Solution Brief

Retail, Health and Life Sciences, Manufacturing

Cisco Brings Edge Compute, Storage and Networking to Retailers

Cisco meets retailers' growing need for retail edge AI and other applications with its Cisco Unified Edge platform that uses compute power of the Intel® Xeon® 6 processor for AI training and inference

The edge is rapidly emerging as the next major arena for computing, becoming the place where consumers interact with AI through real-time inference. The scale of this shift is enormous. By the end of the decade, more than 77 billion edge devices are expected to be in operation, according to Cisco estimates, each producing vast streams of information and requiring immediate, dependable analysis.[1]

Cisco also estimates that by 2027, three-quarters of all enterprise data will originate and be processed at the edge.[1] Several forces are accelerating the shift to edge computing. Chief among them are AI workloads that depend on rapid inference — from real-time fraud monitoring and industrial sensor data to video analytics that enhance manufacturing quality and predict equipment failures, as well as retail use cases that power new in-store experiences.

Retail is a good example of the impact of AI and edge computing. Edge computing can play a big role in making operations more efficient and improving customer experiences. With data processed at the edge, retailers have access to real-time insights that can be turned into actions while a customer is still in the store.

But edge AI applications require high throughput compute performance, higher availability, and more resilient infrastructure. Organizations are also seeking greater efficiency and cost control, which continue to propel the adoption of next-generation edge use cases.

This highly distributed environment introduces a new level of operational difficulty. Managing, deploying, and maintaining remote systems at scale becomes far more complex — and without strong orchestration and management, organizations risk significant data-security exposure.

Edge locations also present inherent physical security challenges. As a result, infrastructure that simplifies rollouts and minimizes the operational burden is increasingly favored. Modern edge platforms focus on automation, keeping configurations consistent and ensuring systems stay updated even when no technical staff are on-site.

This move to the edge is impacting a lot of industries including retailers who can use the technology to improve customer service in geographically dispersed locations using a distributed compute environment.

Cisco, an Intel® Industry Solutions Builders Partner, has recently launched the Cisco Unified Edge platform for these applications. With the compute power of the Intel® Xeon® 6 processor, Cisco Unified Edge offers the compute, storage and networking to serve a wide range of applications.

Integrated Compute, Storage and Networking

Cisco Unified Edge (see Figure 1) is designed with the performance needed for real-time inferencing for agentic and physical AI workloads at the edge. The system combines compute with integrated AI acceleration, networking, storage and security into a unified platform optimized for AI workloads.

The hardware foundation of Cisco Unified Edge is the Cisco UCS® XE9305 chassis, a 3 RU-high chassis with five front-facing flexible slots for compute and networking nodes. The chassis can support up to 120 terabytes of storage via four NVMe E3.S drives and up to two M.2 SATA drives with hardware RAID. The chassis also features redundant power and cooling.

Data processing is done on the Cisco UCS® XE130c M8 Compute Node. The 1 RU, half-width compute nodes feature the Intel® Xeon® 6 SoC with Performance-cores with 12-/20-/32-core options. This node features up to 768 GB of RAM.

AI Compute Provided by Intel® Xeon® 6 Processors

The Intel Xeon 6 SoC with Performance-cores provide real-time CPU-native inference performance for the Cisco Unified Edge compute node. Intel® Advanced Matrix Extensions (Intel® AMX) are built into each Intel Xeon 6 processor core and accelerate inferencing and model training functionality. Intel® AMX significantly enhances inference performance and computational efficiency directly at the processor level eliminating the need for a costly GPU and helping to lower total costs of ownership (TCO).

This enables AI model deployment at the edge, providing the compute that can reduce latency and enhance performance for end-user applications.

Up to five of these compute nodes can be housed inside a Cisco UCS XE9305 Chassis, delivering enterprise-class performance, versatility, and density in a compact edge-optimized form factor.

The Intel® Xeon® 6 processor family has devices with up to 126 Performance cores with support for the latest DDR5 memory and fast inter-socket communications which makes it ideal for AI applications.

Cisco Unified Edge also makes use of the OpenVINO™ toolkit, developed and maintained by Intel, that provides hardware-optimized software acceleration of AI models that is specifically optimized for Intel® architecture, ensuring AI models seamlessly utilize the Intel® processor's integrated AI acceleration features for optimal performance and efficiency.

Zero Touch Security

Cisco Unified Edge offers multi-layer, zero-trust security features to protect data, application and the AI models.

A vital part of the system's embedded data security capability is using Intel® Confidential Computing Solutions to prevent unauthorized access to data by isolating and segmenting the system to limit a hacker's access to data and resources. The Cisco Unified Edge platform is configured to use Intel® Trust Domain Extensions (Intel® TDX) for VM-level isolation and also uses Intel® Software Guard Extensions (Intel® SGX) to provide application-level isolation.

For physical security, the system has built-in anti-tampering features and physical locks. To avoid configuration drift, policy-based templates are in place to ensure the same security features are deployed globally. Audit trail functionality ensures resilience and safeguards compliance as deployments scale.

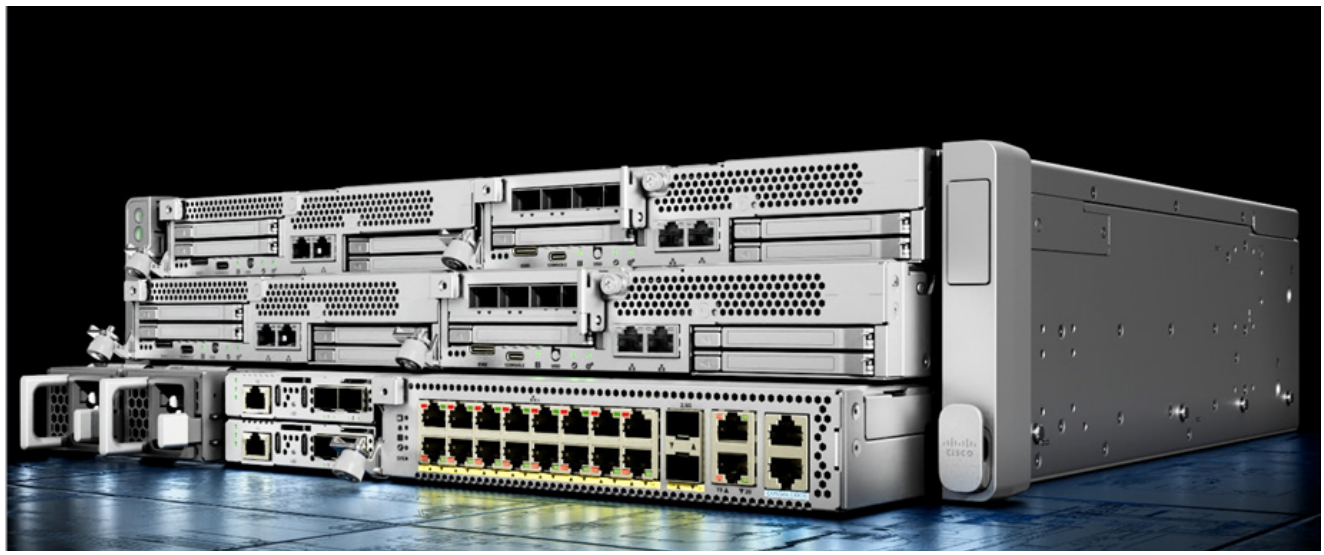

Figure 1. Front view of Cisco Unified Edge platform. The 3RU-high chassis has five slots for compute or networking nodes. It is designed for edge AI applications.

Cisco Unified Edge has a variety of customizable security features, such as the option to protect Kubernetes environments by encrypting the communications between workflows, and the ability to detect and block attacks that are launched to trick, steal or misuse the AI model, among other customizable security features.

Zero-Touch Management

System management and orchestration are provided by Cisco Intersight, which delivers unified operations from edge to cloud. Intersight fleet management enables rapid onboarding and scaling with zero-touch provisioning, automation, and a consistent operating model, simplifying deployment and management across distributed edge sites.

With Cisco-curated blueprints, you can define qualification criteria to match systems to pre-defined configurations, ensuring systems are deployment-ready the moment they're claimed in Intersight. These blueprints enable consistent and repeatable deployments through code, ensuring a single 'golden configuration' can be applied across multiple systems, drastically reducing manual effort and inconsistencies.

Software Ecosystem Approach

Cisco Unified Edge customers can choose their own infrastructure software, which gives them flexibility and integration with existing systems. At launch, Nutanix, Red Hat, VMware, and SUSE have been pre-validated.

Currently, Intel and Cisco are creating a catalog of pre-verified, optimized ISV applications that are certified on the Intel Xeon 6 processor. By working closely with ecosystem partners, customers can customize their systems and shorten the deployment process.

The Edge Solution for Retailers

Modern retail increasingly relies on technology to improve the customer experience and to tightly manage inventory and gain other operational efficiencies. The Cisco Unified Edge platform helps these retailers with workflows that require powerful compute such as:

- AI-Powered Personalization: Drives real-time recommendation engines, personalized promotions, and AI-based customer service bots.
- In-Store Analytics: Processes video feeds and sensor data for foot traffic analysis, shelf monitoring, and loss prevention.
- Supply Chain and Logistics: Optimizes inventory, demand forecasting, and route planning using data analytics.
- Store Operations: Handles localized AI and data processing at the store level, reducing latency and cloud dependency for tasks like smart POS and security.
- Database and Analytics: Powers transactional databases (MySQL, MongoDB) for fast sales reporting and customer data analysis.
- Next-Gen POS and Checkout: Supports faster, more intelligent checkout experiences with integrated AI and analytics.

Cisco Unified Edge delivers AI performance, security, and operational simplicity that retailers need to support these edge workloads. It brings compute power, networking, security, storage, observability, and centralized cloud management together in a scalable, future-ready design.

Conclusion

The rapid rise of edge computing marks a fundamental shift in how organizations deploy and consume AI. As data generation accelerates and real-time inference becomes essential across industries, enterprises need platforms that combine performance, security, and manageability at scale. The Cisco Unified Edge platform, powered by Intel Xeon 6 processors and SOC, brings these capabilities together in an integrated system designed specifically for demanding, distributed AI workloads.

By unifying compute, storage, networking, and security, the platform enables organizations to confidently run mission-critical applications across thousands of locations. Its built-in AI acceleration, hardware-level security protections, and zero-touch provisioning remove many of the operational barriers that have traditionally slowed edge adoption. As a result, businesses can deploy AI-driven services faster, maintain consistent security policies, and scale operations without adding complexity.

While improving retail experiences is a key use case, the Cisco Unified Edge platform can be deployed to solve compute and AI challenges in industrial operations, healthcare, financial and other similar applications. The Cisco Unified Edge platform delivers the high availability and low latency required for next-generation edge use cases. As enterprises navigate the next decade of AI innovation, solutions like this will play a pivotal role in ensuring the edge becomes not just a new deployment location, but a strategic advantage for organizations ready to embrace real-time intelligence.

Learn More

Cisco Unified Edge

Cisco portfolio for retail

Intel® Xeon® 6 Processors Product Brief

OpenVINO™ Toolkit

Intel® Industry Solutions Builders

Retail Builders Community | Intel® Industry Solution Builders

Healthcare and Life Sciences Builders Community | Intel® Industry Solution Builders

Manufacturing - Industrial Builders Community | Intel® Industry Solution Builders

[1] https://blogs.cisco.com/partner/beyond-the-cloud-partners-secure-ai-edge-opportunity-with-cisco

Notices & Disclaimers

Performance varies by use, configuration and other factors.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel optimizations, for Intel compilers or other products, may not optimize to the same degree for non-Intel products.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

See our complete legal Notices and Disclaimers.

Intel is committed to respecting human rights and avoiding causing or contributing to adverse impacts on human rights. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to adverse impacts on human rights.

0126/CG/DJA/PDF Please Recycle 368010-001US