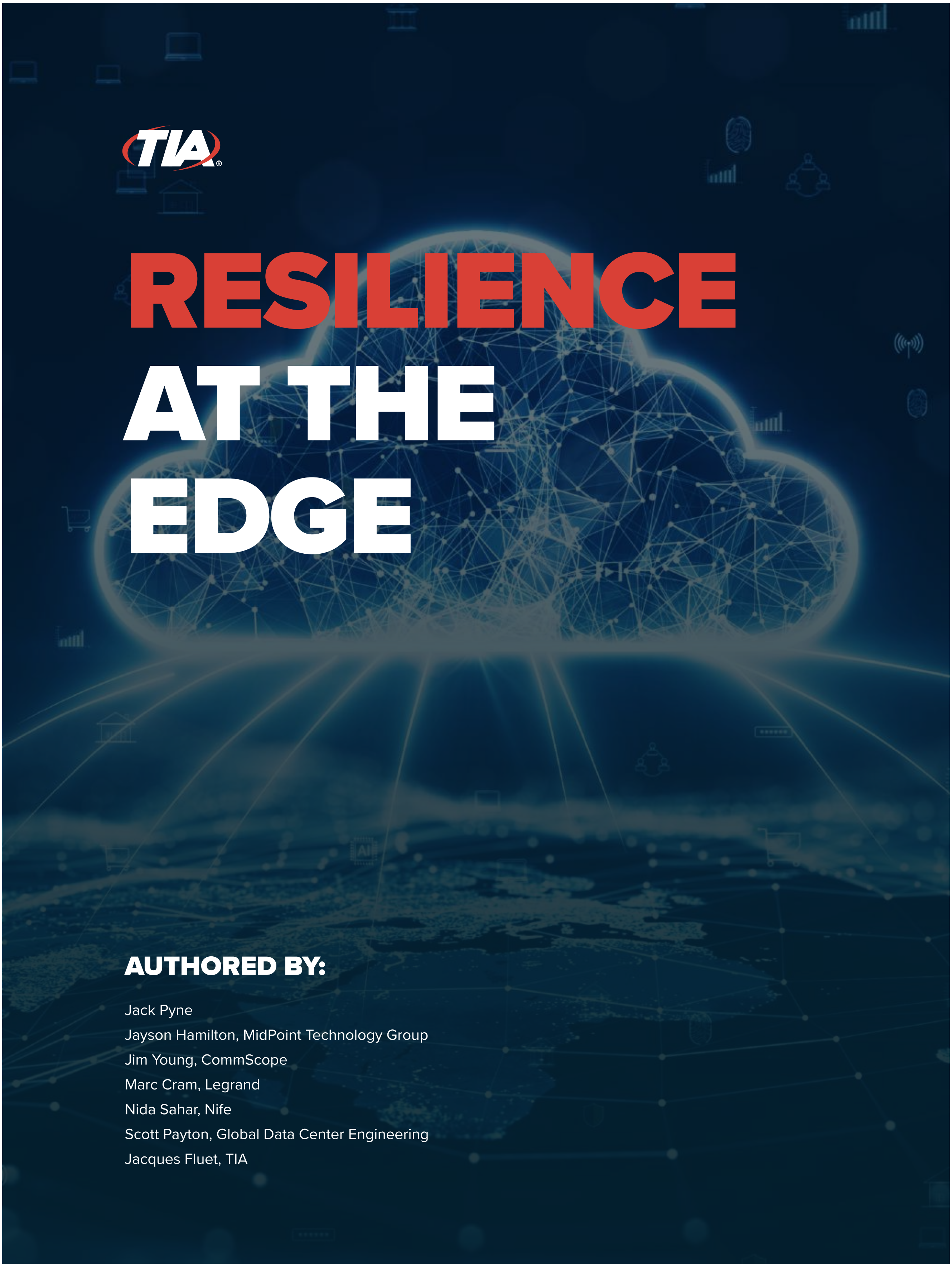


RESILIENCE AT THE EDGE

AUTHORED BY:

Jack Pyne
Jayson Hamilton, MidPoint Technology Group
Jim Young, CommScope
Marc Cram, Legrand
Nida Sahar, Nife
Scott Payton, Global Data Center Engineering
Jacques Fluet, TIA

INTRODUCTION

When planning an Edge Data Center (EDC) deployment, the need to address availability of planned workloads to ensure resiliency is a top priority. The key characteristics of an EDC can be significantly different than those that top the list for larger enterprise or multi-tenant data centers. To help balance the costs and operational aspects of any EDC strategy, this paper outlines a number of critical questions that should be addressed to ensure resiliency at the edge.

HOW AND WHY EDC RESILIENCY IS DIFFERENT FROM TRADITIONAL DC?

The traditional approach to ensuring resiliency is to rely on redundant systems running in parallel that provide ongoing availability to maintain reliable delivery of critical workloads. In establishing edge compute installations and applications, the data center designer will need to know the nature and criticality of the workloads to be performed and determine whether resiliency can be achieved through software that enables the workload to be relocated and reestablished quickly enough in the event of a failure, or if the local infrastructure itself must be resilient. If resiliency through software is possible, it must be determined if there is available capacity in a nearby facility and how quickly that facility can recover the workloads. If redundant infrastructure is needed, both Capex and Opex must be optimized.

Many EDC workloads are reliant on lower latency. Moving workloads closer to users to reduce latency often means that traditional disaster recovery and backup strategies may not meet the transmission latency that applications require. It is therefore important to locate mitigation strategies within distance-limited transmission delay windows.

Service-level agreements (SLAs) are also becoming increasingly complicated as multiple types of services require various levels of performance. Basic telecommunications services typically target 99.999% (i.e., five 9's) availability. Emerging EDC services that enable automated factories, remote healthcare, and intelligent transport systems require even higher availability, but EDCs can also have components able to function properly at much lower levels of availability.

WHAT ARE THE KEY DRIVERS TO EDC RESILIENCY?

AVAILABILITY REQUIREMENTS OF THE SYSTEMS

Given the potentially substantial number of sites and large geographic distribution of EDCs, a standards-based infrastructure for ICT and critical systems, with plug-and-play components compatible with orchestration and management tools is needed.

CRITICAL NATURE OF THE DATA

The criticality of a data set and the need for that data to be stored, archived, or backed up properly depends on the primary purpose of the EDC applications. Applications may be mission critical, revenue critical, or business critical and subject to stringent back up policies based on business continuity requirements or compliance with local and federal regulations. The value of data is often categorized as historical (i.e., research) or immediate (i.e., required for decisions surrounding critical functions).

When determining the value of historical data, the cost of both the uphaul and storage must be considered. The requirements for immediate data must be evaluated in reference to the location of the decisions being made based on that data. The closer to data generation that analysis can occur, the lower the associated uphaul and storage costs. Effort should therefore be made to move the analysis of data closer to the event or application requiring that analysis. While the edge compute model is decoupled from traditional cloud networks, the need for elevated levels of redundancy for data protection to ensure business continuity remains.

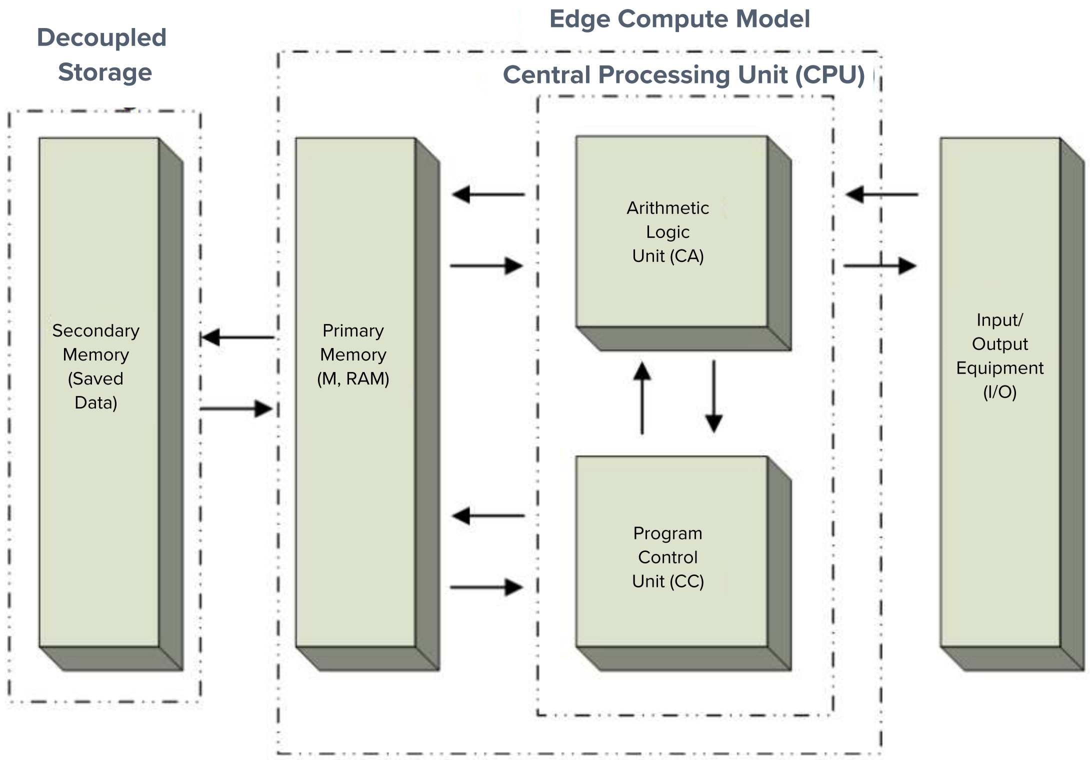

Decoupled Storage
Edge Compute Model
Central Processing Unit (CPU)
Secondary Memory (Saved Data)
Primary Memory (M, RAM)
Arithmetic Logic Unit (CA)
Program Control Unit (CC)
Input/ Output Equipment (I/O)

Figure 1: Storage device on the Edge

It is important to understand that even though the edge compute model moves processing closer to the edge, it still resembles traditional compute models. EDCs are therefore bound by the input and outputs of common storage equipment and the limitations of storage architectures like file, block, and object level when looking to compile data for later use or analysis. Figure 1 demonstrates the edge model with primary and secondary memory as outputs from the central processing unit (CPU).

While primary data runs “in-memory” to enable the localized processing needed to deliver real-time data, the secondary memory is referred to as the “saved data” or any other data outputs to an external device or disk. While secondary data functions might still route back to traditional data centers on the inner edge, the primary data set needs to operate within the far edge’s proximity parameters. This enables the delivery of real-time local data processing to achieve the low application latency required. Ultimately, the criticality of the data set determines how data is stored and protected in the case of an event, even if the intent of the application is to compile data for later use.

LATENCY

To achieve SLAs and meet network performance objectives, determining where and when data is located requires fast, strategic decisions. Table 1 shows typical values of delays as data is transmitted and processed in a networked environment. To meet 5G latency targets, edge computing and first level non-volatile storage will need to be within 50 km of the point of origination or consumption. In many instances, it is the end device or application that decides where best to store the data.

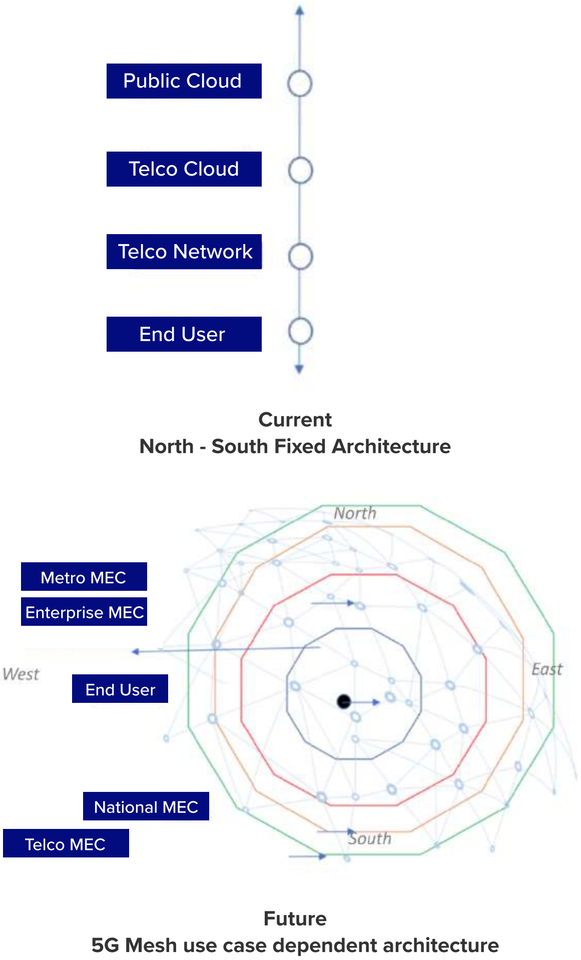
ACTION	AVERAGE LATENCY
3 GHz CPU 1-clock cycle	0.3 ns
Level 1 cache access	0.5 ns
Level 2 cache access	7 ns
Main Memory Reference	100 ns
SSD	150 µsec
HD	10 ms
Ethernet Switch	5-125 µsec
Optical Fiber	1 ms / 200 km

Table 1: Latency of data transaction
https://www.softwareyoga.com/latency-numbers-everyone-should-know/

End-to-end latency, including switching, routing, and processing times, will influence the decisions of the network designer and later the network operator. Numerous nanoseconds that already have cycled through common components processing inside the computer model are often overlooked. Therefore, precision time stamping is required to effectively measure and deliver low-latency applications.

REDUCING NETWORK BACKHAUL/TRAFFIC

One of the key advantages of moving the computing closer to the user is the reduction in backhaul traffic. Edge computing allows for the data to travel with fewer hops, reducing the backhaul link capacity that would have been otherwise required (See Figure 2). Only certain data sets will need to be sent to the cloud over backhaul links.



Current
North - South Fixed Architecture

Future
5G Mesh use case dependent architecture

Figure 2: Centralized vs. Mesh architecture

WHAT TYPE OF RESILIENCY SOLUTIONS ARE AVAILABLE?

REDUNDANT COMPONENTS

Two redundancy approaches are possible—local redundancy and distributed redundancy. The classic approach to resilience for enterprise data centers is to have redundant data centers operating in highly diverse locations (e.g., East-West configuration) that are always running applications in parallel.

5G is built on the concept of abstracting the underlying network and compute hardware from the operating systems. As a means of implementing 5G, wireless network operators will rely on edge data centers having a minimum of available redundant hardware in another location that is capable of taking over a workload in the event a single system fails. Power, cooling, and network infrastructure are likely to have a degree of fault tolerance that will be less than the traditional "five 9's" approach. Instead, these systems will be designed to provide a minimum ride-through time (usually in minutes) until a workload and traffic can be diverted to another nearby edge location. This resembles a slightly modified version of the traditional approach to resilience.

In the event the workloads of a particular edge data center are deemed to be mission critical, relying on nearby infrastructure to pick up the slack in the event of a failure is not an option. In this instance, the edge data center owner/operator is likely to implement onsite power generation systems, battery backup/UPS systems, redundant HVAC systems, and hardened security/access controls.

MESH NETWORKS

Backup at the edge must be local to meet latency requirements (i.e., data flight time consideration). East-West connectivity is therefore needed at the edge between nodes that are working in tandem to provide availability.

It's important to ensure that the underlying outside plant (OSP) infrastructure can accommodate expanded North-South and East-West routes with the ability to add/drop from high-count OSP fiber cables. Long-term life for OSP resources to the edge therefore require a different approach than standard resources with information technology equipment that has a shorter life span.

NETWORK ORCHESTRATION

The number of edge devices and end-user applications is growing. A report published by STL Partners predicts that the number of edge servers will grow from 7.5K in 2022 to 125K in 2025[1]. End-user location determines where the application needs to be deployed. While the applications can be initially deployed at an EDC close to the end user, the application may also need to be mobile as the end-user moves. This applies to several IoT, smartphone, and augmented/virtual reality (AR/VR) headset applications.

A network orchestrator synchronizes and ensures applications are mobile. It takes care of distribution as well as on-demand movement. It is a policy-driven approach to automate the hardware, software, and services required for the application. Software-defined networking (SDN) automates the provisioning, updating, and managing of the resources required to deliver an application via a set of application programming interfaces (APIs) (see Figure 3).

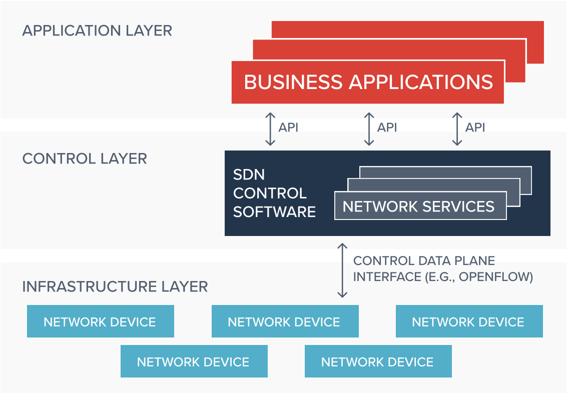


Figure 3: Service Layers

As infrastructure grows and cloud operations illustrate, private and public infrastructure orchestration is prevalent. This means that along with cloud, edge infrastructure is an extended cloud layer. And as the network and workload orchestration progresses, there will be fewer requirements for local redundancy as the overall service orchestration will take care of maintaining the level of performance and availability required for each type of service.

Application mobility is possible by moving applications from one region to another, following the end-user. Applications need to be smaller (microservices) and containerized to help manage, deploy, and scale across the numerous edge devices and locations. The central orchestrator ensures that there is a copy of the container to maintain redundancies. The processed data is moved to a central location.

Once the performance of an application in an edge location has degraded, another location takes over, which impacts the latency but not the uptime of the application itself. Customer SLAs are more stringent, and the backup site needs to meet the SLA requirements as well.

EDC design will have to include a new level of monitoring that can detect early performance degradation for all its sub-systems before failures occur. This data must be available for the service orchestration layer to make the necessary changes to maintain an optimal user experience. With the advances of software and artificial intelligence, real-time service orchestration will increasingly become autonomously managed.

[1] STL Partners, “Forecasting capacity of network edge computing”, https://stlpartners.com/research/forecasting-capacity-of-network-edge-computing/

CONCLUSION

EDCs are different from traditional data centers and have a variety of ways to implement the required resiliency. EDCs must adapt to the time-sensitive performance and availability requirements of their workloads. Through network awareness of both workload contents and criticality, a combination of local and regional resources across a network of EDCs will be able to deliver the computing and storage that meets latency, throughput, and SLA targets while minimizing backhaul traffic.

TIA

BRIEFING PAPER

THANK YOU

TIA DATA CENTER PROGRAM SPONSOR

CAPITOLINE

QUESTIONS? WANT TO GET INVOLVED IN TIA'S DATA CENTER PROGRAM?

EMAIL: DATACENTERINFO@TIAONLINE.ORG

Disclaimer: The information and views contained in this article are solely those of its authors and do not reflect the consensus opinion of TIA members. This article is for information purposes only and it is intended to generate opinion and feedback so that the authors and TIA members can learn, refine, and update this article over time. The Telecommunications Industry Association does not endorse or promote any product, service, company, or service provider. Photos, charts and products used as examples in this paper are solely for information purposes and do not constitute an endorsement by TIA.