Research Report:

Taming the AI-enabled Edge with HCI-based Cloud Architectures

Steve McDowell, Chief Analyst

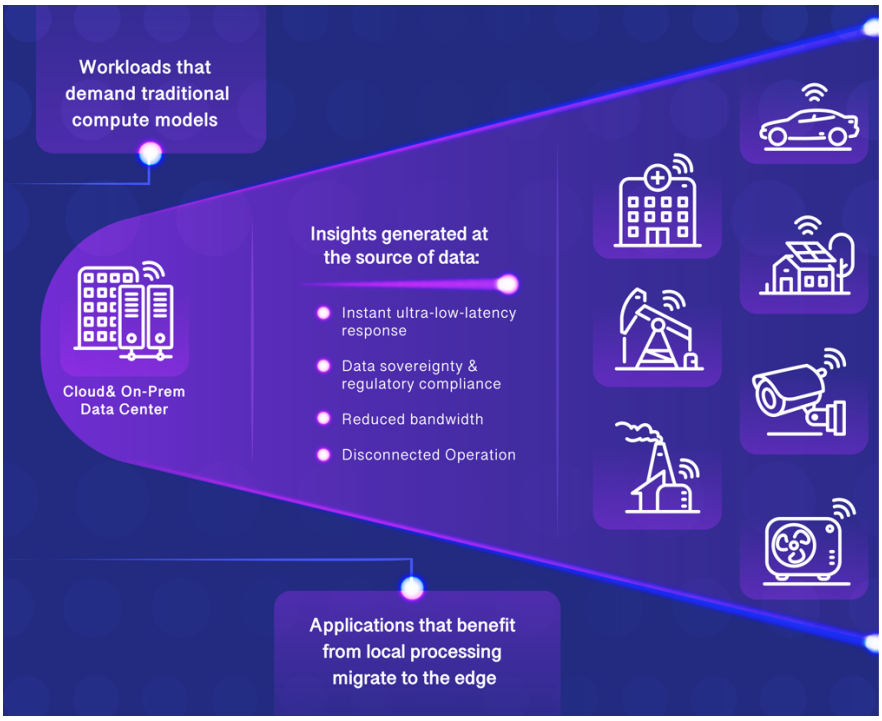
April 2024

This report was commissioned by Nutanix

TAMING THE AI-ENABLED EDGE WITH HCI-BASED CLOUD ARCHITECTURES

The edge computing market has seen a remarkable surge in recent years, a trend poised to carry forward robustly. At the heart of this movement is the escalating demand for low-latency processing and instantaneous data analysis. Industries such as retail, transportation, industrial automation, smart city management, and IoT heavily rely on this rapid-response computing.

Simultaneously, the explosion of IoT and connected devices across homes, businesses, and cities propels edge computing forward. These devices frequently necessitate immediate local data processing, bypassing the need to relay information to distant data centers or cloud servers.



Technological strides in processing capabilities, hardware miniaturization, and next-gen wireless communications like 5G are the technical pillars bolstering edge computing's efficacy and reach. These advancements enable more potent edge computing solutions that are also increasingly accessible.

It's no surprise, then, that the edge market is booming, estimated to grow from $208 billion in 2023 to $317 billion in 2026[1]. This economic potential is recognized across various sectors, with industries such as manufacturing, healthcare, retail, and telecommunications integrating edge computing to refine operational efficiency, enrich customer experiences, and bolster product and service quality.

However, the path to the edge has its challenges. Protocol standardization, security concerns, and the seamless integration of edge computing with existing IT infrastructures pose significant challenges.

WHAT IS THE EDGE? A BRIEF TAXONOMY

Different industries employ IT resources at the edge in other ways, from remote office/branch office (ROBO) environments to the industrial edge. The goal is the same: derive benefit by processing data closer to where it's generated.

Figure 1: Use of Edge Across Industries

[1] IDC Report, 2023 Worldwide Edge Spending Guide

The world of edge computing is necessarily diverse in order to meet the varying needs of different industries. There are numerous ways to describe nearly every aspect of deploying IT resources to the edge, from where the edge begins to how organizations deploy workloads.

The following provides a simple and high-level taxonomy of computing at the edge:

Location	Device Edge	Directly on devices like smartphones, IoT devices, or sensors
	On-Prem Edge	At the user's premises, such as in a factory, retail store, or private network.
	Network Edge	Located at network aggregation points, like cell towers, gateways, or other network nodes.
Type of Edge Device	Constrained Device	Limited in computational power and resources (e.g., sensors, smart bulbs).
	Smart Devices	More capable devices with processing abilities (e.g., smartphones, smart cameras).
	Edge Servers	High-performance servers located at the edge provide significant processing power.
Edge Architectural Layers	Infrastructure	Includes physical devices and network infrastructure.
	Platform	Consists of operating systems, virtualization technology, and management tools.
	Application	Comprises the applications and services running on edge devices or servers.
	Hybrid Cloud	Integration of edge & cloud

Edge Deployment Models	Multi-Access Edge Compute (MEC)	Edge computing in a mobile network environment
	Edge-to-Core	Edge interoperates with traditional data centers (on-premises or cloud) for configuration, backup and recovery, and to exchange data that may be better processed within the core infrastructure.

Understanding this taxonomy is crucial for designing, implementing, and managing edge computing systems. It highlights the diversity of edge computing technologies, architectures, and applications, each with its challenges and opportunities.

EDGE IN A HYBRID MULTI-CLOUD ENVIRONMENT

In many scenarios, edge computing and edge-to-core environments (which may include a hybrid cloud) deliver efficient, scalable, and responsive computing solutions. The edge handles real-time processing and immediate actions. At the same time, the core resources within a cloud or traditional data center environment provide capabilities best served with a hybrid edge-to-core architecture.

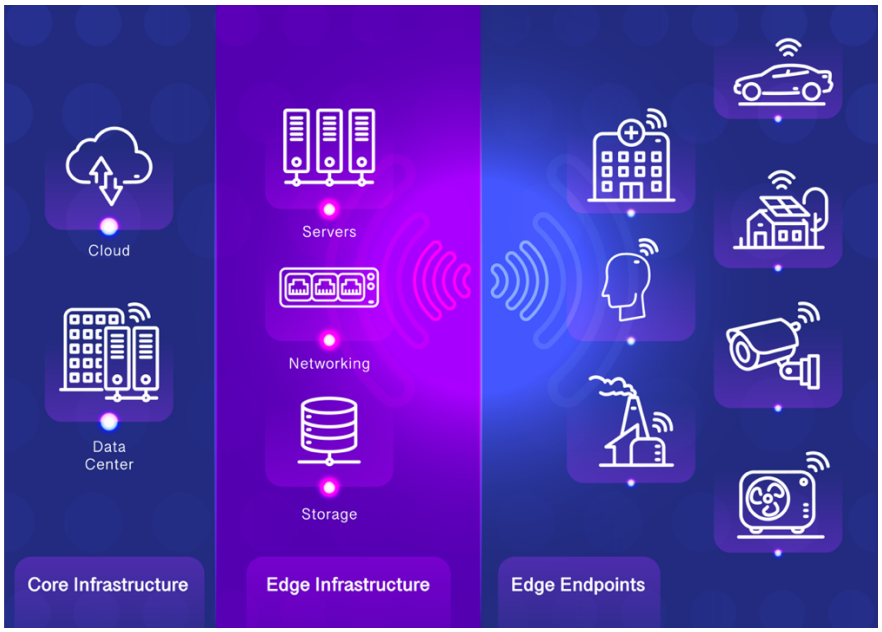


Figure 2: Edge in a Hybrid Cloud Environment

In an edge-to-core environment, data collected and initially processed at the edge is sent to the cloud or traditional on-prem data centers for further analysis, long-term storage, backup and recovery, reporting, compliance, additional computing power, and other capabilities best served in a hybrid edge-to-core environment.

Conversely, management capabilities within the core can send necessary configuration, application updates, remotely processed data, updated AI models, and other centrally controlled items back to the edge for immediate action. Combining edge with core, whether on-prem or cloud, is an efficient, cost-effective, and flexible way to ensure business continuity.

AI AT THE EDGE

Deploying new artificial intelligence (AI) capabilities to the edge disrupts traditional edge architectures. Edge AI enables powerful new applications that require real-time processing and decision-making, making the technology a crucial component in the evolution of edge computing.

Deploying AI at the edge yields several key advantages:

- **Timely Insights**: Edge AI can provide more contextual and timely insights by analyzing data in its immediate environment.
- **Reduced Latency**: By processing data locally, AI at the edge significantly reduces the time to analyze and act upon data. This is crucial for a broad range of real-time or near-real-time decision-making applications, such as autonomous vehicles, industrial automation, vision processing for retail and manufacturing, and even smart healthcare devices.
- **Enhanced Privacy & Security**: Processing data locally means sensitive information does not have to be constantly sent back and forth to a central server, reducing exposure to potential data breaches.
- **Lower Bandwidth Requirements**: Transmitting large volumes of data to the cloud, whether private or on-prem, can be bandwidth intensive. By processing data at the edge, only relevant or summarized data must be sent back to the core, reducing bandwidth requirements and associated costs.
- **Improved Reliability**: Edge AI can operate independently of central servers or cloud services, which is beneficial in environments with unstable network connections. This ensures continuous operation even in the event of network failures.
- **Scalability**: AI at the edge enables scalable solutions by distributing the processing load
- **Energy Efficiency**: Local data processing can be more energy-efficient, as it reduces the energy required for processes as diverse as inference and data transmission, enabling lower-power processing technologies to be deployed at the edge.

Deploying AI to the edge is a natural extension of today's AI workflows. In a true "core to edge" approach, AI-based workloads often begin with a set of foundational models, which are then fine-tuned on traditional servers living in the cloud or data center. These fine-tuned models are then deployed at the edge, where AI-driven insights are most impactful.

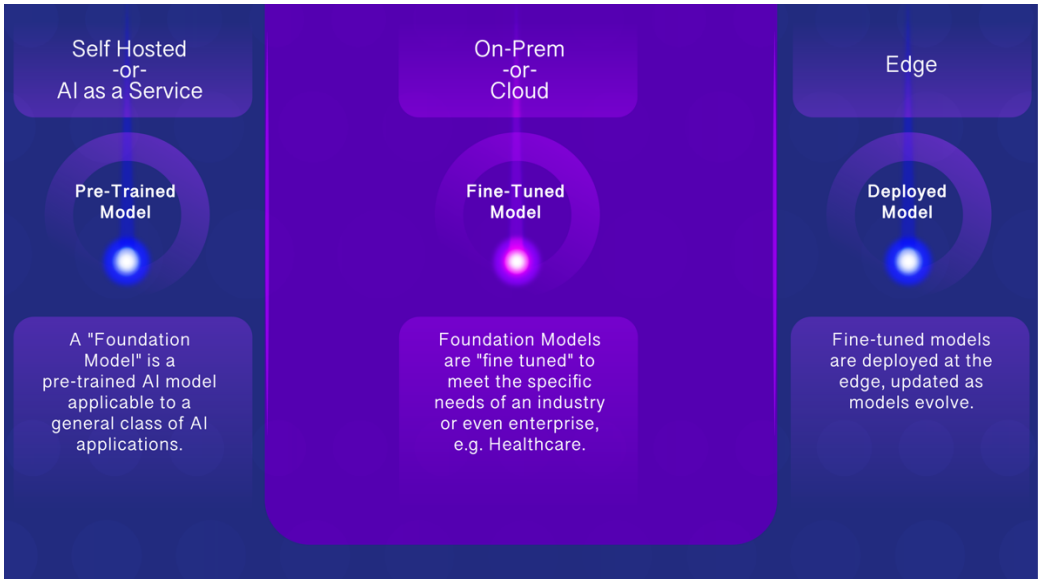


Figure 3: AI Deployment Models

CHALLENGES OF EXTENDING IT TO THE EDGE

While every industry employs edge computing to address specific business challenges and objectives, there's a tremendous amount of commonality in how those resources are deployed and managed.

Managing edge computing environments differs significantly from traditional data centers:

Challenge	Traditional Data Center & Cloud	Edge Deployment
Geographical Distribution	Centralized, allowing easier physical access and centralized management.	Distributed nodes are often located in remote or hard-to-access locations.
Scale & Complexity	Homogenous and concentrated set of resources.	Large number of devices and device types.
Resource Constraints	High-performance, high-capacity resources. Minimal physical limitations.	Often limited processing power, storage, and networking capabilities; includes physical space constraints.
Network Dependency	Stable, high-bandwidth network connections.	Requires the ability to operate with inconsistent and unreliably

		connectivity, often with increased latency and bandwidth when communicating with core resources.
Security & Compliance	Centralized security & compliance, allowing for more controlled and consistent implementations.	Poses unique security challenges due to its distributed nature, increasing the risk of physical and cyber-attacks.
Data Management	Centralized data storage, processing, and management.	Local data processing and decision-making.
Automation & Orchestration	Automation is used, but the scale and heterogeneity are typically less complex.	Requires advanced automation and orchestration tools to manage many distributed nodes and applications efficiently.
Maintenance & Updates	Physical access allows for easier maintenance and updates.	Remote management and updates are essential due to their distributed nature. Over-the-air (OTA) updates and remote troubleshooting are commonly used.
Energy Efficiency & Environmental	Energy efficiency is a concern but on a scale different from distributed edge environments.	More focused on energy efficiency and environmental impact, especially in remote or sensitive locations.
Scalability & Flexibility	Scalability is planned and executed in a controlled and predictable environment.	Must be scalable and flexible to accommodate varying workloads and rapid deployment of new nodes.
Data Protection	Data protection, including backup and recovery, is generally straightforward due to centralized data and robust infrastructure.	Implementing data protection solutions can be complex due to the distributed data and limited resources at edge locations.
Disaster Recovery	The centralized nature of core and cloud deployments focuses on constrained locations with uniform risk factors and recovery strategies.	Disaster recovery must account for multiple, geographically dispersed edge sites, often consisting of many smaller-scale deployments. Due to the volume and variability of nodes, this can complicate disaster recovery. Disaster recovery must also consider the real-time processing and critical nature of edge data, often requiring immediate and localized recovery solutions.

Managing edge computing environments requires dealing with a more distributed, heterogeneous, and dynamic infrastructure than traditional data centers. It demands robust remote management capabilities, advanced automation, and a strong security and network resilience focus.

SIMPLIFYING EDGE COMPUTING WITH AN HCI-BASED SOLUTION

Hyper-converged infrastructure (HCI) is a foundational architecture underpinning a wide range of solutions, including on-prem, hybrid, multi-cloud, and edge-to-core deployment. HCI-based solutions offer several advantages in edge computing environments, addressing many of the unique challenges posed by the distributed nature of edge deployments. Some HCI-based solutions layer additional services and capabilities on top of the core HCI functionality to provide a feature-rich infrastructure platform for workloads at the edge, core, and cloud.

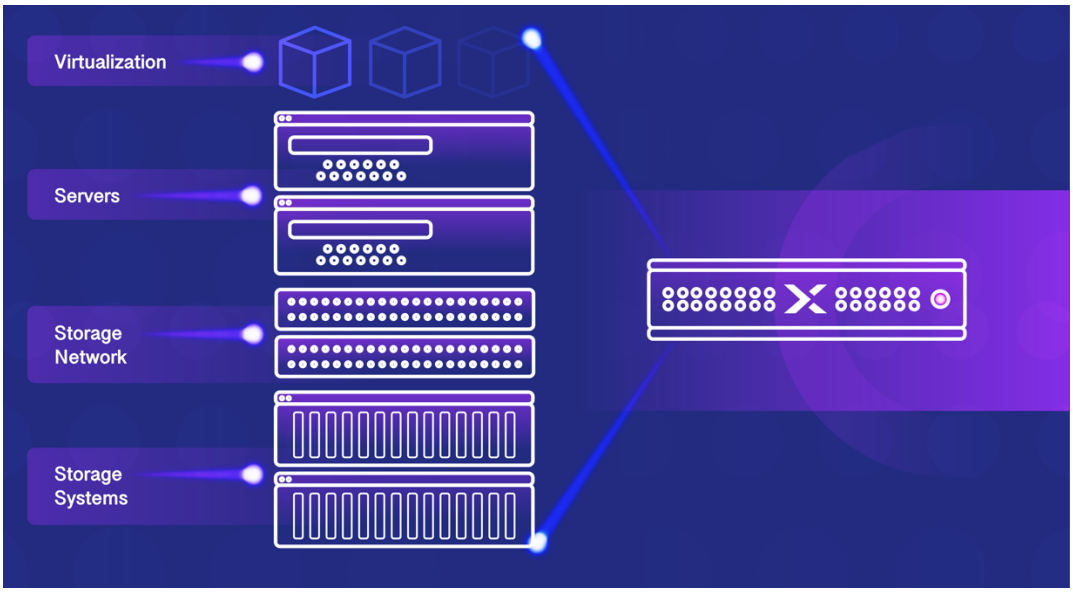


Figure 4: An HCI Architecture Unifies Traditionally Disparate Elements

Let's look at some of how building on the foundation enabled by HCI-like architectures addresses the specific needs of computing at the edge:

- **Simplified Management**: HCI combines compute, storage, and networking into a single system, simplifying the deployment and ongoing management of resources at the edge. This is particularly beneficial in edge computing, where managing numerous distributed nodes can be complex. An HCI-based platform's centralized management tools allow for easier monitoring and management of resources across multiple edge locations while allowing for remote updates, increased security, and comprehensive data and network protection.

- **Reduced Footprint**: Edge computing environments often have space constraints. An HCI's compact and consolidated architecture requires less physical space than traditional infrastructure, making it ideal for edge locations with limited space.
- **Scalability**: HCI-based platform solutions are highly scalable. They allow for easy and rapid resource scaling, which is crucial in edge computing environments where demand may fluctuate significantly. Additional resources, such as GPU-equipped nodes, can be added incrementally without significant infrastructure overhauls.
- **Cost Efficiency**: By consolidating multiple functions into a single solution, an HCI-based platform can reduce the total cost of ownership. It minimizes the need for separate storage, networking, and computing hardware, removes additional solutions like cost governance, automation, hypervisor, and security, and reduces power and cooling requirements, which are crucial for edge locations.
- **Improved Performance**: HCI can provide enhanced performance for edge computing workloads. Its design enables efficient processing and storage capabilities, which is critical for the latency-sensitive applications often run on edge computing infrastructures.
- **Resilience & Availability**: HCI solutions often include built-in disaster recovery and data protection features. This is important for edge computing, where data and applications must be available despite potential connectivity issues to central data centers.
- **Ease of Deployment**: Deploying HCI-based platform solutions is generally straightforward, as they are pre-configured and designed for quick setup. New applications can be remotely deployed at scale using blueprints. This plug-and-play
- **Energy Efficiency:** HCI deployments are designed to be energy efficient, which is beneficial for edge computing deployments, especially in remote locations where energy consumption can be a concern.
- **Flexible Cloud-Native Deployment**: HCI-based platform solutions natively support virtualization and containerized applications, allowing for flexible deployment and management. This aligns well with the dynamic and varied nature of edge computing workloads.
- **Remote Management**: HCI-based solutions often include advanced remote management tools while supporting automation, reducing the need for on-site IT staff at edge locations. This is particularly useful for managing a large number of distributed edge sites.

HCI provides a compact, efficient, and easily manageable infrastructure base with additional software-defined capabilities of some platform solutions that address many challenges associated with edge computing, such as limited space, scalability, and the complexity of managing distributed resources.

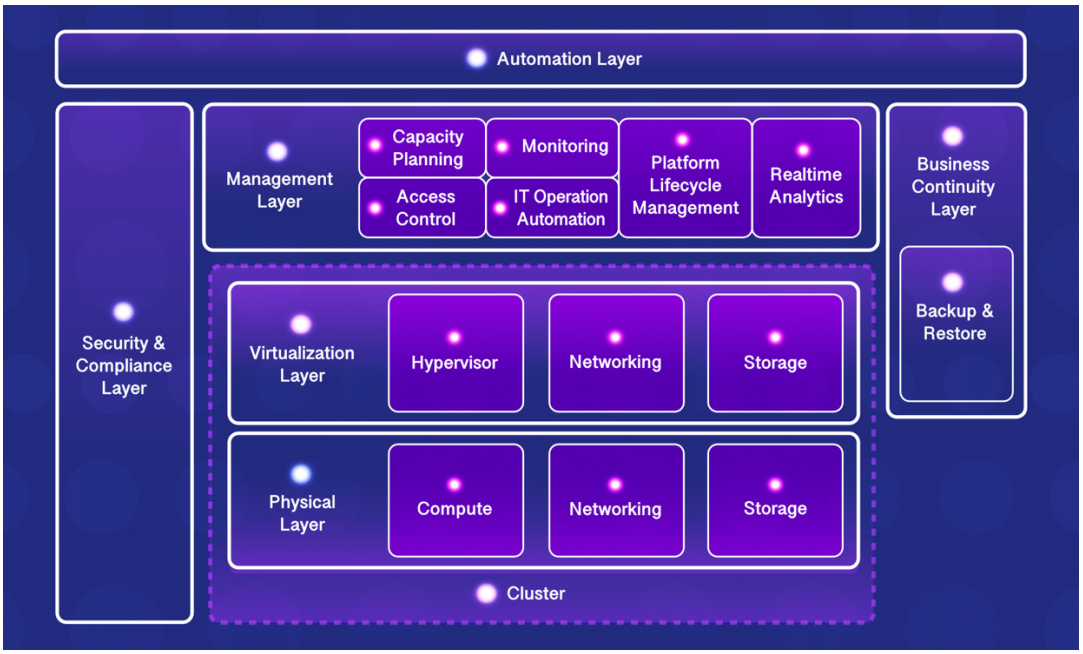


Figure 5: HCI Architecture

SIMPLIFYING EDGE WITH NUTANIX

Nutanix delivers some of the industry's most advanced enterprise cloud platform computing solutions built on HCI. Its broad offerings allow IT organizations to simplify hybrid multi-cloud and edge deployments.

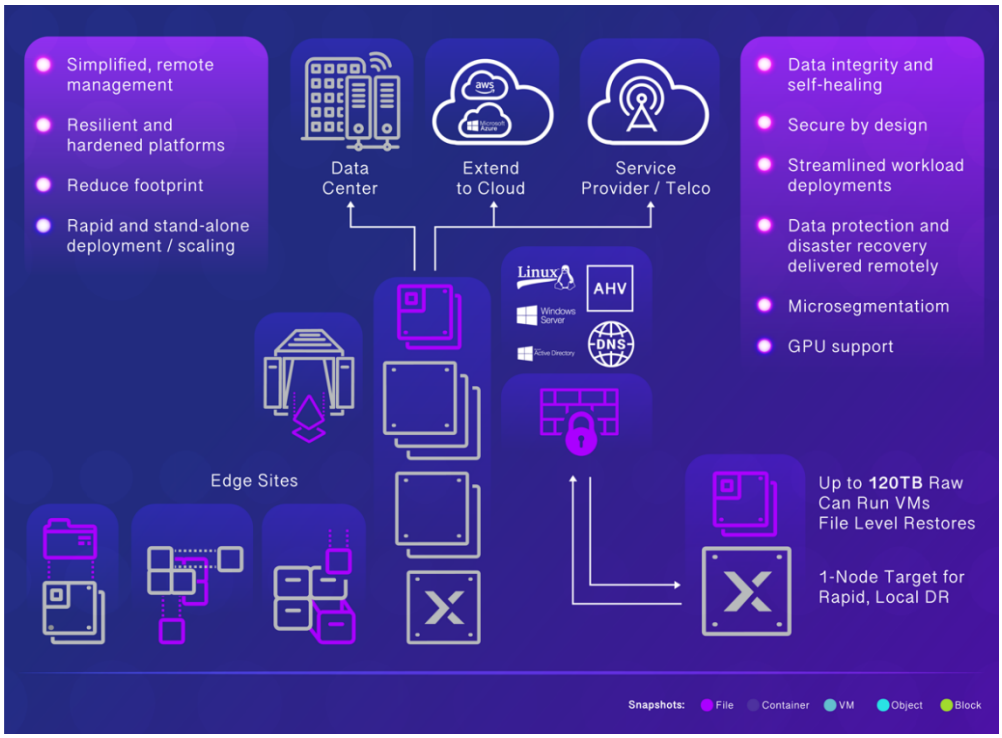


Figure 6: Benefits of the Nutanix HCI Architecture

Nutanix's technology greatly simplifies edge computing in several notable ways:

Challenge	Nutanix
Platform built on Hyper-Converged Infrastructure	Nutanix provides HCI technology, which combines compute, storage, and networking into a single system. This simplification is particularly beneficial for edge computing, where managing separate components can be complex and resource-intensive. HCI's compact form factor is also ideal for edge environments' space constraints.
Simplified Management & Operations	Nutanix offers a centralized management platform, Nutanix Prism and Nutanix Central, simplifying the operation and management of distributed edge environments. It provides a single pane of glass for managing all Nutanix clusters, regardless of location, reducing the complexity and operational overhead of managing multiple edge sites.
Easy Deployment & Scalability	Nutanix designed its solutions for easy deployment and scalability. Edge environments often need to scale quickly or deploy new applications rapidly. IT teams can deploy Nutanix's solutions quickly, scaling as needed without significant downtime or reconfiguration.
Automated Orchestration & Self-Healing Capabilities	Nutanix includes automated orchestration and self-healing capabilities, which are crucial for edge environments with limited IT resources. These features ensure that edge computing

	nodes run optimally and recover from failures without manual intervention.
Enhanced Security Features	Recognizing the security challenges in distributed edge environments, Nutanix incorporates robust security features into its products. These include data-at-rest encryption, micro-segmentation, ransomware detection and remediation, file scanning, and compliance with various security standards, helping to protect sensitive data, whether at rest or flowing across the network.
Data Protection	Nutanix's solution provides increased data protection to the edge with snapshots, clones, replication, and a wide range of backup and recovery options.
Reduced TCO & Energy Efficiency	Nutanix's cloud platform solution reduces the total cost of ownership (TCO) by consolidating hardware and simplifying management. The energy-efficient design of Nutanix systems also minimizes power and cooling requirements, which is essential for edge locations with limited power resources.
Integrated Virtualization & Container Support	Nutanix natively supports virtual machines (VMware and Nutanix's AHV hypervisor) and a spectrum of containerized applications, providing application deployment and management flexibility. This is particularly beneficial for edge computing, which often involves a mix of legacy and modern application architectures.
Remote Monitoring & Support	Nutanix provides advanced remote monitoring and support capabilities, enabling proactive issue resolution, upgrades, and support for edge computing deployments, even in remote locations.
Edge Specific Solutions	Nutanix has developed specific solutions targeting edge computing use cases, ensuring that their offerings are tailored to meet the unique demands of edge environments.
Data Optimization & Local Storage	Nutanix's data optimization features, like data compression and deduplication, maximize local storage efficiency. This is crucial in edge computing, where local storage resources may be limited.

Nutanix helps simplify edge computing through its cloud platform built on HCI technology, which offers centralized management, easy deployment, scalability, automated orchestration, robust security, energy efficiency, and support for virtualization and container technologies.

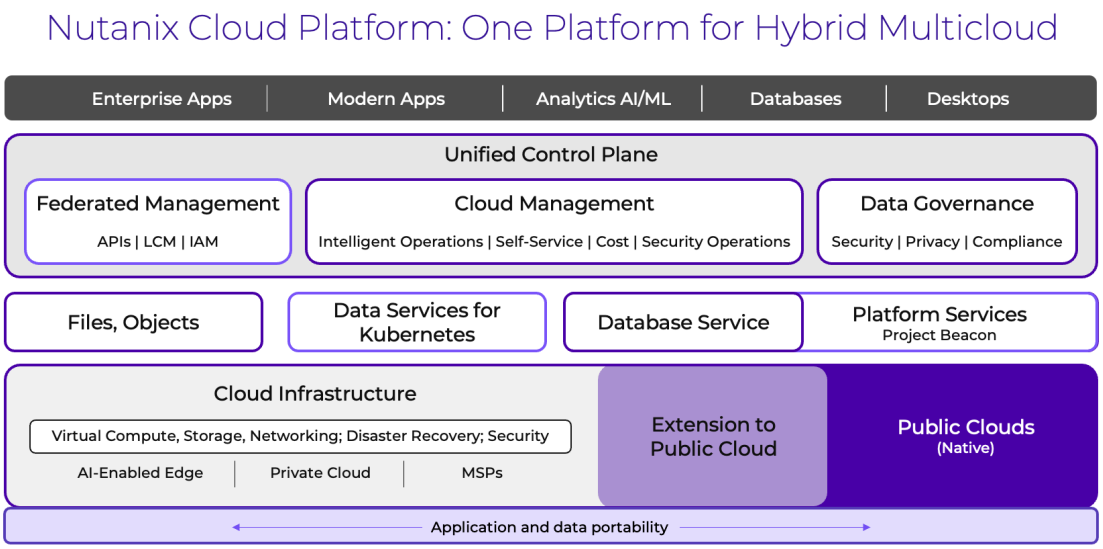


Figure 7: The Nutanix Solution Stack (source: Nutanix)

These features make Nutanix a compelling choice for organizations looking to deploy and manage edge computing infrastructure efficiently and securely.

EDGE AI WITH NUTANIX

Beyond providing the tools for building and managing a hybrid multi-cloud edge infrastructure, Nutanix also provides the tools required to ease the deployment and management of AI and machine learning workloads at the edge.

Here's how Nutanix helps in bringing AI to the edge:

Challenge	Nutanix
Simplified Infrastructure Management	By converging compute, storage, and networking into a single, easy-to-manage platform, Nutanix allows organizations to focus more on AI development than managing complex infrastructure.
Scalability	AI and machine learning workloads often require scalable infrastructure for varying computational intensity. Nutanix's solutions are inherently scalable, enabling easy expansion of resources as AI workloads grow.
Virtualization & Containerization	Nutanix supports a range of virtualization and container platforms, which are crucial for AI and machine learning environments. Containers, in particular, are vital for AI development, as they allow for rapid deployment and scalability of AI models.

Data Management & Storage	Efficient data management and storage are critical for AI applications. Nutanix offers robust data management solutions, including high-performance storage capabilities necessary for the large and often fast-moving datasets used in AI and machine learning, plus file, block, and object support on the same platform.
GPU Support	AI and machine learning workloads frequently require GPUs for efficient processing. Nutanix's platform can integrate GPU resources essential for training complex machine learning models and smaller versions for inferencing at the edge.
Database Management	Nutanix Database Service (NDB) simplifies database management, a key component in many AI and analytics workloads. It enables quick provisioning, cloning, and scaling of database environments, including open-source ones, benefitting data-intensive AI applications.
Automation & Orchestration	Nutanix's solutions include automation and orchestration tools, which can significantly streamline the deployment and management of AI workloads, reducing the time and effort required to maintain these systems.
Security & Compliance	AI workloads often involve sensitive data. Nutanix provides a secure infrastructure with compliance features that ensure data protection, a key consideration for AI and machine learning applications.

Nutanix facilitates AI initiatives by providing a scalable, high-performance, easy-to-manage infrastructure. Its support for high-performance compute and storage, GPU integration, virtualization, containerization, efficient data management, and robust security features make it a suitable platform for developing and deploying AI and machine learning solutions.

IN SUMMARY

The edge computing landscape has significantly grown, driven by the demand for low-latency, on-the-spot data processing across nearly every industry. The use of edge is accelerating as AI upends traditional architectures with its ability to generate instant insights.

Despite the tremendous business benefits of edge computing, it arrives with complex challenges, including managing geographically dispersed nodes, security risks, and integrating with existing IT infrastructure. These challenges can best be addressed with an architecture built around Hyper-Converged Infrastructure, which streamlines management and enhances performance.

Nutanix is one of the most significant players in this field, offering solutions that simplify the management of edge computing through its hybrid multi-cloud platform technology. Nutanix's

solutions are notable for their simplified management, easy deployment, scalability, security, and automated orchestration, which are essential for the distributed nature of edge computing.

Additionally, Nutanix addresses the demand for AI at the edge by providing scalable infrastructure, virtualization and containerization support, and robust data management and storage capabilities.

As the edge computing market flourishes, a cloud platform solution built on HCI, as delivered by companies like Nutanix, is pivotal in overcoming the complexities of edge deployment and is instrumental in harnessing the full potential of edge AI, thus driving forward the next wave of technological innovation.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.