Deploy and Scale Generative AI in Enterprises with Lenovo ThinkAgile VX V4 Systems

Last update: **19 September 2025**

Version 1.0

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

Presents use cases for Lenovo ThinkAgile VX V4 systems with 6th Gen Intel® Xeon® Scalable Processors and VMware vSAN Hyperconverged

Includes benchmark results and Bill of Materials

Cristian Ghetau

LENOVO PRESS

Deploy and Scale Generative AI in Enterprises with Lenovo ThinkAgile VX V4 Systems

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 VX V4 Systems powered by Intel Xeon 6 processors with 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 inference deployment without compromising performance and without investment in expensive GPU accelerators.

Lenovo ThinkAgile VX V4 Systems with Intel 6th Gen Scalable Processors

Lenovo ThinkAgile VX V4 Systems are 2-socket 1U or 2U systems powered by Intel® Xeon® 6 processors (formerly code named "Granite Rapids") provides up to 86 cores and high memory bandwidth to support AI/ML and enterprise workloads. ThinkAgile VX V4 Series systems also support hardware accelerators like Graphics Processing Units (GPUs) and Data Processing Unit (DPU) cards for maximum performance.

Lenovo ThinkAgile VX V4 systems with 6th Gen Intel® Xeon® Processors and VMware vSAN hyperconverged stack is an ideal platform for developing and deploying AI/ML workloads.

ThinkAgile VX630 V4

ThinkAgile VX650 V4

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 matrices, and 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 6 Gen Intel Xeon processors, 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.

Llama 2 and Llama 3 LLM Inference Performance with 6th Gen Intel Xeon Processors

The Generative AI inference testing with Llama 2 7B and Llama 3 8B models was done on Lenovo ThinkAgile VX630 V4 server with 6 Gen Intel Xeon Scalable 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 and to compare the results between the LLM models.

The test and inference serving is targeted on a single node running ESXi 8.0.3 and Ubuntu 22.04.5 LTS 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 VX630 V4 CN
Processor	2 x Intel(R) Xeon(R) 6787P, 86C, 2GHz
Memory	1152GB (12x96GB DDR5 6400 MT/s [6400 MT/s])
NIC	1 x Broadcom 57504 10/25GbE SFP28 4-port OCP Ethernet Adapter
Disk	2 x Micron 480GB M.2 NVMe SSD 4 x Intel 1.92TB 2.5" NVMe SSD
Hyperthreading	Intel® Hyper-Threading Technology Enabled
Turbo	Intel® Turbo Boost Technology Enabled
NUMA nodes	2
BIOS	3.20
BIOS Settings	Performance, Max C-State =C0/C1
ESXi	8.0.3
Guest VM	Ubuntu 24.04.5 LTS
VM Configuration	172 vCPUs 256 GB Memory

	512 GB storage

Table 3. Test Configuration

Workload	LLM Inference
Application	Intel Extension for PyTorch (IPEX) with DeepSpeed
Libararies	IPEX 2.2 with DeepSpeed 0.14; Pytorch 2.2 (public releases)
Script	https://github.com/intel/intel-extension-forpytorch/ tree/v2.2.0%2Bcpu/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)) = 60 · e (Number of inter threads: e=1: run 1 thread per core; e=2: run two threads per physical core) = 1
Model	Llama 2 7B, Llama 3 8B
Dataset	IPEX.LLM prompt.json (subset of pile-10k)
Batch size	1/2/4/8/16
Precision	bfloat16
Framework	IPEX 2.2 (public release), IPEX 2.3 (public release)
# instances	1

Llama 2 7B Performance Results

Figure 1 shows the 2nd token average latency performance with Intel AMX on Intel Xeon 6 processors for Llama 2 7B model. The test with Llama 2 and Intel AMX shows the 2nd token latency for different concurrent requests scenarios (batch sizes 1/2/4/8/16) with input/output token size of 32/256 and 256/256 are well within an acceptable threshold of 100 milliseconds. The 2nd token latency for all the scenarios test with batch size 1/2/4/8 are below the 100ms threshold and the performance for input token size 1024 and 2048 with batch size 16 is well above 100ms.

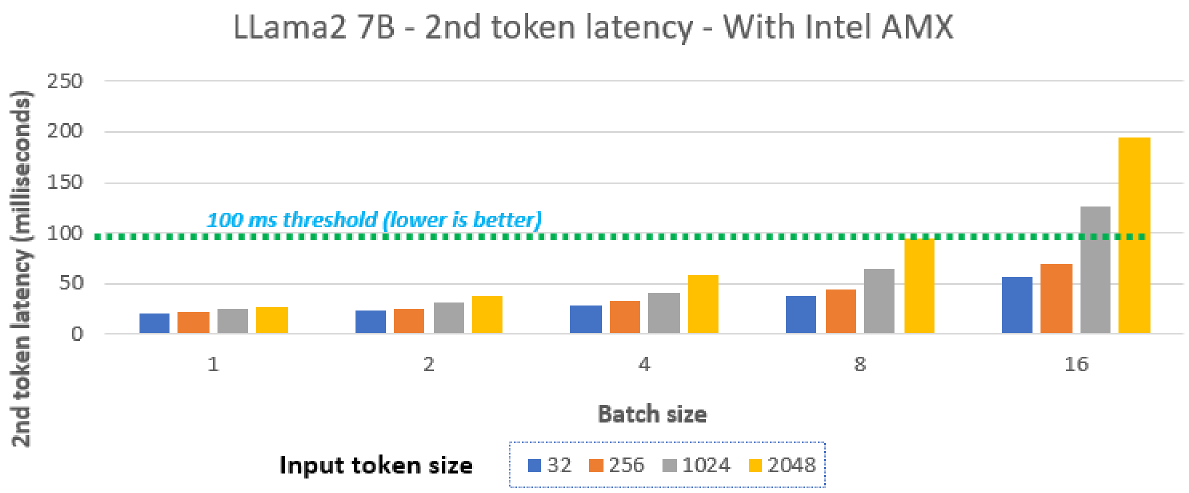

Figure 1. Llama2 7B testing with 4th Gen Intel Xeon CPUs with Intel AMX - 2nd token average latency

Llama 3 8B Performance Results

Figure 2 shows the 2nd token average latency performance with Intel AMX on Intel Xeon 6 processors for Llama 3 8B model. The test with Llama3 and Intel AMX shows significant improvement in 2nd token latency for the scenario with input/output token sizes when compare with Llama 2 7B testing. The 2nd token latency for different concurrent requests scenarios (batch sizes 1/2/4/8/16) with input/output token size of 32/256, 256/256 and 1024/256 are below an acceptable threshold of 100 milliseconds.

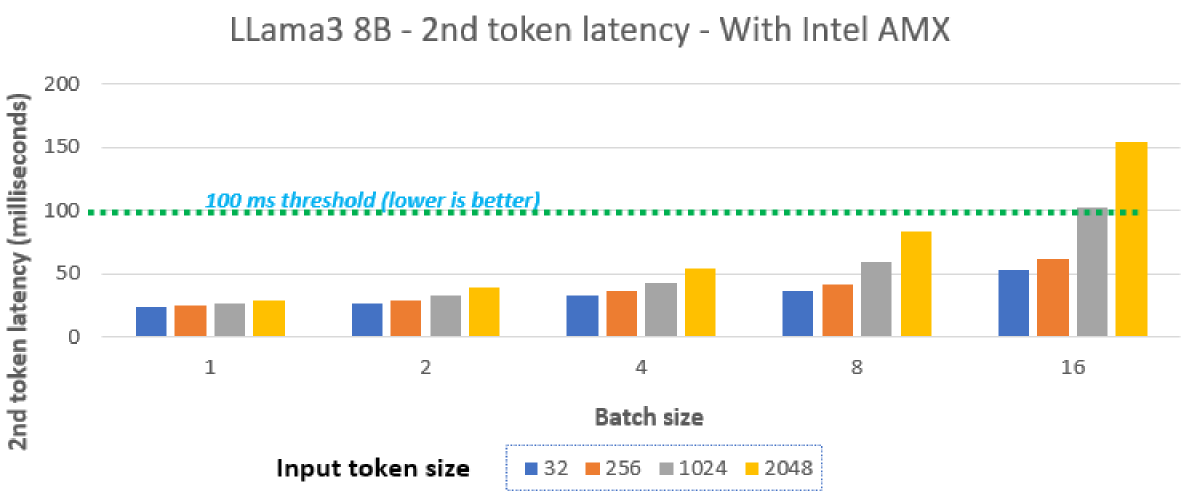

Figure 2. Llama 3 8B testing with 6th Gen Intel Xeon CPUs with Intel AMX - 2nd token average latency

Table 4 below shows the results for the comparison of inference latency and 2nd token average latency performance with Intel AMX on 6th Gen Intel Xeon Scalable processors for Llama 2 7B and Llama 3 8B models.

The scenario tested is with input/output token size 32/256 and the 2 token latency for batch size 1 is within acceptable threshold of 100 milliseconds and it shows considerable performance increase can be achieved with INT8 quantization even with 2048 input tokens and the batch size of 8. The Llama 3 8B shows up to 40% reduction in inference latency for larger input token size 2048 when compare with Llama 2 7B model.

Table 4. Llama 2 7B and Llama 3 8B testing with 6th Gen Intel Xeon CPUs

LLM	Llama 2 7B			Llama 3 8B		
Test Scenario	Time (s)	Inference latency (ms)	2nd token Latency (ms)	Time (s)	Inference latency (ms)	2nd token Latency (ms)
bfloat16 (input token size =32, output token size=256, batch-size=1)	9.5	9609	37	8.6	8722	34
INT8 (Quantization) (input token size =32, output token size=256, batch-size=1)	7.2	7191	28	6	6178	24
INT8 (Quantization) (input token size =2048, output token size=256, batch-size=8)	28	29000	94	25	25000	83

Llama 3 models provide higher accuracy and better reasoning capabilities and reduced latency consistently than Llama2 model on the ThinkAgile VX V4 systems with Intel AMX. It is recommended to use high clock speed CPU to reduce latency further and define tradeoff criteria case by case to achieve service level objectives by reducing batch or context size and using balanced hardware configuration with 6th Gen Intel Xeon processors to achieve better ROI.

Bill Of Materials: Lenovo ThinkAgile VX650 V4

Part number	Product Description	Qty
7DG6CTO1WW	Server : ThinkAgile VX650 V4	1
C68E	ThinkAgile VX650 V4 24x2.5" Chassis	1
BVGL	Data Center Environment 30 Degree Celsius / 86 Degree Fahrenheit	1
B0W3	XClarity Pro	1
C5QR	Intel Xeon 6520P 24C 210W 2.4GHz Processor	1

C3QR	ThinkSystem 2U V4 Performance Heatsink	1
BYTJ	ThinkSystem 32GB TruDDR5 6400MHz (2Rx8) RDIMM	16
5977	Select Storage devices - no configured RAID required	1
BT2G	vSAN ESA	1
BYRM	AF-0	1
C2BS	ThinkSystem 2.5" U.3 7500 PRO 3.84TB Read Intensive NVMe PCIe 4.0 x4 HS SSD	3
C46P	ThinkSystem 2U V4 8x2.5" NVMe Backplane	1
C26V	ThinkSystem M.2 RAID B545i-2i SATA/NVMe Adapter	1
BQ1Y	ThinkSystem M.2 5400 PRO 480GB Read Intensive SATA 6Gb NHS SSD	2
BLA3	SW stack for ThinkAgile VX Deployer	1
C1YK	ThinkSystem SR650 V4/SR630 V4 x16 OCP Cable Kit	1
BPPW	ThinkSystem Broadcom 57504 10/25GbE SFP28 4-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
C3RD	ThinkSystem 2U 6056 20K Performance Fan Module	5
C4S2	ThinkSystem SR650 V4 Processor board,BHS,DDR5,Santorini,2U	1
C7Y8	ThinkSystem SR650 V4 System I/O Board	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 VX630 V4 Hyperconverged System
https://lenovopress.lenovo.com/lp2134.pdf

Lenovo ThinkAgile VX650 V4 Hyperconverged System
https://lenovopress.lenovo.com/lp2135.pdf

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

Trademarks and special notices

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, ThinkSystem, ThinkAgile, ThinkCentre, ThinkVision, ThinkVantage, ThinkPlus and Rescue and Recovery are trademarks of Lenovo.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.