Solution Brief

Industry 4.0
Edge Computing

NetApp, Supermicro Deliver Low-Latency Thick Edge Computing

NetApp® virtualization, data management software enables next-gen AI container-based workflows to run with legacy VMs. Supermicro servers optimize storage, convergence using 3rd generation Intel® Xeon® Scalable processors.

The next generation of on-premises edge computing technologies for increasingly intelligent services is ramping up, and momentum is growing for both higher throughput computing and networking technologies.

Artificial intelligence (AI), internet of things (IoT), data security, and use cases in manufacturing, health care, transportation, retail, and others rely on ultra-low-latency computing to generate and process data to meet the needs of real-time applications.

For example, in manufacturing, data from sensors analyzed in near–real time can help enable robots to more efficiently run a production line. In health care operations, such data can help medical professionals improve how they share and collaborate on data for patient care and make faster decisions to support positive outcomes. In transportation, real-time data can help enable better flow of traffic and help improve the safety of road conditions. In retail, checkout-free stores or drive-up windows enable a more personalized customer experience. The use cases for low-latency computing are wide and vast.

Thick Edge Servers and Convergence

Edge computing describes servers located outside a data center or telecom central office designed to deliver low-latency computing by offering cloud services close to where data is created. Industry analyst firm Gartner predicts that by 2025, 75 percent of enterprise data will be processed outside of a traditional centralized data center.[1]

For some use cases, where data volumes exceed the capacity or cost effectiveness to move to a data center, or applications are highly sensitive to latency, edge servers need to be located at or near a customer location. The first on-premises applications to become popular were universal customer premises equipment (uCPE) servers that were cost-effective systems for branch office networking. In this application, single-function networking appliances are replaced by virtualized network functions converged on a single system.

The advent of 5G and the success of cloud technologies (see Figure 1) has driven the need for “thick edge” servers that provide a cloud computing infrastructure for compute intensive workloads.[2]

Thick edge servers decentralize data, moving compute resources closer to the source for reduced transport times and very low latency.

Digital transformation empowered by edge computing requires convergence between information technology (IT), operational technology (OT), and communications technologies (CT) at the thick edge. Industry analyst firm IDC predicts that more than 90 percent of new operational processes will be deployed on edge infrastructure by 2024.[3] AI applications are a big part of this convergence, as high latencies can hinder applications requiring real-time data processing that are enabled by AI.

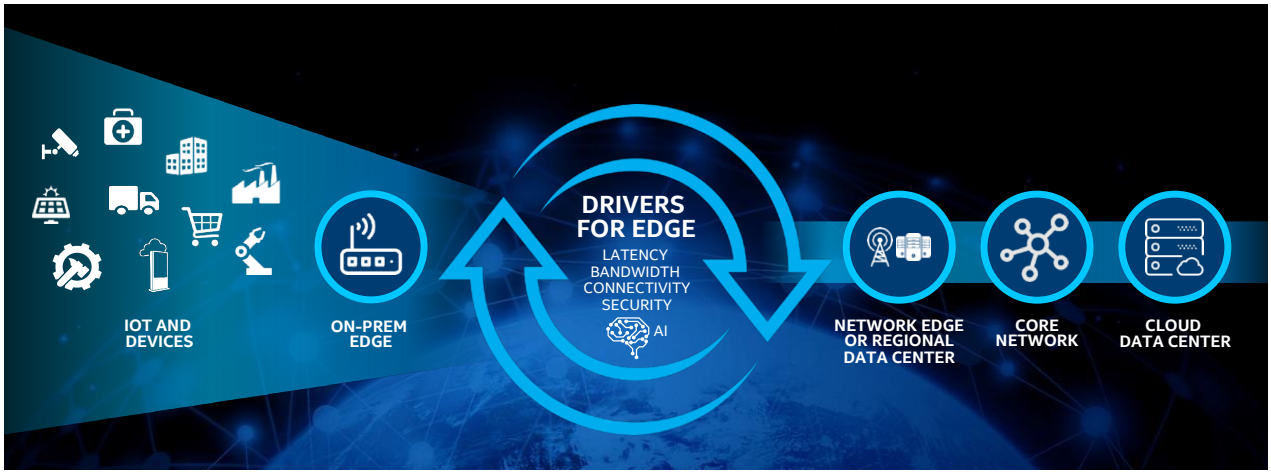

Figure 1. Drivers for edge networking.

Supporting Cloud Security Services

There are numerous proof points that illustrate the value of IT and OT convergence. Incorporating operational insights directly with enterprise business systems can yield better informed decisions for managing costs. For example, traditional OT analysis of equipment maintenance strategies, real-time inventory optimization, and vendor and partner management can yield cost savings when integrated into an IT analytics program.

These systems, however, have historically been siloed and separate, and their convergence requires new attention to security. Gartner notes[1] that a challenge introduced by deploying on-premises servers is the potential to increase cybersecurity threats due to an expanded attack surface. This makes delivering cloud security services very important.

Secure access service edge (SASE) is an emerging service that brings together software-defined wide area networks (SD-WAN) with network security functions. First deployed as a cloud service, SASE is now deployable from a thick edge cloud server.

SASE security functions may include zero trust network access (ZTNA), firewall as a service (FWaaS), secure web gateway (SWG), and cloud access security broker (CASB). Most functions are delivered "as a service." SASE uses factors such as identity (person, device type, edge location, etc.), real-time context, and security policies to deliver the security function. Gartner, who coined the term SASE, predicts that by 2024 at least 40 percent of enterprises will have plans to adopt SASE.[4]

Targeting this application, Supermicro, an Intel® Network Builders ecosystem partner, has collaborated with NetApp to build a high-performance thick edge server. The solution utilizes NetApp storage with AI expertise running on Supermicro's high-performance, Intel® architecture-based servers.

NetApp Enables AI for Thick Edge

NetApp offers an AI processing system designed for enterprises that are deploying advanced AI workloads to support digital transformation and industry 4.0 initiatives. These enterprises are facing the recurring challenge of deploying these systems without increasing complexity, driving up support costs, adding yet another unique process to maintain, or disrupting existing operations. Meeting this simplification challenge requires platform reconciliation through the use of virtualized and containerized applications running on Intel architecture servers.

When combined with dynamic load balancing, this virtualized infrastructure offers a scalable infrastructure that can re-allocate resources depending on application needs. Virtualization can also enable this collection of applications to be run over a heterogeneous cluster of edge deployed servers that provide necessary compute, GPU, storage, and networking resources. This infrastructure enables both analytics and IT infrastructure resource nodes as required to better manage the systems and the data flows.

The server clusters can be built using legacy and new hardware, with newer edge nodes delivering more compact physical nodes with an enhanced capability to scale to handle the increasing demands of evolving AI applications.

NetApp is designed to process massive AI data flows without bottlenecks. Using NetApp's proven AI technology enables more efficient data collection, accelerated AI workloads, and smoother cloud integration. Key features of the NetApp AI solution include:

- **Streamlined data flows:** NetApp has features that streamline the flow of data reliably and speed up analytics, training, and inference with a data fabric that spans from edge to core to cloud.
- **Deliver the right performance and scalability:** Deep learning (DL) training routines demand massive amounts of compute power. Faster image training can cut down on overall compute costs while accelerating AI innovation and productivity.
- **NetApp AI key benefits:** Helps reduce risk with flexible, validated solutions that feature reduced design complexity and guesswork for fast implementation and comes configured to streamline deployment.

The AI technology offers nondisruptive system growth with an integrated pipeline that intelligently manages data from edge to core to cloud. The technology unifies AI workloads to eliminate infrastructure silos and to flexibly respond to business demands.

NetApp Manages Thick Edge Data Storage

NetApp enables the storage environment with powerful data management capabilities, proven storage efficiencies, leading cloud integration, and security features that span flash, disk, and cloud storage with the ability to integrate future technologies as well. Support for multiple storage architectures, including hardware storage systems, software-defined storage (SDS), and the cloud is built into the NetApp solution.

NetApp's FabricPool functionality is built into the system enabling automatic tiering of cold data and hot data. FabricPool works with cloud-based external object stores to create a capacity tier for cold (inactive) data, while hot (active) data is kept on SSDs in the server. This maintains high performance, reduces storage costs, and frees up space on existing NetApp storage systems in order to consolidate more workloads.

For storage efficiency, NetApp has inline data compression, deduplication, and compaction features that work together to reduce storage costs and maximize data storage. Additional efficiency technology includes support for NetApp Snapshot, a thin provisioning, replication, and cloning technology.

NetApp Is Powered by Supermicro

To power these advanced AI workloads and deliver the performance storage required, Supermicro is providing Intel architecture-based thick edge hardware that offers performance for AI inferencing and training workloads and for other low-latency thick edge applications. The Supermicro E403 family of servers are the foundation of NetApp's IoT edge strategy. The servers feature a range of cost-performance ratios and provide the ability to implement AI functions with GPU/VPU cards and introduce PCIe version 4.0 for higher speed connectivity between acceleration cards and the CPU.

With Supermicro edge servers, enabling these functions in the same form factor reduces environmental changes, and lowers deployment risk while improving capacity and capabilities.

For optimal performance, NetApp software can run on the Supermicro E403-12P-FN2T, which features a single-socket 3rd generation Intel Xeon Scalable CPU with up to 32 cores, three PCI-E 4.0x16 slots, two 10 Gigabit Ethernet ports, four 2.5" internal SATA drive bays and four USB 3.0 and two USB 2.0. The server is a three-slot compact chassis edge server measuring 4.3" high x 10.5" wide x 16" deep. The system includes 8 DIMM slots with up to 2 TB DRAM.

For cost-sensitive applications, the Supermicro E403-9D family of high-performance edge servers offers the low-power and high-performance benefits of Intel Xeon D processors. The edge servers measure 4.3" x 10.5" x 16" and include two 10GBase-T LAN ports, two SFP+ 10G ports, and nine GbE LAN ports. The edge servers also feature a dedicated Intelligent Platform Management Interface (IPMI) for management and monitoring capabilities. The servers can be expanded through two PCI-E 3.0 x16 slots, two PCI-E 3.0 x8 slots, and one PCI-E 3.0 x16 slot.

NetApp Industry 4.0 Use Case — Weld Integrity Analytics

Industrial applications are ideal for low-latency services served from a thick edge server. One example is manufacturing using multiple types of welding and assembly techniques. Certain types of welding require a high degree of post-weld quality, and using AI for quality assurance provides confidence of weld integrity for critical parts.

Welding manufacturing for automotive, aircraft, and other metal-to-metal applications using semi-automatic welding processes is ideal for integrating robotics. Robotics allow for ultra-fast processing and flexible positioning, and don't expose humans to fumes and other hazardous welding conditions.

As part of a pilot program with Supermicro, a customer leveraged the NetApp solution across hundreds of robotic systems across many factories. For each robot, there are multiple workloads managing simultaneous AI functions in real time. The systems are configured in a high availability (HA) mode enabling 24/7/365 production capability.

Figure 2. Data flow for robotic welding thick edge server application.

3rd Generation Intel® Xeon® Scalable Processors

- **Flexibility from the edge to the cloud,** bringing AI everywhere with a balanced architecture, built-in acceleration, and hardware-based security.
- **Part of a complete set of network technology from Intel,** including accelerators, Ethernet adapters, Intel® Optane™ persistent memory, FlexRAN, Open Visual Cloud, and Intel® Smart Edge.
- **Engineered for modern network workloads,** targeting low latency, high throughput, deterministic performance, and high performance per watt.
- **Enhanced built-in crypto-acceleration** to reduce the performance impact of full data encryption and increase the performance of encryption-intensive workloads.
- **Hardware-based security** using Intel® Software Guard Extensions (Intel® SGX), enhanced crypto processing acceleration, and Intel® Total Memory Encryption.[5]

The NetApp solution proved it could deliver the AI processing needed for this customer. Adding additional AI workloads onto a single server running multiple inferencing models saves not just capital costs and networking needs but also improves the real-time weld quality and flexibility. The thick server along with the inferencing, removes the need to validate the quality through multiple human inspection checks where potential flaws cannot be detected by the human eye. The result is very fast production with improved quality.[6]

Conclusion

Exciting new AI, IoT, data security, and other use cases across a range of vertical industries rely on ultra-low-latency computing to generate and process data to meet the needs of real-time applications. Decentralizing data for processing at the thick edge is emerging as a way to support low-latency requirements.

Digital transformation of organizations to empower thick edge computing requires convergence between IT, OT, and CT across key workloads such as AI or security. NetApp and Supermicro have built a high-performance thick edge solution that utilizes NetApp storage and AI expertise running on Supermicro's high-performance, Intel architecture-based servers to meet the needs of these converged environments and workloads.

Learn More

Intel® Network Builders

Supermicro

NetApp

3rd Gen Intel® Xeon® Scalable processors

Notices & Disclaimers

[1] https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders/
[2] https://project.linuxfoundation.org/hubfs/LF%20Edge/StateoftheEdgeReport_2021.pdf
[3] https://enterprisersproject.com/article/2020/12/8-edge-computing-trends-2021
[4] https://blogs.gartner.com/andrew-lerner/2020/01/06/networking-predictions-2020-edition/
[5] These technologies are not supported when using Intel® Optane™ persistent memory
[6] Data provided by NetApp, June 2021.

Intel technologies may require enabled hardware, software or service activation.
No product or component can be absolutely secure.
Your costs and results may vary.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

0621/DO/H09/PDF ♻ Please Recycle 347430-001US