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

HEAVY READING
WHITE PAPER

The Distributed Cloud:
Infrastructure at the Edge

A Heavy Reading white paper produced for Intel

AUTHOR: ROZ ROSEBORO, PRINCIPAL ANALYST, HEAVY READING

INTRODUCTION

Communications service providers (CSPs) are transforming their networks to improve their service agility and operational efficiency. Distributed cloud architectures, in which compute, network, and storage resources are deployed in numerous locations outside of a centralized data center, will play an important role in the next generation of services that CSPs wish to deploy. The distributed cloud will present some new challenges, but it will also offer the promise of a variety of benefits, including lower costs and increased revenue. CSPs are progressing in a similar way as enterprises did years ago, starting with open source and virtualization and moving on to the cloud and its attendant automation.

Most CSPs agree that cloud-based architectures and principles will be critical to managing the sustained growth of data traffic. A majority of respondents to a recent Heavy Reading survey of CSPs indicated they had adopted or were planning to adopt a distributed cloud strategy. Equally important will be CSPs' ability to deliver new services. With the distributed cloud, CSPs can offer Internet of Things (IoT)-ready services as well as position themselves to support third-party applications and potentially enter adjacent markets to create new revenue streams. Note that these new services can be delivered even before 5G is deployed. However, 5G will require a distributed cloud be in place – and it will accelerate network transformations. Low latency services, which require processing closer to the end user, are expected to be a significant part of CSPs' 5G service portfolio.

The term "the edge" is used differently in different contexts and can describe locations as varied as a device, customer premises, a mobile base station, or a central office. Some definitions of the edge focus on the location, while others capture the architecture or the environment. CSPs appear to be taking all of these considerations into account as they deploy the distributed cloud.

With the opportunity to deploy infrastructure in so many different locations, CSPs' challenge will be to understand the criteria needed to optimize that infrastructure. Some of the key considerations include performance (e.g., latency and bandwidth), availability, security, integration, and data size/volume. Ideally, CSPs will consider their current and future needs to maximize the useful life of their distributed cloud infrastructure.

CSPs value different attributes in the different locations that comprise the distributed cloud. This highlights the fact that they will run a variety of workloads, each with its own requirements, in environments that have their own constraints. Scalability and security are highly rated attributes across many locations, as are high availability and performance, but their relative importance differs. In terms of environment, CSPs will need to account for more remote locations that may face more challenging conditions, like dust and heat, requiring more ruggedized form factors than are needed in centralized data centers.

A host of open source projects are working on solutions for edge deployments. Some, like Akraino, StarlingX, Central Office Re-architected as a Data Center (CORD), and Cloud Central Office (CloudCo), focus on creating a new edge software stack. Others, like Kubernetes and Open Networking Automation Platform (ONAP), provide management and orchestration (MANO) capabilities. The Open Compute Project (OCP) offers open source hardware, and the Open Network Edge Services Software (OpenNESS) project provides an open source reference toolkit to help users move applications from the cloud to the network and on-premises edge.

NETWORK TRANSFORMATION AND THE DISTRIBUTED CLOUD

What Is Distributed Cloud?

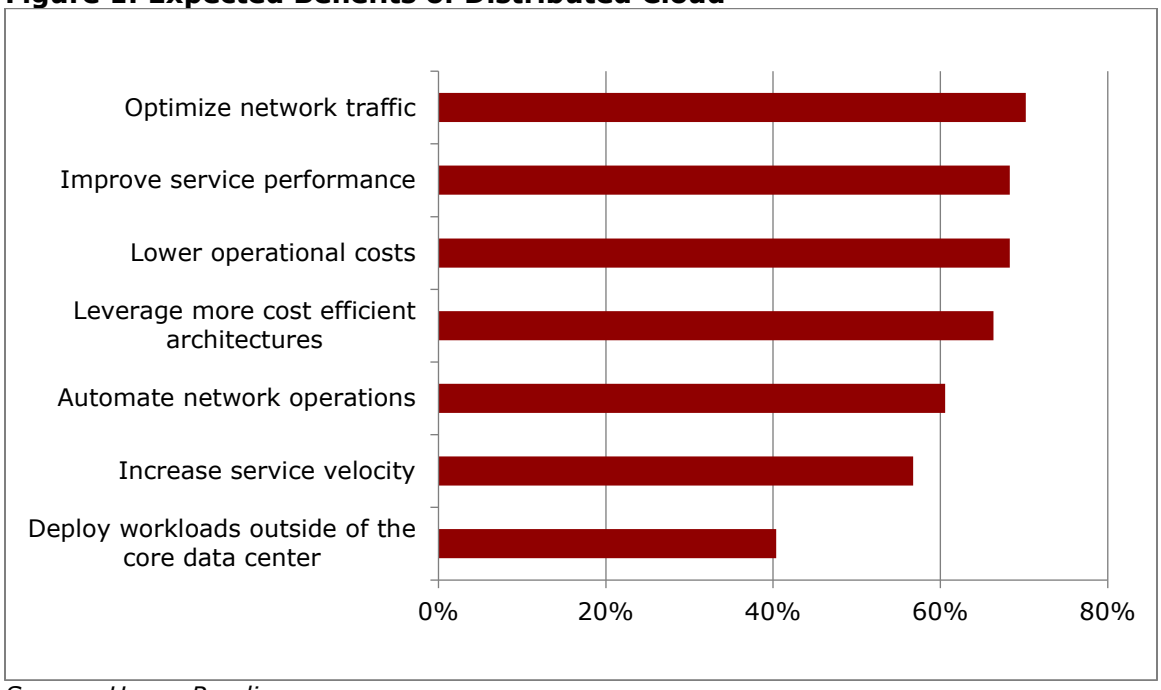
Because of the different possible definitions of "the edge" (as discussed later in this paper), Heavy Reading more often refers to the phenomenon of putting cloud infrastructure outside of a centralized data center as the "distributed cloud." Having resources in different types of locations presents different types of challenges and opportunities, along with a variety of benefits. As discussed later, each requires different attributes from its cloud infrastructure. CSPs are deploying distributed cloud infrastructure both to lower costs by using more efficient architectures and approaches and to increase revenue by deploying new types of services as well as improving the quality of existing services.

More Efficient Architecture

To improve their service agility and operational efficiency, CSPs are transforming their networks. They are walking the same path many enterprises did years ago, when they began consuming open source solutions such as Linux, virtualized their servers and storage, and then moved on to the cloud. Virtualization is usually the first step, with network functions virtualization (NFV) already a key strategic initiative for many CSPs. While it offers some benefits over the traditional approach of proprietary, purpose-built solutions, cloudification is where the most benefits can be achieved.

The pooling of resources and more highly automated operations that comes with cloud-based architectures and approaches delivers significant cost savings through more flexible, efficient, and agile operations. Note that the term "pooling" is used here intentionally. In the context of cloud, pooling moves one step beyond resource "sharing," as is done with hardware virtualization. Virtual resources appear unlimited to an application, as all resources are pooled and seen as available.

Figure 1: Expected Benefits of Distributed Cloud

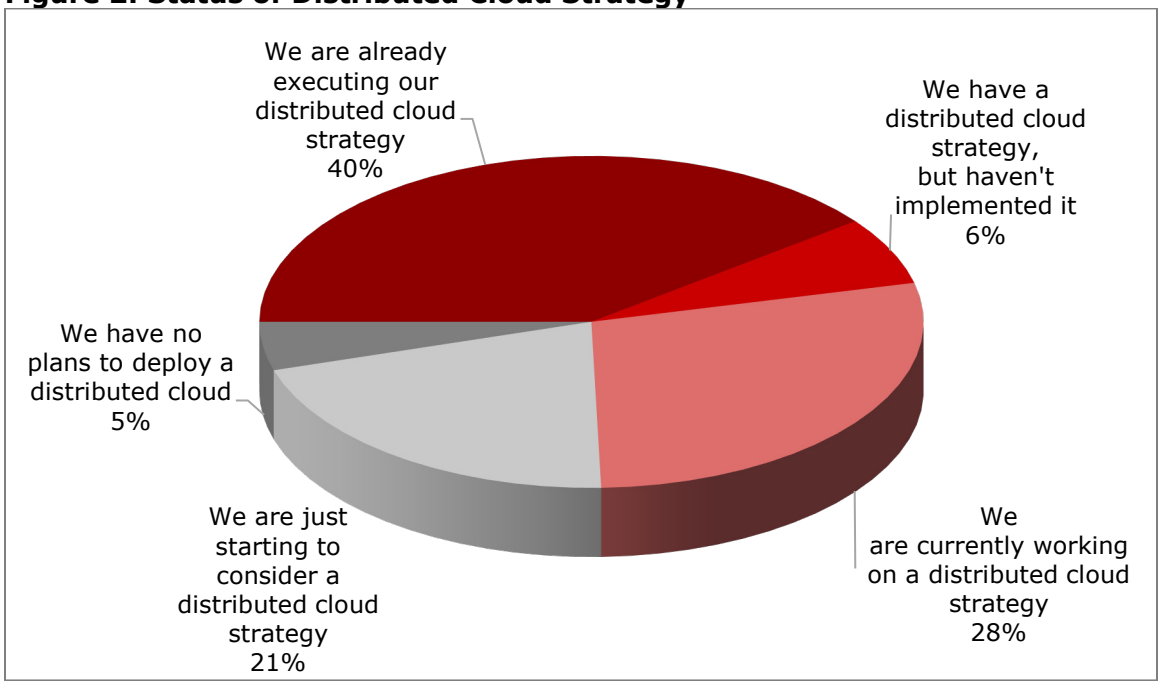


Source: Heavy Reading

Many CSPs have said that the only way they can efficiently manage the continued explosive growth in data traffic is to leverage cloud principles and architectures. While hyperscale was the initial model used by the public cloud giants, it is not well-suited for all use cases. Multi-cloud and distributed cloud approaches are more appropriate in some cases, with the latter already being leveraged to move and manage content in content delivery networks (CDNs).

Figure 2 shows that nearly all CSPs in the Heavy Reading survey referenced earlier either have or are working on a distributed cloud strategy.

Figure 2: Status of Distributed Cloud Strategy

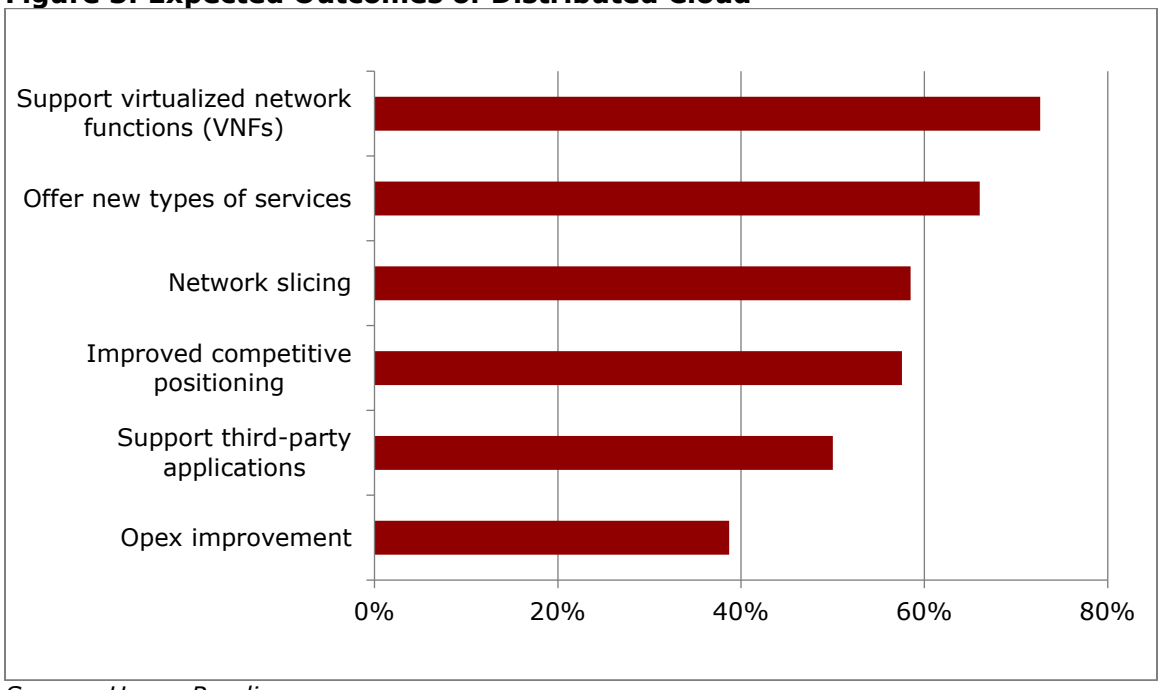


Source: Heavy Reading

Support New Services to Drive Revenue Growth

Just as important as lowering costs is increasing revenue. As shown in **Figure 3** below, CSPs believe that distributed cloud architectures will help them deliver new services like IoT, as well as support third-party applications driven by key use models and vertical markets. They will also be better positioned to address opportunities in adjacent markets and establish new partnerships. The potential exists for CSPs to extend and/or re-invent their businesses in the same way that Amazon did in evolving from a bookseller to a public cloud provider and entertainment/content producer.

Figure 3: Expected Outcomes of Distributed Cloud



Source: Heavy Reading

5G Will Accelerate Transformation

CSPs are particularly keen to get fiber and next-generation 5G deployed in their networks so they can begin to offer new, compelling services. 5G promises higher speeds, ultra-reliability, greater bandwidth, and most notably, lower latency than existing mobile networks. Having resources at the edge will be critical for 5G, since low latency connectivity will not matter much if the compute and storage needed to support a latency-sensitive service are thousands of miles away in a hyperscale data center.

To better manage backhaul costs, CSPs will not want to pay to backhaul terabytes of traffic that could be processed locally. Edge processing opens the door to services and use models that were not practical before, including augmented and virtual reality (AR/VR), connected car, real-time automation, and autonomous robots. Existing services, such as smart cities, content delivery, video, and gaming, will also benefit from the attributes 5G technologies bring.

In a 5G world, billions of devices and things will be connected, which means data traffic growth will continue unabated. Only cloud infrastructure, with its associated automation capabilities, will give CSPs the agility to operationalize 5G services. Much is touted about the lower latency and greater capacity that comes with 5G, but equally critical will be the increased intelligence, scalability, agility, and security it will provide.

DEFINING THE EDGE

What Does the Edge Mean?

The edge means many things to many people, so it is important to define this early on in any discussion. Edge computing can be summarized as the placement of data center-grade networks, compute, and storage closer to endpoint devices to improve service capabilities, optimize total cost of ownership, comply with data locality requirements, and reduce service latency. The primary focus of this paper is the On-Premise Edge, which includes the Enterprise Edge and the Network Edge, as shown in **Figure 4** below.

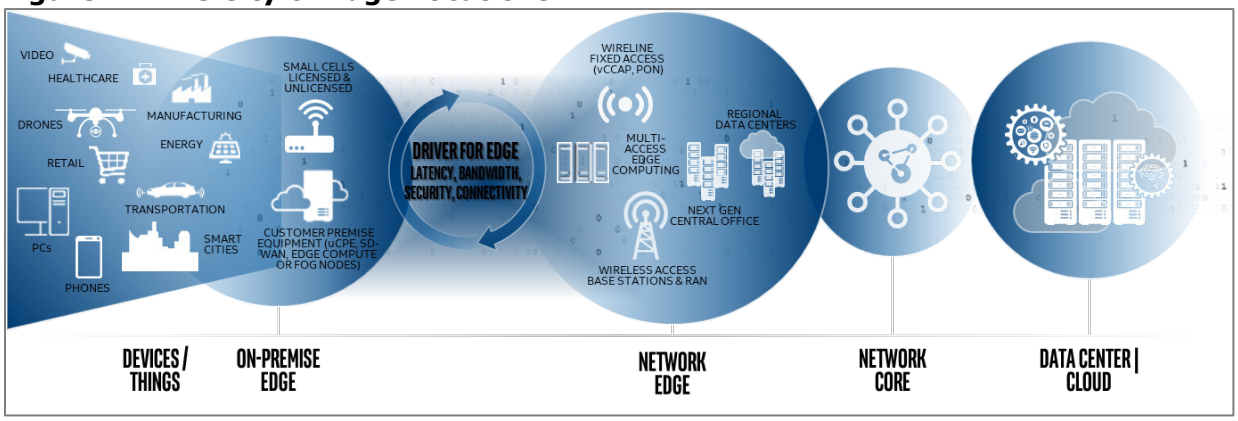
Heavy Reading recently surveyed CSPs to determine how they were defining the edge and found nearly equal levels of agreement (around half of respondents for each) for the following definitions:

- The edge describes a location representing where processing resources exist farthest away from a centralized data center.
- The edge describes an architecture where processing resources are distributed outside of a centralized data center.
- The edge describes an environment with space, power, and other constraints that affect equipment selection and management approaches.

Where Is the Edge?

The edge varies according to the use case and could even be customer-dependent. **Figure 4** illustrates the diversity of edge locations.

Figure 4: Diversity of Edge Locations



Source: Intel

Note: This paper is not intended to provide detail around the Devices/Things or the Data Center/Cloud.

OPTIMIZING THE INFRASTRUCTURE

Because CSPs now have a number of choices and requirements as to where to place workloads, their challenge will be to understand the criteria necessary to provide and optimize the infrastructure for each type of new workload. The more important criteria to consider include performance (latency and bandwidth), security, integration, and data size/volume (e.g., how much data must be moved and how long it would take). CSPs must account for both current and future needs so the infrastructure they deploy today can grow with them over time. The expected useful life of next-generation CSPs' edge infrastructure will be shorter than previous generations of largely telco-type equipment, but it will certainly be longer than the 3 to 5 years expected in an enterprise environment.

Performance and Availability

Dramatically reducing latency is the only way services like AR/VR and connected car will be feasible, making this an important determinant of workload placement. The shorter the distance and the fewer equipment hops between where traffic is generated and where it is processed, the lower the latency. **Figure 5** shows latencies calculated by British Telecom in its 5G network latency modeling. In its testing, "content [was] served from the same location as the User Plane Function (UPF) node."

Figure 5: Modeled Latencies by Location

UPF Location	Access	Aggregation	Core
Number of Sites	1,200	106	10
Transport Latency (1-way)*	0.6 ms	1.2 ms	4.2 ms
Estimated 5G Latency	9.2 ms (Enhanced Mobile Broadband [eMBB])	10.4 ms (eMBB)	16.4 ms (eMBB)
	2.2 ms (ultra-reliable low latency communication [URLLC])	3.4 (URLLC)	9.4 ms (URLLC)

Source: BT

*Assumptions:

- Latency figures based on 95th percentile of transmission delay (i.e., 95% of cell sites are within this) and overhead for IP
- 5G RTT assumes 8 ms overhead for 5G New Radio and Next-Gen Core (eMBB case), 1 ms for URLLC (as per 3GPP 5G)

Performance is critically important for real-time applications where latency back to the core does not meet the applications control requirements, so control and processing must be handled locally. Another aspect of performance is how quickly resources can be made available for an application when they need them, including when the application needs to scale to support increased demand at a particular location. This is important for services that have unpredictable or intermittent resource requirements, such as a mobile service that needs to scale well beyond normal capacities to support people attending a sporting event. Today, networks must be overprovisioned for such a circumstance, but the application of NFV allows this right-sizing approach.

Data availability is related to CSP network performance, which will vary from CSP network to CSP network due to the differing configurations already deployed and is equally important. It defines how the edge provides access to data when needed, at the right time, regardless of the size of that data. Resources at the edge may help to cache content to enable better service scalability.

Security

Security encompasses numerous dimensions. Platform security, including hardware root-of-trust-based verified chain of platform boot and attestation, helps ensure trust in the underlying infrastructure. Platform security ensures that all firmware and operating systems installed on the platform is authorized for deployment, protecting against firmware rootkits and botnets. Physical security and access control of the premises itself – which may be deployed in unmanned locations – are important. Using platform security for privacy, data sovereignty, and addressing regulatory requirements is important and includes hardware protection for sensitive cryptographic credentials (e.g., customer private keys as part of hardware-based cryptographic accelerators).

For enterprises that do not want to host their own IT platforms, CSPs can offer a secure edge perimeter using virtual CPE (vCPE). These edge services cloud platforms are multi-tenant environments that require hardware-enforced cryptographic isolation between different tenants' workloads and between the tenant workloads and the software infrastructure (e.g., operating systems, virtual machine manager/hypervisor).

Integration

CSPs have not, and most likely will not, move all of their applications to a cloud. This means that as they develop new applications, they must consider how closely integrated they need to be to existing ones. Often, the work it would take to re-architect an old application for the cloud cannot be justified, so it remains on-premises or in a network point of presence. The potential integration challenge may drive CSPs to keep new applications colocated with existing applications despite their suitability for another location.

Data and Traffic Volume

The size of the workload is another important consideration. CSPs, even those that own their own fiber, aim to minimize transport costs. Therefore, they would rather process large workloads as close to the point they are generated as possible. Doing so allows them to backhaul only the most relevant information to a centralized data center.

IMPACT OF DISTRIBUTED ARCHITECTURES ON INFRASTRUCTURE

What Attributes Are Important at the Edge?

Heavy Reading research shows that CSPs value different attributes for cloud infrastructure depending on the location. This reflects that they expect to run different types of workloads at each location, as well as different constraints those locations must account for. As shown in **Figure 6**, while the important attributes are similar, the rankings are notably different.

Figure 6: Ranking of Cloud Infrastructure Attributes by Location

Attribute Rank	Base Station	Central Office/Node	Centralized Data Center
1	Energy Efficiency	Scalability	Scalability
2	Security	Security	Elasticity
3	Manageability	Manageability	High performance
4	High performance	High availability	Energy efficiency
5	High availability	High performance	High availability

Source: Heavy Reading

As noted above, CSPs need high performance in their cloud infrastructure regardless of where it is deployed. Hardware acceleration will likely play a critical role. Heavy Reading research shows that CSPs expect to use a mix of SmartNICs, graphics processing units (GPUs), application-specific integrated circuits (ASICs), and field-programmable gate arrays (FPGAs) to support a host of new use cases, with SmartNICs and GPUs as the most frequently cited. This is in addition to expected continual evolution in network throughput, memory, storage technologies, and more.

Given the massive geographic coverage of CSPs, the locations where CSPs are looking to deploy cloud infrastructure can be quite different from the traditional data center facilities – any of thousands of on-premises, base stations, or central offices. Some may be hard-to-reach locations, making remote operations a critical capability. Some may be in dusty, dirty environments or have high vibration, shock, which makes ruggedized hardware a necessity. All of this is to say that distributed cloud infrastructure must support a range of form factors that were not required for cloud infrastructure or traditional data center infrastructure currently used today.

What Role Will Open Source Play at the Edge?

In addition to the attributes described above, Heavy Reading research shows that containers and open, disaggregated hardware will play a role at the edge, with 97% of CSPs saying it is "very" or "somewhat" important to their distributed cloud infrastructure strategies. Numerous open source projects are developing solutions for the edge, with some of the most important described in **Figure 7** below.

Figure 7: Open Source Projects Addressing Distributed Cloud Infrastructure

Project Name	Description
Akraino	Developing a software stack of open source components to support "high availability cloud services optimized for edge computing systems and applications." Its focus is "on creating blueprints … of validated hardware and software configurations against defined use case and performance specifications."
Central Office Re-architected as a Data Center (CORD)	Using NFV, software-defined networking (SDN), and cloud architectures to build a next-gen, common platform to deliver fixed and mobile access services from a central office.
Cloud Central Office (Cloud CO)	Using NFV, SDN, and cloud technologies to redefine access and aggregation network architectures for central office infrastructure.
Kubernetes	Is "an open-source system for automating deployment, scaling, and management of containerized applications."
Open Compute Project (OCP)	Delivering open source designed data center hardware that meets four core tenants: efficiency, scalability, openness, and impact.
Open Network Edge Services Software (OpenNESS)	The Open Network Edge Services Software (OpenNESS) project provides an open source reference toolkit that to help users move applications from the cloud to the network and on-premises edge.
Open Networking Automation Platform (ONAP)	Is "a comprehensive platform for real-time, policy-driven orchestration and automation of physical and virtual network functions … to rapidly automate new services and support complete lifecycle management."
StarlingX	Developing a "cloud infrastructure software stack for the edge" by leveraging existing cloud technologies to orchestrate services.

Source: Heavy Reading

CONCLUSION

Distributed cloud architectures will play an important role in CSPs' network transformations. Having cloud infrastructure available in numerous locations outside of centralized data centers will allow CSPs to deliver next-generation services while also reaping the operational and cost benefits that cloud principles provide. Today, CSPs can deliver services like IoT using distributed cloud infrastructure, and the coming 5G networks will rely upon them – helping to accelerate CSPs' transformation.

The distributed cloud will be made of numerous edge locations, from the customer premises to mobile base stations to central offices, each of which has its benefits and constraints. CSPs will need to consider attributes such as performance, availability, security, integration, and data and traffic volume to optimize the infrastructure for a myriad of workloads. A host of open source projects, including Akraino, ONAP, OCP, and OpenNESS, are actively developing and delivering solutions to help CSPs build the robust distributed cloud infrastructure needed to drive current and future services.