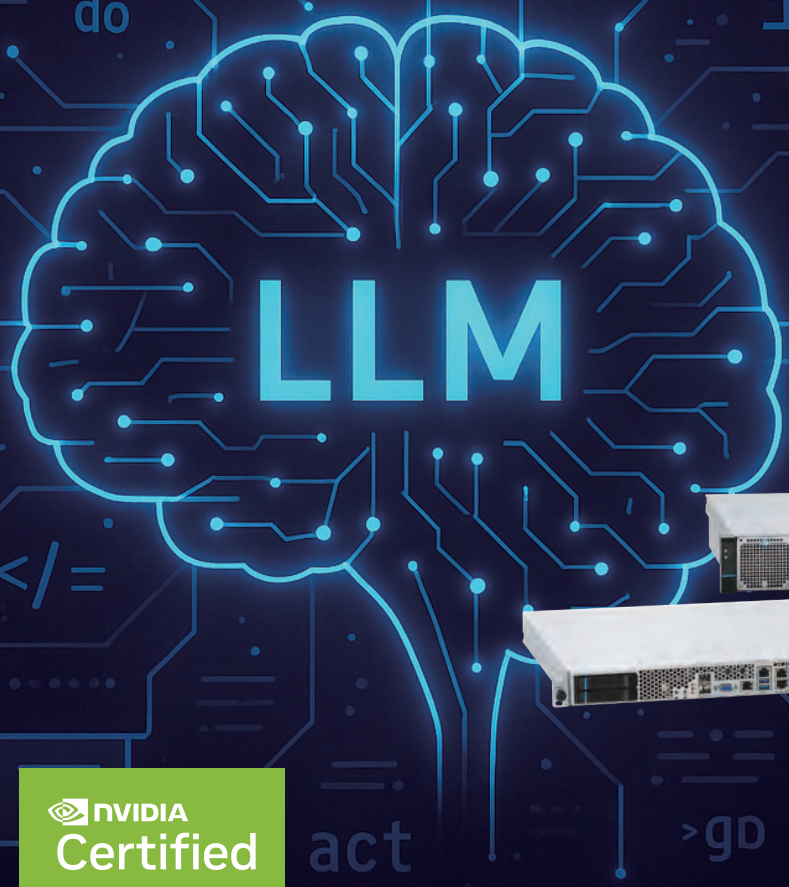
Accelerate LLMs at the Edge

Maximize Savings in TCO, Time, Space, Risk



AXE

NVIDIA Certified

ADLINK
LEADING EDGE COMPUTING

Full Edge Data Processing, Zero Cloud Dependency

As enterprise AI moves from centralized cloud to on-premises edge computing, the demand for powerful, flexible, and secure GPU servers grows. ADLINK AXE Series fulfills this need by enabling full AI processing at the edge, delivering real-time insights with maximum privacy—without cloud dependency.

Most AXE Series servers are NVIDIA Certified, guaranteeing reliable, compatible, and optimized performance for AI workloads, having passed rigorous testing for seamless operation, fast deployment, and stability.

AXE Series also accelerates edge deployment of LLMs—such as DeepSeek and Llama—for low-latency, secure AI services. Additionally, it significantly reduces total cost of ownership (TCO), development time, space, and data leak risks, empowering efficient AI applications in medtech, smart manufacturing, and on-premises agentic AI.

Flexible Expansion

Supports up to **4** full-length,dual-slot PCIe x16 GPUs — scale without chassis swaps, reducing upgrade costs

Superior Performance

High tokens/s, low latency on mainstream LLMs, shortening tuning and rollouts time

Short Chassis

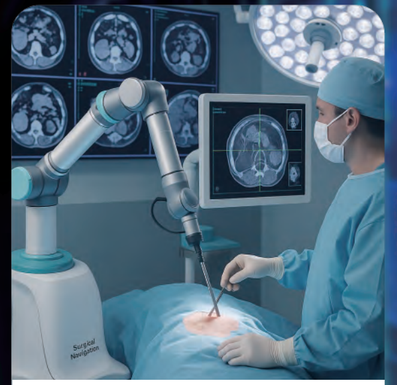
Short-depth chassis design saves up to **50%** of space, comparing to conventional servers

100% On-Prem

Keeps data on-site—no cloud dependency, minimizing risk of leaks

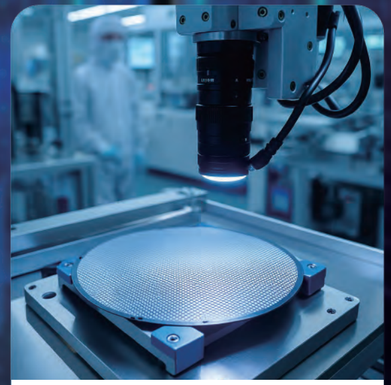
Applications

MedTech



- Protects patient data on-site for HIPAA compliance
- Enables real-time AI diagnostics without cloud delays

Smart Manufacturing



- Cuts downtime and IT costs with on-site AI
- Delivers fast, precise defect detection for higher yield

On-Premises Agentic AI

- Enables instant updates of on-prem models (e.g., DeepSeek, Llama) for agile workflows
- Keeps proprietary data secure, avoiding cloud exposure

Micro LED AI AOI

Challenges

Micro LED sizes are typically <100μm, with mainstream dimensions ranging between 30μm and 85μm, and the potential to shrink below 10μm. For an 8K display (7680x4320x3 RGB), around 100 million LEDs are needed, making precision and efficiency in mass transfer a major challenge.
If the mass transfer yield rate of Micro LED is 99.5%, the number of defective wafers on an 8K display will be as high as 497,664.
Inspecting such a large number of wafers in a short period of time is a difficult task.

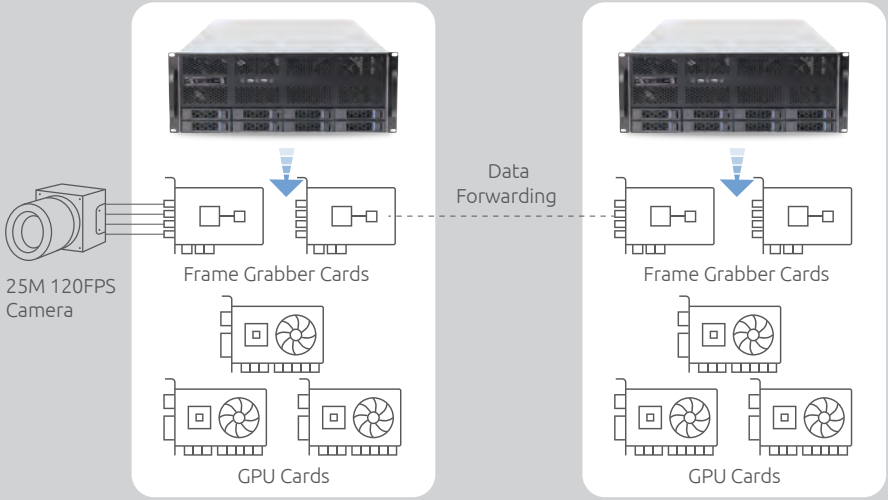
Solution

ADLINK data forwarding technology integrates the ADLINK AXE server, frame grabber card, and GPU cards to provide a high-speed AI AOI solution, solving the bottleck in image processing speed.
This solution is flexible, with the AXE server supporting up to 3x PCIe x16 FHFL double-deck GPU cards and 2x PCIe x8 frame grabber cards. Depending on the project scales, it can extend the GPU processing capability by adding compactly designed AXE servers.

High Speed AI AOI

AI AOI enhances traditional AOI by using machine learning and deep learning algorithms to improve the accuracy, speed, and adaptability of defect detection processes in manufacturing, particularly in semiconductors. It reduces false positives and negatives, increases inspection speed, and enhances overall production efficiency. Take Micro LED as an example: as the size gets smaller and the quantity increases, the inspection process becomes more complicated and is time consuming.

Diagram



Factory AI Knowledge Base

Challenges

In manufacturing project management, all sensitive cost and product data must stay inside the plant. Yet supplier capability surveys remain highly manual: managers sift through multiple document versions, compare revisions, and contact different departments for confirmation. This makes the process slow, error-prone, and fragmented. Meanwhile, legacy racks and tangled cabling undermine system stability, adding risk to already complex workflows.

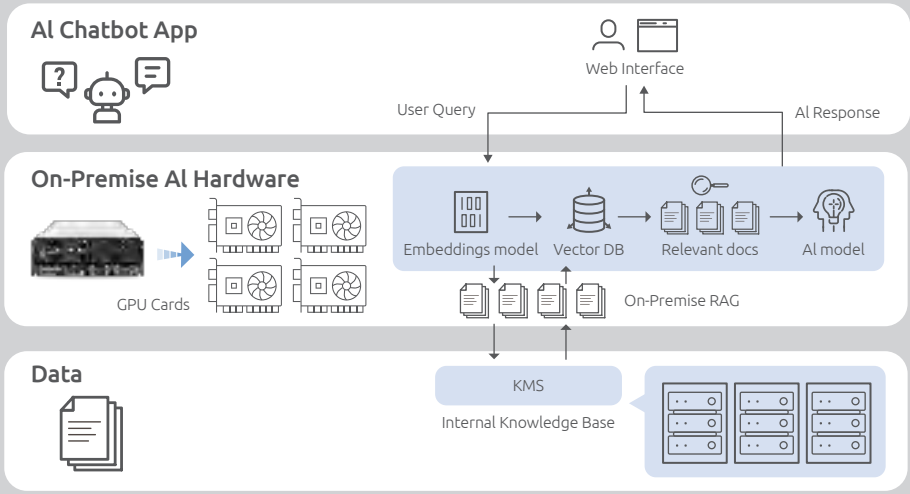
Solution

We deploy short-depth edge servers to boost flexibility and airflow, reducing instability from cluttered cabling. Running 100% on-prem preserves network segmentation and ensures audit compliance. Benchmarks on mainstream LLMs show excellent performance—high tokens per second and low latency—so sales and product teams receive a reliable SLA that supports fast, stable operations.

On-Premise AI Solution

On-premise AI applications run AI models locally, ensuring high data privacy, security, and low latency. They allow enterprises to protect sensitive data while benefiting from fast AI inference, suitable for real-time image processing and natural language tasks.

Diagram



CT / MRI

Challenges

In MedTech, a key challenge is the need for compact edge AI servers that can be deployed in constrained clinical environments.
Since applications often involve surgery or diagnostics, image recognition must be both highly accurate and real-time to support critical decisions.
Modalities such as CT and MRI particularly demand precise, timely image processing to ensure clinical reliability and patient safety.

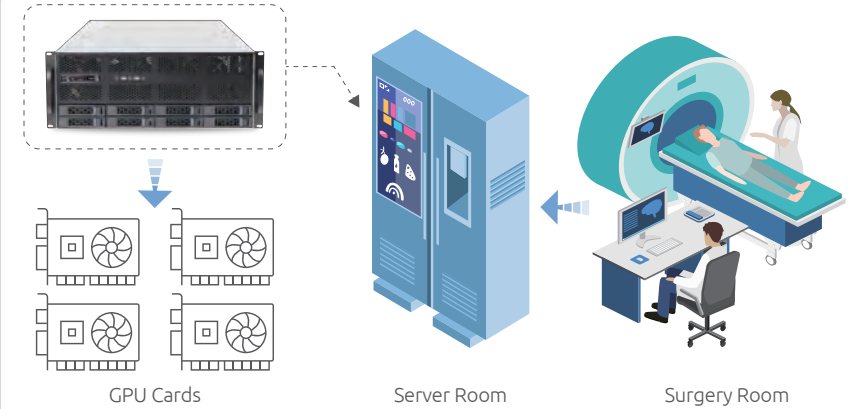
Solution

The ADLINK AXE server, with up to 4 GPUs, provides an efficient solution by enabling rapid processing of complex medical images. It handles large datasets effectively, supports the alignment of multiple scans for 3D reconstruction, and enhances model interpretability, addressing key computational challenges in medical image analysis and reconstruction. The compact design can also be embedded into healthcare equipment.

Medical Image Processing, Reconstruction & Analysis

Medical image processing, reconstruction and analysis play a vital role in healthcare by enhancing diagnosis and treatment through advanced imaging techniques. Image processing improves image quality, while reconstruction transforms 2D images into 3D views, enabling clearer anatomical visualization. Analysis algorithms detect and measure features like tumor size, supporting clinical decisions in systems such as CT, MRI, and C-ARM, ultimately improving accuracy and delivering personalized patient care.

Diagram



AI GPU Servers

NVIDIA Certified

AXE-7400SR

AXE-7220SR

AXE-7120SR

AXE-6120IL

AXE-5100R7

	AXE-7400SR	AXE-7220SR	AXE-7120SR	AXE-6120IL	AXE-5100R7
U/Depth (mm)	4U / 650	2U / 450	1U / 420	1U / 420	1U / 440
CPU Type	4th/5th Gen Intel® Xeon® Scalable	4th/5th Gen Intel® Xeon® Scalable	4th Gen Intel® Xeon® Scalable	Intel® Xeon® D-1700 Series Processors	AMD Ryzen 7000 Processors
Max GPU Support	4× double-deck GPUs	2× FHFLDW Gen4 GPUs	2× FHFL GPUs Gen4 ×16	1× FHFL GPU Gen3 ×16	1× FH, 3/4LSW GPU Gen5 ×8
Max Memory	2TB DDR5	768GB DDR5	384GB DDR5	192GB DDR4	128GB DDR5
Storage	6x 2.5"/3.5" SATA HDD 1x M.2 NVMe SSD	4x 2.5" SATA SSD 2x M.2 SATA/NVMe SSD	2× M.2 SATA/NVMe SSD	2x 2.5" SATA SSD 2x M.2 NVMe	2x 2.5" SATA SSD 2x M.2 NVMe/SATA
Networking	2× 10G, 2× GbE, 1× IPMI	2× 25G SFP28, 2× GbE	2× 25G SFP28, 2× GbE	8× 10G SFP+, 2× GbE	8× GbE, 1× GbE (IPMI)
Redundant Power	Yes (2× 1600–2000W, 1+1)	Yes (2× 800–1600W, 1+1)	Yes (2× 550W, 1+1)	Yes (2× 450W, 1+1)	Yes (2× 300W, 1+1)
Target Use Case	• MedTech: CT/MRI • Manufacturing: Micro LED	• MedTech: Ophthalmology instant image diagnose • Manufacturing: SMT for PCB components	• MedTech: Bone X-rays for orthopedics • Manufacturing: Real-time detection of scratches	• MedTech: Batch scanning of slides and detection of cancer cells • Manufacturing: Individual inspection stations for contamination	• MedTech: Portable diagnostic devices • Manufacturing: PCB connectors or small electronic parts detection