Deploy and Scale Generative AI in Enterprises with Lenovo ThinkSystem SR650 V3

Solution Brief

Get Started with the Infrastructure You Know

Generative AI seems to be everywhere and it's an inflection point that companies can't afford to miss. It has the potential to transform and reinvent virtually every aspect of business from customer experience to business operations to employee engagement, yet it can often be a daunting task to determine just how to get started. But it doesn't have to be complicated. Companies looking to start their journey on Generative AI can extend their existing infrastructure with the Lenovo ThinkSystem SR650 V3, accelerated by 4th Gen Intel® Xeon® processors, and achieve revolutionary business impacts without having to invest in dedicated (and often costly) GPU accelerators.

Solution and Testing Overview

Intel testing has shown the Lenovo ThinkSystem SR650 V3, with 4th Gen Intel Xeon processors, delivers a highly performant, scalable solution for Generative AI. A latency of 100ms or less is a response time perceived as instantaneous for most conversational AI and text summarization applications. Intel's testing 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, like Generative AI that require optimized hardware architecture. With flexible storage and networking options, the ThinkSystem SR650 V3 can easily scale for changing needs. The ThinkSystem SR650 V3 supports one or two 4th Gen Intel Xeon processors. With built-in Advanced Matrix Extensions (AMX), 4th 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 hot-swap 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 the Lenovo ThinkSystem SR650 V3 with 4th Gen Intel Xeon processors was performed by Intel and validated by Lenovo. A variety of batch sizes were used to simulate concurrent users and token lengths between 32-1024 represent a typical enterprise chatbot scenario.

As demonstrated with LLAMA 2, 7B and 13B parameters, the Lenovo ThinkSystem SR650 V3 with 4th Gen Intel Xeon processors helps achieve less than 100ms next token latency from batch size 1 to batch size 16 for Generative AI inference across input token lengths 32 to 1024.

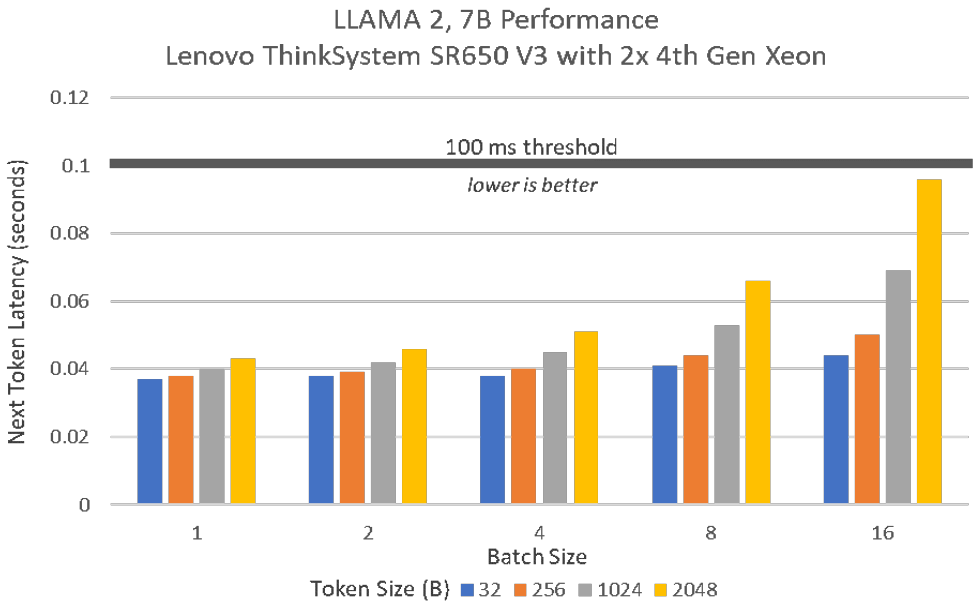

Figure 2. Llama 2, 7B Performance on Lenovo ThinkSystem SR650 V3 with 2x 4th Gen Intel Xeon CPU using DeepSpeed (AutoTP)

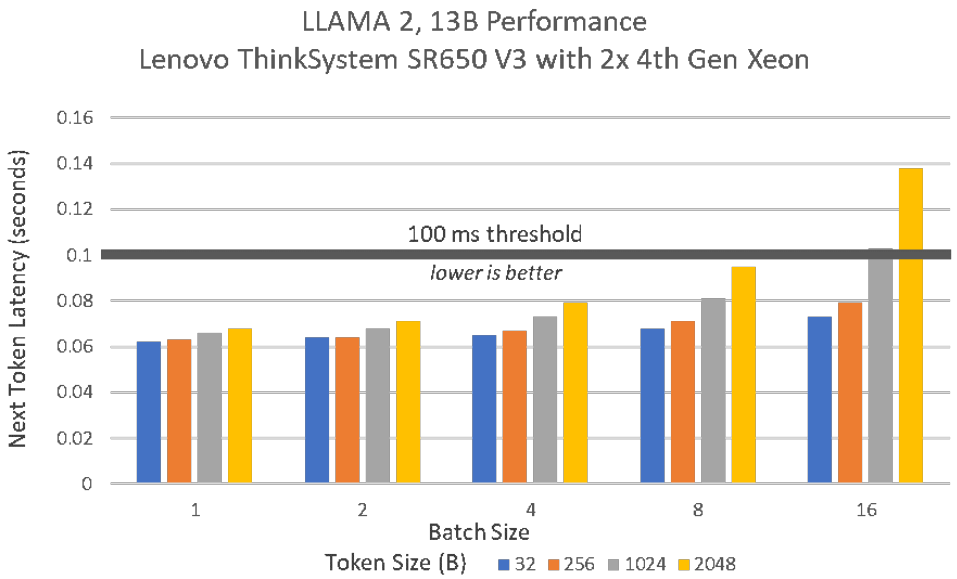

Figure 3. Llama 2, 13B Performance on Lenovo ThinkSystem SR650 V3 with 2x 4th Gen Intel Xeon CPU using DeepSpeed (Auto TP)

Configuration Details

Tested by Intel as of September, 2023

Table 1. Hardware Configuration

Server	Lenovo ThinkSystem SR650 V3
Processor	2x Intel Xeon Platinum 8462Y+ processors
Sockets	2
Cores per Socket	32
Hyperthreading	Intel® Hyper-Threading Technology Enabled
CPUs	128
Intel Turbo Boost	Enabled
Base Frequency	2.8GHz
NUMA Nodes	2
Installed Memory	1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s])
NIC	1x ThinkSystem Broadcom 57508 100GbE QSFP56 2-Port PCIe Ethernet Adapter, 1x Ethernet Controller E810-XXV for SFP
Disk	10x 3.2TB Intel SSDPF2KE032T1O, 1x 1TB Micron_7450_MTFDKBA960TFR
BIOS	ESE114R-2.14
Microcode	0x2b0004b1
OS	Red Hat Enterprise Linux 8.8 (Ootpa)
Kernel	4.18.0-477.21.1.el8_8.x86_64

Table 2. Other Configuration Details

Software Configuration	Pytorch Llama2 model BF16 precision
Framework /Toolkit	Torch 2.2.0.dev20230911+cpu IPEX 2.2.0+git880fda9/llm_feature_branch Deepspeed 0.10.2+f15e6d48 Transformers 4.31.0
Topology or ML Algorithm	meta-llama/Llama-2-7b-hf, meta-llama/Llama-2-13b-hf
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)	Token Length 32/128/1024/2048 (in); Token Length 32 (out)
Precision	BF16
Warmup Steps	10
Num Interations	100
Batch Size	1, 2, 4, 8, 16
Beam Width	1 (greedy search)
Input Token Size	32, 256, 1024, 2048
Output Token Size	32

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

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- ThinkSystem SR650 V3 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, LP1859, was created or updated on December 18, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1859
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1859.

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.