VERTIV.

DEFINING FOUR EDGE ARCHETYPES AND THEIR TECHNOLOGY REQUIREMENTS

Introduction

Over the last several years, "edge computing" has become one of the most talked about trends in IT, and for good reason. Grand Valley Research projects a **CAGR of 41 percent for edge computing** between 2018 and 2025. Nearly every industry is recognizing the limitations of supporting users and emerging technologies through centralized IT infrastructures and is pushing storage and computing closer to users and devices.

That shift is becoming necessary because of the increased connectivity of devices and people and the huge volumes of data they generate and consume. According to the **Cisco Visual Networking Index** global IP traffic reached 1.2 zetabytes in 2016. By 2021, it will almost triple, reaching 3.3 zetabytes. Also by 2021, Cisco projects the number of devices connected to IP networks to be three times the global population. That represents more than 23 billion connected devices in just three years. **Other firms are making similar projections**: by 2020, Gartner projects 20.8 billion connected devices, IDC 28.1 billion and IHS Markit 30.7 billion.

A large portion of this IoT data will be mobile sensor data that must be transmitted on wireless or mobile networks rather than wired Internet connections, putting a strain on mobile network infrastructure. **Mobile IP traffic is projected to increase seven-fold by 2021**, double the pace of the growth in fixed IP traffic.

The changes in the compute and storage infrastructure required to support the smart and connected future, particularly at the local level, will be profound.

However, when you explore the information available today about edge computing, you discover that few, if any, resources exist that provide a comprehensive view of the edge ecosystem. A close analysis of the market reveals a wide variety of current and emerging use cases and, while they share some similarities based on the broad definition of edge computing, they are also distinct in some significant ways.

Vertiv edge experts, in conjunction with an independent third party consulting firm, analyzed the use cases that comprise the edge ecosystem to develop a better understanding of these differences and their implications for the supporting infrastructure. As a result of this analysis, we have identified four main archetypes for edge applications:

- Data Intensive
- Human-Latency Sensitive
- Machine-to-Machine-Latency Sensitive
- Life Critical

This paper presents a description of each archetype with examples of the most impactful use cases, along with an overview of their connectivity requirements to local, metro and regional hubs, which represent the edge transmission layer and core and are sometimes differentiated as edge, fog and cloud computing.

Understanding Edge Use Cases

To identify the four archetypes, it was first necessary to understand the use cases for edge technology. The Vertiv research team identified and reviewed more than 100 use cases for edge technology and refined this initial list to the 24 that will have the greatest impact on IT infrastructure for more detailed analysis.

That analysis looked at the performance requirements of each use case in terms of latency, availability and projected growth as well as security requirements such as the need for encryption, authentication and regulatory compliance. Also evaluated was the need to integrate with existing or legacy applications and other data sources and the number of potential locations required to support the use case.

Most importantly, the team studied the data characteristics of each use cases and found that the applications that underpin each have a data-centric set of workload requirements in addition to their requirements for availability and security. These include data volume, how data is accessed, data transmission requirements, data integrity and data analytics. This data-centric approach, filtered through requirements for availability and security, is central to understanding and categorizing the requirements of various use cases.

A list of the 24 use cases, organized by archetype, can be found in Figure One.

The Edge EcoSystem

DATA INTENSIVE	MACHINE TO MACHINE LATENCY SENSITIVE	LIFE CRITICAL	HUMAN LATENCY SENSITIVE
• Restricted Connectivity • Smart Cities • Smart Factories • Smart Home/Building • HD Content Distribution • High-Performance Computing • Virtual Reality • Oil and Gas Digitization	• Smart Security • Smart Grid • Low-Latency Content Dist. • Arbitrage Market • Real-time Analytics • Defense force simulation	• Digital Health • Connected/Autonomous Cars • Drones • Smart Transportation • Autonomous Robots	• Web Site Optimization • Augmented Reality • Smart Retail • Natural Language Processing

Figure 1: *Archetypes*

Archetype One: Data Intensive

Bandwidth	Latency	Availability	Security
High	Medium	High	Medium

The Data Intensive Archetype represents use cases where the amount of data makes it impractical to transfer over the network directly to the cloud, or from the cloud to the point-of-use, because of data volume, cost or bandwidth issues.

Probably the most widely discussed example of a data-intensive edge application is high-definition content delivery. **In 2016, video accounted for 73 percent of all IP traffic and that is expected to grow to 82 percent by 2021** as streaming video and virtual reality continue to grow. Major content providers, such as Amazon and Netflix, are actively partnering with colocation providers to expand their delivery networks and bring data-intensive video streaming closer to users to reduce costs and latency.

Already, **35 percent of the content accessed by a North American Internet user is sent from the municipal area where the user is located.** That is projected to increase to 51 percent by 2021 as content providers continue to extend their networks to the edge. Yet this represents only the first wave of core-to-edge computing. As the demand for high-definition video continues to grow, local hubs will increasingly support the current metro hubs to further reduce bandwidth costs and latency issues.

Another prime example of the Data Intensive Archetype is the use of IoT networks to create smart homes, buildings, factories and cities. A 2018 survey by 451 Research and Vertiv found that while only 33 percent of the 700 organizations surveyed had broadly deployed IoT, 56 percent indicated that at least 25 percent of their IT capacity currently supports IoT. Despite IoT still being in its early stages, organizations are already struggling to manage the volume of data being generated.

In this case, the challenge is the opposite of the one presented by high-definition content delivery. Rather than moving data closer to users, these applications must move the huge amounts of data generated by devices and systems at the source to a central location for processing. This will require the evolution of an edge-to-core network architecture.

IoT and the Industrial Internet of Things (IIoT) represent a mesh of sensors that generate huge volumes of data each hour. This data supports a "sense-infer-react" loop that enables visibility into and control of everything from home appliances to industrial equipment. Only a subset of this data is transmitted to a local, regional or cloud data center for further processing, which means massive amounts of compute will be required at the extremity of the edge to enable devices and systems to make decisions and act on the data provided by sensors.

The simplest of these applications, the smart home, must support multiple data-intensive devices and systems, including entertainment, HVAC systems, and security.

Data Intensive
According to IHS Markit, **the world market for connected home devices will grow from over 100 million units in 2017 to about 600 million units in 2021.**

Smart cities and factories take the data challenges inherent in smart homes and amplify them. Many cities are already piloting or evaluating smart city technology to improve traffic flows, support emergency services, and reduce costs.

Smart factories, which leverage the convergence of IoT, cyber-physical systems and cloud computing to allow manufacturers to use real-time data to increase efficiency, reduce costs and adapt to changes in demand, is being promoted as the next industrial revolution. According to McKinsey, factories and other production environments have the potential to realize the biggest financial impact from the application of IoT. They predict the IIoT will generate an **economic value of between $1.2 trillion and $3.7 trillion** by 2025. This value will come from new energy efficiencies, labor productivity, inventory optimization, and improved worker safety. But realizing it will require robust local infrastructure.

In the oil and gas industry, digitization has already created vast improvement in the efficiency of exploration and extraction processes, but has also introduced huge data management challenges. A single drilling rig can generate terabytes of data every day.

Other use cases that fall into the Data Intensive Archetype include virtual reality, high-performance computing and environments with restricted connectivity, such as areas where recovery operations are taking place following a natural disaster or cyberattack.

What all of these use cases have in common is the need to move large volumes of data to users where it can be consumed, or from devices and systems where it is generated to a central repository.

Archetype Two: Human-Latency Sensitive

Bandwidth	Latency	Availability	Security
Medium	High	Medium	Medium

The Human-Latency Sensitive Archetype covers use cases where services are optimized for human consumption. As the name suggests, speed is the defining characteristic of this archetype.

The challenge of human latency can be seen in the customer-experience optimization use case. In applications such as e-commerce, speed has a direct impact on the user experience; web sites that are optimized for speed using local infrastructure translate directly into increased page views and sales.

Human-Latency Sensitive
Google has found that adding a 500 millisecond delay to page response times resulted in a 20 percent decrease in traffic while Yahoo observed that a 400 millisecond delay caused a 5 to 9 percent decrease in traffic.

This effect also extends to payment processing. Amazon found that a 10 millisecond delay in payment processing caused a 1 percent decrease in retained revenue. Centralized approval via password took, on average, 7 seconds. A move to local processing brought the time down to 600 milliseconds, an improvement of 6,400 milliseconds with each 100 milliseconds potentially resulting in an extra 1 percent of retained revenue.

Another emerging example of a human-latency sensitive application is natural language processing. Voice is likely to be the primary form of interaction with everyday IT applications in the future. Natural language processing for Alexa and Siri is currently performed in the cloud. However, as the volume of users, applications and languages supported increase, it will be necessary to migrate these capabilities closer to users.

Other human latency use cases identified include smart retail, such as the cashier-less Amazon Go stores, and immersive technologies such as augmented reality where small latency lags can mean the difference fun and nausea.

In each case, delays in delivering data directly impact a user's technology experience, as with language processing and augmented reality, or a retailer's sales and profitability as with web site optimization and smart retail. As these use cases grow, so too will the need for local data processing hubs.

Archetype Three: Machine-to-Machine Latency Sensitive

Bandwidth	Latency	Availability	Security
Medium	High	High	High

The Machine-to-Machine Latency Sensitive Archetype covers use cases where services are optimized for machine-to-machine consumption. Because machines can process data much faster than humans, speed is the defining characteristic of this archetype. The consequences for failing to deliver data at the required speeds can be even higher in this case than in the Human-Latency Sensitive Archetype.

For example, the systems used in automated financial transactions, such as commodities and stock trading, are latency sensitive. In these cases, prices can change within milliseconds and systems that don't have the latest data when needed cannot optimize transactions, turning potential gains into losses.

M2M Latency Sensitive
According to a study by the Tabb Group, a broker could lose **as much as $4 million in revenues per millisecond** if its electronic trading platform was 5 milliseconds behind the competition.

Smart grid technology also falls into this archetype. This technology is being deployed in the electrical distribution network to self-balance supply and demand and manage electricity use in a sustainable, reliable and economic manner. It enables distribution networks to self-heal, optimize for cost and manage intermittent power sources, assuming the right data is available at the right time.

Other Machine-to-Machine Latency Sensitive applications include smart security systems that rely on image recognition, military war simulations, and real-time analytics.

Archetype Four: Life Critical

Bandwidth	Latency	Availability	Security
Medium	High	High	High

The Life Critical Archetype encompasses use cases that directly impact human health and safety. In these use cases, speed and reliability are paramount.

Probably the best examples of the Life Critical Archetype are autonomous vehicles and drones, which provide great benefits when they operate as designed; however, if they make bad decisions, they can endanger human health.

Autonomous vehicles have progressed faster than many expected, with a number of automotive and technology companies already actively testing vehicles today. Most of these vehicles have a human in the driver's seat ready to override automatic controls if problems are experienced to minimize the risk to human health. But, in the near future, driverless delivery vehicles and transport systems will be on the road. If these systems don't have the data they need when they need it, the consequences could be disastrous.

The same is true of drones. We could easily be looking at a future where hundreds of delivery drones are flying over a city at any given time.

Life Critical
Large e-commerce and package-delivery companies, such as Amazon and DHL, are already experimenting with drones for package delivery.

The increased use of technology in healthcare also represents a Life Critical Archetype. Electronic health records, cyber medicine, personalized medicine (genome mapping) and self-monitoring devices are reshaping healthcare and generating huge volumes of data.

Other examples include smart transportation and autonomous robots. The transportation and logistics industries are looking at data-centric solutions to improving driver and passenger safety, fuel efficiency, and asset management. Technology in this space will include intelligent transportation systems, fleet management and telematics; guidance and control systems; passenger entertainment and commerce applications; reservation, toll and ticketing systems; and security and surveillance systems.

Technology Requirements for Local and Regional Hubs

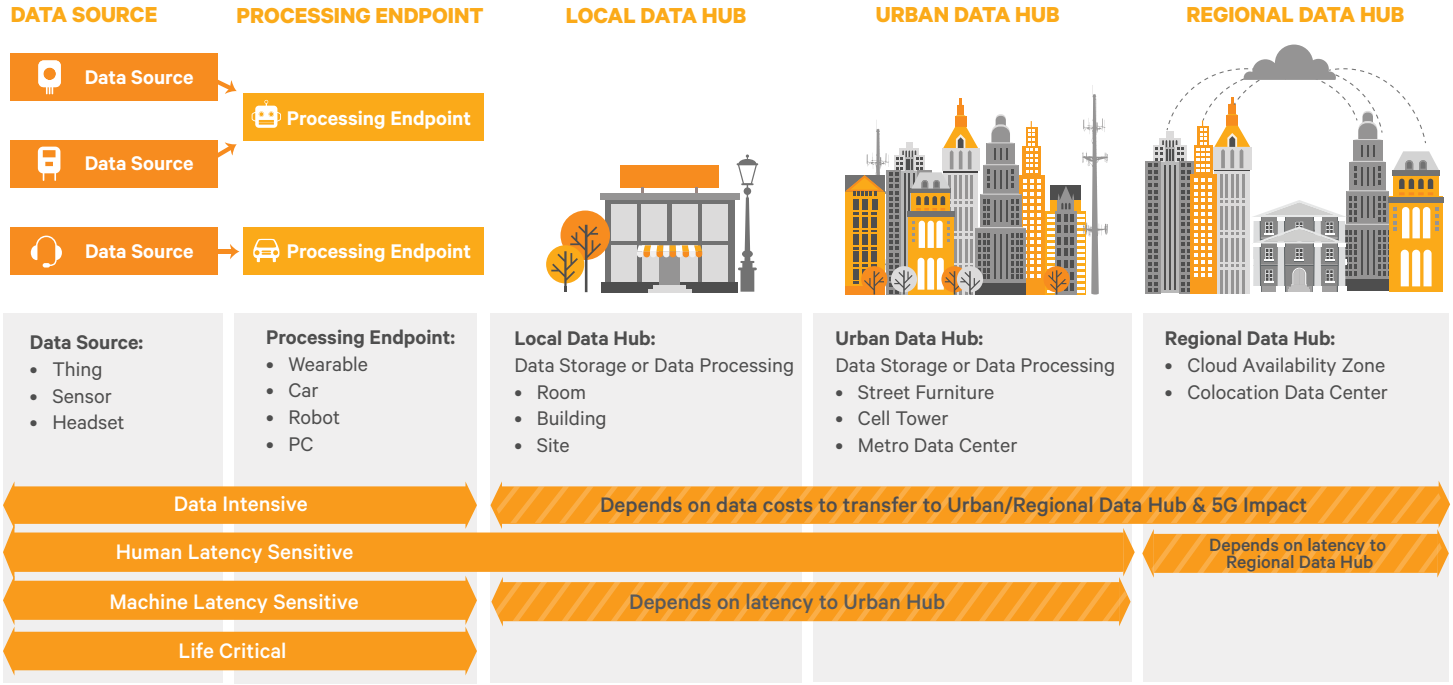
The infrastructure required to support these current and established use cases consists of four layers of storage and compute in addition to the communications infrastructure required to move data between the layers.

At the source, there is typically the device that generates or consumes data and a processing endpoint. The device could be a sensor monitoring anything from the powered status of a lamp, the access to a door, the temperature in a room or other desired information. The processing endpoint may be as simple as the PC or tablet a consumer is streaming video to, or could be the microprocessors embedded in automobiles, robots or wearable devices. These components are application-dependent and are typically designed in by the equipment manufacturer or retrofitted to existing devices.

Every archetype will also require a local data hub, which provides storage and processing in close proximity to the source. In some cases, the local hub may be a freestanding data center. More commonly, it will be a rack- or row-based system providing 30-300 kW of capacity in an integrated enclosure that can be installed in any environment.

These rack- and row-based enclosure systems integrate communication, compute and storage with appropriate power protection, environmental controls and physical security. For archetypes that require a high degree of availability, such as Machine-to-Machine Latency Sensitive and Life Critical, the local hub should include redundant backup power systems and be equipped to enable remote management and monitoring. Many uses cases will also require data encryption and other security features within the local hub.

For all archetypes except Life Critical, the local hub will require the ability to connect to a metro and/or regional hub, which will provide longer-term data storage and support capabilities such as machine learning.



The metro hub leverages the existing telco infrastructure to support the local hub with longer-term data storage and more robust processing capabilities. The regional hub is likely to be a cloud data center operating in the same region as the local hub.

For both the metro and regional hubs, modular designs capable of easily scaling beyond the initial design spec should be considered to account for unexpected surges in demand. These facilities should also be designed to scale in terms of density. Image-intensive applications, such as virtual reality, and processing-intensive applications, such as analytics and machine learning, will likely require rack densities that exceed the typical 10 kW design specification. In virtually all cases, these hubs should provide the same or higher level of redundancy and security as the local hub.

Moving Forward

By identifying the workload needs for the twenty-four use cases discussed, four leading archetypes emerged that can guide decisions regarding infrastructure and configuration requirements for the use cases analyzed as well as those that will emerge in the coming years. Vertiv will build on this initial archetype work to further define specific technology requirements and configurations for each archetype.

VertivCo.com | **Vertiv Headquarters,** 1050 Dearborn Dr. Columbus, OH 43085, USA

SL-11490 (R04/18)