immediate actions include simple trials using colocation and edge-specific networking capabilities, or simply placing remote location or branch office compute functions in a standardized enclosure (for example, "data center in a box").
- Some applications, such as client-facing web properties and branch office solutions, will be simpler to integrate and deploy, while data thinning and cloud interconnection will take more planning and experimentation to get right.
- We are beginning to see viable offerings from hyperscale cloud providers in extending their programming models and management systems to on-premises and edge-located devices, complementing their mostly centralized computing model with a distributed analog.
- For distributed applications requiring a consistent, global infrastructure, with less emphasis on IoT or unique physical endpoints, consider an edge infrastructure as a service provider, such as Cloudflare or NetActuate.
- Enterprises must also become familiar with an emerging "Edge-IN" application model, in which edge gateways and hubs serve as the linchpins for deploying heterogeneous, multicloud and multiendpoint applications. These are often based on open-source frameworks and technologies, such as containers and orchestration systems like Kubernetes.

Business Impact: Edge computing has quickly become the decentralized complement to the largely centralized implementation of public cloud. Edge computing solves many pressing issues, such as unacceptable latency and bandwidth limitations, given a massive increase in edge-located data. The edge computing topology enables the specifics of IoT, digital business and distributed IT solutions, as a foundational element of next-generation applications of all kinds.

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Akamai; Amazon; Cisco; Cloudflare; HPE; IBM; Microsoft; Vapor IO; Verizon; ZEDEDA

Recommended Reading: “The Edge Completes the Cloud: A Gartner Trend Insight Report”

“Top 10 Strategic Technology Trends for 2019: The Empowered Edge”

“The Future Shape of Edge Computing: Five Imperatives”

“How Edge Computing Redefines Infrastructure”

Micro Data Centers

Analysis By: David Cappuccio; Philip Dawson

Definition: A micro data center is modular or containerized and smaller than a computer room — usually no more than a rack of equipment or two — and typically one rack or less. All required facilities and IT functionalities (such as uninterruptible power supply, servers, storage, networking and cooling) are contained in the micro data center. It is designed to handle specific needs (for example, accumulating sensor data or small remote office support) at distributed locations, and is typically managed remotely from a large data center.

Position and Adoption Speed Justification: For more than 20 years, small computer rooms (as small as a closet) were subsumed into larger, consolidated data centers or their workloads migrated to external cloud providers — a trend we see continuing unabated. Yet, not all IT capabilities need to be (or should be) centralized. Placing workloads local for business units, or for factors such as latency or site-specific operations, are common reasons for deploying micro data centers. Additionally, with the distributed nature of the Internet of Things (IoT), there could be a need to accumulate and process data locally before a subset of that data is sent to a larger data center for additional analysis.

Micro data centers are typically based on more mature IT and facilities technology, although new features, including hyperconvergence, packaging, containerization and remote management, have been added. Trends, such as the distributed nature of the IoT and edge computing, mean that micro data centers will be much more widely deployed than they are now. They can be located virtually anywhere, including as stand-alone centers in appropriate containers (for environmental protection), typically found in branch offices. Micro data centers use any available means for communication but are typically connected to an office LAN/WAN.

User Advice: Our advice to infrastructure and operations leaders is to:

- Design micro data centers to run autonomously, but to be controlled centrally.
- Create a standard, self-contained solution designed for easy deployment, simple replacement, and remote monitoring and management.
- Focus on maximizing operational independence, while minimizing IT skills and risks at remote sites.

Business Impact: This functional capability widely exists (such as at retail sites and bank branches), but micro data centers differ through their standardization, integration, remote management and enhanced security. These solutions promise significantly lower costs and enhanced manageability. We view micro data centers as transformative to supporting digital business development, with a particular focus on the IoT, and expect them to be deployed in the thousands (and possibly tens of thousands) across most enterprise verticals.

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Huawei; Panduit; Rittal; Schneider Electric; Vertiv; Zellabox

Recommended Reading: "Competitive Landscape: Providers Strive for an Edge in the Micro Modular Data Center Market"

Secure Access Service Edge (SASE)

Analysis By: Joe Skorupa; Neil MacDonald

Definition: Secure access service edge (SASE, pronounced "sassy") delivers multiple capabilities such as SD-WAN, SWG, CASB, NGFW and zero trust network access (ZTNA).

SASE supports branch office and remote worker access. SASE is delivered as a service, and based upon the identity of the device/entity, combined with real-time context and security/compliance policies. Identities can be associated with people, devices, IoT or edge computing locations.

Position and Adoption Speed Justification: SASE is driven by enterprise digital business transformation: the adoption of cloud-based services by distributed and mobile workforces; edge computing and business continuity plans that must include flexible, anywhere, anytime, secure remote access. While the term originated in 2019, the architecture has been deployed by early adopters as early as 2017. By 2024, at least 40% of enterprises will have explicit strategies to adopt SASE, up from less than 1% at year-end 2018.

By 2023, 20% of enterprises will have adopted SWG, CASB, ZTNA and branch FWaaS capabilities from the same vendor, up from less than 5% in 2019. However, today most implementations involve two vendors (SD-WAN + Network Security), although single vendor solutions are appearing. Dual-vendor deployments that have deep cross-vendor integration are highly functional and largely eliminate the need to deploy anything more than a L4 stateful firewall in the branch office. This will drive a new wave of consolidation as vendors struggle to invest to compete in this highly disruptive, rapidly evolving landscape.

SASE is in the early stages of market development but is being actively marketed and developed by the vendor community. Although the term is relatively new, the architectural approach (cloud if you can, on-premises if you must) has been deployed for at least two years. The inversion of networking and network security patterns as users, devices and services leave the traditional enterprise perimeter will transform the competitive landscape for network and network security as a service over the next decade, although the winners and losers will be apparent by 2022. True SASE services are cloud-native — dynamically scalable, globally accessible, typically microservices-based and multitenant. The breadth of services required to fulfill the broad use cases means very few vendors will offer a complete solution in 2020, although many already deliver a broad set of capabilities. Multiple incumbent networking and network security vendors are developing new or enhancing existing cloud-delivery-based capabilities.

User Advice: There have been more than a dozen SASE announcements over the past 12 months by vendors seeking to stake out their position in this extremely competitive market. There will be a great deal of slideware and marketecture, especially from incumbents that are ill-prepared for the cloud-based delivery as a service model and the investments required for distributed PoPs. This is a case where software architecture and implementation matters

When evaluating SASE offering, be sure to:

- Involve your CISO and lead network architect when evaluating offerings and roadmaps from incumbent and emerging vendors as SASE cuts across traditional technology boundaries.
- Leverage a WAN refresh, firewall refresh, VPN refresh or SD-WAN deployment to drive the redesign of your network and network security architectures.
- Strive for not more than two vendors to deliver all core services.
- Use cost-cutting initiatives in 2020 from MPLS offload to fund branch office and workforce transformation via adoption of SASE.
- Understand what capabilities you require in terms of networking and security, including latency, throughput, geographic coverage and endpoint types.
- Combine branch office and secure remote access in a single implementation, even if the transition will occur over an extended period.
- Avoid vendors that propose to deliver the broad set of services by linking a large number of products via virtual machine service chaining.
- Prioritize use cases where SASE drives measurable business value. Mobile workforce, contractor access and edge computing applications that are latency sensitive are three likely opportunities.

Some buyers will implement a well-integrated dual vendor best-of-breed strategy while others will select a single vendor approach. Expect resistance from team members that are wedded to appliance-based deployments.

Business Impact: SASE will enable I&O and security teams to deliver the rich set of secure networking and security services in a consistent and integrated manner to support the needs of digital business transformation, edge computing and workforce mobility. This will enable new digital business use cases (such as digital ecosystem and mobile workforce enablement) with increased ease of use, while at the same time reducing costs and complexity via vendor consolidation and dedicated circuit offload.

COVID-19 has highlighted the need for business continuity plans that include flexible, anywhere, anytime, secure remote access, at scale, even from untrusted devices. SASE's cloud-delivered set of services, including zero trust network access, is driving rapid adoption of SASE.

Benefit Rating: Transformational

Market Penetration: 1% to 5% of target audience

Maturity: Emerging

Sample Vendors: Akamai; Cato Networks; Cisco; Citrix; iboss; Netskope; Open Systems; Palo Alto Networks; VMware; Zscaler

Recommended Reading: "The Future of Network Security Is in the Cloud"

"Magic Quadrant for Cloud Access Security Brokers"

"Market Guide for Zero Trust Network Access"

"Market Trends: How to Win as WAN Edge and Security Converge Into the Secure Access Service Edge"

"Quick Answer: Cost Effectively Scaling Secure Access While Preparing for a Remote Workforce"

5G

Analysis By: Sylvain Fabre

Definition: 5G is the next-generation cellular standard by the 3rd Generation Partnership Project (3GPP). The standard targets maximum downlink and uplink throughputs of 20 Gbps and 10 Gbps respectively, latency below 5 milliseconds and massive scalability. New system architecture includes core slicing as well as wireless edge.

Position and Adoption Speed Justification: Seventy-three operators have announced 5G rollouts (Source: Global mobile Suppliers Association [GSA], April 2020), just under 9% (up from 5% one year ago) of mobile networks.

3GPP Release 16 freeze date has been postponed due to the COVID-19 pandemic, with a freeze target date of mid-2020.

5G encompasses a range of 3GPP standards focused on different functionality:

- R15: Extreme broadband (5G NSA and then 5G SA)
- R16: Augmentations for Industrial IoT (massive IoT, slicing and security improvements)
- R17: Augmentations for wider ecosystem expansion (freeze target date end of 2021)
- R18: Additional augmentations (e.g., extra territorial 5G systems, railway smart station services)

Due to this phased introduction, and the time required from the vendors' ecosystem to build standard compliant networks and grow silicon and device availability, Gartner expects the full potential for 5G use cases to materialize first in 2022.

Use of higher frequencies and massive capacity, will require very dense deployments with higher frequency reuse. Here we see regional differences, whereby mmWave will be leveraged in the U.S. and South Korea, but may not see initial adoption elsewhere.

Gartner expects many 5G deployments to initially focus on islands of deployment, without continuous national coverage.

Less than 45% of CSPs globally will have launched a commercial 5G network by 2025. Uncertainty about the nature of the use cases and business models that may drive 5G is currently a source of uncertainty for many CSPs, enterprises, and technology and service providers (TSPs). Gartner estimates that 5G capable handset penetration will reach 50% in 2023 in Western Europe, and could be a little faster in North America.

We are seeing different dynamics by regions, where in many parts of Africa for example, 5G would not be the next step to lower bandwidth services, and handset cost may be an inhibitor for lower income subscribers. Adoption is more aggressive in APAC and NAR, with Europe cautiously enthusiastic — and the developing world lagging.

User Advice: TSP product managers should:

- Focus mobile infrastructure planning on LTE, LTE-A, LTE-A Pro, small cells and heterogeneous networks (HetNets), as part of a planned transition toward 5G.
- Ensure backward compatibility to preceding generation (LTE) devices and networks. This is necessary because 5G coverage may be limited, so new 5G devices need to be able to seamlessly transition to 4G fallback infrastructure for uninterrupted service.
- Focus on related architecture initiatives — such as software-defined network (SDN), network function virtualization (NFV), CSP edge computing, distributed cloud architectures and cloud native containerization, as well as end-to-end security in preparation for 5G.
- Provide solutions where new frequency allocations (preferably) should be used for the latest technology — 5G — to benefit from lower cost per byte, higher bandwidth and more capacity.
- Help CSPs refine generic services to vertical-focused solutions (B2B) for 5G.
- Have a clear understanding of specific verticals and their use cases for more effective consultative selling of their 5G solutions.
- Build their ecosystem of partners to target verticals more effectively with 5G.

Enterprise business leaders should:

- Identify use cases that definitely require the high-end performance of 5G; these may be few or even nonexistent for many verticals.

- Evaluate the multiple IoT alternatives available that may prove adequate, more available and more cost-effective than 5G for many use cases (e.g., low-power wide-area [LPWA] such as NarrowBand Internet of Things [NB-IoT], long-range [LoRa], Wireless Smart Ubiquitous Networks [Wi-SUN]).
- Clarify the level of complexity involved in operating a private 5G network.
- Evaluate options for CSPs or other providers to be involved in running the 5G network.

Business Impact: Gartner Enterprise 5G Surveys indicate that vertical use cases with 5G would be first motivated by operational cost savings. Another driver is agility — in particular, in oil and gas and manufacturing.

In addition, the vertical users for 5G appear to value lower latency from ultrareliable and low-latency communications (URLLC) and expect 5G to outperform rivals in this area.

With massive machine-type communications (mMTC), scenarios of very dense deployments can occur, supported by the 5G target of 1 million connected sensors per square kilometer.

5G enables, principally, three technology deployment and business scenarios, which each support distinct new services, and possibly new business models (such as latency as a service):

- Enhanced mobile broadband (eMBB) supports high-definition video.
- mMTC supports large sensor and IoT deployments.
- URLLC covers high availability and very low latency use cases, such as remote vehicle/drone operations.

URLLC and mMTC will be implemented after eMBB. Only eMBB addresses the traditional mobile handset requirement of ever higher throughput. URLLC addresses time critical industrial applications such as automation, with latency around 1ms over a limited range for a limited number of connections — where reliability and latency requirements surpass bandwidth needs. Finally, mMTC addresses the scale requirements of IoT. Apart from some smart city scenarios, mMTC may not be required in most locations for some years, with NB-IoT and other LPWA such as LoRa being sufficient for a while.

Benefit Rating: High

Market Penetration: 5% to 20% of target audience

Maturity: Early mainstream

Sample Vendors: Cisco; Ericsson; Huawei; NEC; Nokia; Samsung; ZTE

Recommended Reading: “Market Guide for 5G New Radio Infrastructure”

“Assessing 5G Mobile Technology for Organizations”

“How to Select 5G NSA/SA Migration Paths”

“Forecast: Communications Service Provider Operational Technology, 1Q20 Update”

“Market Trends: Strategies Communications Service Providers Can Use to Address Key 5G Security Challenges”

“Reduce Privacy Risks When Using 5G Products and Services”

mmWave

Analysis By: Bill Ray

Definition: mmWave is the frequency band between 30GHz and 300GHz, where the wavelength varies between 10 mm and 1 mm, respectively. However, nearby bands including 24 and 28GHz are also commonly referred to as “mmWave.” The 5G mobile telephony standard extends into several mmWave bands, which are also used for point-to-point wireless communication and satellites links.

Position and Adoption Speed Justification: The addition of mmWave bands to the 5G standard has focused a great deal of attention, and investment, on what was (hitherto) a band used only by specialists and niche applications. The U.S., in particular, has been active in deploying 5G networks into mmWave and China has announced plans to allocate mmWave frequencies to its domestic operators. mmWave is attractive because it is largely unoccupied, so it can provide very high capacity, but it’s poor penetration and propagation makes deployment of blanket coverage expensive and complicated.

The WiGig standard (802.11ad and 802.11ay) uses mmWave at 60GHz but has achieved little commercial success despite being available for more than a decade. The limited range, and the increasing speed of Wi-Fi, prevented WiGig from being adopted outside a few niche applications (such as wireless VR headsets). However, the adoption of mmWave in 5G has triggered investment in, and mass production of, mmWave antennas and transceivers, reducing the cost of equipment capable of operating in the band.

This innovation is further driven by the release of large blocks of mmWave into the public domain. In 2019 the U.K. regulator Ofcom has released 2.24GHz of spectrum around 24GHz for unlicensed use indoors, while the U.S. regulator Federal Communications Commission (FCC) released 14GHz of spectrum at 60GHz (including the band used by WiGig). These allocations are intended to spur development of new standards, and new applications, for high-speed, short-range, wireless connections.

Investment in 5G is creating new antenna designs and beamforming techniques which will benefit all applications of mmWave, increasing the utility while decreasing the cost of deployment. Phased-array mmWave antennas capable of accurately directing radio signals are a necessity for the next generation LEO satellite constellations. The technology can also be used for highly-accurate location tracking and for using reflected signals to improve the coverage of mmWave systems.

The publicity around 5G has driven mmWave to the top of the Hype Cycle very quickly, and it is now descending as 5G customers find that the bandwidth they get does not match with what was promised by network operator. This, combined with spurious concerns over the safety of mmWave, will push the technology into the trough of disillusion over the next year or two.

User Advice: Anyone looking for point-to-point, or point-to-multipoint, links should include mmWave in the technologies being evaluated. The limited range, and highly-directional, nature of mmWave means that license-free (or light license) systems can operate without fear of interference, and at much lower cost, than licensed systems. Campus networks should consider mmWave point-to-multipoint systems for backhaul of private LTE cells or Wi-Fi access points, especially where clear line of sight is available.

Equipment to run such links is already available but is also developing quickly as enhancements created for 5G networks filter into the rest of the industry. Users should expect to see costs falling rapidly over the next couple of years (following a similar path to the drop in pricing of equipment using microwave frequencies, once that was incorporated into the 4G standard).

Business Impact: The availability of mmWave bands provides plenty of bandwidth. In many instances mmWave will be able to replace wired connections, with ultra-reliable connectivity (using simultaneous connections to multiple access points) making it as reliable (if not more reliable) than wired connectivity. This will enable production lines to reduce the cable loom, increasing flexibility and reducing maintenance costs. Stationary equipment is ideally suited for highly directional mmWave signaling, but with beamforming and other optimization techniques mmWave rapidly becomes suitable for robots and other mobile endpoints.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Metawave; Qualcomm; Renesas Electronics; Texas Instruments

Recommended Reading: “Market Trends: mmWave Opportunities Outside Smartphones”

“Don’t Expect 5G to Replace Wired Access WANs Anytime Soon”

Sliding Into the Trough

Edge Servers

Analysis By: Thomas Bittman

Definition: Edge servers collect data, deliver content, and perform analytics and decision making close to data producers (e.g., sensors and cameras) and data consumers (e.g., people and actuators). They have broader and more general capability than gateway servers, but are less powerful or multitenant than micro data centers. Edge servers connect to enterprise or cloud data centers, and gateways that are connected to many local endpoints.

Position and Adoption Speed Justification: Edge serving has been in existence for decades in one form or another. For example, edge servers are used by content delivery networks (CDNs) to perform content caching, outbound delivery and low-level, end-user interactions. Industrial programmable logic controllers (PLMs) can read sensors and take simple actions when certain triggers are met. However, such traditional edge servers lack the performance to do sophisticated analysis, and often lack the capacity to handle large volumes of data. For some use cases, edge servers are mature. For others, edge servers are still evolving.

Edge serving is evolving rapidly, driven by the explosion of data being generated at the edge by the industrial Internet of Things (IoT) and immersive experiences. There will not be enough compute power, storage capacity or network bandwidth in the cloud or central data storage facilities to handle the volumes of data that will be generated by these devices in the future. The need to locally handle large volumes of data, make real-time decisions and execute machine learning algorithms is increasing. Hence, the edge server must become ever-more powerful, and will morph into micro or even macro edge data centers that can be deployable to almost any remote location.

These edge computing devices will need to gather the data, perform network protocol translation and preprocess the data before passing only the valuable data to an enterprise or cloud data center. They will also perform intense analytics processing and storage that has traditionally been performed in the cloud or at a more central location. They must also be manageable at extreme scale, operate in diverse environments that might lack traditional infrastructure, and be able to provide robust and modern digital and physical security capabilities.

Gartner predicts that demand for edge servers in support of industrial IoT and immersive experiences (e.g., augmented or mixed reality in the workplace, school or mall) will grow dramatically through 2021. By year-end 2021, more than 50% of large enterprises will deploy at least one edge-computing use case to support IoT or immersive experiences, versus fewer than 5% in 2019. By 2022, more than 50% of enterprise-generated data will be created and processed outside the data center or the cloud.

User Advice: Edge server requirements will be unique for each situation; however, the ability to evolve as requirements evolve is important. Because the data being collected and the actions that need to be performed are evolving rapidly, choose edge servers that can be deployed rapidly and are extensible in the field to match changing requirements. Expect this market to evolve rapidly, avoid lock-in where possible and plan for technology changes in a few years in many cases. Edge servers are only as good as the data analytics that they support, so give preference to IoT frameworks that are mature and have broad support. Security requirements must be an upfront design goal in any edge server deployment.

Business Impact: Edge servers will become a critical part of most enterprises' infrastructure topologies — either owned by the enterprise or leveraged as-a-service through cloud or telecom providers. The center of gravity of digital business will expand from central processing in the enterprise and the cloud to real-time processing and engagement at the edge. Edge servers will be important enablers of new digital business capabilities.

Benefit Rating: High

Market Penetration: 1% to 5% of target audience

Maturity: Early mainstream

Sample Vendors: ADLINK; Cisco; Dell; Eurotech; Hewlett Packard Enterprise (HPE); Lenovo

Recommended Reading: “Why and How I&O Should Lead Edge Computing”

“How to Overcome Four Major Challenges in Edge Computing”

“Exploring the Edge: 12 Frontiers of Edge Computing”

Edge Video Analytics

Analysis By: Alan Priestley

Definition: Edge video analytics is the real-time analysis for object detection/avoidance and event recognition of video data, either executed within a video camera or edge computing system. Edge video analytics systems have the ability to run deep neural network (DNN) algorithms within either the camera or edge computing system.

Position and Adoption Speed Justification: Growth in deployment of video cameras for a wide range of applications, such as surveillance, security, factory inspection systems, facial recognition (security authentication, personalized retail experiences) and autonomous vehicles creates a massive volume of data that must be analyzed and utilized for decision-making purposes. For many applications, the data analysis must be automated and is time-critical, requiring processing at point and time of data capture.

Edge video analytics systems place analysis and decision making either in the video capture system or an edge computer system capable of processing data captured from multiple video sources. This deployment of video analytics into edge systems and devices is being enabled by the development of a range of semiconductor devices such as VPUs, GPUs and ASICs that can execute the DNN algorithms necessary to process the captured video streams. A number of the major cloud service providers are also developing services, APIs and chips (for example, Google’s Tensor Processing Unit [TPU]) to support edge video analytics. The use of higher resolution cameras, with consequent higher data volumes, drives the need for high-performance edge video analytics systems.

User Advice: The number of deployed video cameras is rapidly increasing and enterprise architecture and technology innovation leaders must evaluate a range of criteria to determine where best to implement the necessary analytics capabilities, either at the edge or within a central data center (cloud or on-premises). Some of the criteria that must be considered include:

- The volume of captured video data
- The need for real-time analysis and decision making
- Complexity of the DNN algorithms being proposed versus the capabilities of the processing hardware available in the edge systems and endpoint devices

- Maturity of the DNN algorithms and associated video analytics applications
- Data communications availability

Business Impact: Video cameras are already widely used for surveillance and monitoring applications, however, the data is often stored and analyzed at a later date — often by humans. Video analytics enables real-time analysis and decision making based on the captured video data. The ability to place this analytics capability within edge systems will enable a wider range of use cases, bringing efficiencies to many deployments — such as real-time traffic monitoring and control, factory inspection systems and facial-recognition-based security systems. The growth in deployment of autonomous vehicles also requires real-time object detection systems and the ability to interpret the real-world environment surrounding the vehicle.

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: Herta; Huawei; Intel; IntelliVision; Mobileye; NVIDIA; Xilinx

Recommended Reading: “Forecast Analysis: AI Neural Network Processing Semiconductor Revenue, Worldwide”

“Market Trends: Edge Computing Creates New Equipment Design Challenges”

“5 Questions a Product Manager Must Ask When Creating an AI-Enabled Edge Product Strategy”

“Exploring the Edge: 12 Frontiers of Edge Computing”

“Top 10 Strategic Technology Trends for 2020: Empowered Edge”

Industrial IoT Gateways

Analysis By: Santhosh Rao

Definition: An industrial IoT (IIoT) gateway is a bridge between a field network and the IoT platform or, sometimes, a business application. IIoT gateways can be viewed as data aggregation points for field devices or wireless networks. They provide local storage and compute capabilities, as well as user interfaces (UIs) for data processing and system management.

Position and Adoption Speed Justification: Today, most IIoT gateways are designed to offer four basic functions: data aggregation, data normalization, basic event filtering and data forwarding to the IoT platform. However, Gartner expects gateway hardware vendors to develop gateways and edge servers with various form factors that enable higher processing capabilities and address a broader range of use cases. We expect the next generation of gateways to include ASICs, field-programmable gate arrays (FPGAs) and graphics processing units (GPUs). They will include chipsets specifically designed for vision processing and artificial intelligence (AI) and advanced cooling capabilities. From a software standpoint, containers are increasingly being viewed as the software building block for IIoT gateways. Attributes such as modularity, portability, low overhead and isolation make containers attractive choices for application and device management in the gateway.

User Advice: IT leaders must create a roadmap for the current and future role of gateways in the IoT architecture, and invest in a combination of low-end IoT gateways and intelligent IoT gateways to address various use cases. They must standardize on gateways that can be managed and programmed by preferred IoT platforms and can provide software development kits for custom integration and management. They must aim to extend the usable life of gateways and reduce the need for new investments by choosing gateways that are modular and can accommodate new peripherals. I&O leaders must select gateways that are certified for international, as well as country and industry-specific, safety standards, and provide an adequate level of overall system security.

Business Impact: IIoT gateways are viewed as a critical component of any industrial IoT project as today's gateway's begin their transition from systems that aggregate, filter and forward data to systems that can provide localized and near-real time insights. Manufacturing, automotive and utility sectors are increasingly deploying IoT gateways as part of the Industry 4.0 project initiatives to aggregate data from assets at the plant to address use cases such as predictive maintenance, production quality control and personnel safety. IoT gateways are being installed in buildings to connect HVACs and lighting systems to IoT platforms in order to forecast and better manage energy needs. Gateways play a critical role in smart city initiatives to address a broad range of use cases such as traffic management, water quality management, smart parking and various others.

Benefit Rating: High

Market Penetration: 20% to 50% of target audience

Maturity: Early mainstream

Sample Vendors: Advantech; Cisco; Dell; Efftronics; Eurotech; Hewlett Packard Enterprise; Logic Supply; Moxa; SECO

Micro-OS (IoT)

Analysis By: Martin Reynolds

Definition: A micro-OS (IoT) is an embedded real-time operating system (RTOS) that provides the minimum resources needed to support a computing device. The footprint may be as small as 256 bytes. Using a micro-OS in a product improves software security and reliability over monolithic code, and makes modifications and enhancements easier.

Position and Adoption Speed Justification: Micro-OSs have existed for many years, but increasingly sophisticated microcontrollers have rendered them somewhat obsolete. A few dollars of extra cost could support a stripped-down Linux. However, the proliferation of IoT devices and the need to drive costs down and reliability up create new opportunities for these small operating systems.

The technology is well-developed, with multiple vendors, open-sourced or available code, and sizes range from a few hundred bytes to a few megabytes. Amazon has integrated FreeRTOS into its IoT portfolio to promote its IoT initiatives.

Compared to, say Arduino, a micro-OS brings full multithreading capability; this feature makes it possible to run multiple functions independently of each other. The micro-OS, therefore, increases code reliability by creating some separation between functions. The more structured code also makes it easier to port the code to a new architecture.

Compared to Linux, even the smallest Linux systems require hundreds of megabytes of storage and a 32-bit processor. However, micro-OSs represent a significant step up in capability and complexity, while decreasing cost and risk relative to a monolithic system design.

Of particular note are the controllers manufactured by Espressif Systems, which integrate a 32-bit microcontroller core and a Wi-Fi interface into a single chip. This device is available in tiny single board computers for less than $6 in single unit quantities, supported by a richly customized FreeRTOS toolset. As a micro-OS platform, it brings advanced wireless connectivity without the burden of a full operating system.

User Advice: For small IoT devices where cost and power consumption are critical, use a low-power microcontroller supported by a micro-OS. A micro-OS is also useful for core security functions, where the microcontroller provides access and control for security features of a larger system, in a manner similar to that of a Trusted Platform Module (TPM).

In many cases, a micro-OS will fit inside the memory on the chip. This capability can significantly increase product security, as it is difficult to track code execution or download the binary code.

Although monolithic code can do the job, a micro-OS adds flexibility and reliability. A micro-OS is particularly important where the device must run several functions in parallel.

If you are developing an IoT product, look to systems that run a micro-OS to give you a semiconductor system cost well under $10, even with a wireless interface. The development tools are inexpensive, and enable your teams to prototype IoT devices with both lower risk and cost than associated with more sophisticated Linux systems.

If you are using a cloud-based system, trial the inexpensive microcontrollers offered by Amazon and Microsoft. These microcontrollers are well secured and managed at the hardware level.

Business Impact: Low-cost, low-power microcontrollers, in conjunction with a micro-OS, can reduce power consumption and increase software reliability of edge devices. These are critical attributes for a profitable IoT project and can significantly reduce operating costs. For example, an 8-bit system with a micro-OS might deliver several years of battery life, compared with months for a similar function implemented with an advanced OS. Software update costs and battery replacement costs are important factors in the cost of an IoT project, as they scale with the number of endpoints.

Benefit Rating: Moderate

Market Penetration: 1% to 5% of target audience

Maturity: Adolescent

Sample Vendors: Amazon Web Services; Espressif Systems; Micrium; Microchip Technology; MicroPython; TinyOS

LPWA

Analysis By: Tim Zimmerman; Aapo Markkanen; Bill Menezes

Definition: Low power wide area (LPWA) network technologies — referred also as LPWANs — are a set of wireless wide-area network solutions designed for IoT applications that are characterized by their long signal range, low cost, and low throughput. The category comprises technologies such as LoRA, Sigfox and Cat-M/NB-IoT. Communications are typically over 5 miles/10 km. and involve small amounts of data at data rates of between 0.3 Kbps to 50 Kbps between low-powered devices such as sensors that have battery life of several years.

Position and Adoption Speed Justification: The number of IoT devices continues to explode and use cases are not limited to in-building or close proximity applications. The adoption of LPWAN has been slow because some of the solutions are vendor-specific and proprietary, for example, Sigfox. Coverage, slow standardization (3GPP) and cost of data have also been significant impediments to faster adoption. With the availability of LTE-M as part of the 3GPP specification Release 13 and the reduced complexity of NB-IoT (no gateway required), we expect proprietary solutions to become niche or special purpose. The 3GPP 5G specifications include support for NB-IoT and LTE machine type communications, with enhanced features in Rel. 17 slated for late 2020 ratification.

User Advice: The LPWA category involves a variety of technologies that aim to address similar use cases but have some inherent differences in comparison to each other. Gartner's current view of the most important technologies can be summarized as following:

- The compatibility of Cat-M/Cat-M1 allows it to be deployed over existing LTE networks. End users benefit from the technology's relatively seamless mobility, but its comparatively high-power consumption limits applicability for battery-operated devices.
- NB-IoT is the standard that is enjoying the greatest level of support and adoption in the 3GPP ecosystem, making it the current frontrunner among the available LPWA technologies.
- LoRa's adoption as a public, operator-driven technology has been low, but it has seen growing uptake in the domain of private enterprise networks.
- Sigfox has seen limited traction, considering its low-cost, high-volume proposition.

Business Impact: Differing use cases and technologies will affect what technology is selected and how it will be implemented, please see below:

- Organizations with large numbers of geographically dispersed sensor should consider LPWA for sensor communications including smart metering, smart outdoor lighting or remote sensors used for monitoring applications such as agriculture, construction or utilities.

- Organizations must continue to monitor data plan pricing to assure that any solutions provide the required ROI.

Benefit Rating: Moderate

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: AT&T; Deutsche Telekom; Huawei; Semtech; Sequans Communications; Sierra Wireless; Sigfox; Telefónica; Verizon; Vodafone

Recommended Reading: "Magic Quadrant for Managed IoT Connectivity Services, Worldwide"

"Critical Capabilities for Managed IoT Connectivity Services, Worldwide"

Internet of Things

Analysis By: Alfonso Velosa; Benoit Lheureux

Definition: The Internet of Things (IoT) is a core building block for digital business and digital platforms. IoT is the network of dedicated physical objects that contains embedded technology to communicate and sense or interact with their internal states and/or the external environment. IoT comprises an ecosystem that includes assets and products, communication protocols, applications, and data and analytics.

Position and Adoption Speed Justification: Gartner's CIO Survey 2020 shows that IoT is regarded by CIOs as one of the top five game-changing technologies, with enterprises vary widely on their IoT adoption depth and maturity. Enterprises on a global basis have ongoing IoT-enabled initiatives for use cases ranging from incremental benefits (for example, asset optimization or compliance reporting) to transformative benefits (for example, product as a service or guaranteed asset uptime). The more developed use cases center on fleet management and industrial equipment maintenance, where ROI is calculated from cost optimization such as reducing maintenance and fuel costs. Many enterprises are now exploring employee and citizen safety solutions using IoT enabled capabilities. Finally, Gartner's 2019 IoT Survey indicates that while enterprises expect a 3-year payback on average for their IoT projects, 42% expect payback in less than 2 years. In the 2020 economic downturn, many clients are pushing for even shorter project paybacks.

The hype has decreased from the highs in 2016 through 2019; we reflect this by moving the profile's position into the trough. Enterprises continue to address cost, complexity and scaling challenges implementing IoT-enabled business solutions, as well as increased adoption of contact-less monitoring solutions, drone inspections, etc. driven by the 2020 pandemic. Challenges include end-to-end integration complexity, the need to bridge cultural divides between IT and operations, confusing vendor marketing, especially as they increasingly shift to IoT-enabled business solutions, security concerns, and the 2020 pandemic disruption on IoT project schedules.

User Advice: CIOs should take action to address IoT concerns across the following areas:

- Business: Measure and deliver IoT value based on digital and strategic business objectives. If you are still experimenting, use a proof of business value approach. Build business cases with project payback of less than 18 months to account for implementation challenges and cultural resistance. Add employee and customer safety to your priority list of IoT projects and capabilities.
- Management: Build or contribute to an IoT center of excellence (COE) composed of IT, operational and business personnel. Use the COE to drive global alignment on best practices, alignment to business objectives, budgeting and people allocation. Remember that IoT is really about business process transformation, so focus on culture change first and technology second to ensure success.
- Architecture: Ensure the architecture teams focuses on both the IT and operational technology portfolio as well as the need to manage a multi-IoT platform approach. Ensure analytics and applications are part of the conversation.
- Skills: Invest in business and architecture skills to support project ideation and prioritization, as well as technical skills for IoT platforms, integration, analytics and security. Drive learning via projects with short-term outcomes, and include business leaders, IT leaders, and front-line workers.
- Vendors: Assess and select providers on how they lower project risk for your enterprise via their vertical market expertise, technical capabilities (including best-of-breed partners) and trained professional services partners. Ensure your vendors integrate into your IT infrastructure.
- Governance: Establish accountability, participation, predictability and transparency policies for IoT — addressing sponsorship, budgets, digital ethics, data ownership and rights to monetize IoT data, etc...
- Risk: Scan for threats from enterprises in your ecosystem who may use IoT capabilities to damage or limit your differentiation and competitiveness.

Business Impact: As an evolutionary business impact, IoT will impact most enterprises' internal operations, differentiation, competitive position, and product strategies. Connected things will help lower costs, drive revenue, and improve enterprise processes in these types of usage scenarios:

Optimization of a range of business processes:

- Cost optimization: Lower operating costs for energy reduction, maintenance minimization, minimizing inventory spoilage, lowering theft
- Operations optimization: Better productivity, increased efficiency, logistics and coordination
- Optimize assets: Asset utilization, health monitoring, reliability, predictive maintenance
- Conserve resources: Energy efficiency and pollution reduction

New revenue strategies:

- Generate revenue via improved products, contractual services, usage-based pricing, and monetizing IoT data

- Increase engagement: Improved experiences of consumers, citizens and others in order to improve loyalty and increase customer lifetime value

Safety focus:

- Drive employee and citizen safety by monitoring and checking people's health, shifting to over-the-air updates to avoid in person visits, fall monitoring for the elderly and remote workers.

Benefit Rating: Transformational

Market Penetration: 5% to 20% of target audience

Maturity: Adolescent

Sample Vendors: ABB; Alibaba Cloud; Altizon; GE Digital; Hitachi Vantara; Tencent; Vodafone

Recommended Reading: "Predicts 2020: As IoT Use Proliferates, So Do Signs of Its Increasing Maturity and Growing Pains"

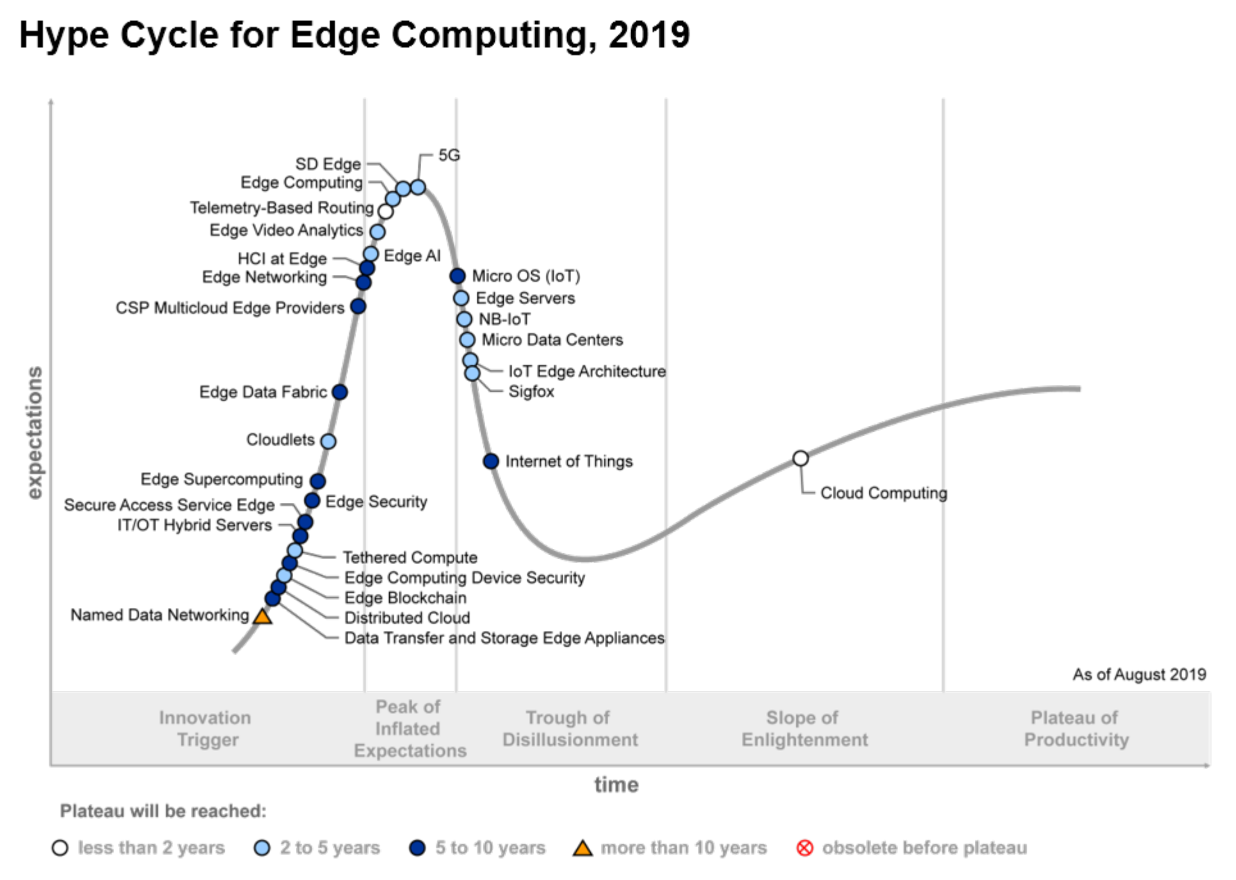
"Toolkit: Enterprise Internet of Things Maturity"

"Survey: Manufacturers See Quick Return on IoT Projects"

"Forecast: Enterprise and Automotive IoT, Worldwide"

Appendixes

Figure 3. Hype Cycle for Edge Computing, 2019



Source: Gartner
ID: 403314

Hype Cycle Phases, Benefit Ratings and Maturity Levels

Table 1. Hype Cycle Phases

Phase	Definition
Innovation Trigger	A breakthrough, public demonstration, product launch or other event generates significant press and industry interest.
Peak of Inflated Expectations	During this phase of overenthusiasm and unrealistic projections, a flurry of well-publicized activity by technology leaders results in some successes, but more failures, as the technology is pushed to its limits. The only enterprises making money are conference organizers and magazine publishers.
Trough of Disillusionment	Because the technology does not live up to its overinflated expectations, it rapidly becomes unfashionable. Media interest wanes, except for a few cautionary tales.
Slope of Enlightenment	Focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the technology's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools ease the development process.
Plateau of Productivity	The real-world benefits of the technology are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generations. Growing numbers of organizations feel comfortable with the reduced level of risk; the rapid growth phase of adoption begins. Approximately 20% of the technology's target audience has adopted or is adopting the technology as it enters this phase.
Years to Mainstream Adoption	The time required for the technology to reach the Plateau of Productivity.

Source: Gartner (July 2020)

Table 2. Benefit Ratings

Benefit Rating	Definition
Transformational	Enables new ways of doing business across industries that will result in major shifts in industry dynamics
High	Enables new ways of performing horizontal or vertical processes that will result in significantly increased revenue or cost savings for an enterprise
Moderate	Provides incremental improvements to established processes that will result in increased revenue or cost savings for an enterprise
Low	Slightly improves processes (for example, improved user experience) that will be difficult to translate into increased revenue or cost savings

Source: Gartner (July 2020)

Table 3. Maturity Levels

Maturity Level	Status	Products/Vendors
Embryonic	▪ In labs	▪ None
Emerging	▪ Commercialization by vendors ▪ Pilots and deployments by industry leaders	▪ First generation ▪ High price ▪ Much customization
Adolescent	▪ Maturing technology capabilities and process understanding ▪ Uptake beyond early adopters	▪ Second generation Less customization
Early mainstream	▪ Proven technology ▪ Vendors, technology and adoption rapidly evolving	▪ Third generation ▪ More out-of-box methodologies
Mature mainstream	▪ Robust technology ▪ Not much evolution in vendors or technology	▪ Several dominant vendors
Legacy	▪ Not appropriate for new developments ▪ Cost of migration constrains replacement	▪ Maintenance revenue focus
Obsolete	▪ Rarely used	▪ Used/resale market only

Source: Gartner (July 2020)

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

Understanding Gartner’s Hype Cycles

The Edge Completes the Cloud: A Gartner Trend Insight Report

4 Steps to Successful Edge Computing Deployments

Leading the Edge: Gartner’s Initial Edge Hardware Infrastructure Forecast

Minimize the Network Risk of Edge Computing

Cool Vendors in Edge Computing

Top 10 Strategic Technology Trends for 2020: Empowered Edge

Cloud and Edge Infrastructure Primer for 2020

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp