Telecommunications Industry Association

TIA POSITION PAPER

Edge Data Centers

www.TIAonline.org • @TIAonline

Patent Identification

The reader's attention is called to the possibility that compliance with this document may require the use of one or more inventions covered by patent rights.

By publication of this document, no position is taken with respect to the validity of those claims or any patent rights in connection therewith. The patent holders so far identified have, we believe, filed statements of willingness to grant licenses under those rights on reasonable and nondiscriminatory terms and conditions to applicants desiring to obtain such licenses.

The following patent holders and patents have been identified in accordance with the TIA intellectual property rights policy:

» No patents have been identified

TIA shall not be responsible for identifying patents for which licenses may be required by this document or for conducting inquiries into the legal validity or scope of those patents that are brought to its attention.

Authors & Contributors

Authors:

David Sowry	CommScope
Dharmesh Jani	Flex
Don Duet	Independent
Frank Yang	CommScope
Harry Smeenk	TIA
James Young	CommScope
Phillip Marangella	EdgeConneX
Robert Bunger	Schneider Electric

Contributors:

Edward van Leent	EPI
Marta Soncodi	TIA
Rebecca Hunter	Crown Castle
Tom Craft	CommScope

Scope and Purpose

The next wave of technology innovation is already here with new applications transforming the way we live, work, and travel. The huge adoption of these new services drives exponential growth in the total demand for data. We must provide more data capacity and higher computing speeds if we hope to keep up. The sheer scale and scope of the gap we face demands that we re-think the way we have traditionally organized the design and deployment of networks and data centers. As many hands make light work, deploying many smaller distributed data centers seem the most viable solution. These facilities are often called "Edge Data Centers" (EDCs).

The Telecommunications Industry Association's (TIA) EDC Working Group, is providing this Positioning Paper as an overview defining development, implementation, and operational aspects of EDCs. EDCs seem to contrast with the opposite trend, consolidation of Data Center resources and public cloud. In fact, both types of DCs will be needed in the future. Machine-generated traffic might well dwarf current internet growth. Gartner research analysts forecast that one million new IoT devices will be sold every hour by 2021 sparking the huge growth of machine-to- machine data communications [1]. Independent of IoT, the 5G transformation of the telecom network will enable next-gen Augmented Reality (AR), Smart Factories, Smart Cities, and even Autonomous Vehicles, provided that the network can scale to support all these new applications.

EDCs offer a network strategy to serve the next wave of services while also delivering the growing capacity that these services will require. EDCs deployed at a large scale create a large supporting network with new characteristics that are designed to suit the emerging "Edge" service requirements. EDCs will be different from large centralized data centers that support the centralization of computing (think Hyperscale). Edge data will range from small clusters of "Edge Cloud" resources located on a street light, to a few racks located in a shelter at the base of a cell tower, inside buildings or in a building behind the corner gas station, to a self-contained ready-to-drop and plug-in container.

Where and how data services will be deployed will suit a new business model where the economy of scale's relationship with competitive advantage is broken and proximity becomes the dominant factor. Smaller capacities delivered by EDCs will be co-located with the service demand. While this seems to be a compelling new model realizing, it will require a new approach to Data Centers.

An architecture that introduces many thousands more EDCs will need to find ways to optimize capital costs, operational complexity and speed to market – and most importantly minimize the environmental impact of EDCs. Bell Labs [2] predicted that 60% of servers would be placed in an Edge Data Center by 2025. There is no doubt that industry guidelines and standards will speed successful design deployments while maintaining an appropriate eye on the cost of the EDC resources. We will highlight the major differences between a traditional data center and an edge data center.

Topics to be discussed include the challenges of a highly distributed geographical edge data center architecture, automation, security, and operational risk management. Our view of the edge as an extension of the current core networks and data centers is addressed. In addition, this paper will review how open standards, multi-tenancy, and standards-based certification, among other factors, need to be translated to the edge in a manner that supports mass geographic distribution.

This paper focuses on Edge Data Center (EDC) infrastructure. The associated drivers of EDCs – edge computing, mobile edge computing, cloud computing, fog computing, for example – are beyond the scope of this paper however background information is included in Appendix A.

Executive Summary

Higher capacity, lower latency and reduced network costs are enabled by moving computing and data storage closer to the end-user. Centralized data centers tucked away many cities from the user will need to be augmented with Edge Data Centers located much closer, literally blocks away.

While there are many similarities between the Traditional Data Center and the Edge Data Center, there are many new considerations presented that require industry attention to develop best practices and standards. The table below summarizes many of the key differences to be addressed. Further explanations for each are provided within this paper.

Traditional Data Center vs. Edge Data Centers

Category	Traditional Data Center (DC)	Edge Data Center (EDC)	Comments
		Planning	
Location	Site selection is well planned and is strategically located	Anywhere, downtown to remote, not readily available, hazardous; latency may force a compromise	A large multi-story DC vs a EDC which could be a 1/2 rack in a cabinet
Building Types	Usually newly purpose-built DC facility with fully integrated systems & support; potentially multi-tenant facility	Many options: 1) modified cell tower shelter; 2) modified cabinet; 3) Drop and Plug Shelter; 4) Drop and Plug Cabinet; 5) IDF closet in a Building; 6) Co-located in a Central Office or Data Center; 7) In box on a light pole, etc.	There are many more possibilities for EDCs and most of these will require site preparation. Scale-out will favor standardized drop & plug designs.

Category	Traditional Data Center (DC)	Edge Data Center (EDC)	Comments
Natural Hazards	Addressed through normative references	Lightning suppression, seismic migration, others site-by-site basis	EDC will have many unique requirements driven by the need to support specific geo-location requirements.
Proximity to end-users or connected devices	Could be very far, even cities or counties away	Typically very close, determined by latency requirements	Latency is a key driver for EDC locations.
Application	General purposes; could support thousands of applications	Few; specific and latency sensitive applications	EDCs provide smaller cloud capacity to support varied applications.
Latency	Higher, due to distance from end-user	Specific latency requirements	RTT <20ms may require EDC. EDCs can address specific business needs vs more generic, centralized DCs.
Design			
Availability	Mission-critical data, applications, and services per ANSI/TIA-942 standards	Smaller scale, still mission-critical, with reduced time added to data, applications, and services. Modified ANSI/TIA-942 standards.	Critical systems, power, cooling, network capacities and others will need to be balanced against scale-out requirements.
Standard References	As defined in ANSI/TIA-942	Most ANSI/TIA-942 references will apply, but others will have to be added.	EDC extensions to ANSI/TIA-942 will deal with scale-out considerations.

Category	Traditional Data Center (DC)	Edge Data Center (EDC)	Comments
Redundancy	Redundancy levels are based on business requirements and ANSI/TIA-942 standards.	EDCs will balance redundancy with parallel system deployment.	Next closet EDC may provide redundancy. Network and power designs will need to follow a parallel approach.
Size	Varies by design, typically larger than 100s, or 100,000s sq. ft.	Smaller, perhaps a few racks	Think small for EDCs - For every 1 DC there could be 100s or even 1,000s EDCs
Physical Security	Physically secured, access is controlled; built-in and near impossible to move	Easy to deploy, anchored or bolted in place. May require civil works to secure site.	Most EDCs are expected to be unmanned. EDCs will require surveillance measures, anchoring, alarms, etc.
Power	Purpose-built systems, multi-grid fed power distribution and backup	Repurposed existing power systems; redundancy trade-off to site and business needs	EDCs may use AC or DC systems. Existing plant may be suitable. Site locations and size may preclude power redundancy.
Cooling	ASHRAE A1, trending to less restrictive A2 or A3 ranges with better PUE	Site specific requirements; mix of traditional mechanical and "free" cooling systems	Best practice PUE design will apply to EDCs.
Cabling Systems	Complicated, multi-room, multi-floor, multi-types of cabling	Smaller remote systems still require high availability due to higher operational support costs with remote/unmanned sites.	Cable redundancy to match overall site design. Asset management and monitoring to address operational cost and security concerns.

Category	Traditional Data Center (DC)	Edge Data Center (EDC)	Comments
Lighting Systems	Exterior and interior considerations are normally part of the plan.	Remote unmanned changes the game, ex. integrated sensors on lighting.	
Deployment			
Deployment	Lengthy processes may even take multiple years.	Much quicker, could even be standardized to reduce to delivery to weeks/months.	More EDCs need. Must be deployable in faster, scalable, and replicable method.
Testing	Defined testing requirements (i.e. electrical backup, others)	Could be factory tested.	Notable differences
Operation and Maintenance			
Operational Risk Management	Cost of failure is high.	Lower failure cost; risk to high availability services (ex. AV) overrides cost.	Availability differences; designed to suit the services supplied
Maintenance	Controlled and very processed	Will need to define new practices due to remote and large numbers.	Adding sensors and monitoring systems to EDC means lower OPEX, higher availability, and lower MTTR.
Automation	Dedicated automation applications in place today	Virtualization enables automation of operations	Automation will be a core design and operational consideration for EDCs.

Edge Computing

Driving the Need for Edge Data Centers

The TIA Edge Data Center Working Group defines edge computing as the delivery of computing solutions (applications and services) at the logical extremes of the network edge, closer to the end-user. Reducing the distance data must traverse provides much greater network bandwidth. For example, vast amounts of IoT or machine-to-machine data can be processed by edge computing and only the most relevant and valuable information that can justify a trip back to a centralized data center consumes those scarce network resources.

Defining the Edge

Edge computing provides services in a context of reduced network costs and application latency.

Centralized data centers or hyperscale cloud provider models deliver large-scale resources and gain advantage via economies of scale. Parties share capacity, critical infrastructure, applications, cloud services, staff, and peering in one location.

This model will be enhanced by adding EDCs to support application requirements that cannot be supported over long network links from centralized DCs. This will overcome barriers such as communication latency and long-distance transmissions costs and will make it possible to support IoT and many of the next generation low-latency technologies.
It is important to keep in mind that EDCs relocate the geographic extent of the network and computing infrastructure; it is complementary to the current computing design and network infrastructure. The edge and the core form a holistic infrastructure where the needs and requirements of applications including latency, performance, security and cost determine the location of the resources supporting them.

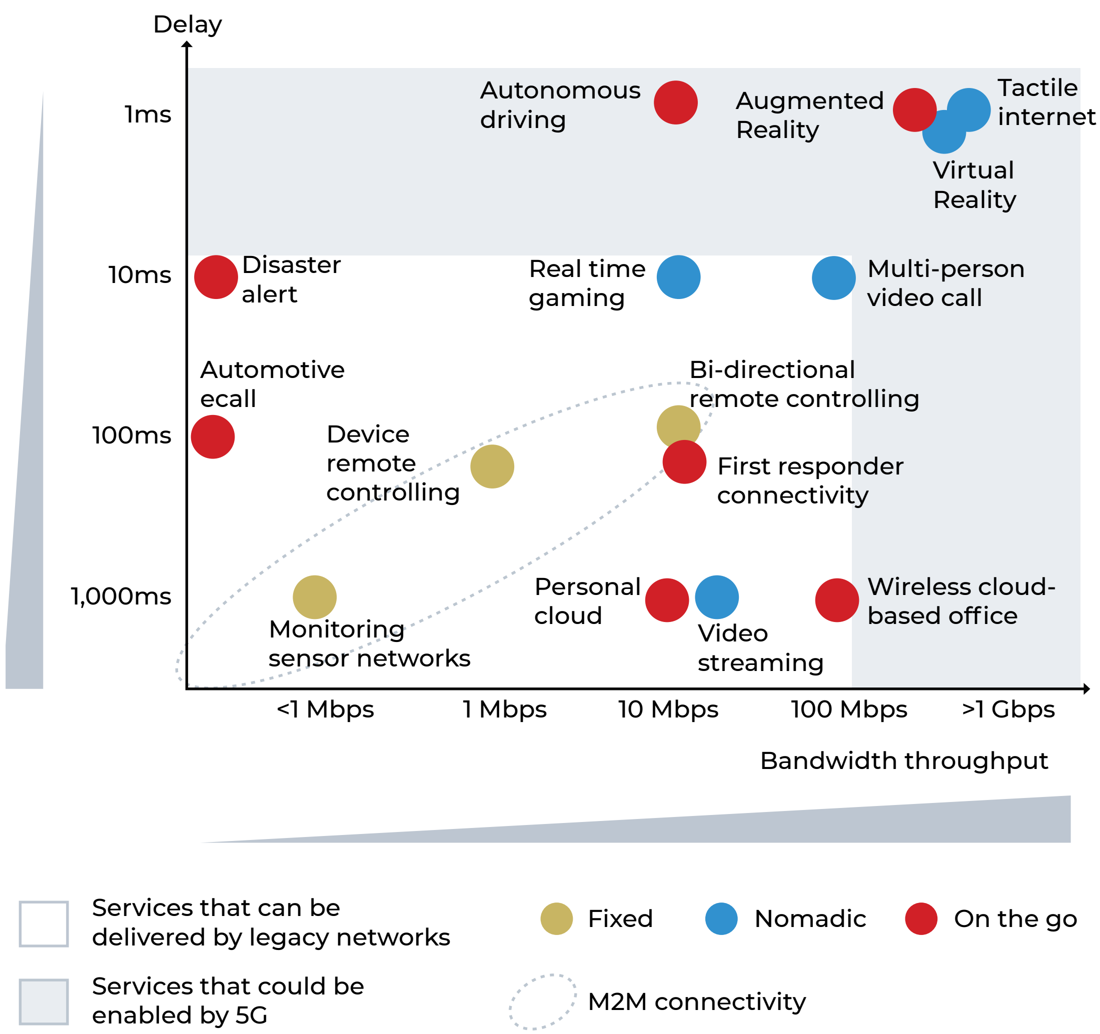
What is an Edge Data Center?

EDCs share many attributes with larger, more traditional DC facilities however they are designed to support widely distributed (often cloud based) services. The location of EDCs is driven by different business cases – application latency, network capacity/cost are common drivers – otherwise most services would simply follow the trend and end up in centralized facilities. EDC design will require a new balance between redundancy and availability. Operating many distributed EDCs also implies a potentially large impact in overall energy consumption. TIA standards will provide guidance to responsibly implement EDCs at scale.

The Importance of Latency

Today, many modern network functionalities, such as content streaming and interactive entertainment, are already constricted by latency more than bandwidth. Next generation applications, such as connected cars, AR/VR and drone technology require even higher constraints on application latency.

High-reliability, low-latency networks rely on the ability of the network to deliver consistent low latency and high capacity. Many applications, such as self-driving cars, medical technology, drones and various public safety systems, rely on guaranteed response times. Augmented reality and virtual reality (AR/VR) also require low latency to provide an acceptable user experience.



Various Application Latency Response Times and Those Further Enabled by 5G—Source; GSMA [3]

Application latency increases as the network distances increase. Long-distance transmissions inherently have too much latency to support many of the emerging applications. The way to solve this problem is to move the data center closer to the end-user, thus Edge Data Centers.
The Importance of Edge Computing to 5G Telecommunications Infrastructure
According to the Ericsson Mobility Report 2018 [4], 5G (fifth-generation mobile network) will kick off with enhanced mobile broadband as its first use case. By the end of 2023, they predict there will be 1 billion 5G subscriptions, accounting for approximately 20 percent of mobile data traffic.

With 3GPP 5G Standards [5] completed in June 2018, operators around the world have publicly announced that they will begin providing 5G services between late 2018 and mid-2019. The first 3GPP smartphones supporting 5G are expected in early 2019.
5G will have a service-based architecture, meaning that network functions can be delivered as service components. The 5G network has evolved far beyond the traditional mobile broadband of networks past.

5G also represents a material refactoring of the underlying design and physical hardware for the telecommunications radio architecture and network. Proprietary radio network hardware now shifts onto white-box x86 or ARM-based servers. This important change means that 5G services will be implemented as Edge Computing on cloud-based hardware in EDCs. In this context, 5G will become a major demand driver for the distributed data centers since 5G propagation is reduced compared to 4G LTE and will require many more antennas and more EDCs to support the growth and scale of 5G deployments.

5G mobile networks will use EDCs (as Mobile Edge Computing) to provide efficient local data services (i.e. local content caching or shunting internet traffic that reduces backhaul traffic and core network loads). EDCs may also help redirect edge traffic away from the carrier networks to local public internet networks (local bypass), making better use of scarce carrier network resources.

Unequivocally, 5G represents the underlying technological roadmap underpinning the deployment of a distributed and locally integrated mobile network, as well as representing a first-generation edge application.

Implications and Outcomes

As exemplified, computing at the edge represents a significant opportunity for innovation while generating important new requirements for critical infrastructure services (such as provided by Edge Data Centers) to enable edge applications. There are three principal criteria that the TIA Edge Data Center Working Group considers essential to the evolution of the edge infrastructure services:

Distributed Scale

Latency drives the need for mass distribution of application and services which, in turn, drives the need for a new generation of highly distributed data center services. This implication likely necessitates a move from 1,000s to 10,000s or even 100,000s of geographic locations in which computing, application and data services will be required. The density requirement will be driven by the latency and bandwidth constraints of edge applications, their functions, and design.

New Data Flow

Data and flows of data will shift from principally download (downlink) to bi-directional (uplink to the core and downlink to the edge). The network capacity must be designed with new distributed compute and storage solutions. Network capacity/topology will change to enable an equal bi-directional information flow given the large amount of data created at the edge of the network which will generate information that will in turn will be distributed to other nodes or uplinked to core facilities.

Increased Criticality

As computing and data facilities are geographically dispersed, significant questions concerning physical and logical security must be answered. The geographic scale and smaller form factors have not been fully considered in the design standards used today. Critical edge applications may impact life safety, for example, and require end-to-end “dial tone” level resiliency and deterministic performance. Higher investments are required to support the operation and management of these critical services; however this might be prohibitively costly for other edge applications. The concept of availability is a complex subject that new EDC standards will need to address.

Edge Data Center Considerations

Critical infrastructure services play a foundational role in the development and delivery of any data center. Centralized data centers may exceed 100,000 square feet. In contrast, EDCs support a wide variety of distributed services with appropriate availability and investment costs. EDCs require new design and operating procedures.

In this section, the unique issues of the edge and the important design differences between an EDC and a traditional data center are explored.

Location

Latency requirements reduce the effective coverage area on an EDC. As a result, many more EDC sites will have to be selected. Important questions remain regarding the future of EDCs and how the historical “scale up” design of current data centers can adapt to the “scale out” needs of the edge.

Traditional data center sites are selected in large part based on geographical considerations, including power availability and cost, customer demand, real-estate price, and physical and security risk factors. However, for the edge data center, location will also be dictated by the latency requirements of the service it supports and access to network resources.

Edge Data Center Location Examples include:

- A cabinet located on a street corner or near other utility equipment
- Re-utilization of existing macro cell site shelters
- Placement of new shelters / cabinets at macro-cell sites
- Co-location within Central Offices
- In buildings / Smart Buildings
- In factories to support Industrial IoT
- Behind or co-located with gas stations, drugstores and other businesses
- Integrated with community / neighborhood mailbox

As a result, EDC site selection will need to be adapted to include locations with less than ideal climate profiles, population density (urban), and macro operational risk factors, including proximity to flood zones, fault lines, and airports. EDC location criteria must also satisfy security and availability requirements through new and innovative network and facility designs.

Physical and Logical Security

Protection of computing equipment and information from physical and logical threats is one of the most important technical challenges we face today. Deploying many EDCs will increase data vulnerability due to the larger number of access points and risk vectors. It is imperative that security be integrated into the design of the EDC to control unnecessary exposure to risk. Given the assumption of principally unmanned data centers supporting the scale of the edge, it will be important for the design of these facilities to consider identity and authorization as an essential basis of control.

Operational Risk Factors and Considerations

Availability and MTTR are critical factors in EDCs. Given that EDC sites are remote and unmanned, strategies to minimize service outages and excessive operating expense are important design criteria.

Developing a data center design that is fit for purpose has been well documented by TIA through the ANSI/TIA-942 standard [6]. EDCs follow much of the same best practices, but additional factors must be considered to ensure availability and performance consistency.

These factors include, but are not limited to:

- Edge data centers and supporting applications may operate in a parallel fashion allowing for lower cost and resiliency at each individual site while maintaining availability requirements.
- Hazardous location may necessitate additional risk mitigation to ensure availability requirements.
- Advanced monitoring and automation to meet MTTR and operational cost SLAs.

Critical Infrastructure Considerations

The critical infrastructure in any data center includes cabling, power, cooling, and the enclosure as defined by architectural and structural requirements. The well-established and proven rating systems, such as the ANSI/TIA-942 rating system [6], provide standard reference designs and applicable availability ratings. EDCs that operate in parallel, with equal opportunity to provide the same services mitigate the failure of a single node, provided that the service hand-off is similarly automated and reliable.

The table below outlines some of the differences, by subsystem, between traditional and edge data centers:

Traditional and Edge Data Center Comparison

	Traditional Data Center (DC)	Edge Data Center (EDC)
Power	» Purpose-built utility / site electrical feeds » Multiple grid feed for redundancy » Diesel generators provide long-term energy source during utility outage	» May leverage existing power infrastructure repurposed for edge computing » May lack redundant power grid feed » Diesel generator backup may not be feasible due to space, noise, or pollution restrictions
Cooling	» Purpose-built » Cooling capability and redundancy built by design	» May leverage existing power infrastructure repurposed for edge computing » Cooling design options may be limited due to the size or location of data center
Connectivity	» Redundant connectivity is typical » Design for performance	» Redundant connectivity is desirable but depends on applications and site restrictions » Design for low latency
Facility	» Purpose-built facility or room » Dedicated or MTDC » Flexibility in site location, to reduce hazards » Typically manned, with varying levels of monitoring and automation	» Purpose-built facility, leverage existing facility, or purpose-built enclosure » Dedicated or MTDC » Latency and location of compute needs drive EDC location » More likely to be unmanned, with remote monitoring and automation

Environmental Operating Considerations

Traditional data center operators tightly control the data center operating environment. An EDC might be small, unmanned, and located in the desert among oil pipes – obviously challenged to provide the significantly controlled, monitored and restricted operating environment of the traditional data center. Modern connectivity, storage, and compute hardware are routinely able to operate in a relaxed ASHRAE A3 environment [7]. Traditional DCs often operate with overly restrictive environmental A1 limits. Relaxing the operating environment restrictions greatly reduces the requirement for mechanical cooling, lowering PUE and power requirements for cooling.

Use of alternate "free" cooling systems should be considered where practical. Minimizing the impact on EDCs is an important design objective considering the large anticipated number of these facilities.

Cabling Systems

If present, EDC network cabling systems are also critical to service availability. As EDCs may be remote and numerous consistent design and operating procedures enhance service agility and MTTR. Automated Infrastructure Management (AIM) can be used to track connectivity and inventory status of this important foundational resource.

Automation, Monitoring and Support

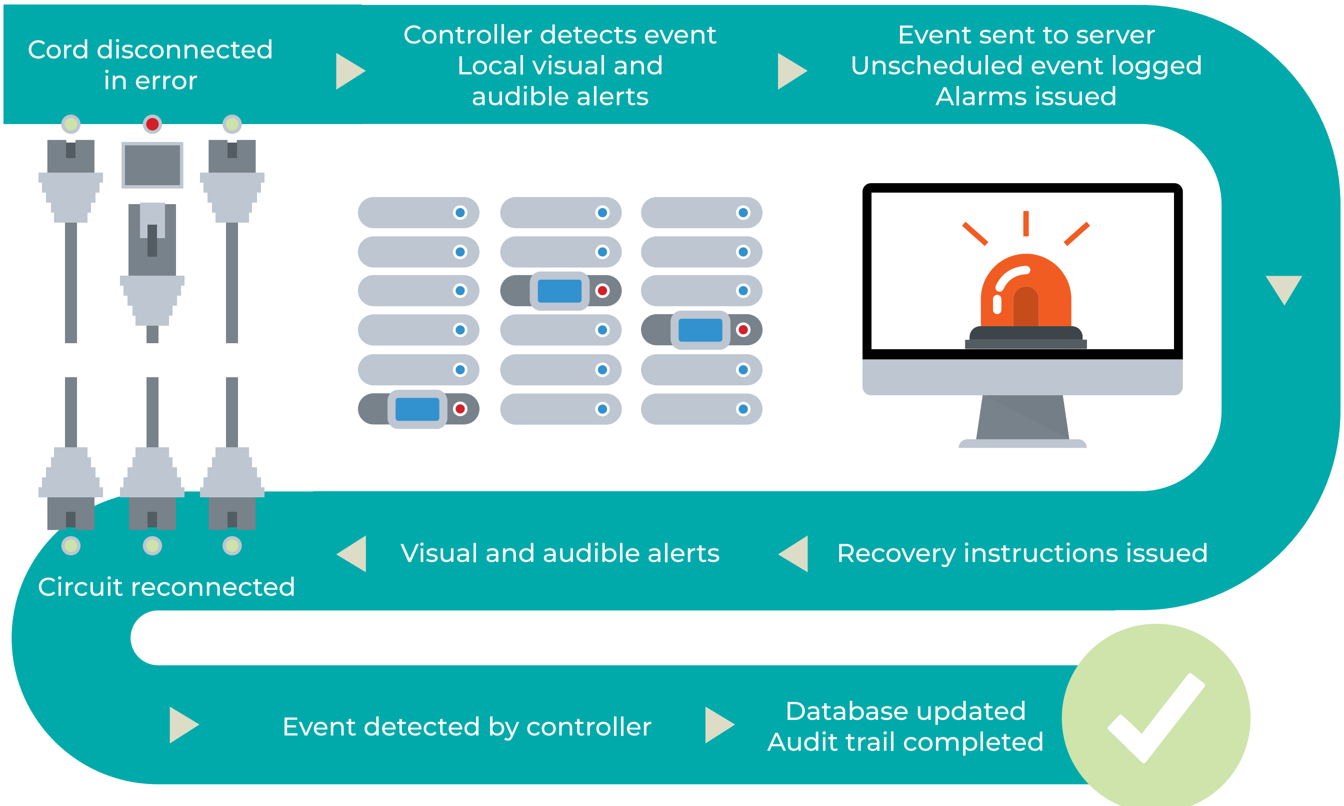
To meet the scale and needs of the edge, investments in integrated monitoring will be an integral component for managing operational risk.

Historically, most organizations have viewed facilities monitoring as independent from the monitoring of applications and computing.

Considering the anticipated scale of the edge, with remote locations supporting performance-sensitive applications, monitoring at the edge will become more holistic, spanning both the facilities and the classic application domains to comprehensively understand and maintain the expected SLAs.

One anticipated pain point for EDC management is the fiber and copper cabling infrastructure, which historically has often been unmanaged. Standards-based Automated Infrastructure Management (AIM) systems add coverage to physical networks in the EDC. AIM is specified in the ISO/IEC 18598 standard [8]. The standard defines AIM as an "integrated hardware and software system that automatically detects the insertion or removal of cords,

[and] documents the cabling (connectivity) infrastructure including connected equipment enabling management of the infrastructure and data exchange with other systems."



AIM Overall System—Source; CommScope [9]

As illustrated above, AIM systems enable EDC management personnel to monitor, manage and optimize the connected environment in real time, enhancing the ability to:

- Plan and execute changes to the connectivity.
- Troubleshoot connectivity issues in real time.
- Discover and track the location of connected devices.
- Manage and monitor capacity and asset information.

Intrinsic benefits on the connectivity infrastructure are enabled by functionality within the AIM system and include:

- Accurate and automatic documentation to replace error-plagued manual tracking.
- Change management to help reduce the cost of moves, adds and changes.
- Incident management that can decrease downtime and mean-time-to-resolution.
- Capacity management that enables higher port utilization and improves planning.

Conclusions on Edge Data Centers

While we are still early in the adoption phase of edge computing and have much to learn, our research has led to the following three conclusions;

- Increased scale and scope coupled with reduced latency dependent, broadly distributed critical services will require many EDCs
- Automation, parallel architecture and standards will be needed to meet availability requirements
- Energy efficient design and operations must be a high priority especially in view of the anticipated scale of EDC deployments

The ICT industry should expect to see massive deployment of EDCs in the foreseeable future. The Edge does not stand alone but is connected forming an interoperable fabric extending from device to core to cloud. The importance of connectivity and the bridging of wireless and wired networks in a distributed, consolidated manner, will be an essential value for enabling ICT service provider and enterprise data center managers to drive further innovation at the edge.

TIA's Edge Data Center Working Group is continuing its work to establish standards that will guide edge data center design and development and allow certification for interoperability.

Acronyms and Abbreviations

5G	Fifth-generation mobile network
AIM	Automated Infrastructure Management
AR	Augmented Reality
AV	Automated Vehicle
CO	Central Office
DC	Data Center
EDC	Edge Data Center
ICT	Information and Communications Technology
IIoT	Industrial Internet of Things
IoT	Internet of Things
ITS	Information Technology Services
M2M	Machine-to-Machine
MTDC	Modified Total Direct Cost
MTTR	Mean Time to Repair
PUE	Power Usage Effectiveness
RTT	Round Trip Time
SLA	Service Level Agreement
TIA	Telecommunications Industry Association
VR	Virtual Reality

References

1. Gartner Research, Forecast: Internet of Things – Endpoints and Associated Services, 2017

2. The Future X Network: A Bell Labs Perspective by Marcus K. Weldon (2016-03-01), CRC Press

3. GSMA, Unlocking Commercial Opportunities, From 4G Evolution to 5G

4. Ericsson Mobility Report, 2018

5. 3GPP Technical Specifications Group, Release 15

6. ANSI/TIA-942 – Telecommunications Infrastructure Standard for Data Centers

7. ASHRAE: American Society of Heating, Refrigerating, and Air Conditioning Engineers

8. ISO/IEC 18598 – Automated infrastructure management (AIM) systems – Requirements, data exchange and applications

9. "Ready, Set, AIM! - The Automated Infrastructure Management Standard is Now Published", CommScope 2016

Appendix A: Edge Data Centers Glossary

Note: TIA collaborates with the Linux Foundation in maintaining the Open Glossary of Edge Computing. Terms included in this paper are listed below. For the full list of Open Edge Compute Glossary terms refer to the official repository: **https://github.com/State-of-the-Edge/glossary**.

3G, 4G, 5G

3rd, 4th, and 5th generation cellular technologies, respectively. In simple terms, 3G represents the introduction of the smartphone along with their mobile web browsers; 4G, the current generation cellular technology, delivers true broadband internet access to mobile devices; the coming 5G cellular technologies will deliver massive bandwidth and reduced latency to cellular systems, supporting a range of devices from smartphones to autonomous vehicles and large-scale IoT. Edge computing at the infrastructure edge is considered a key building block for 5G.

Automated Infrastructure Management (AIM)

An integrated hardware and software system that automatically detects the insertion or removal of cords, [and] documents the cabling (connectivity) infrastructure including connected equipment enabling management of the infrastructure and data exchange with other systems. (From ISO/IEC 18598)"

Autonomous Vehicle

A vehicle capable of navigating roadways and interpreting traffic-control devices without a driver actively operating any of the vehicle's control systems. In the case of the autonomous vehicle, the contextual edge becomes an integrated component of in-vehicle computing and applications that provide autonomous and extreme low-latency processing. These on-vehicle capabilities further integrate into a tiered ecosystem with a tradeoff between higher capacity and aggregation versus lower absolute latency."

Centralized Data Center

A large, often hyperscale physical structure and logical entity which houses large compute, data storage and network resources which are typically used by many tenants concurrently due to their scale. Located a significant geographical distance from the majority of their users and often used for cloud computing.

Cloud Computing

A system to provide on-demand access to a shared pool of computing resources, including network servers, storage, and computation services. Typically, utilizes a small number of large centralized data centers and regional data centers today.

Co-Location

The process of deploying compute, data storage and network infrastructure owned or operated by different parties in the same physical location, such as within the same physical structure. Distinct from Shared Infrastructure as co-location does not require infrastructure such as an edge data center to have multiple tenants or users. ADD: Also distinct from multi-tenancy as co-location doesn't involve sharing of instanced software.

Data Center

A purpose-designed structure that is intended to house multiple high-performance compute and data storage nodes such that a large amount of compute, data storage and network resources are present at a single location. This often entails specialized rack and enclosure systems, purpose-built flooring, as well as suitable heating, cooling, ventilation, security, fire suppression and power delivery systems. May also refer to a compute and data storage node in some contexts. Varies in scale between a centralized data center, regional data center and edge data center.

Decentralized, Distributed Architecture OR Distributed Edge Cloud Architecture

Distribution of computing power and equipment at the edge of the network spread across different physical locations to provide closer proximity to the use cases of said compute. Components are presented on different platforms and several components can cooperate with one another over a communication network in order to achieve a specific objective or goal. A distributed system is a system whose components are located on different networked computers, which then communicate and coordinate their actions by passing messages to one other."

Edge Computing

The delivery of computing capabilities to the logical extremes of a network in order to improve the performance, operating cost and reliability of applications and services. By shortening the distance between devices and the cloud resources that serve them, and also reducing network hops, edge computing mitigates the latency and bandwidth constraints of today's Internet, ushering in new classes of applications. In practical terms, this means distributing new resources and software stacks along the path between today's centralized data centers and the increasingly large number of devices in the field, concentrated, in particular, but not exclusively, in close proximity to the last mile network, on both the infrastructure and device sides.

Edge Data Center

A data center which is capable of being deployed as close as possible to the edge of the network, in comparison to traditional centralized data centers. Capable of performing the same functions as centralized data centers although at smaller scale individually. Because of the unique constraints created by highly-distributed physical locations, edge data centers often adopt autonomic operation, multi-tenancy, distributed and local resiliency and open standards. Edge refers to the location at which these data centers are typically deployed. Their scale can be defined as micro, ranging from 50 to 150 kW of capacity. Multiple edge data centers may interconnect to provide capacity enhancement, failure mitigation and workload migration within the local area, operating as a virtual data center.

Edge Cloud

Cloud-like capabilities located at the infrastructure edge, including from the user perspective access to elastically-allocated compute, data storage and network resources. Often operated as a seamless extension of a centralized public or private cloud, constructed from micro data centers deployed at the infrastructure edge.

Fog Computing

A distributed computing concept where compute and data storage resource, as well as applications and their data, are positioned in the most optimal place between the user and Cloud with the goal of improving performance and redundancy. Fog computing workloads may be run across the gradient of compute and data storage resource from Cloud to the infrastructure edge. The term fog computing was originally coined by Cisco. Can utilize centralized, regional and edge data centers.

High Availability

The ability of a system to maintain above average levels of uptime through design and resiliency characteristics. At the edge high availability is even more important due to the unique requirements of the use cases and the compute support necessary. Due to the distributed nature of edge data systems, they will be able to provide the high availability required."

Internet of Things (IoT)

A large increase in machine-to-machine communication will produce large increase in bandwidth requirements. Machine response times are many times faster that human interactions creating even more opportunities to benefit from ultra-low latency communication. Data traffic patterns will change with local peer-to-peer relationships and localized IoT data/control systems generating data that does not need to traverse backhaul links that will be hard pressed to carry the flood of traffic to come. Distributed EDCs will transform large volumes of raw data into valuable information that can then be transferred to more centralized tiers where required.

Latency

In the context of network data transmission, the time taken by a unit of data (typically a frame or packet) to travel from its originating device to its intended destination. Measured in terms of milliseconds at single or repeated points in time between two or more endpoints. A key metric of optimizing the modern application user experience. Distinct from jitter which refers to the variation of latency over time. Sometimes expressed as Round-Trip Time (RTT).

Latency Critical Application

An application that will fail to function or will function destructively if latency exceeds certain thresholds. Latency critical applications are typically responsible for real-time tasks such as supporting an autonomous vehicle or controlling a machine-to-machine process. Unlike Latency Sensitive Applications, exceeding latency requirements will often result in application failure.

Security Risk Vectors

By their nature, edge data centers have significantly more security risk vectors than traditional data centers, as the number of edge data centers maintained increases the number of attack points significantly more than increasing the size of a traditional data center. In order to mitigate this edge data center operators will have to consider the specific defenses they will implement.

Shared Infrastructure

The use of a single piece of compute, data storage and network resources by multiple parties, for example two organizations each using half of a single edge data center, unlike co-location where each party possesses their own infrastructure.

Unmanned Edge Data Centers

Due to their size and distributed nature, most edge data centers are expected to be unmanned, creating new security (surveillance, anchoring, alarms, etc.), maintenance (additional monitoring systems and sensors), operational (increased automation), and resiliency considerations that do not exist with manned data centers.

Telecommunications Industry Association

www.TIAonline.org