Deploy and Scale Generative AI in Enterprises with Lenovo ThinkSystem SR650 V3 on Red Hat OpenShift, Accelerated by Intel

Solution Brief

Simplified Deployment, Scalability, and Management

Getting started with Generative AI can seem complicated, but the good news is that companies don't have to start from scratch. They can extend their existing infrastructure with the Lenovo ThinkSystem SR650 V3, accelerated by 5th Gen Intel® Xeon® processors, and achieve powerful business impacts without having to invest in dedicated (and often costly) GPU accelerators. Deploying Generative AI use-cases on a cluster with Red Hat® OpenShift® container platform allows for ease of deployment, usability, and scalability with containers and services managed by Kubernetes. Red Hat OpenShift makes the most of Lenovo's AI optimized hardware, since it supports hardware acceleration for inference use cases, a broad ecosystem of AI/ML and application development tools, and integrated security and operations management capabilities. AI models can be updated frequently to improve accuracy by integrated DevOps capabilities in OpenShift.

Solution and Testing Overview

Testing has shown the Lenovo ThinkSystem SR650 V3, with 5th Gen Intel Xeon processors, delivers a highly performant, scalable solution for Generative AI. A next token latency of 100ms or less is a response time perceived as instantaneous for most conversational AI and text summarization applications. Test results demonstrated this solution could successfully meet that target and provide the necessary performance to support a variety of use cases, including real-time chatbots.

The Lenovo ThinkSystem SR650 V3 offers high performance, storage, and memory capacity to tackle complex workloads that require optimized hardware architecture - like Generative AI. With flexible storage and networking options, the SR650 V3 can easily scale for changing needs. The ThinkSystem SR650 V3 supports one or two 5th Gen Intel Xeon processors. With built-in Intel® Advanced Matrix Extensions (AMX), 5th Gen Intel Xeon processors deliver high performance on cutting-edge AI models.

Enterprises may require multiple Generative AI models to perform different tasks, including image creation, synthetic data generation, and chatbots. Generative AI models can require a large amount of storage. The ThinkSystem SR650 V3 can support many Generative AI models in a single 2U server with its tremendous amount of storage and flexibility. With three drive bay zones, it supports up to 20x 3.5-inch or 40x 2.5-inch hotswap drive bays.

The ThinkSystem SR650 V3 offers energy-efficiency features to save energy and reduce operational costs for Generative AI workloads. These features include advanced direct-water cooling (DWC) with the Lenovo Neptune Processor DWC Module, where heat from the processors is removed from the rack and data center using an open loop and coolant distribution units, resulting in lower energy costs, high-efficiency power supplies with 80 PLUS Platinum and Titanium certifications, and optional Lenovo XClarity Energy Manager, which provides advanced data center power notification, analysis, and policy-based management to help achieve lower heat output and reduced cooling needs.

Figure 1. Lenovo ThinkSystem SR650 V3

Results

The Generative AI testing on up to 4x Lenovo ThinkSystem SR650 V3 servers with 5th Gen Intel Xeon processors was performed by Intel and validated by Lenovo. The setup involves an 8-node cluster which includes 3x control plane nodes and 4x worker nodes powered by Red Hat OpenShift. Red Hat OpenShift is an industry leading hybrid cloud application platform powered by Kubernetes. Using Red Hat OpenShift, Lenovo and Intel showcases a platform for deploying, running, and managing applications with ease-of-use and scalability in mind. Red Hat OpenShift delivers a consistent experience across public cloud, on-premises, hybrid cloud, or edge architecture.

A variety of batch sizes were used to simulate concurrent users and input token lengths between 256-2048 represent a typical enterprise chatbot scenario. The workloads chosen for these tests were the Llama 2 model and the Falcon model, both of which are available from HuggingFace. The software stack and model-specific parameters are documented in Table 2 below.

Based on our initial profiling of the workloads, we chose 'network-latency' as the system setting profile. As mentioned in Red Hat Enterprise Linux Performance Tuning Guide documentation, the network-latency profile optimizes for low latency network tuning. It is based on the latency-performance profile. It additionally disables transparent hugepages, NUMA balancing and tunes several other network related sysctl parameters. Figure 2 below shows for LLAMA 2 13B parameters, 1x Lenovo ThinkSystem SR650 V3 with 5th Gen Intel Xeon processors helps achieve less than 100ms next token latency from batch size 1 to batch size 8 for Generative AI inference across input token lengths 256 to 1024.

As demonstrated below with Falcon 40B parameters, given the model size and architecture, 4x Lenovo ThinkSystem SR650 V3 with 5th Gen Intel Xeon processors helps achieve less than 100ms next token latency from batch size 1 to batch size 4 for Generative AI inference across input token lengths 256 to 2048.

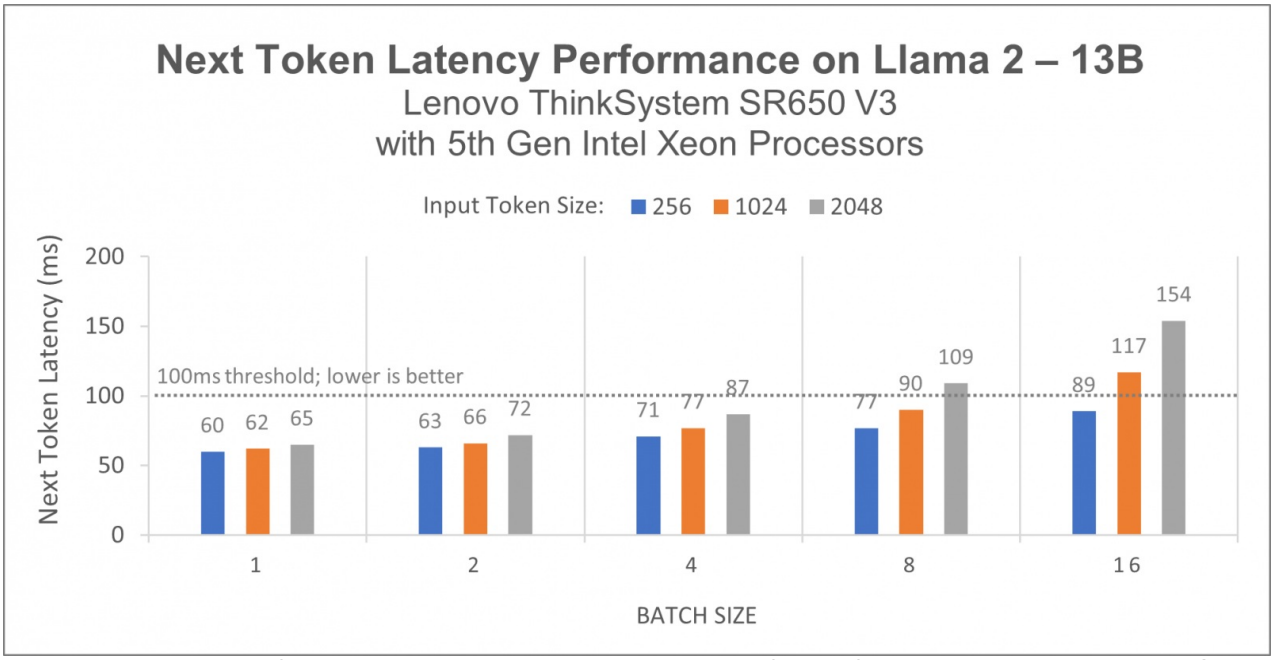

Figure 2. Llama 2-13B performance with BF16 precision on Lenovo ThinkSystem SR650 V3 on 1 node with 2x 5th Gen Intel Xeon CPU using DeepSpeed (AutoTP)

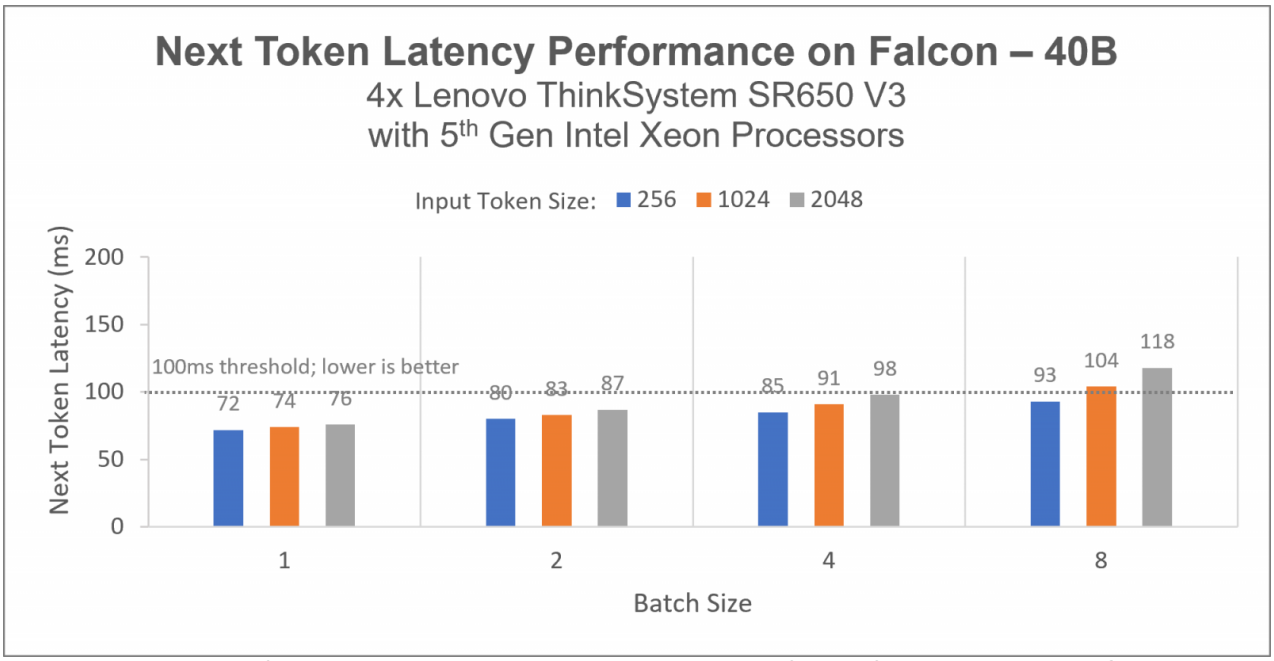

Figure 3. Falcon-40B performance with BF16 precision on 4x Lenovo ThinkSystem SR650 V3 with 2x 5th Gen Intel Xeon CPU using DeepSpeed (AutoTP)

Configuration Details

Tested by Intel as of January 2024

Table 1. Hardware Configuration

Nodes	3x Control Plane Nodes	4x Worker Nodes
Server	Intel Corporation S2600WFT	Lenovo ThinkSystem SR650 V3
Processor	2x Intel Xeon Gold 6248 processors	2x Intel Xeon Platinum 8562Y+ processors
Sockets	2	2
Cores per Socket	20	32
Simultaneous Multithreading (SMT)	Intel Hyper-Threading Technology Enabled	Intel Hyper-Threading Technology Enabled
CPUs	80	128
Intel Turbo Boost	Enabled	Enabled
Base Frequency	2.5GHz	2.8GHz
NUMA Nodes	2	2
Installed Memory	384GB (12x32GB DDR4 2933 MT/s [2934 MT/s])	512GB (16x32GB DDR5 5600 MT/s [5600 MT/s])
NIC	2x Ethernet Controller E810-C for QSFP, 2x Ethernet Connection X722 for 10GBASE-T	2x Ethernet Controller E810-XXV, 4x I350 Gigabit Network Connection, 2x Ethernet Controller E810-C for QSFP
Disk	1x 894.3G INTEL SSDSC2KB96, 1x 3.5T INTEL SSDPF2KX038TZ	2 x 894.3Gb nvme0n1 Micron_7450_MTFDKBA960TFR, 1 x 2.9Tb nvme2n1
BIOS	SE5C620.86B.02.01.0010.010620200716	ESE122N-3.10
Microcode	0x5003604	0x21000161
OS	Red Hat Enterprise Linux CoreOS 414.92.202312011602-0 (Plow)	Red Hat Enterprise Linux CoreOS 414.92.202312011602-0 (Plow)
Kernel	5.14.0-284.43.1.el9_2.x86_64	5.14.0-284.43.1.el9_2.x86_64

Table 2. Other Configuration Details

Software Configuration	Meta LLAMA-2-13B, Tiiuae Falcon 40B BF16 precision
Framework /Toolkit	Torch==2.1.0, intel_extension_for_pytorch==2.1.0, accelerate==0.25.0, sentencepiece==0.1.99, protobuf==4.25.1, datasets==2.15.0, transformers==4.31.0, wheel==0.42.0, neural-compressor==2.3.1, TorchCCL--branch v2.1.0+cpu, mpi4py==3.1.4, Deepspeed --branch gma/run-opt-branch
Orchestration	RHOS 4.14 (Kubernetes v1.27.8)
Topology or ML Algorithm	meta-llama/Llama-2-13b-hf, tiiuae/falcon-40b
Dataset	LaMBDa License: Creative Commons by 4.0
Compiler	gcc version 12.3.0 (GCC)
Libraries	oneDNN v3.2, oneccl-bind-pt 2.1.0+cpu
Dataset (size, shape)	prompt.json Tokens Length: 256, 1024 and 2048; Output Token Length: 256
Precision	BF16
Warmup Steps	10
Num Interations	100
Batch Size	1, 2, 4, 8, 16
Beam Width	1 (greedy search)
Input Token Size	256, 1024, 2048
Output Token Size	256

Conclusion

Lenovo offers AI optimized infrastructure, including the SR650 V3, for energy-efficient, high-performance computing to address large language models and any other AI workload challenge. To see this and other Lenovo AI solutions, please visit https://www.lenovo.com/ai

For customers looking to adopt AI faster, Lenovo's AI Discover Center of Excellence (COE) provides access to AI experts, workshops, and best practices. We can help you develop optimized solutions that enable you to extract valuable business insights from your data quickly, responsibly, and ethically. Contact the AI Discover COE directly here.

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.

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/ai

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, LP1902, was created or updated on February 28, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1902
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1902.

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®
Lenovo Neptune®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

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