Deloitte.

intel.

Solution Brief

Data and Communications Equipment Maintenance

Serving AI Assistant for Field Operations

Generative AI enabled real-time maintenance and repair diagnosis solution deployed on Intel® Xeon® 6 Edge Server and Intel® Arc™ Pro GPU

Introduction

The emergence of Large Language Models (LLMs) and Transformer architectures has ushered in the era of Generative AI, representing a paradigm shift beyond traditional pattern recognition. These massive models, containing billions of parameters, excel not only in conversational AI but also in sophisticated text processing tasks including information extraction and content synthesis. Vision-Language Models (VLMs), a specialized subset of LLMs, extend these capabilities to video and image processing, creating powerful multi-modal AI systems that seamlessly integrate visual and linguistic understanding.

Modern edge hardware equipped with built-in AI acceleration now readily supports the execution of sophisticated LLM models locally, eliminating dependency on datacenter or cloud infrastructures. This technological advancement opens unprecedented possibilities for Generative AI based use cases at the edge where actions take place. In particular, the convergence of Convolutional Neural Network (CNN) based visual intelligence with LLM reasoning capabilities enables powerful AI solutions that deliver real-time intelligent decision-making across many vertical sectors like retail, healthcare, manufacturing, energy, and transportation.

Deloitte Consulting LLP and Intel have collaborated to develop a minimum viable product solution that demonstrates the application of Generative AI at the edge using readily available, off-the-shelf edge hardware available from Supermicro. It is a multi-modal AI Assistant solution providing real-time guidance to field technicians when performing diagnosis and repair at remote locations. This paper details the solution architecture and its implementation, which is built on Intel CPU and GPU hardware along with Intel® software components.

Use Case: AI Assistant for Field Operations

Consider the scenario illustrated in Figure 1: A field technician arrives at a remote telecommunications facility to diagnose a failing network appliance. Instead of relying solely on manual troubleshooting, the technician leverages an AI-powered digital consultant accessible through a mobile device. The process begins with the technician capturing images or video of the status indicators on the appliance's control panel, then submitting this visual data to the remote AI assistant for immediate analysis. Within moments, the AI Assistant processes the information and delivers step-by-step diagnostic recommendations, enabling the technician to pinpoint and resolve the issue in minutes — a task that may traditionally require many more minutes of manual investigation.

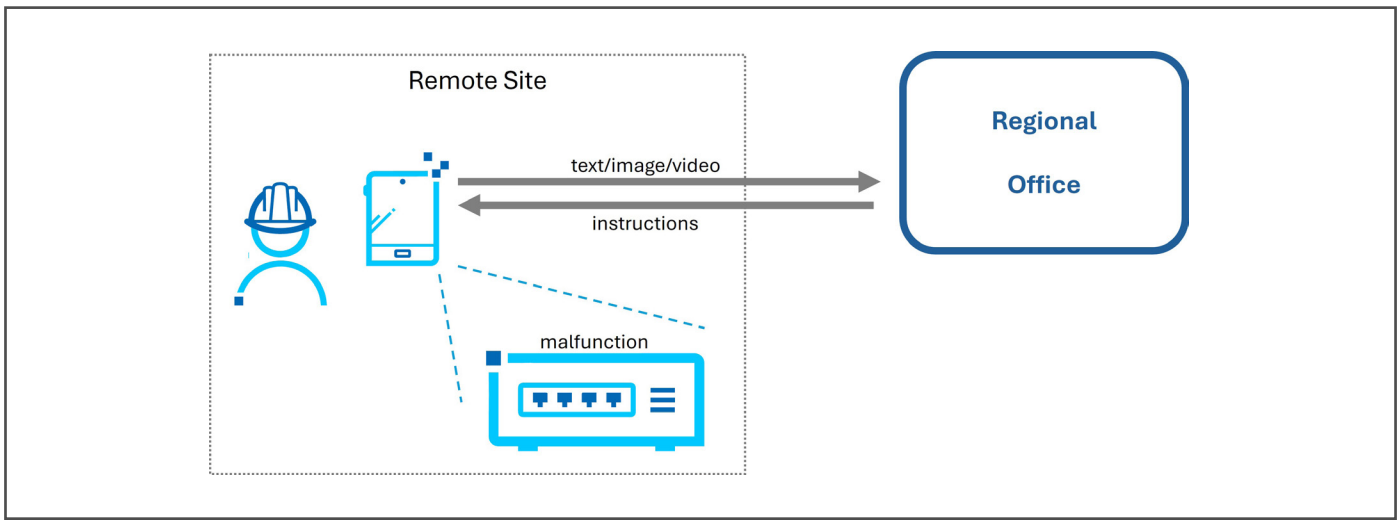

Figure 1. Using generative AI at edge to assist with remote-site field operations.

Architecture

Figure 2 presents the application's functional architecture and mapping to underlying hardware resources. The application is made up of a multi-modal and multi-model composite pipeline. At the front-end is a web-based interface that enables technicians to interact through text inputs and media uploads.

On the server side, an incoming query will trigger a two-step process: issue identification following by remediation generation. In the first step, received media undergoes preprocessing and is analyzed with a computer vision (CV) model: RF-DETR, and interpreted with a Vision Language Model (VLM): Llama 3.2 11b–vision instruct. The results are used to query a Postgres database containing historical context data.

The retrieved options are then returned and presented to the technician for issue identification. In the second step, the identified issue along with the original prompt is fed into a Retrieval Augmented Generation (RAG) construct, composed of an embedding model for parsing, a large language model (LLM) for reasoning, and a Chroma vector store of data built from technical documentation and troubleshooting guides. The LLM model, based on Llama 3.1 3b-language instruct, processes the original prompt in conjunction with relevant text retrieved from the vector store according to the identified issue.

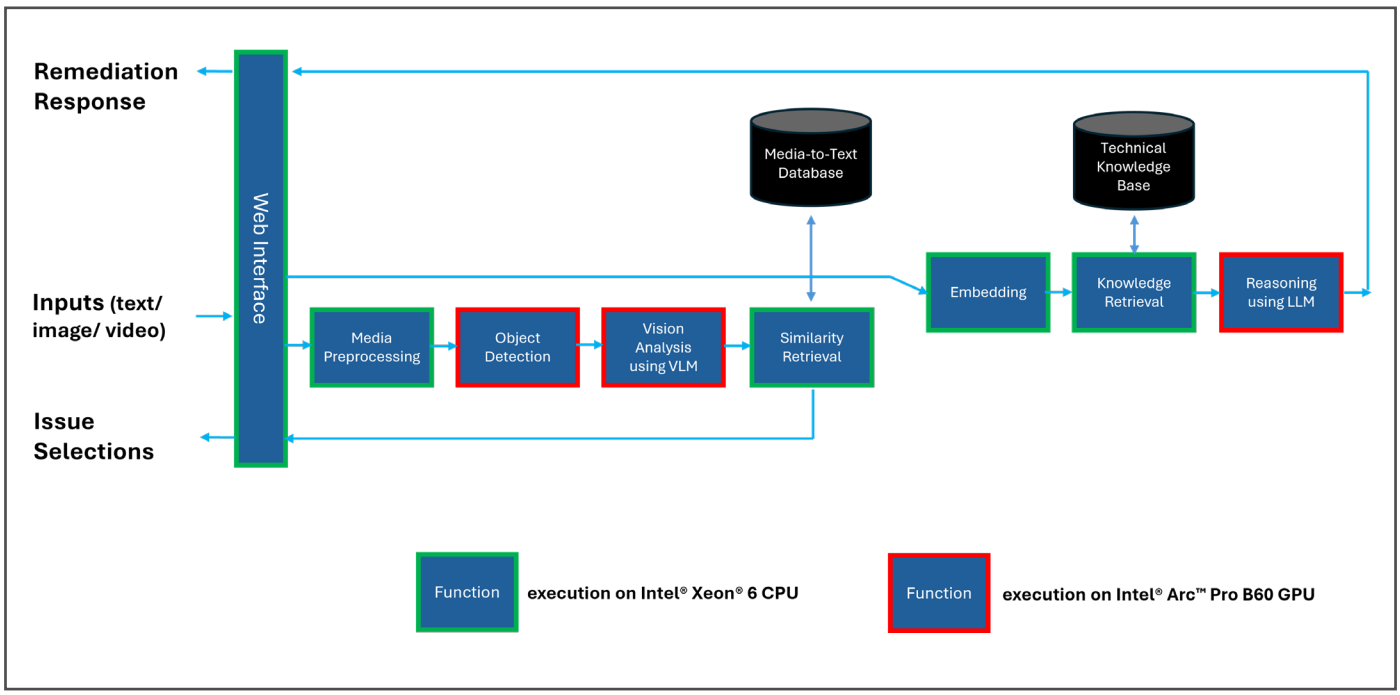

Figure 2. Functional architecture of the Smart Field Operations application.

Supermicro SYS-212B-N2T Edge Server Configuration	
CPU	Intel® Xeon® 6731P (32 cores)
Memory	512GB DDR5 (Dual Channel)
GPU	2 x Intel® Arc™ Pro B60 Cards (with 24GB GDDR6 memory)
OS	Ubuntu 24.04

Figure 3. Configuration of the edge server for deploying the Smart Field Operations application.

Implementation: Edge Hardware

The application is deployed on an edge server located within a regional office. It is a server from Supermicro's SYS-212B-N2T line in a compact 2U chassis with front I/O access and full-height full-length PCIe Gen5 slots, well-suited for edge AI deployment. As detailed in Figure 3, the server configuration includes a single Intel® Xeon® 6731P processor, two Intel® Arc™ Pro B60 GPU cards, and 512GB of system memory.

The Intel Xeon 6731P server processor, from the Intel® Xeon® 6 family, has 32 P-cores at close to 4GHz turbo frequency. With excellent power efficiency, it is utilized here to host databases, and handle support functions such as control logic, user interface, media processing, and data retrieval tasks.

On the other hand, AI inference workloads — including CV, VLM, embedding model and LLM model execution — are offloaded to the two Intel Arc Pro B60 GPUs cards. This GPU is architected for acceleration of parallel neural network computation, achieving close to 200 TOPS at INT8 precision. Additionally, each card is equipped with 24GB of graphics memory, meeting the needs of sizable VLM/LLM models.

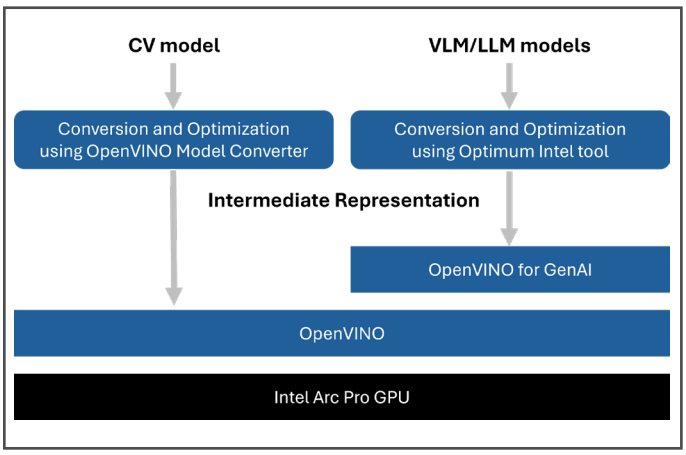

Figure 4. Using OpenVINO tools for model conversion, pipeline instantiation, and model execution.

Implementation: Edge AI Optimization

Figure 4 depicts the software process for AI Inference on the edge server, powered by the OpenVINO framework. First, the models are pre-converted to Intermediate Representation (IR) format and pre-optimized to ensure computational efficiency. While the CV model is passed to directly to OpenVINO runtime, VLM/LLM pipeline and inference are facilitated through the OpenVINO-for-GenAI layer on top of the OpenVINO runtime. The runtime handles loading and executing the IR-formatted models on the Intel Arc Pro GPU.

Performance

Deloitte conducted a performance evaluation of the implementation, which a focus on processing efficiency at both the model and pipeline levels.

First, according to their measurements of individual model execution on a single Intel Arc B60 GPU, the VLM and LLM components each achieved sub-second time-to-first-token latency while maintaining robust token-per-second throughout. It demonstrated the GPU's capability to deliver real-time AI inference for these types of models.

Second, at the pipeline level for issue identification and remediation reasoning, their findings confirmed that in both cases the edge server configuration could readily process and respond to multiple parallel requests with only modest increases in response time. This performance scalability demonstrated the architectural synergy between Intel Xeon 6 CPU and Intel Arc Pro GPU, creating a robust platform for deploying complex Generative AI functions at the edge.

Scaling

The AI Assistance for Field Operations application was deployed and tested on an Intel Xeon server with Intel Arc Pro GPUs. However, the implementation architecture supports flexible scaling across diverse hardware configurations with minimal software modifications. This is made possible with OpenVINO, which provides a unified AI inference runtime targeting Intel client and server processors and accelerators.

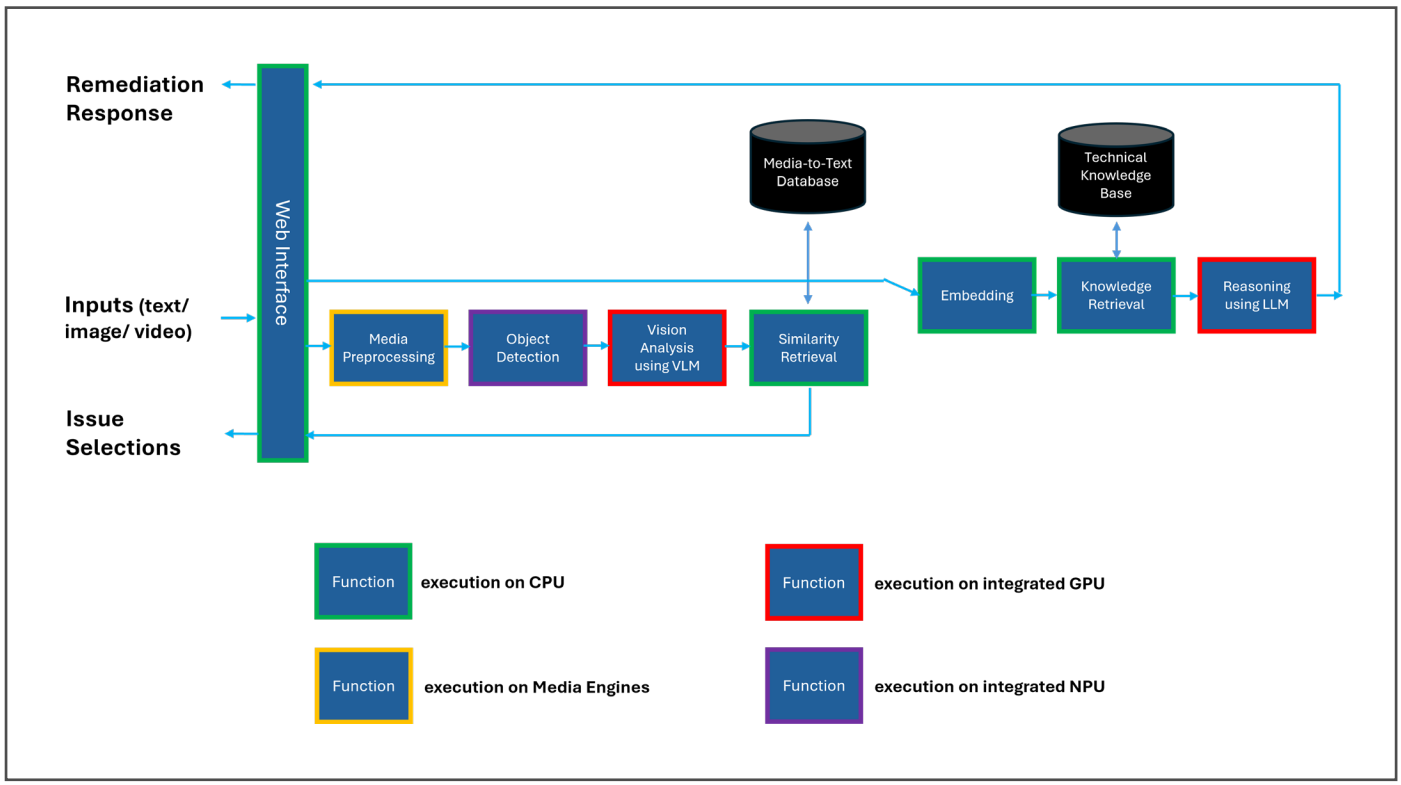

Figure 5. Possible mapping of workloads on Intel® Core™ Ultra processor.

On client platforms, OpenVINO can harness multiple compute engines, including CPU, GPU, and NPU. For example, as shown in Figure 5, if deploying the AI Assistance for Field Operations application on an Intel® Core™ Ultra platform, media pre-processing can be handled by media engines on the integrated GPU, and AI model execution can be offloaded to the integrated GPU and NPU, while using CPU resources exclusively for database hosting, web interface operations, and data retrieval.

On Intel Xeon server platforms without discrete GPU presence, OpenVINO can execute AI models on the CPU leveraging its Advanced Matrix Extensions (AMX) capability.

For performance scale-up, client or server systems can incorporate one or more Intel Arc Pro GPU cards to handle and speed up heavy AI workloads.

Conclusion

The AI Assistant for Field Operations solution empowers service teams to harness digitized expert knowledge, enabling accelerated issue diagnosis, higher first-time-fix rates, reduced repeat service visits, and more consistent execution across field operations. It demonstrates the transformative potential of deploying sophisticated Generative AI capabilities at the edge using commercially available Intel®-based hardware, such as Supermicro's SYS-212B-N2T rackmount server. By successfully integrating Vision Language Model with Large Language Model in a unified multi-modal architecture, this implementation showcases how modern edge computing can deliver enterprise-grade AI performance without relying on cloud computing. The solution's ability to process complex visual and textual data locally while maintaining fast response time represents a significant advancement in edge AI deployment.

OpenVINO's unified inference runtime enables seamless adaptation across deployment scenarios — from compact client devices with integrated GPU and NPU to high-performance servers with multiple discrete GPUs. This scalability, combined with demonstrated performance metrics, validates edge-based Generative AI for mission-critical field operations.

As organizations prioritize data sovereignty and operational independence, this AI Assistant provides a blueprint for intelligent edge applications that operate reliably across diverse environments. By leveraging readily available hardware and an open software stack, the solution delivers performance scalability with significant total cost of ownership (TCO) advantage.

Learn More

Intel® Xeon® 6 Processors

Intel® Arc™ Pro B-series GPU

Supermicro Edge AI Servers

[1]As used in this document, "Deloitte" means Deloitte Consulting LLP, a subsidiary of Deloitte LLP. Please see http://deloitte.com/us/about for a detailed description of the company's legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

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.

0826/KL/DJA/PDF Please Recycle 369370-001US