Deploy and Scale Generative AI in Enterprises with Intel AMX and Lenovo ThinkAgile HX V4 and FX V4

Last update: **18 December 2025**

Version 1.0

Highlights Lenovo on-prem infrastructure solutions with Nutanix as influencer of AI adoption rate

Presents use cases for Lenovo ThinkAgile HX V4 and FX systems with 6th Gen Intel® Xeon® Processors and Nutanix software

Includes benchmark results and Bill of Material

Cristian Ghetau

LENOVO PRESS

Deploy and Scale Generative AI in Enterprises with Intel AMX and Lenovo ThinkAgile HX V4 and FX V4

Abstract

Lenovo is guided by a principle of enabling smarter technology and AI for all and becoming the most trusted partner in intelligent transformation. These principles drive our commitment to expedite innovation by empowering global partners and customers to develop, train, and deploy AI at scale across various industry verticals with utmost safety and efficiency. Enterprise adoption of AI is increasing, and many adopters are successful in getting ROI and seeing tangible business value. The early adoption of Generative AI across industries shows transformation in workforce to improve productivity and efficiency, generating creative content, extracting information from a variety of documents, and integrating with other AI/ML use cases.

Lenovo on-prem infrastructure solutions influence AI adoption rate with a choice of servers, storage, accelerators, and AI software to address training and inference performance, costs, data sovereignty, and compliance. Lenovo ThinkAgile HX V4 and FX V4 systems with 6th Gen Intel® Xeon® Processors and Nutanix software stack is an ideal platform for developing and deploying AI/ML workloads. Intel Xeon Scalable processors with powerful, integrated AI accelerators can address fine tuning and inferencing performance objectives while reducing system complexity and deployment and operational costs for greater business return. The solution empowers CPU-based AI/ML deployment without compromising performance and without investment in expensive GPU accelerators.

ThinkAgile HX V4 and FX V4 Systems with Intel Xeon 6 Processors

Lenovo ThinkAgile HX650 V4 2U, FX650 V4 2U, HX630 V4 1U and FX630 V4 1U are hyperconverged solutions with Nutanix powered by Intel Xeon 6 processors are optimized for AI workloads and **Accelerated by Intel** offerings. Lenovo ThinkAgile HX V4 and FX V4 systems support up to 86 cores per socket with 6th Gen Intel Xeon processors.

ThinkAgile HX V4 and FX V4 systems are factory-integrated, pre-configured ready-to-go integrated systems built on proven and reliable Lenovo ThinkSystem V4 servers that provide compute power for a variety of workloads and applications and are powered by industry-leading hyperconverged infrastructure software from Nutanix.

Intel Optimized AI Libraries & Frameworks

Intel provides a comprehensive portfolio of AI development software including data preparation, model development, training, inference, deployment, and scaling. Using optimized AI software and developer tools can significantly improve AI workload performance, and developer productivity, and reduce compute resource usage costs. Intel® oneAPI libraries enable the AI ecosystem with optimized software, libraries and frameworks. Software optimizations include leveraging accelerators, parallelizing operations, and maximizing core usage.

Intel® Advanced Matrix Extensions (Intel® AMX)

Intel® AMX is a new set of instructions designed to work on matricesand it enables AI fine-tuning and inference workloads to run on the CPU. Its architecture supports bfloat16 (training/inference) and int8 (inference) data types and Intel provides tools and guides to implement and deploy Intel AMX. The Intel AMX architecture is designed with two components:

1. **Tiles**: These consist of eight two-dimensional registers, each 1 kilobyte in size, that store large chunks of data.
2. **Tile Matrix Multiplication (TMUL)**: TMUL is an accelerator engine attached to the tiles that performs matrix-multiply computations for AI.

Refer more information about Intel AMX here.

With integrated Intel AMX on Intel Xeon 6processors, many AI inferencing and finetuning workloads, including many Generative AI use cases, can run optimally.

Intel AI software and optimization libraries provide scalable performance using Intel CPUs and GPUs. Many of the libraries and framework extensions are designed to leverage the CPU to provide optimal performance for machine learning and inference workloads. Developers looking to leverage these tools can download the AI Tools from AI Tools Selector.

Table 1: Intel AI optimization software and development tools

Software/Solution	**Details**
Intel® oneAPI Library	· Intel® oneAPI Deep Neural Network Library (oneDNN) · Intel® oneAPI Data Analytics Library (oneDAL) · Intel® oneAPI Math Kernel Library (oneMKL) · Intel® oneAPI Collective Communications Library (oneCCL)
MLOPs	Cnvrg.io is a platform to build and deploy AI models at scale
AI Experimentation	SigOpt is a guided platform to design experiments, explore parameter space, and optimize hyperparameters and metrics
Intel® Extension for PyTorch	Intel Extension for PyTorch extends PyTorch with the latest performance optimizations for Intel hardware, also taking advantage of Intel AMX
Intel Distribution for Python	· Optimized core python libraries (scikit-learn, Pandas, XGBoost) · Data Parallel Extensions for Python. · Extensions for TensorFlow, PyTorch, PaddlePaddle, DGL, Apache Spark, and for machine learning

	· NumPy, SciPy, Numba, and numba-dpex.
Intel® Neural Compressor	This open-source library provides a framework-independent API to perform model compression techniques such as quantization, pruning, and knowledge distillation, to reduce model size and speed up inference.

Lenovo and Nutanix simplify and accelerate AI inference deployments

The partnership between Lenovo and Nutanix offers compelling solutions to bring hybrid cloud AI to any organization. Lenovo's ThinkAgile HX650 V4 systems powered by 6th Gen Intel Xeon processors with GPT-in-a-Box™ Nutanix Validated Design (NVD) provide turnkey AI solutions for organizations wanting to implement Generative Pre-trained Transformer (GPT) capabilities while maintaining control over data and applications.

These NVDs are architected and fully tested bundled solutions, including hardware, software, and services which are pre-validated and can be pre-configured to accelerate the deployment of AI initiatives. The solution enables customers to quickly launch every layer of the stack, delivering consistent and verified results. Rather than starting from scratch, customers are provided a simple, proven recipe for success. These solutions include support for several popular large language models, including Llama and Falcon.

Llama3 LLM Inference Performance with 6th Gen Intel Xeon Processors

The Generative AI inference testing with Llama3 8B model was done on Lenovo ThinkAgile HX650 V4 server with 6th Gen Intel Xeon processors by Intel. The test was carried out with different input token sizes 32/256/1024/2048 with varying batch sizes of 1-16 to simulate concurrent requests with static output token size 256. The objective of the testing is to validate different scenario's performance with acceptable latency of less than 100ms latency.

The test and inference serving is targeted on a 4-node Nutanix 7.3 cluster with AHV 10.3 and Ubuntu 22.04.4 guest virtual machines. The model performance can be scaled out by using multiple nodes, but it is not in scope of the current version.

Table 2. Test Hardware and virtual machine configuration

Server	Lenovo ThinkAgile HX650 V4 CN
Processor	2 x Intel(R) Xeon(R) 6767P, 64C, 2.4GHz
Memory	16 x 64 GB ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM
NIC	1 x ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-port PCIe Ethernet Adapter

Disk	2 x ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD 2 x ThinkAgile HX 2.5" U.2 VA 1.92TB Read Intensive NVMe PCIe 4.0 x4 HS SSD
Hyperthreading	Intel® Hyper-Threading Technology Enabled
Turbo	Intel® Turbo Boost Technology Enabled
NUMA nodes	2
BIOS	1.20
BIOS Settings	Performance, Max C-State =C0/C1
NUTANIX	7.3
Hypervisor	AHV 10.3
CVM	16 VCPU, 64GB RAM
Guest VM	Ubuntu 24.04.5 LTS
VM Configuration	240 vCPUs 384 GB Memory 512 GB storage

Table 3. Test Configuration

Workload	LLM Inference
Application	Intel Extension for PyTorch (IPEX) with DeepSpeed
Libraries	IPEX 2.3.100 with DeepSpeed 0.18.2; Pytorch 2.6 (public releases)
Script	https://github.com/intel/intel-extension-for-pytorch/tree/release/2.3/examples/cpu/inference/python/llm
Test Run settings	· warm-up steps = 5 · steps = 50 · -a flag (Max number of threads (this should align with · OMP_NUM_Threads)) = 240 · e (Number of inter threads: e=1: run 1 thread per core; e=2: run two threads per physical core) = 1
Model	Llama3 8B
Dataset	IPEX.LLM prompt.json (subset of pile-10k)
Batch size	1/2/4/8/16
Precision	bfloat16
Framework	IPEX 2.3 (public release)
# instances	1

Llama3 8B Performance Results with Intel AMX

Figure 1 shows the 2nd token average latency performance with Intel AMX on Intel Xeon 6 processors for Llama3 8B model.

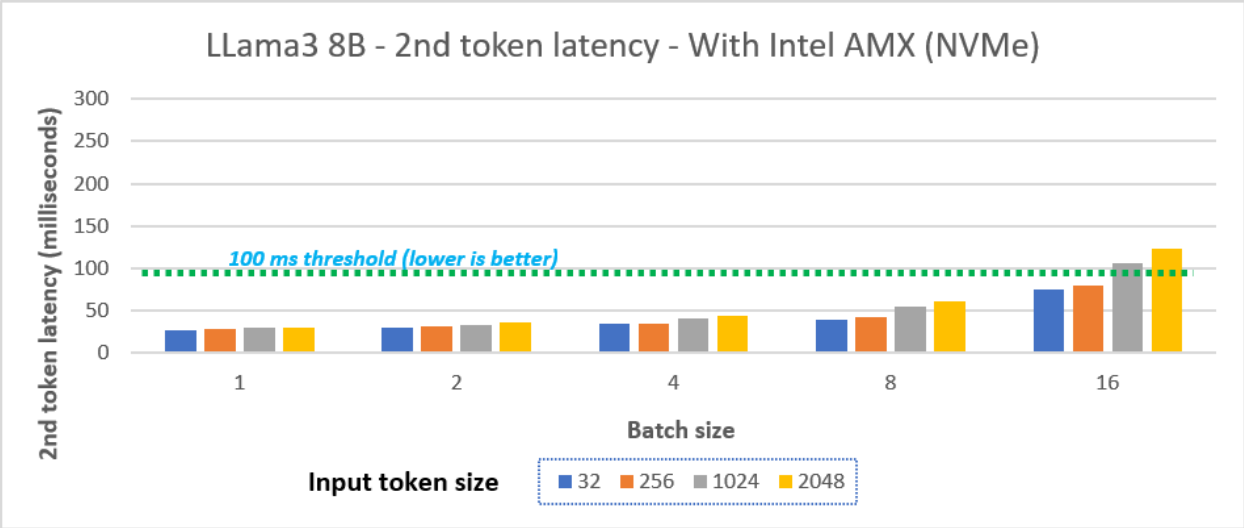

Figure 1. Llama3 8B testing with Intel Xeon 6 CPUs with Intel AMX (NVMe storage) - 2nd token average latency

Table 4 below shows Llama3 2nd token average latency performance comparison between 4th Gen Intel Xeon Scalable processors and Intel Xeon 6 on ThinkAgile HX platform. For Llama3 performance results with 4th Gen Intel Xeon SP on ThinkAgile HX V3, refer https://lenovopress.lenovo.com/lp2190.pdf

Intel Xeon 6 processors provide ~20-50% improvement than 4th Gen and 2x improvement for larger batch size and input tokens.

Table 4. Llama3 latency performance between 4th Gen and 6th Gen Intel Xeon Processors

Batch size	Input tokens	Xeon 8468, 48C latency(ms)	Xeon 6767P latency(ms)
1	32	30	27
1	256	40	28
1	1024	45	29
1	2048	50	30
2	32	45	30
2	256	50	31
2	1024	55	32
2	2048	70	35
4	32	50	32
4	256	60	35

4	1024	70	41
4	2048	95	43
8	32	72	39
8	256	88	42
8	1024	112	54
8	2048	145	60
16	32	110	75
16	256	125	79
16	1024	180	106
16	2048	260	124

On Lenovo ThinkAgile HX650 V4 and FX V4 with 6thGen Intel® Xeon® (AMX enabled), Llama3 8B delivers sub-100 ms 2nd-token latency up to batch-size 8 across all tested context lengths (32 to 2048 tokens). Batchsize 8 is the best steady-state throughput point for long prompts (~2K), while batch-size 16 adds throughput for ≤1K-token prompts at the cost of crossing the 100 ms target for longer contexts.

It is recommended to use high clock speed CPU to reduce latency and define tradeoff criteria case by case to achieve service level objectives by reducing batch or context size and balanced hardware configuration

Bill of Material - ThinkAgile HX650 V4 All Flash Configuration with Nutanix

Table 5. Bill of Materials

Part number Feature code	**Product Description**	**Qty**
7DG4CTO1WW	Server: Lenovo ThinkAgile HX650 V4 Hyperconverged System	1
C6TQ	ThinkAgile HX650 V4 Base	1
B15S	Nutanix Software Stack on Nutanix AHV	1
BVKV	Nutanix Cloud Platform (NCP) Pro Software License with Mission Critical Support	1
C5QY	Intel Xeon 6767P 64C 350W 2.4GHz Processor	2
C0TQ	ThinkSystem 64GB TruDDR5 6400MHz (2Rx4) RDIMM	16
C26V	ThinkSystem M.2 RAID B545i-2i SATA/NVMe Adapter	1
C46P	ThinkSystem 2U V4 8x2.5" NVMe Backplane	2
B0SW	Nutanix Flash Node Config	1

Part number Feature code	Product Description	Qty
C2BR	ThinkSystem 2.5" U.3 7500 PRO 1.92TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	6
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BN2T	ThinkSystem Broadcom 57414 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1
C0U3	ThinkSystem 2000W 230V Titanium CRPS Premium Hot-Swap Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
C2DJ	ThinkSystem Advanced Toolless Slide Rail Kit V4	1
C3RG	ThinkSystem SR650 V4 Left Rack Latch with USB/MiniDP	1
C3RD	ThinkSystem 2U 6056 20K Performance Fan Module	6
BVKV	Nutanix Cloud Platform (NCP) Pro Software License with Mission Critical Support	1
SCJC	XClarity One - Managed Device, Per Endpoint w/1 Yr SW S&S	1

Accelerated by Intel

To deliver the best experience possible, Lenovo and Intel have optimized this solution to leverage Intel capabilities like processor accelerators not available in other systems. Accelerated by Intel means enhanced performance to help you achieve new innovations and insight that can give your company an edge.

References:

Lenovo ThinkAgile HX650 V4 Hyperconverged System
https://lenovopress.lenovo.com/lp2133

Intel AI Development Software
https://www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/development-software.html