Deploy and Scale Generative AI models using TorchServe in Enterprises

Solution Brief

Deploy and Scale Generative AI in Enterprises

With Generative AI transforming and reinventing every aspect of business today, enterprises are constantly looking for the latest technologies to deploy and scale models efficiently to meet customer needs. From customer experience to business operations to employee engagements, choosing the right AI models, infrastructure, and underlying hardware is critical now for businesses to get the best outcome. Intel and Lenovo are continuing to meet their customer needs by enabling solutions that cater to different use cases.

Gen AI model inference at low latency supporting concurrent users is critical for enterprises to meet customer needs. Companies looking to start their journey on Generative AI can extend their existing infrastructure with the latest Lenovo ThinkSystem SR650 V3, accelerated by 5th Gen Intel Xeon Scalable processors, and achieve revolutionary business impacts leveraging processors ideal for mixed enterprise workloads.

Solution Overview

Intel testing has shown the Lenovo ThinkSystem SR650 V3, with the Xeon Scalable processors, delivers a highly performant, scalable solution for Generative AI across a variety of use cases, including offline content creation and real-time chatbots with low latency (a target latency of ~100ms)

This platform 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. It supports one or two Intel Xeon processors per node. With built-in Intel Advanced Matrix Extensions (Intel AMX), Intel Xeon CPUs deliver high performance on cutting-edge AI models.

Enterprises will 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. 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.

To optimize model performance and deployment, consider integrating TorchServe, a powerful platform for serving PyTorch models, with the SR650 V3. This combination accelerates model inference and simplifies the management of complex AI workloads.

The ThinkSystem SR650 V3 offers 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 testing, performed by Intel in March of 2024, leveraged online inferencing using TorchServe and the latest release of Intel Extension for PyTorch (IPEX) 2.2.0. With the setup using Red Hat OpenShift, topology and CPU and memory managers were configured to maximize performance. CoreOS and additional services were assigned a share of the total cores. The remaining cores were equally distributed across the two pods which deployed two instances of TorchServe. In this current platform, it translates to 128 total cores, of which 4 cores are assigned to CoreOS and other services. Refer to the Configuration details section for the configuration of the server.

Deploying two instances of TorchServe rather than a single instance yielded better performance results on a single node. A load balancer was further provisioned using Route exposing round robin logic to help equally balance the requests from multiple users to the two instances of TorchServe. This ensured for every concurrent user testing scenario, each instance receives exactly half of the requests which makes the 2nd token latency the same, ~±1ms, on both instances.

Testing the setup by varying the number of concurrent users accessing the TorchServe instances endpoint, we captured the LLM inference times, including next-token latency, first token latency, and the total round-trip time (RTT) for each scenario of concurrent users. Batch size per TorchServe instance is dynamic and is handled by a load balancer. The load balancer that serves all incoming inference requests passes them to TorchServe instances in round-robin way. With 2 TorchServe instances, the number of user requests processed per instance per benchmark iteration is [num_of_users]/2. These requests, after reaching their TorchServe instance within the maxBatchDelay window (5000 ms), are dynamically grouped into a single batch to optimize processing efficiency and resource utilization. The maximum size of this dynamic batch was set to 64 (on each instance) in our test scenarios.

The graph below shows the single node online inference performance for both INT8 and BFloat16 datatypes accelerated via Xeon processors with built-in AI acceleration with Intel AMX. With the models in the X-axis and Number of concurrent users supported on the Y-axis, we scale the input token size from 256, 1024 and 2048 in Z-axis. Targeting a next-token latency of ~<100ms, we validate how many concurrent users can be served by the model serving setup described above on a single node. The models tested are Llama2-7B and Llama2-13B, with optimizations enabled by Intel for BF16 and INT8 quantized precisions.

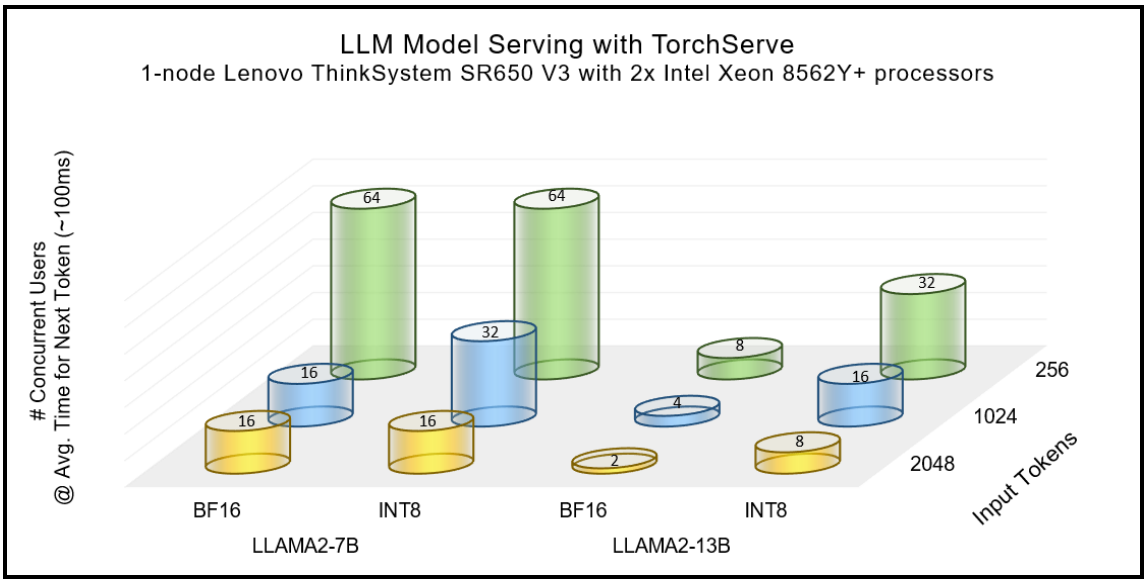

Figure 1. LLM Model Serving with TorchServe

With Lenovo ThinkSystem SR650 and Intel Xeon 8562Y+ processors, online inference using TorchServe on Red Hat OpenShift infrastructure provides for ease of use in deployment and scaling to meet concurrent user needs.

- LLAMA2-7B and LLAMA2-13B (BF16 and quantized) models can support up to 64 and 32 concurrent users @ next token latency requirements (~<100ms) for input token size 256, output tokens: 256, Greedy Search.
- LLAMA2-7B and LLAMA2-13B (BF16 and quantized) models can support up to 32 and 16 concurrent users @ next token latency requirements (~<100ms) for input token size 1024, output tokens: 256, Greedy Search.

Modeling for a Daily Active User (DAU) chatbot for an enterprise, the data below is based on the number of chatbot sessions that can be processed over a 24-hour period, using Llama2-7B INT8 concurrency data for 1024 input tokens and 256 output tokens, assuming next token latencies of ~100ms. Each session consists of four separate prompts to mimic a chatbot conversation. Based on provisioning for a single server, an average of 2.3% of the DAUs would be able to access the chatbot simultaneously at a given time without experiencing additional latency, supporting ~1500 Daily active users in any given 24-hour period*.

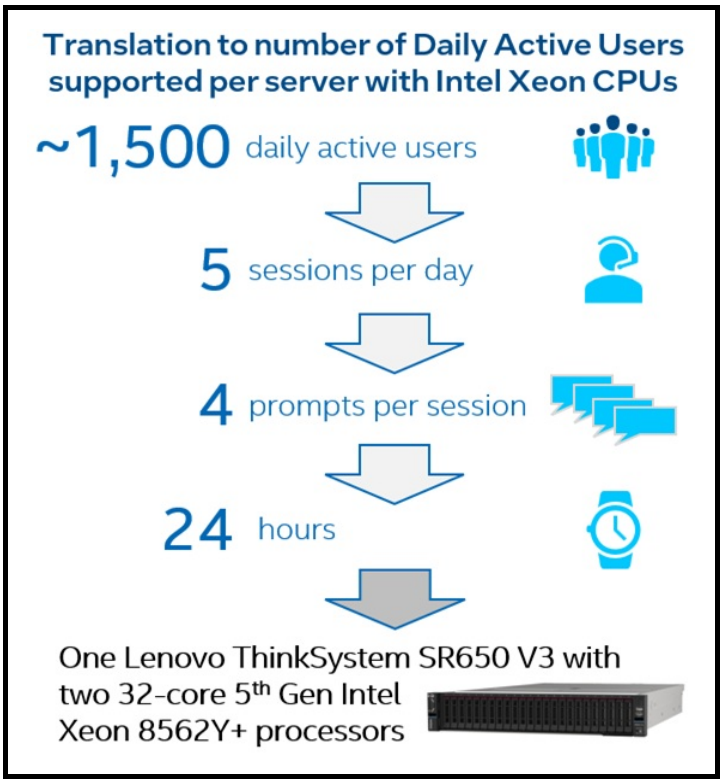

Figure 2. Number of daily active users per server

* A further explanation for the DAU scenario chart: Each one of your 1500 DAUs will query the chatbot 5 times over the course of a 24-hour day. Similar to how you interact with a chatbot, very rarely do you get your answer in a single prompt, so there will be some back-and-forth. To make it a more realistic analysis, we're assuming each query (or session) will consist of 4 separate prompts to mimic an actual back-and-forth with the user and the chatbot. You can then calculate the actual server runtime by taking 1500 DAUs, multiplying it by 5 queries and 4 prompts per query – in this example that's 30,000 total prompts in a workday. You then multiply that result by the total latency time for each prompt, using both the first and next token latency times. Once you have the total server runtime (in hours), you can then divide by 24 hours to calculate the number of servers needed to perform all the queries for that workday.

Conclusion

In conclusion, the deployment and scaling of Generative AI models using TorchServe on Lenovo ThinkSystem SR650 V3, powered by 5th Gen Intel Xeon processors, offer enterprises a powerful, flexible, and efficient solution for meeting their AI-driven goals. The combination of high-performance hardware, advanced AI acceleration, and robust infrastructure capabilities ensures that businesses can handle diverse workloads and large-scale user demands with ease.

This setup not only provides the necessary computational power and efficiency but also supports energy-saving features, reducing operational costs and environmental impact. As enterprises continue to integrate Generative AI into their operations, leveraging such innovative solutions will be key to maintaining competitive advantage and driving transformative business outcomes.

Configuration Details

The following table lists the server and OS configuration.

Table 1. Server and OS configuration

Parameter	Detail
Server	Lenovo ThinkSystem SR650 V3
Processor	Intel Xeon Platinum 8562Y+ processor
Microarchitecture	EMR_MCC
Sockets	2
Cores per Socket	32
Hyperthreading	Intel Hyper-Threading Technology Enabled
CPUs	128
Turbo	Intel Turbo Boost Technology Enabled
Base Frequency	2.8GHz
All-core Maximum Frequency	3.8GHz
Maximum Frequency	4.1GHz
NUMA Nodes	2
Installed Memory	512GB (16x32GB DDR5 5600 MT/s [5600 MT/s])
NIC	1x ThinkSystem Intel E810-DA2 10/25GbE SFP28 2-Port OCP Ethernet Adapter, 1x ThinkSystem I350-T4 PCIe 1Gb 4-Port RJ45 Ethernet Adapter
Disk	2x ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD, 1x ThinkSystem 2.5" U.2 P5620 3.2TB Mixed Use NVMe PCIe 4.0 x4 HS SSD
BIOS	ESE124B-3.11
Microcode	0x21000200
OS	Red Hat Enterprise Linux CoreOS 414.92.202312011602-0 (Plow)
Kernel	5.14.0-284.43.1.el9_2.x86_64

The following table lists the software configuration.

Table 2. Software configuration

Parameter	Detail
SW versions	cmake-3.20.2, findutils-4.6.0, zip2-1.0.6, gcc-8.5.0, gcc-c++-8.5.0, gcc-toolset-12-12.0, gcc-toolset-12-runtime-12.0, git-2.39.3, gperftools-devel-2.7-9.el8, libatomic-8.5.0, libfabric-1.18.0, procps-ng-3.3.15, python3-distutils-extra-2.39, python39-3.9.18, python39-devel-3.9.18, python39-pip-20.2.4, unzip-6.0, wget-1.19.5, which-2.21, java-17-openjdk-17.0.10.0.7, intel-oneapi-openmp-2023.2.1, torch 2.2.1+cpu, ninja 1.11.1.1, accelerate 0.25.0, sentencepiece 0.1.99, protobuf 4.25.1, datasets 2.15.0, transformers 4.35.0, wheel 0.42.0, torchserve 0.9.0, intel_extension_for_pytorch 2.2.0, oneccl_bind_pt 2.2.0+cpu
Orchestration	RedHat OpenShift, Kubernetes Version: v1.27.8+4fab27b
LLM Models	LLaMA 7B v2, LLaMA 13B v2
Dataset	LAMBADA, License: Creative Commons by 4.0
Compiler	gcc version 12.3.0 (GCC)
Dataset	Token Lengths: 256/1024/2048 (in); 256 (out)
Precision	BF16, INT8
Warmup steps	10
Num Iterations	60
Number of Users	1,2,4,8,16,32,64
Beam Width	1 (greedy search)
Batch size per inference server (TorchServe instance)	max 64. Actual batch size processed per instance on a given scenario depends on the load balancer (Round Robin) that passes requests among the available instances. Requests are then dynamically grouped into larger batch on each instance to optimize processing efficiency. Standard batch size per instance: [Num of users]/[Num of instances]

TorchServe Configuration	minWorkers: 1
	maxWorkers: 1
	maxBatchDelay: 5000
	responseTimeout: 720
	parallelType: "tp"
	deviceType: "cpu"
	int8_enabled: False
	torchrun:
	  nproc-per-node: 1
	  OMP_NUMBER_THREADS: 30

Note: Performance varies by use, configuration, and other factors. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. No product or component can be absolutely secure. Your costs and results may vary. Intel technologies may require enabled hardware, software, or service activation.

For more information

For more information, see the following resources:

- Artificial Intelligence solutions from Lenovo:
  https://www.lenovo.com/ai
- ThinkSystem SR650 V3 datasheet:
  https://lenovopress.lenovo.com/datasheet/ds0143-lenovo-thinksystem-sr650-v3
- Intel AI Performance:
  https://www.intel.com/content/www/us/en/now/ai-performance.html
- Intel Extension for PyTorch
  https://intel.github.io/intel-extension-for-pytorch/
- TorchServe with Intel Extension for PyTorch
  https://intel.github.io/intel-extension-for-pytorch/cpu/2.3.100+cpu/tutorials/performance_tuning/torchserve.html
- Llama-2-7b-hf
  https://huggingface.co/meta-llama/Llama-2-7b-hf
- Llama-2-13b-hf
  https://huggingface.co/meta-llama/Llama-2-13b-hf

Authors

David Ellison is the Chief Data Scientist for Lenovo ISG. Through Lenovo's US and European AI Discover Centers, he leads a team that uses cutting-edge AI techniques to deliver solutions for external customers while internally supporting the overall AI strategy for the World Wide Infrastructure Solutions Group. Before joining Lenovo, he ran an international scientific analysis and equipment company and worked as a Data Scientist for the US Postal Service. Previous to that, he received a PhD in Biomedical Engineering from Johns Hopkins University. He has numerous publications in top tier journals including two in the Proceedings of the National Academy of the Sciences.

Ajay Dholakia is a Principal Engineer, Master Inventor, AI Leader and Chief Technologist for Software & Solutions Development with Lenovo ISG. His current focus is on solution architectures in the areas of AI / ML, Generative AI, Data Analytics, Edge Computing, and Blockchain. In his more than 30 years with Lenovo, and IBM before that, Ajay has led diverse projects ranging from research and technology to product development, as well as business and technical strategy. Ajay holds more than 60 patents and has authored over 60 technical publications including a book. He received PhD in Electrical and Computer Engineering from N.C. State University and MBA from Henley Business School.

Mishali Naik, Ph.D. is a Sr. Principal Engineer in Intel's Data Center and AI (DCAI) organization. She is currently leading Enterprise Solutions Engineering for the Market Readiness organization. Her interests include AI, computer systems and distributed architecture, application-level performance analysis and optimization, integrated HW-SW solutions and co-design, as well domain-specific customization.

Abirami Prabhakaran is a Principal Engineer in Intel's Data Center and AI (DCAI) organization. Part of the Market Readiness team, she is the solution architect for end-to-end AI solutions. Her focus includes enablement and performance optimizations of AI and analytics use cases, distributed infrastructure performance and power optimization.

Edward G Groden is an AI Sales Enabling Manager in Intel's Sales and Marketing Group and works with worldwide internal and external partner sales organizations to deliver content, training, and tools to help drive Intel AI platform sales.

Andy Morris is an industry veteran in cloud computing and AI/M, and is currently leading Enterprise AI marketing at Intel.

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, LP1998, was created or updated on August 7, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1998
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1998.

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.

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.