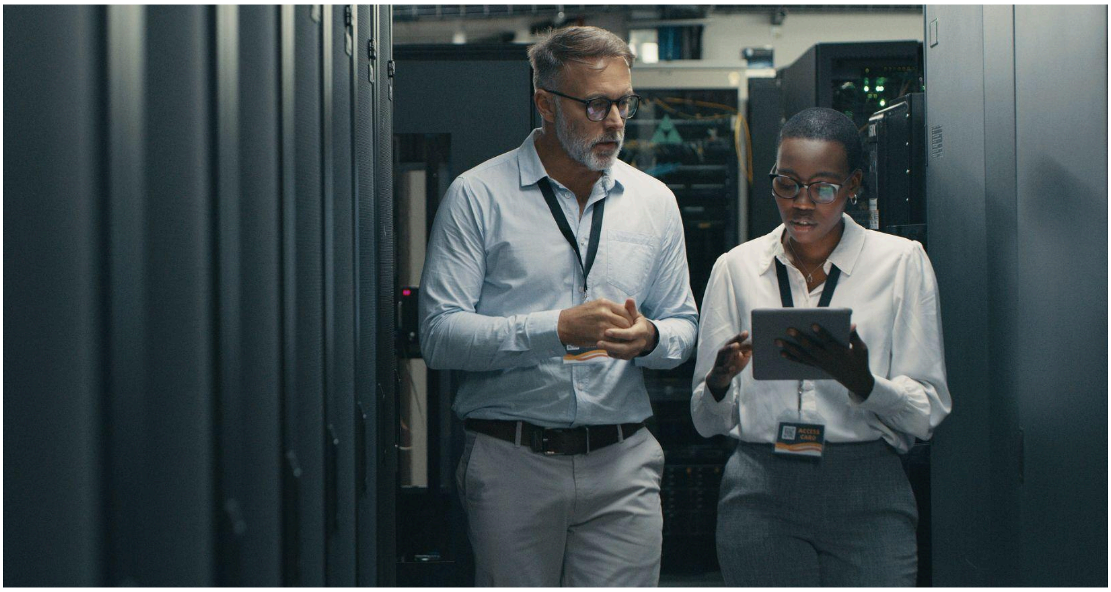
Canonical WHITEPAPER Q4 2025

Enabling AI and networking at the edge

Solution Brief: Druid, Airspan and Ecrio
running on Canonical MicroCloud and Kubernetes
on the Dell PowerEdge XR8000 with Intel Xeon Scalable processors



Contents

Executive Summary

As artificial intelligence (AI) and real-time analytics are being adopted by a growing number of organizations to enhance safety, efficiency, and automation, the need for robust edge computing solutions has surged. Industries such as manufacturing, healthcare, and mining rely on distributed data processing to meet their operational demands. However, deploying AI and networking workloads at the edge presents brand new challenges, such as ensuring low latency, scalability, and security in diverse and remote environments.

The combination of Intel® Xeon® Scalable processors, Dell PowerEdge XR8000, Ubuntu, MicroCloud, Canonical Kubernetes, Druid Software Raemis™ Core Network Platform, Airspan AirVelocity gNB and Control Platform (ACP), and Ecrio iota-e addresses these challenges by delivering a comprehensive edge computing and communication platform. Intel® Xeon® Scalable processors are engineered to handle the complex and performance-intensive demands of AI workloads at the edge. The Dell XR8000 provides robust hardware optimized for AI and networking workloads, while Ubuntu offers a secure and adaptable operating system environment that ensures software stability and broad hardware compatibility. At the heart of the software architecture is Canonical MicroCloud, which provisions lightweight virtualized infrastructure on each hardware sled. Canonical Kubernetes ensures long-term stability, and low-latency performance with enterprise-grade container orchestration. Druid's Raemis™ and Airspan's RAN solution extend secure Private 5G connectivity to edge locations with simplified operational management. Ecrio iota-e adds ultra-reliable, real-time human-to-machine communications capabilities directly at the edge, further enhancing situational awareness, responsiveness, and multimedia support for AI-enabled mission-critical applications.

By deploying this integrated solution, businesses can improve operational efficiency, enhance safety in industrial environments, and drive innovation with AI-powered insights, all while minimizing costs and improving security.

Introduction

The rapid growth of AI-driven applications and the increasing demand for low-latency data processing are driving the need for robust edge computing solutions. Industries such as manufacturing, healthcare, mining, and retail increasingly rely on AI and real-time data analytics to improve safety, efficiency, and productivity. Edge computing plays a pivotal role by bringing compute power closer to where data is generated, enabling faster insights, decision-making, and automation. However, edge environments pose unique challenges, such as harsh operating conditions, limited space, and diverse hardware requirements.

Deploying AI and networking solutions at the edge requires a cohesive set of technologies that work together to ensure security, performance, and operational simplicity. The combination of Intel® Xeon® Scalable processors, Dell PowerEdge XR8000, Canonical Kubernetes and MicroCloud infrastructure software, Druid's Raemis™ Core Network Platform, Airspan's RAN solution, and Ubuntu offers a powerful and reliable foundation. Adding to this, Ecrio iota-e introduces a crucial layer of embedded real-time communication services, enhancing responsiveness and enabling actionable rich, context-aware interactions between people, machines, and systems at the edge.

This integrated stack delivers the tools and infrastructure necessary to support mission-critical edge deployments. This whitepaper examines each component of the solution in detail, moving through the hardware, software, and orchestration layers, and finally explores how they come together to enable real-world use cases across multiple industries.

Challenges of AI and networking at the edge

Deploying AI and networking solutions at the edge presents multiple unique challenges.

Low Latency

Achieving low latency is crucial for edge environments, where data must be processed rapidly to support mission-critical applications. Industrial automation systems, for example, rely on low-latency computing to ensure robotic arms react in milliseconds to avoid production faults. Similarly, in healthcare, patient monitoring systems must transmit alerts immediately to ensure rapid medical intervention.

These use cases cannot rely solely on extensive compute resources and applications located in distant cloud environments. Instead, compute power and applications running close to the data source are required. Local hardware must be powerful enough to handle real-time processing, and its foundation software stack must be optimized to ensure responsiveness and efficient resource utilization.

Security risks

Security is a major concern for edge environments. Unlike data centers, edge nodes are often located in less secure or remote environments, making them more susceptible to physical tampering and network intrusions. Since edge devices frequently collect and process sensitive data, the risk of data breaches is heightened.

To address these concerns, robust encryption protocols, secure device management frameworks, and strict access control policies must be implemented. Additionally, maintaining a proactive security posture with frequent software updates and patch management is crucial to mitigate emerging vulnerabilities.

Hardware constraints

Deploying AI workloads at the edge introduces unique hardware challenges. Edge devices often operate with limited CPU, GPU, and storage capacity, while still being expected to handle complex AI inference models. Moreover, AI workloads tend to be resource-intensive, requiring optimized hardware to ensure efficient performance.

Balancing performance and hardware efficiency demands strategic hardware deployment. Solutions such as modular server architectures, hardware accelerators, and EPA (Enhanced Platform Awareness) features in Canonical Kubernetes help to address these challenges.

Cost and power efficiency

Operating AI workloads at the edge presents financial and energy efficiency challenges. Deployments must minimize power consumption to ensure sustainable operations in energy-constrained environments. In manufacturing plants, for example, servers must efficiently process data without exceeding strict power budgets.

Solutions like the Dell XR8000 use energy-efficient Intel Xeon processors that dynamically adjust power usage based on workload requirements, thereby reducing operational costs. CPU native inference without the need for AI accelerators like GPU results in lower total cost of ownership (TCO) and sustainable solutions. Canonical Kubernetes further reduces resource waste by optimizing workload distribution, ensuring computing power is allocated efficiently.

Limited space and harsh conditions

Edge environments – such as mining sites, manufacturing plants, and outdoor installations – often have limited physical space for server hardware. Additionally, these locations may expose devices to harsh environmental conditions, like extreme temperatures, dust, and vibrations. Ruggedized hardware like the Dell XR8000 is designed to withstand extreme environments, making it ideal to address the challenges presented by edge computing environments. Its modular design also ensures flexible hardware deployment in compact spaces, allowing enterprises to deploy scalable edge infrastructure where needed most.

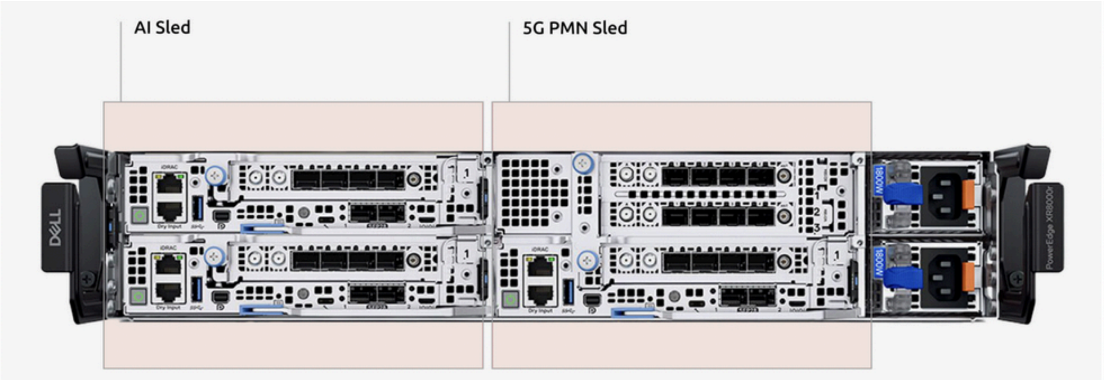


Fig.1: Edge AI powered by XR8000 equipped with two XR8620t sleds

Dell PowerEdge XR8000 for edge AI and networking

The Dell PowerEdge XR8000 is purpose-built to meet the rigorous demands of modern edge computing environments. In edge use cases – where latency, resilience, and real-time performance are non-negotiable – this platform stands out for its ability to deliver high compute density and rugged reliability in a compact form factor.

Designed for deployment in environments where space is limited and conditions are harsh (for example, factory floors, outdoor installations, or remote industrial sites), the XR8000 features a short-depth chassis and modular sled design. This makes it highly adaptable to physical constraints while still supporting scalable, multi-node configurations. The system is also built to withstand extreme temperatures, dust, and vibration, making it well-suited for mission-critical edge operations.

From a compute perspective, the XR8000 is powered by Intel Xeon Scalable processors, delivering the performance needed to support demanding AI inference, data analytics, and network processing workloads at the edge. Its architecture supports diverse connectivity options –including wired, Wi-Fi, and Private 5G networks – ensuring seamless integration into various edge networking environments.

In terms of efficiency, the platform supports right-sized sled options and energy-optimized processing, enabling organizations to scale intelligently without unnecessary power consumption. Combined with Dell's long-standing expertise in enterprise IT infrastructure, the XR8000 empowers edge deployments to be both performance-driven and cost-effective, paving the way for scalable innovation across industry verticals.

To further understand the processing capabilities that drive this platform, this whitepaper will now explore the role of Intel Xeon Scalable processors in enabling efficient AI inference at the edge.

Intel Xeon scalable processors: built for AI at the edge

Intel Xeon Scalable processors are engineered to handle the complex and performance-intensive demands of AI workloads at the edge. These processors combine compute power, hardware-accelerated AI features, and energy efficiency to deliver exceptional inference performance in environments where space, power, and connectivity are limited. With support for Intel® Advanced Vector Extension 512 (AVX-512), Intel® Advanced Matrix Extensions (AMX), and Intel® Deep Learning Boost (DL Boost), they accelerate AI workloads directly on the CPU, eliminating the need for additional accelerators or discrete GPUs in many edge deployments. AI application developers can leverage these CPU features using Intel® OpenVINO™ Toolkit.

Intel's OpenVINO toolkit is an open-source, free software suite that accelerates deep learning AI inference on Intel hardware and other compatible platforms. It optimizes AI models from various frameworks for deployment on a range of Intel CPUs, GPUs, NPUs, and other accelerators to improve speed, reduce model footprint, and lower latency for applications in computer vision, large language models, and generative AI.

Intel AMX improves the performance of deep learning (DL) training and inference, making it ideal for workloads like natural language processing (NLP), recommender systems, and image recognition. Intel AMX architecture consists of two components:

- The first component is tiles. Tiles consist of eight two-dimensional registers, each 1 kilobyte in size. They store large chunks of data.
- The second component is Tile Matrix Multiplication (TMUL). TMUL is an accelerator engine attached to the tiles that performs matrix-multiplication computations for AI.

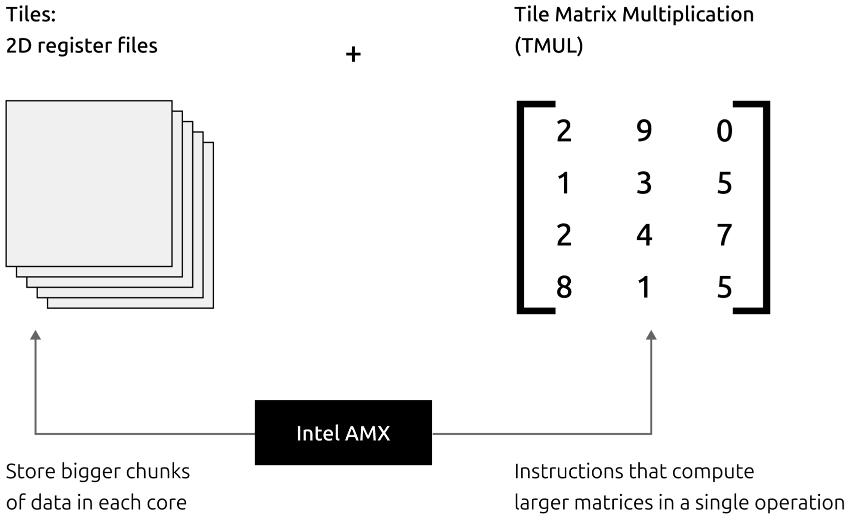


Fig.2: Intel AMX - Matrix multiplications for deep learning and AI workloads

The latest generations of Intel Xeon Scalable processors, including the Edge Enhanced SKUs and those with integrated Intel vRAN Boost, offer built-in features for optimizing low-latency, high-throughput applications. These include Time Coordinated Computing (TCC) and support for Time-Sensitive Networking (TSN), which are essential for deterministic performance in real-time industrial and telco scenarios. Additionally, the processors offer high memory bandwidth, scalable core counts, and integrated security features, making them ideal for deploying secure and responsive AI inference models at the edge.

Intel's sustained commitment to performance leadership in real-world AI applications is demonstrated by its ability to handle diverse workloads such as computer vision, natural language processing, and anomaly detection directly on Xeon CPUs. This versatility makes Intel Xeon Scalable processors a cornerstone of edge architectures that require flexibility, power efficiency, and consistent performance under demanding conditions. Ubuntu provides the optimized software environment needed to fully utilize Dell's XR8000 hardware innovations and capabilities of the Intel Xeon Scalable processors.

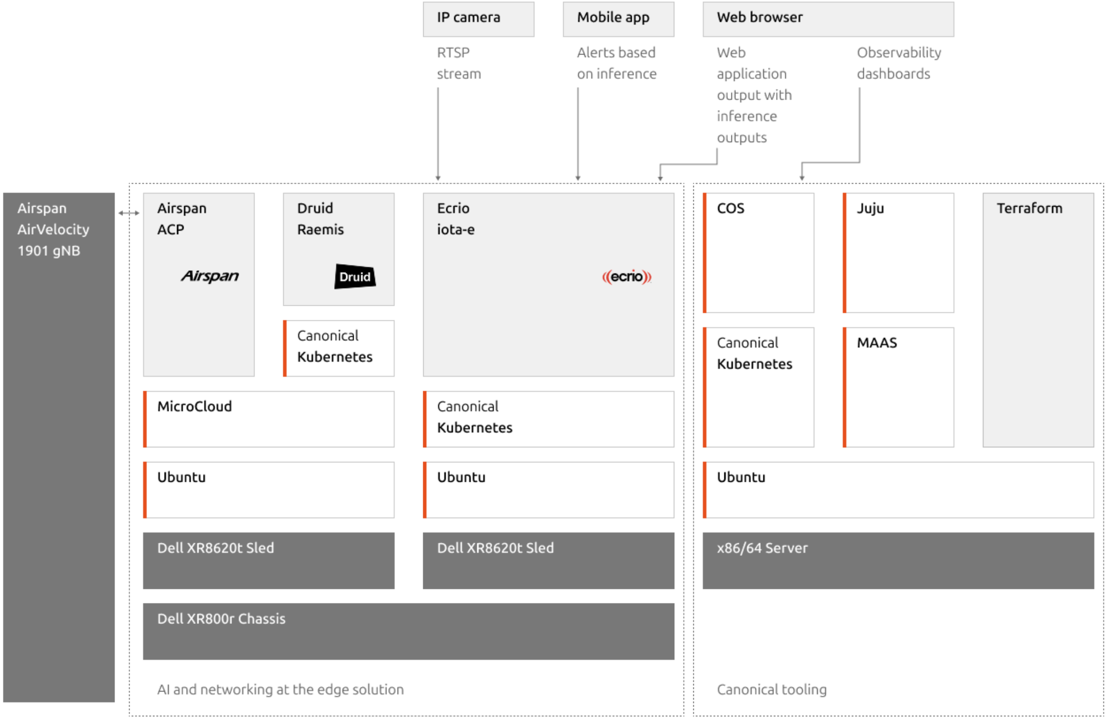


Fig. 3: Edge AI Canonical software stack and associated support tooling

Ubuntu: A robust software foundation for AI at the edge

Ubuntu stands as a leading Linux distribution trusted by AI developers, cloud providers, and edge solution architects worldwide. In edge computing environments, where performance and reliability are paramount, Ubuntu provides a solid operating system foundation, optimized for AI inference workloads. Its extensive ecosystem of drivers and packages ensures compatibility with a wide array of hardware platforms and accelerators.

Canonical has optimized Ubuntu to align with the evolving capabilities of Intel Xeon Scalable processors, ensuring seamless integration with hardware features designed for edge AI performance. Through this alignment, Ubuntu enables support for advanced AI instruction sets and edge-focused innovations while maintaining the stability and openness expected from a leading Linux distribution. These enhancements contribute to robust and efficient inference execution, particularly in edge environments constrained by power, space, and cost limitations.

In addition, Ubuntu's predictable release cadence and commitment to long-term support make it ideal for industrial and telco applications where stability and maintainability are non-negotiable. With a lightweight footprint and real-time kernel support, Ubuntu is uniquely positioned to serve as the operating system of choice for AI inference at the edge. Ubuntu Pro further extends Ubuntu’s core strengths with critical security, compliance, and operational support.

Ubuntu Pro: secure, compliant, and enterprise-ready

Ubuntu Pro extends the capabilities of the Ubuntu distribution with advanced security, compliance, and support features tailored to enterprise and telco environments. It includes Expanded Security Maintenance (ESM), which delivers critical CVE patches and kernel updates for up to 12 years, and ensures that even older packages remain secure and supported.

Livepatch is another key feature of Ubuntu Pro, enabling kernel updates to be applied without system reboots. This is especially valuable at the edge, where uptime is crucial and physical access to nodes may be limited. Canonical also provides automated compliance tools that simplify achieving standards such as CIS, FIPS, and DISA-STIG, which are commonplace in highly regulated industries.

In addition to technical hardening, Ubuntu Pro includes 24/7 enterprise-grade support through phone and ticketing systems, offering peace of mind for organizations running critical workloads. Finally, the inclusion of Canonical's Real-Time Ubuntu kernel ensures ultra-low latency performance and deterministic task execution, meeting the stringent demands of AI, 5G, and industrial automation use cases.

These enterprise capabilities provide a strong operational and security foundation, from which MicroCloud and Canonical Kubernetes deliver scalable orchestration at the edge.

MicroCloud: lightweight virtualization for network-centric edge deployments

To support the increasing complexity and modularity of 5G Core and RAN architectures, the solution integrates Canonical MicroCloud to provide lightweight, secure virtualization at the edge. MicroCloud is Canonical's distributed cloud platform that brings the ease of public cloud operations to on-premise and far-edge environments. It is simple to deploy, resilient by design, and optimized for resource-constrained hardware.

In the solution's design, MicroCloud is deployed directly on the network sled of the Dell XR8000, where it provisions and manages two virtual machines. The first virtual machine hosts Canonical Kubernetes, which runs the containerized Druid Raemis 5G Core components. The second virtual machine runs the Airspan Control Platform, which provides configuration and monitoring of the Airspan indoor 5G gNB. This clean separation of concerns ensures both workload isolation and operational flexibility, while maintaining efficient use of available compute resources.

Through MicroCloud, the platform benefits from fast provisioning times, integrated high availability, and centralized management of virtual workloads. It eliminates the need for traditional heavyweight virtualization platforms, reducing both operational complexity and footprint.

Canonical MicroCloud adds an essential layer of abstraction and reliability for edge environments that require a hybrid of containerized and virtualized workloads. Its inclusion in the architecture ensures the platform is ready for complex 5G deployments that span diverse infrastructure layers.

Canonical Kubernetes for secure, reliable edge deployments

Canonical Kubernetes is engineered to meet the rigorous demands of edge computing with an emphasis on long-term support, operational efficiency, and high performance. It provides a consistent and securely designed Kubernetes distribution, which simplifies cluster management and accelerates the deployment of containerized applications at the edge. Built on upstream Kubernetes, Canonical's implementation introduces production-grade enhancements while maintaining full adherence to open standards.

A key differentiator is Canonical Kubernetes' Long-Term Support (LTS) model. While most Kubernetes distributions support only the last two releases ("n-2") for roughly eight months, Canonical extends this window to up to 12 years. This extended lifecycle includes continuous maintenance, security patches, and CVE fixes, which is vital for telco and enterprise environments, where edge nodes may be difficult to access physically and have strict uptime requirements. This allows operators to adopt Kubernetes with confidence, knowing that they can maintain stability and control upgrade cycles in alignment with their business needs.

Canonical Kubernetes also brings performance and reliability enhancements specifically tailored for edge workloads. It supports Enhanced Platform Awareness (EPA), enabling optimized scheduling and resource allocation based on hardware capabilities such as CPU pinning, NUMA awareness, and SR-IOV networking. This results in deterministic performance for latency-sensitive AI/ML and network functions.

Canonical Kubernetes supports advanced Container Network Interfaces (CNIs) such as Cilium and Multus, providing critical networking capabilities for AI-driven edge use cases. Cilium leverages eBPF (extended Berkeley Packet Filter) for high-performance, low-latency networking, fine-grained security policies, and deep observability, all of which are essential for modern, distributed AI workloads. Its capabilities help ensure that edge applications can scale securely and efficiently across Kubernetes clusters.

Multus enhances Canonical Kubernetes with multi-homing support, allowing edge workloads to connect to multiple network interfaces. This is particularly useful in AI at the edge scenarios that require dedicated data and control planes, or separate interfaces for video streaming, telemetry, and inference workloads. Together, these CNIs enable Canonical Kubernetes to orchestrate complex AI and networking topologies while ensuring optimal performance, isolation, and flexibility in constrained edge environments.

Together, these attributes make Canonical Kubernetes an ideal foundation for edge deployments, enabling scalable, secure, and cost-efficient operations across a variety of verticals.

This robust Kubernetes foundation also serves as the platform on which Druid’s Core is deployed. Canonical Kubernetes helps operators to simplify the deployment, scaling, and lifecycle management of 5G Core software, making it especially effective for mobile private network use cases that demand both operational agility and telco-grade reliability.

Druid Software Raemis™ core network platform

The Raemis™ technology platform is a software-based mobile core purpose-built for private enterprise and mission-critical networks. Designed for IT and Operational Technology teams and solutions providers, it combines telco-grade performance with the agility and simplicity required for enterprise deployments. This delivers the control, flexibility, and performance needed to connect operations, automate processes, and enable next-generation applications.

The platform supports full virtualization and cloud-native deployments such as Kubernetes enabling highly scalable and resilient operations. When deployed at the network edge, Raemis delivers ultra-low latency and secure local processing for time-sensitive workloads, supporting AI-driven applications such as real-time video analytics, industrial automation, autonomous systems, and other time-critical communications.

Raemis also includes advanced capabilities including network slicing which gives enterprises the ability to create tailored services for specific needs from IoT to mission-critical communications. Its orchestration tools (SPO) allow service providers to deploy, monitor, and manage multiple customer networks from a single interface. Meanwhile, distributed network manager (DNM) ensures multi-site deployments of private networks with each site running its own core to maintain full local service even during backhaul disruptions. The result is always-on connectivity that keeps hospitals operational, production lines moving, and critical field teams connected when it matters most.

Airspan RAN solution

AirVelocity 1901 is part of Airspan's gNB product family which enables operators to provide 5G services. The AirVelocity 1901 is an enhanced indoor 5G NR gNB product that is to be installed indoors. AirVelocity 1901 is easy to install on a suspended ceiling or wall. The effective coverage is achieved by using a high gain high quality antenna. The unit can be powered by either PoE++ or AC power (using an AC/DC converter).

Airspan Control Platform (ACP) is a scalable, easy-to-deploy, and easy-to-use comprehensive management solution that manages Airspan's 4G-LTE and 5G-NR network portfolio. It is a client-server application, with 'always-on' server components implemented as Linux services and a front-end web application. It offers full CBRS compliance, incorporating a Domain Proxy service within its application.

Ecrio iota-e: enabling intelligent edge communications among people and machines

Ecrio's iota-e software solution is a lightweight, cloud-native platform designed to bring intelligent communications capabilities to edge devices. It enables real-time, reliable voice, video, and data services through a fully embedded IMS and 5G Core stack that operates within constrained edge environments. In addition, the Digital Twin Engine package within iota-e offers a unique capability that enables communications between people and machines such as cameras, sensors, drones and robots. Built to integrate seamlessly with AI applications, iota-e supports use cases that demand instant situational awareness, low-latency actionable alerts, and high-quality multimedia processing close to the source of data.

By embedding communication services directly into the edge platform, iota-e eliminates the dependency on centralized network infrastructure for critical real-time functions. This approach enhances resilience and performance while reducing backhaul congestion and latency. The Digital Twin Engine package leverages Intel OpenVINO™ for efficient and reliable AI/ML models for object recognition and vision analytics. In addition, the package supports NLP technology to offer voice-enabled command and control of IIoT devices such as cameras and sensors. The iota-e has been deployed in industrial safety systems, remote healthcare platforms, and smart city applications – ensuring timely, context-aware communication between machines, sensors, and humans. Ecrio's solution is also best suited for similar industrial verticals such as mining, ports, manufacturing and oil & gas industries. The iota-e seamlessly works with popular collaboration and messaging platforms such as WebEx and Teams for IT/OT convergence. This allows IT personnel to command-and-control IIoT devices in the field.

Iota-e’s compatibility with Canonical Kubernetes and support for deployment on Intel-based platforms, including those using Ubuntu and Dell XR8000, positions iota-e as a key enabler for intelligent, connected edge systems that combine AI inference with real-time communications. Ecrio’s solution offers people to people and people to machine communications on the same platform, taking the AI/ML based inferences to actionable tasks.

Use cases across industries

Edge computing is transforming operations in diverse industries by enabling real-time AI-powered decision-making directly at the point of data generation. With the combination of Dell XR8000 and the supporting Canonical stack, organizations are deploying intelligent systems that can rapidly process information, improve safety, and optimize performance across a range of settings.

Worker safety in industrial environments

In high-risk industrial settings such as manufacturing plants, mining operations, or construction zones, protecting workers is a top priority. Traditional safety monitoring methods often fall short due to latency in alerting systems or lack of real-time visibility. Edge-based AI inference running on Dell XR8000 infrastructure enables real-time video analysis and sensor data processing, allowing safety protocols to be enforced immediately on-site.

For example, vision-based AI models can detect whether personnel are wearing the required personal protective equipment such as helmets and high-visibility vests. If a worker is identified without proper gear, the system can issue alerts to supervisors instantly. In parallel, proximity sensors and cameras can monitor movement patterns around dangerous machinery and trigger warnings if someone enters a restricted zone. With Canonical Kubernetes orchestrating these inference workloads, the solution ensures rapid response times and consistent performance under variable operating conditions.

These capabilities not only reduce the risk of injuries but also foster a culture of safety, driven by intelligent automation that adapts in real time to evolving on-site dynamics.

Healthcare safety and patient monitoring

In hospital environments, patient safety and care continuity are critically dependent on timely data and staff responsiveness. Deploying AI solutions at the edge within healthcare facilities allows systems to monitor patients continuously without requiring data transmission to centralized cloud infrastructure.

For instance, AI models deployed on-site can analyze visual and physiological data to determine whether patients are correctly positioned for examination or transport, preventing incidents like falls or pressure injuries. These models can also ensure that medical staff are present in critical care areas or respond quickly when a patient's condition changes.

Running these workloads on Canonical Kubernetes within the Dell XR8000 platform ensures data privacy, regulatory compliance, and consistent low-latency performance. Hospitals benefit from improved operational efficiency and patient safety without compromising the integrity of sensitive medical information.

Smart manufacturing and logistics

Modern manufacturing lines and logistics centers rely on precision, speed, and traceability. Deploying AI models at the edge enables businesses to analyze video, barcode, and sensor data in real time to streamline operations and detect anomalies before they disrupt workflows.

On the factory floor, AI-driven quality control systems can identify defects in parts, improper tool usage, or irregularities in machine behavior. These insights are instantly processed and acted upon locally, avoiding costly delays or downtime. In logistics environments, AI solutions assist in tracking inventory by recognizing and cataloging items as they move through various stages of fulfillment, ensuring end-to-end visibility and reducing errors.

The combination of Canonical Kubernetes and Dell XR8000 provides a high-performance foundation that supports these continuous, high-frequency inference tasks. With scalable infrastructure and optimized software orchestration, smart manufacturing and logistics operations can meet their performance goals while adapting to changing market demands.

Conclusion

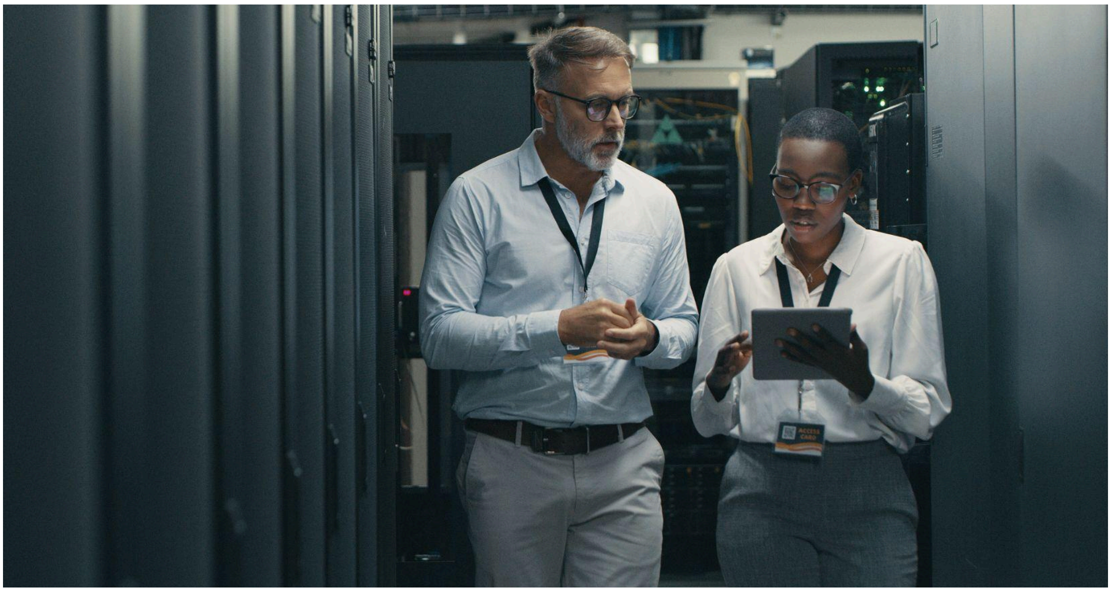
The integration of Dell's PowerEdge XR8000, Ubuntu, MicroCloud, Canonical Kubernetes, Druid's Raemis™ Core Network Platform and Ecrio iota-e provides a robust and comprehensive solution for deploying AI and networking at the edge. This combination effectively addresses the unique challenges of edge computing, including low latency, security risks, hardware constraints, cost and power efficiency, and limited space. Through the use of Intel Xeon Scalable processors and a secure, enterprise-ready software foundation, organizations can achieve enhanced operational efficiency, improved safety, and innovative AI-powered insights.

This integrated stack enables a variety of critical use cases across industries, from ensuring worker safety in industrial environments to enhancing patient monitoring in healthcare and streamlining operations in smart manufacturing and logistics. The modular and rugged design of the Dell XR8000, combined with the long-term support and advanced orchestration capabilities of the Canonical infrastructure stack, ensures scalable, reliable, and cost-effective edge deployments. Ultimately, this solution empowers businesses to bring compute power closer to the data source, enabling faster decision-making, automation, and innovation in real-time.

Learn more

If you're interested in building secure, high-performance edge platforms that combine AI, 5G, and container orchestration, we can help. Canonical provides a fully supported open source stack, from bare metal to Kubernetes and MicroCloud, optimised for edge workloads and network functions.

Contact us to discuss how Canonical can support your edge AI and private 5G initiatives.



Canonical