AI/ML Workload Solutions on Lenovo ThinkSystem V4 Servers

Solution Brief

Enabling AI at any scale

The evolution of artificial intelligence / machine learning (AI/ML) provides different value streams across enterprise business and drives significant business impact. Growing adoption of generative AI in the enterprise is driving the need for more hardware and accelerators which increase Total Cost of Ownership. The Lenovo ThinkSystem V4 server portfolio powered by Intel® Xeon® 6 processors with E-cores enables right-sized AI compute with efficient, secure, workload optimized solutions for all classical machine learning and enterprise private AI use cases.

In September 2024, Lenovo will start shipping the ThinkSystem SD520 V4 1U half width servers followed by the ThinkSystem SR630 V4 1U rack servers in November 2024 equipped with 6th generation Intel Xeon 6 processor family with E-Cores (codename Sierra Forest). These E-core processors have a scalable architecture with a higher number of cores to support AI/ML and enterprise workloads. The SD520 V4 servers are built with Neptune Liquid cooling modules to provide efficient cooling on the CPUs.

Figure 1. Lenovo ThinkSystem SD520 V4

Figure 2. Lenovo ThinkSystem SR630 V4

Intel Xeon 6 processors with E-cores are enhanced to deliver density-optimized compute in the most power-efficient manner. Xeon processors with E-cores provide best-in-class power-performance density, offering distinct advantages for cloud-native and hyperscale workloads:

- 2.5x better rack density and 2.4x higher performance per watt.
- Support for 1S and 2S servers, with up to 144c per CPU and TDP as low as 200W.
- Modern instruction set with robust security, virtualization and AVX with AI extensions.

Intel Xeon 6 processors with E cores are designed to be more power efficient and consume 30-40% less power than 5th Gen processors when servers are utilized at 40-60%. It dramatically reduces power and cooling costs and 6th Gen processors provide up to 20% more performance than previous generation processors which increases consolidation ratio for any workloads. The compute intensive AI/ML workloads benefit greatly from Xeon 6 architecture with Intel Optimized AI software libraries, and the solution reduces rack, power, and cooling cost to achieve better Return on Investment.

Table 1. Intel Xeon 6 Processors with E-Cores

Feature	Sierra Forest SP (E-Core)	Sierra Forest AP (E-Core)
Sockets	1S, 2S	1S, 2S
Max Cores	64 to 144	192 to 288
TDP	205W to 330W	350W to 500W
DIMMs	12	12
Accelerators	Next Gen Quick Assist Technology, Dynamic Load Balancer (DLB) br>2.5, Data Streaming Accelerator (DSA) 2.0 64GB/s, In Memory br>Analytics Accelerator (IAX) 2.0 64GB/s, Advanced Matrix Extensions	Next Gen Quick Assist Technology, Dynamic Load Balancer (DLB) br>2.5, Data Streaming Accelerator (DSA) 2.0 64GB/s, In Memory br>Analytics Accelerator (IAX) 2.0 64GB/s, Advanced Matrix Extensions
AVX Support	AVX2 (2x128)	AVX2 (2x128)
FP16, BF16	Fast Upconvert	Fast Upconvert

Features to optimize AI/ML use cases

Lenovo ThinkSystem SR630 V4 and SD520 V4 servers have the following features to optimize AI/ML use cases:

- Sub NUMA Clustering (SNC) feature can provide improved performance for Resnet50.
- E-core 64-144c provide more energy efficiency and ideal for inference workloads and SMBs.
- Optimization support for AVX2-128 (VNNI/Int8 & Bfloat16), Accelerator ISA(AiA), 5G ISA.
- Fast upconvert for FP16 and BF16.
- Memory support for DDR5-6400 MT/s.
- GPU support - SR630 V4 (Up to 3 single width 75W GPUs) and SD520 V4 (1 single width 75W GPU).

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, 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 AI software and optimization libraries provide scalable performance using Intel CPU and GPU. Many of the libraries and framework extensions are designed to leverage CPU to provide optimal performance for machine learning and inference workloads. Intel Xeon 6th Gen Scalable processors with E-cores are compatible with many Intel Optimized AI Libraries and tools and provide ecosystem for model development and deployment for enterprise-wide use cases.

Table 2. Intel AI optimization software and development tools

Software / Solution	Details
Intel oneAPI Library	• Deep Neural Network Library • Data Analytics Library • Math Kernel Library • Collective Communications Library
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 AI Reference Models	Repository of pretrained models, sample scripts, best practices, and step-by-step tutorials for many popular open source, machine learning models optimized to run on Intel https://github.com/intel/models
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
AI Model Optimization Intel® Neural Compressor	Support for models created with PyTorch, TensorFlow, Open Neural Network Exchange (ONNX) Runtime, and Apache MXNet

Deploy and Scale Generative AI Workloads with ThinkSystem V4 Systems

Lenovo ThinkSystem V4 systems with Intel Xeon 6 processors with E-cores and low-end GPU accelerators provide a cost effective infrastructure solution to scale your AI deployment and Generative AI use cases. With higher core counts, power efficiency and Optimized AI software, many AI/ML classical use cases and inference workloads can seamlessly run on the CPU without need for expensive GPU accelerators.

XClarity One Powered by AIOps

Lenovo ThinkSystem V4 servers are supported by the XClarity One platform, a hybrid cloud-based unified systems management solution. XClarity One provides three predictive failure analytics engines to swiftly identify potential issues and minimize system downtime while increasing accuracy.

Why Lenovo

Lenovo is a US$70 billion revenue Fortune Global 500 company serving customers in 180 markets around the world. Focused on a bold vision to deliver smarter technology for all, we are developing world-changing technologies that power (through devices and infrastructure) and empower (through solutions, services and software) millions of customers every day.

For More Information

To learn more about this Lenovo solution contact your Lenovo Business Partner or visit: https://www.lenovo.com/us/en/servers-storage/solutions/ai/

References:

Lenovo ThinkSystem SD520 V4: https://lenovopress.lenovo.com/ds0184

Lenovo ThinkSystem SR630 V4: https://lenovopress.lenovo.com/ds0185.pdf

Intel AI Development Software: https://www.intel.com/content/www/us/en/newsroom/news/intel-unveils-future-generation-xeon.html

Intel Unveils Future Generation Xeon Architecture: https://learn.microsoft.com/en-us/sql/sql-server/what-s-new-in-sql-server-2022?view=sql-server-ver16

Related product families

Product families related to this document are the following:

- ThinkSystem SD520 V4 Server
- ThinkSystem SR630 V4 Server

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1976, was created or updated on June 21, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1976
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1976.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
Neptune®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

Intel® and Xeon® are trademarks of Intel Corporation or its subsidiaries.

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