Hewlett Packard Enterprise | intel

Solution brief

HIGH-PERFORMANCE, ENERGY-EFFICIENT AI INFERENCE AT THE EDGE

Speed AI results with HPE Edgeline and Intel Arctic Sound-M ATS-M75 and ATS-M150 PCIe accelerators

Sample use cases

- AI visual inference
- Media AI analytics
- Machine learning (ML)
- Simulation and visualization
- Video transcoding and streaming
- Virtual desktop infrastructure
- Mobile cloud gaming
- PC cloud gaming
- Health and life sciences: medical imaging
- Manufacturing/automotive: quality assurance
- Financial services: fraud detection
- Retail loss prevention
- Warehouse robotics and inventory monitoring
- Safety and security

WHAT IS AI INFERENCING AT THE EDGE?

Artificial intelligence (AI) is often considered the domain of high-performance computing (HPC) clusters or supercomputers running in the data center and cloud to help data scientists develop and train deep learning (DL) models.

AI inferencing at the edge refers to deploying trained AI models outside the data center and cloud—where the data is created and can be acted upon quickly to generate business value. These edge AI solutions place the compute infrastructure closer to the source of the data, and closer to the systems and people who need to make data-driven decisions in real time.

While growth in AI development and training remains robust, industry projections show that the AI inference market will grow rapidly to tens of billions of dollars by the mid-2020s, with most of the growth in AI inference happening at the edge.

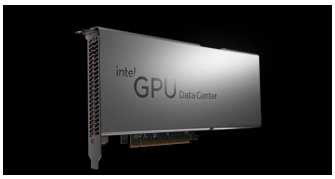
AI inference workloads span an array of use cases across many industry verticals, such as healthcare, financial services, and manufacturing. These workloads are often compute- and data-intensive, requiring high-speed AI hardware. Thus, edge AI inference hardware is optimized to deliver capabilities not always found in the data center, such as:

- **Speed**—Processing data closer to the source, edge computing greatly reduces latency. The result is higher speeds that enable real-time use cases.
- **Security**—Critical data does not need to be transmitted across different systems. User access to the edge device can be very restricted.
- **Scalability**—Edge devices are endpoints of an AI system that can grow without performance limitations. This allows you to start small, with minimal costs, and scale as needed.

HPE Edgeline

- HPE Edgeline EL8000 and EL8000t System enclosures
- HPE Edgeline e920, e920d, e920t Server Blades, with Intel Xeon Scalable processors

hpe.com/edgeline



Intel ATS-M75

- PCIe Gen4
- HHHL form factor
- 75-watt TDP

Intel ATS-M150

- PCIe Gen4
- FH 3/4L form factor
- 150-watt TDP

intel.com

The use of accelerators that are specialized for AI workloads can further increase the speed of an AI model, enabling the server to run AI inference tasks more efficiently than with a conventional processor alone.

HPE and Intel® are working together to tightly integrate and optimize performance of the HPE Edgeline EL8000 Converged Edge Systems with the new Intel Arctic Sound-M ATS-M75 and ATS-M150 PCIe accelerators.

SOLUTION COMPONENTS

HPE Edgeline Converged Edge Systems
HPE Edgeline Converged Edge Systems put enterprise-class compute, storage, networking, security, and systems management at the edge. Built on the same technology as data center systems, HPE Edgeline delivers enterprise IT capabilities in a ruggedized, compact form factor designed for the harsh operating environments found at the edge.

The tested solution is built with HPE Edgeline EL8000 and EL8000t systems, which are purpose-built with an optimized size, weight, and power profile to deliver unprecedented levels of compute, storage, and networking performance at the edge.

The EL8000 is a compact, toolbox-sized (17" deep, 5U, half-rack width) bladed system that supports up to four independent server blades clustered together using dual-redundant chassis integrated switches. It has a maximum capacity of 8x 350 TOPS. The EL8000t is also compact (17" deep, 2U, full-rack width) and supports two independent server blades.

The EL8000/EL8000t systems leverage the HPE Edgeline e920 family server blades, which use the same Intel® Xeon® Scalable processors as mainstream data centers. This enables fast local SSD storage along with support for Intel Arctic Sound-M ATS-M75 and ATS-M150 PCIe accelerators.

Intel Arctic Sound-M ATS-M75 and ATS-M150 PCIe accelerators
Intel Arctic Sound-M ATS-M75 and ATS-M150 PCIe accelerators bring the industry's first hardware-based AV1 encoder into a GPU to provide 30% bandwidth improvement and includes the industry's only open-source media solution. The media AI analytics supercomputer enables leadership transcode quality, streaming density, AI visual inference, content creation, mobile cloud gaming, PC cloud gaming, simulation and visualization, ML, and virtual desktop infrastructure (VDI).

ENABLE AI AT THE EDGE

Contact your authorized HPE sales representative to find out how you can start speeding AI inferencing workloads at the edge, today.

LEARN MORE AT

hpe.com/ai

Hewlett Packard Enterprise

a50006385ENW