Solution Brief

Solution Provider, Independent Software Vendor

Delta Electronics Delivers GenAI/ LLM for Network Security AI with Intel® Arc™ GPU

Delta DV-5002I, based on Intel® Xeon® Scalable processors and Intel® Arc™ A-series Graphics, is optimized for Large Language Model (LLM) and Artificial Intelligence (AI)-based network security workloads.

Network security is a critical component in a corporate environment, where corporations are looking for new network security AI solutions to provide better detection of cybersecurity threats, and responses to incidents. Network security and LLM use cases can effectively complement each other against fast-changing cybersecurity threats.

Human-computer interactions have changed with the introduction of Generative Artificial Intelligence (GenAI) and LLM. The enthusiasm and rapid development are driving LLMs to become a transformative tool for creativity, problem-solving, enhancing productivity, and personalized assistants for corporations, researchers and individual contributors. Publicly available trained models, such as Meta's LlaMA models, allow users to access state-of-the-art AI models to build applications, to solve problems, and to explore different creative projects. Network security with LLM could be the new AI-driven methods to enhance threat detection, response and perform logs, alerts and patterns to identify unusual activities.

Delta Electronics, an Intel® Industry Solution Builders Network Builders Community member, has developed the Delta DV-5002I. This network security appliance enables AI-powered network security servers to protect digital assets and to help ensure secure communication within and between networks.

Delta Model DV-5000I/5002I Delivers LLM with Intel® Arc™ GPU

The Delta DV-5000I/5002I is a 1U/2U rackmount server-based solution for network security AI deployment. The main difference between DV-5000I and DV-5002I (Figure 1 shows DV-5002I) is the form factor of the expansion PCIe slot. DV-5000I can support half-height, half-length (HHHL) PCIe cards, and DV-5002I can support full-height, full-length (FHFL) PCIe cards in a 2U chassis. The system is based on Intel® Xeon® Scalable processors and Intel® Arc™ A-series Graphics. There are 8x 10GmG networking ports, 8x 10G SFP+ connections, 4x 25G SFP28 and 2x 100G QSFP28 networking ports. It also supports up to 2x 1300W redundant power supplies and 6x redundant cooling fans. The Delta DV-5000I supports up to 2T DDR5 memory capacity and a flexible PCIe 5.0 expansion slot to support various commercial 100G/400G Ethernet high-performance GPU PCIe cards, fulfilling various applications such as firewall, VPN, and security gateway.

Figure 1. Front and back views of Delta DV-5002I Outlook

Performance from 5th Gen Intel® Xeon® Scalable Processors and Intel® Arc™ A-Series Graphics

For the Delta DV-5000I/DV-5002I, Delta chose the 5th Gen Intel® Xeon® Scalable processors to optimize demanding AI, mainstream data, multi-cloud computer, and network and storage workloads. With support for higher memory speeds and enhanced memory capacity, the processors deliver advanced performance, energy-efficient compute, enhanced memory capabilities, hardware-enhanced security and workload acceleration. It is designed for AI by providing greater performance to customers deploying AI workloads across cloud, network and edge use cases. The Delta DV-5002I server with 5th Gen Intel Xeon Scalable processors come with Intel® Advanced Matrix Extensions (Intel® AMX), Intel® SSE4.2, Intel® Advanced Vector Extensions (Intel® AVX), Intel® Advanced Vector Extensions 2 (Intel® AVX2) and Intel® Advanced Vector Extensions 512 (Intel® AVX-512) instruction sets for AI acceleration in all processor cores.

The Delta DV-5000I/DV-5002I supports a flexible PCIe 5.0 expansion slot to support high-performance GPU PCIe cards. The Delta team chose the Intel Arc A-Series Graphics to offload GenAI/LLM workload to the GPU and utilize Intel® LLM Library for PyTorch* (IPEX-LLM). The Intel Arc A-Series Graphics for Desktops delivers up to 32 Xᵉ-cores for faster AI inferencing in network security AI workloads. The Intel® Xᵉ Matrix Extensions (Intel® XMX) Engines offer significant AI performance to meet different network security AI workload requirements. The combination of Intel processors and Intel GPUs is well-suited for GenAI/LLM workloads for network security server applications.

The IPEX-LLM is built on top of Intel® Extension for PyTorch and contains all the latest performance optimizations for Intel hardware. With the Xᵉ-cores and Intel XMX AI accelerations on Intel discrete GPUs, PyTorch models can take advantage of IPEX-LLM optimization and run LLM models with high-performance acceleration.

The collaboration between Intel and Delta brings additional possibilities in network security AI applications. The Delta DV-5002I servers are designed with scalability in mind. The design supports easy network upgrades and expansions for businesses, ensuring efficient cost management for different network security AI workloads. LLM could be one potential application for productivity improvement in a corporation.

Delta DV-5002I Performance tests

The Delta team has measured the AI inference performance of Delta DV-5002I in typical LLM use cases: Chatbots and different LLM models benchmarking. The test is performed with IPEX-LLM. The LLM models chosen were Meta-LLaMa families, and the same steps can be applied for different popular LLM models such as Mistal-7B, Phi2 and other IPEX-LLM supported models. The models chosen were measured with the first token latency and second+ average token latency for token generations in all-in-one benchmark and another LLM chatbot reference use cases with PrivateGPT available in the IPEX-LLM repository. Delta Electronics team has been testing based on this setup:

Hardware Environment	
CPU:	Intel® Xeon® Gold 6538N processor
RAM:	256GB
GPU:	Intel® Arc™ A770 Graphics 16GB

Software Environment	
OS:	Ubuntu 22.04.5
Kernel:	6.4.7
GPU driver:	24.22.29735.27

The all-in-one benchmarking with IPEX-LLM allows users to test the model benchmarks and to record the results in a CSV format. Users can provide chosen LLM models for benchmarking and related information in the config.yaml file available in the directory. For more information, refer to: https://github.com/intel/ipex-llm/tree/main/python/llm/dev/benchmark/all-in-one.

The Delta team performed the benchmark script (run-arc.sh) on the Delta DV-5002I with Intel Arc A770 Graphics and observed the benchmarking results, as shown in Table 1 below.

Table 1. All-in-one benchmark with IPEX-LLM based on the chosen LLM models

Platform	Framework	Model	Precision	Input/output tokens	1st token avg latency (ms)	2+ avg latency (ms/token)
Delta DV-5002I	IPEX-LLM 2.20b20250218	meta-llama/Llama-2-7b-chat-hf	FP16	1024-128	351.63	40.77
			FP8		311.72	22.26
			INT8		314.14	26.33
			INT4		321.90	21.80
Delta DV-5002I	IPEX-LLM 2.20b20250218	meta-llama/Meta-Llama-3.1-8B-Instruct	FP16	1024-128	297.96	50.24
			FP8		333.29	26.66
			INT8		334.59	32.15
			INT4		329.60	24.52
Delta DV-5002I	IPEX-LLM 2.20b20250218	meta-llama/Llama-3.2-1B	FP16	1024-128	62.41	18.82
			FP8		68.19	13.35
			INT8		68.35	15.47
			INT4		67.45	12.63
Delta DV-5002I	IPEX-LLM 2.20b20250218	meta-llama/Llama-3.2-3B	FP16	1024-128	142.95	29.54
			FP8		163.46	22.67
			INT8		163.11	24.11
			INT4		159.05	17.59

The LLM AI model benchmarking can be reduced to 22.26ms for FP8 precision from 40.77ms in FP16 precision in Meta-llama/Llama-2-7b-chat-hf for 2+ average latency per token. That is an 83.15% performance improvement with IPEX-LLM FP8 optimization. The FP8 data is based on the reduction of LLM model's weights and activations. It allows for more efficient memory usage and faster computations, without significantly impacting model quality. The newer meta-llama/Llama-3.2-1B model is also supported in IPEX-LLM for smaller parameters that introduced better average latency in first and 2+ tokens during the benchmarking. The Delta team also performed one of the IPEX-LLM reference applications evaluation, PrivateGPT which leverages local LLMs running on Intel GPU as shown in Figure 2. PrivateGPT is a production-ready AI project that allows users to ask questions about documents using the power of LLMs, even in scenarios without an Internet connection.

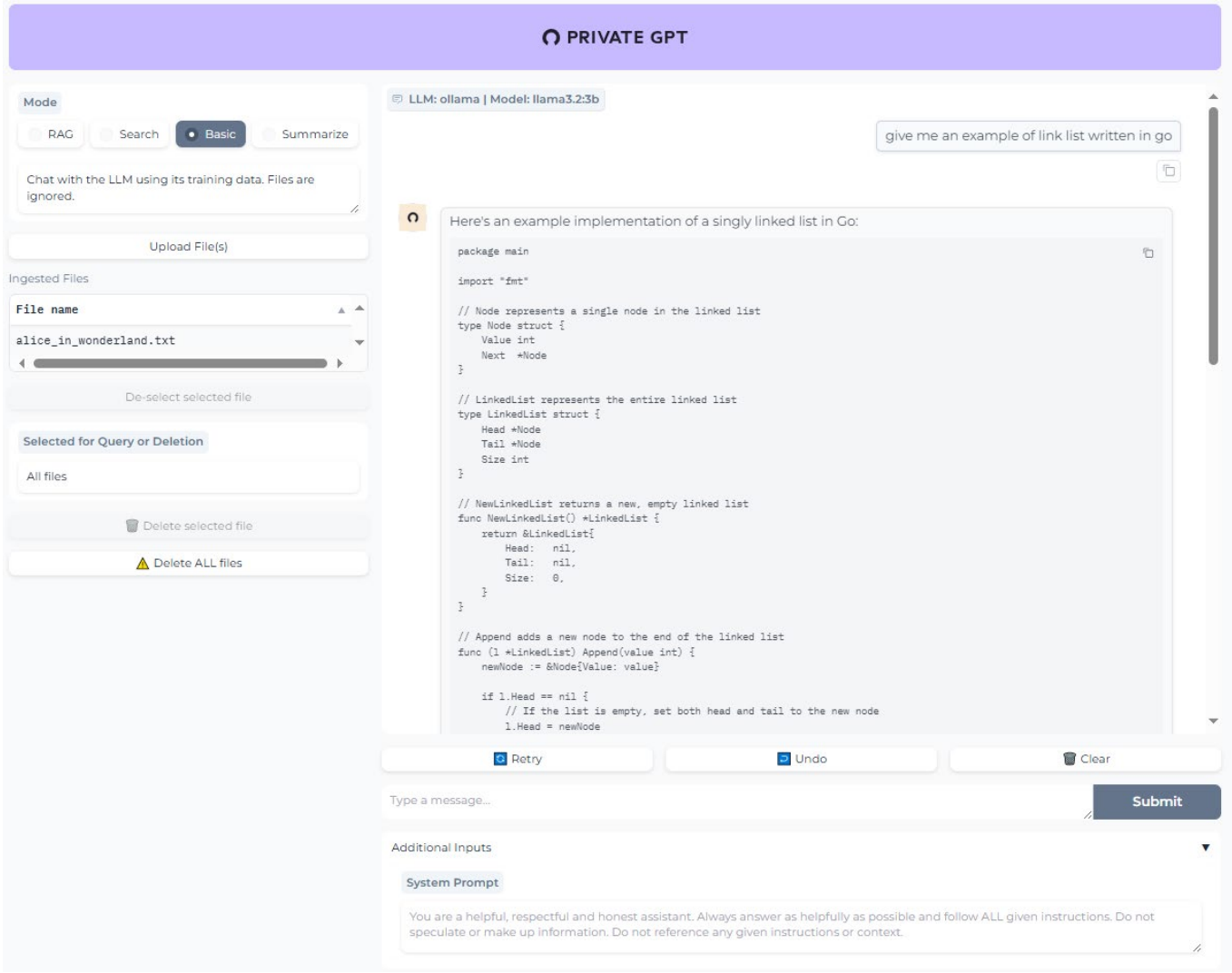

Figure 2. PrivateGPT use cases with Intel® Arc™ A770 Graphics

In this example, Delta is using llama 3.2 3B in PrivateGPT to provide an example of a link list written in Golang. This PrivateGPT supports Ollama on Intel Arc GPU for running local LLMs. The PrivateGPT also supports chat with the model (Chatbot) and chat over documents/retrieval augmented generation (RAG) that has been uploaded to the system. The RAG allows users to receive responses from the model based on the uploaded vectorized content.

In addition to the reference chatbot application of PrivateGPT, the Delta team provides a dashboard review for CPU and GPU hardware utilization. The dashboard is based on the Grafana dashboard, and it serves as a gateway to monitor real-time metrics by viewing all the telemetry data by aggregating data from multiple sources into a single dashboard as shown in Figure 3.

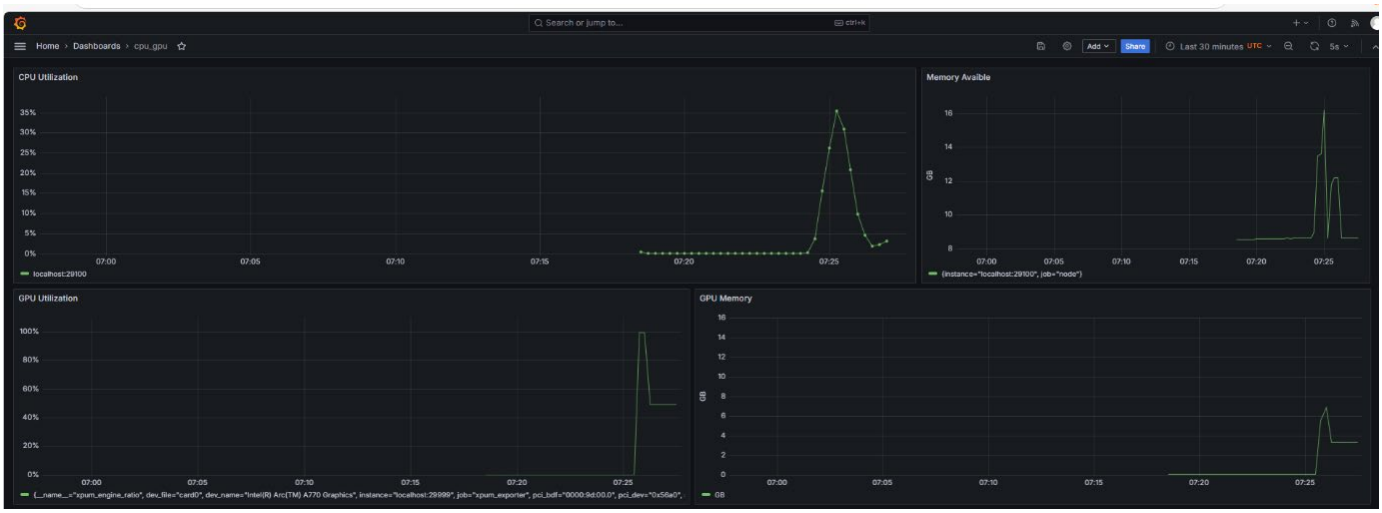

Figure 3. Dashboard view for network security AI LLM use cases with Intel CPU and GPU

The dashboard shows the CPU utilization, system memory, GPU utilization, GPU memory while running the LLM AI inferencing on Intel Arc GPU. The dashboard is essential for data-driven environments, especially in network security AI environments. A dashboard is useful for tracking traffic alerts, network security incidents, monitoring system health and performance in real-time, and can be tailored to meet the requirements of different use cases, teams and objectives.

Conclusion

The IPEX-LLM application test results shown in this paper demonstrate that the Delta DV-5000I/DV-5002I rackmount server is a good choice for network security AI use cases, especially in LLM applications. The Delta DV-5000I/DV-5002I can benefit from Intel AMX, Intel AVX-512 and other instruction set extensions in Intel Xeon Scalable processors and Intel Xᵉ Matrix Extensions in Intel Arc A-Series Graphics. All network security AI use cases discussed in this paper are available in a single package NetSec Software package offered by Intel. Contact your Intel representative for more information.

Learn more

- Delta Home Page
- Delta DV-5000I
- Intel® Xeon® Processors
- Intel® Industry Solution Builders

0425/KF/MK/PDF 365110-001US