NEC's AI Supercomputer: One of the Largest in Japan to Support Generative AI

KITANO Takatoshi

Abstract

In recent years, generative AI has dramatically evolved due to increases in the number of parameters and the volume of training data. Furthermore, the multimodalization of generative AI, which encompasses language, images, and audio, requires significantly increased computational power. Therefore, a system that efficiently and stably conducts large-scale distributed training using numerous GPUs is essential for developing foundation models for generative AI. This paper introduces the system architecture of NEC's AI supercomputer that supports the training of generative AI and future initiatives.

Keywords
generative AI, large language model, AI supercomputer, data center network, distributed file system, Kubernetes, deep learning framework

1. Introduction

In recent years, generative AI has made dramatic progress due to increases in the number of parameters and the volume of training data. Furthermore, the advancement of multimodal generative AI, which incorporates not only language but also images and audio, requires significantly increased computational power. Because large language models have now emerged that would take over 350 years to train on a single graphics processing unit (GPU), distributed training systems that can efficiently and reliably leverage multiple GPUs have become indispensable for advancing generative AI technologies.

NEC anticipated the increasing importance of computational power with advancing AI sophistication and multimodalization. In March 2023, NEC began full operation of an AI supercomputer equipped with 928 NVIDIA A100 Tensor Core GPUs, making it the largest in the industry in Japan[1)]. This paper describes the system architecture of NEC's AI supercomputer that enables fast and stable learning of large-scale generative AI foundation models from two perspectives: the computing infrastructure and the software infrastructure that constitute the AI supercomputer[2)]. Then, we explain the computing infrastructure, which consists of the advanced server computer architecture and the innovative network architecture. Next, we discuss the software infrastructure for execution and operational monitoring centered on Kubernetes. Finally, we explain future initiatives.

2. Architecture of the AI Supercomputer's Computing Infrastructure

2.1 Computer architecture

NEC's AI supercomputer comprises 928 NVIDIA A100 80GB Tensor Core GPUs and 116 GPU servers. The computer architecture and network architecture of the AI supercomputer's computing infrastructure have been designed to enable high-speed distributed training by scaling out such a large number of GPUs. Next, we explain the details.

2.1.1 Scaling-out method for GPU servers

In NEC's AI supercomputer, the combination of 16 GPU

servers and multiple sets of leaf switches is designated as the unit for scaling out, and this unit is referred to as the computing unit (CU). All CUs have the same configuration, and scaling out of GPU servers is achieved by arranging CUs (**Fig. 1**). Multiple CUs can be scaled out to any size by connecting them through spine switches and they also have remarkably high maintainability due to unified hardware and software.

2.1.2 Hardware topology supporting high-speed distributed training

To perform high-speed distributed training with multiple GPUs, a hardware topology that reduces communication latency between GPU servers is important. To achieve low-latency communication, one network interface card (NIC) is placed under the PCI switch to which the GPU belongs, minimizing the physical distance between the NIC and the GPU. In total, four NICs are arranged in this manner. This enables communication with low latency and a broadband using GPUDirect RDMA, bypassing the CPU (**Fig. 2**).

In addition, the network is divided into three separate networks: management, computing, and storage. For each network, a total of six NICs are placed to separate communication traffic: one for management, four for computing, and one for storage. This separates the communications for management and storage, preventing communication latency between GPU servers from worsening.

In this way, by arranging the NICs so that communication latency is reduced for parameter exchanges in the deep learning framework, high-speed distributed training can be done even for large-scale generative AI foundation models.

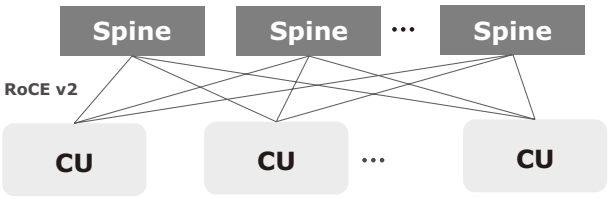

Fig. 1 Scaling out GPU server clusters with CUs.

2.2 Network architecture

As for the network architecture, low-latency and high-bandwidth communication is achieved by integrating GPU servers and various high-speed switches in an end-to-end manner, using a spine-leaf architecture (**Fig. 3**). Next, we explain the details.

2.2.1 Architecture for high-speed communication for distributed deep training

In distributed deep training, numerous parameters must be communicated between GPU servers for each iteration. Also, because foundation models for generative AI have a large number of parameters, low-latency and high-bandwidth communication is required to improve distribution efficiency.

Therefore, it is essential to enable GPUDirect RDMA in an end-to-end manner, using RoCE v2, for communication. For this reason, all servers are connected by using NVIDIA's Spectrum SN3700 high-speed Ethernet switches, which support 200 Gbps ultrahigh-speed Ethernet, and NVIDIA ConnectX-6 low-latency interconnects. Also, the network configuration adopts a two-layer spine-leaf architecture to minimize the number of hops between GPU servers, thereby realizing low-latency communication.

Thanks to these innovations, communication between

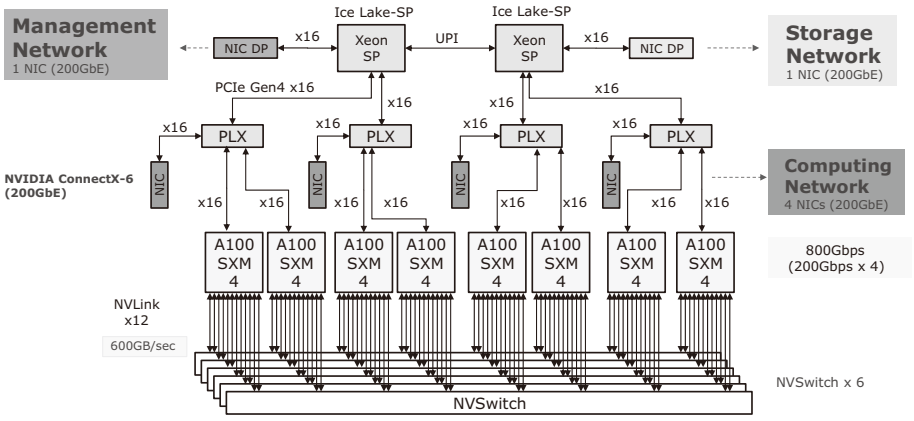

Fig. 2 NIC placement for high-speed and stable distributed training.

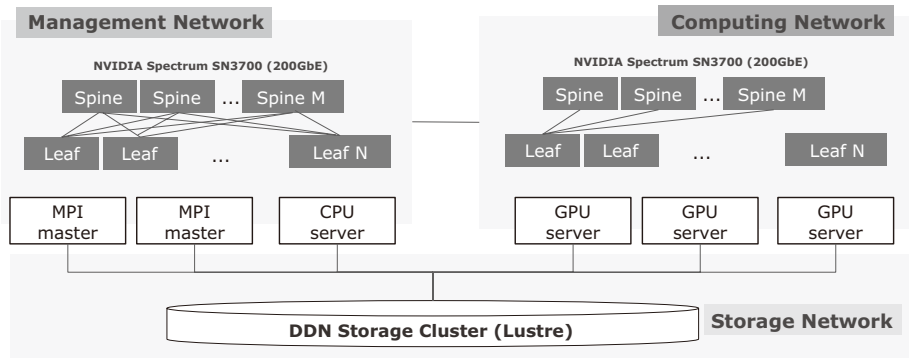

Fig. 3 High-speed communication using spine-leaf architecture.

GPU servers can be achieved with low latency and a high bandwidth, enabling high-speed distributed training even with hundreds of GPUs.

2.2.2 Use of data center network technologies

In NEC's AI supercomputer network, data center network technologies are used to construct the AI supercomputer (**Fig. 4**).

The overall network architecture adopts IP Clos and uses EVPN as the control plane and VXLAN as the data plane. In addition, RoCE v2 over VXLAN is used to enable high-speed communication between servers. Also, BGP is used as the routing protocol, and BGP unnumbered is used to facilitate IP management.

By leveraging a variety of data center network technologies, we have realized a high-performance network that enables large-scale distributed training while improving convenience and reducing network management costs.

2.3 Storage architecture

To develop multimodal generative AI, diverse modal datasets, including text, images, videos, and audio, are required and necessitate high-speed broadband access. To achieve this, a distributed file system capable of supporting high input/output operations per second (IOPS) and broadband access is necessary, and this enables simultaneous access from multiple GPU servers and the use of large volumes of language data and large datasets containing tens of millions of images. NEC's AI supercomputer uses Lustre, a distributed file system with a strong track record in the world of high-performance computing (HPC). Next, we explain the configuration of a large-scale storage system that uses Lustre and the method of scaling out storage to enable high-speed access to large datasets.

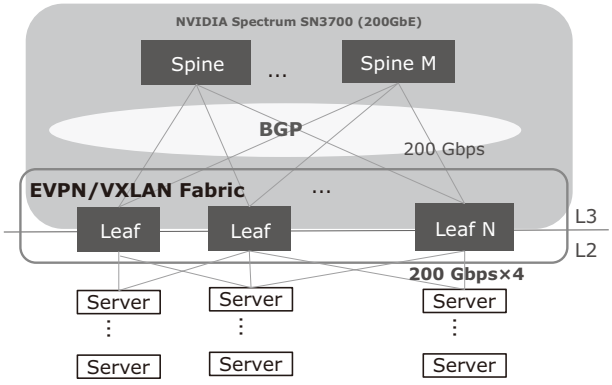

Fig. 4 Use of data center network technologies.

2.3.1 Configuration of the storage system supporting large datasets

The storage system of the AI supercomputer consists of two components: the high-speed area consisting of many NVMe (nonvolatile memory express) solid-state drives (SSDs) for high-performance IOPS applications such as logs and the large-capacity area consisting of hard drive disks (HDDs) used to store large-scale training datasets. By dividing the storage according to the input/output characteristics, high-speed data access is achieved. Also, the storage servers that make up the storage system are equipped with multiple 200 Gbps NICs, making RoCE v2 available even for data access to enable broadband communication.

2.3.2 Method of scaling out storage

In large-scale file systems, metadata processing tends to be a bottleneck because of the large number of files. Therefore, we enhance metadata processing performance by scaling out multiple metadata servers (MDSs)

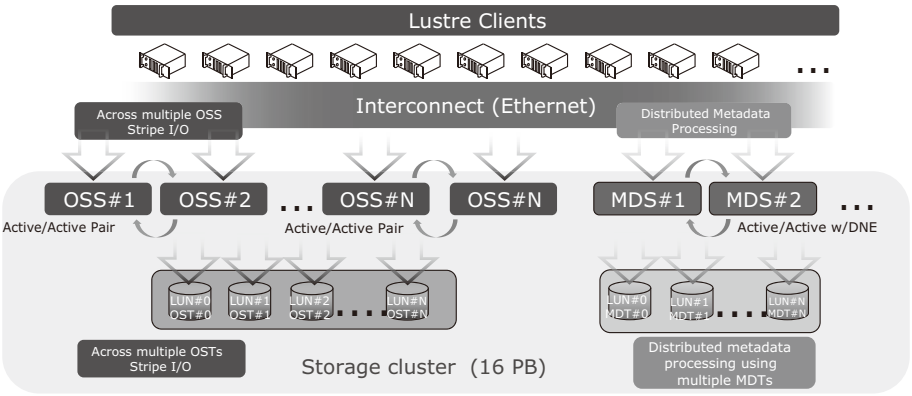

Fig. 5 Storage architecture supporting high-performance I/O.

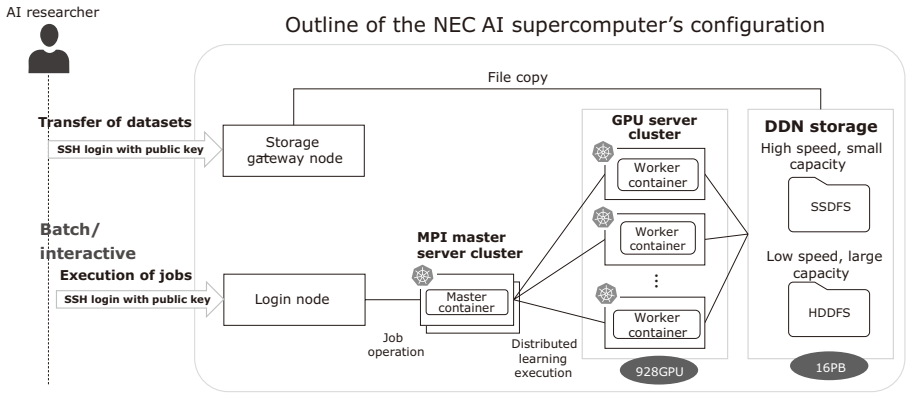

Fig. 6 Execution platform for software on the AI supercomputer.

(**Fig. 5**). Also, to prevent user data access from being concentrated on specific MDSs, we have developed a program that automatically arranges directories so that each user can use different MDSs and thus distribute the I/O load on metadata. These efforts enable high-speed data access in large-scale file systems without bottlenecks.

3. Software Architecture Supporting Advanced AI

3.1 Execution platform for software using Kubernetes

To enable researchers to efficiently train cutting-edge AI without the need to construct a complex deep learning environment, Kubernetes, a container platform, is installed as the core software and is extended to construct the execution platform for software (**Fig. 6**). Specifically, the authentication and authorization mechanism as well as the security and other mechanisms are extended to suit NEC's systems. In addition, a Kubernetes job scheduler has been optimized for the physical configuration of the AI supercomputer's network topology, enabling efficient distributed training.

By installing Kubernetes at the core of the software platform, we have achieved a cutting-edge research and development environment for AI that enables flexible expansion and efficient distributed training in accordance with organizational requirements.

3.2 Deep learning container environment supporting advanced AI research

A wide variety of deep learning frameworks is used for AI training, and updates are frequent due to the rapid advancement of AI technology. In addition, NEC conducts research and development in a variety of AI fields such as image recognition, video analysis, language modeling, optimization, and control. Therefore, in accordance with NEC's research and development, the latest deep learning environments are packaged as containers and provided to NEC researchers.

As a result, researchers can use cutting-edge deep

learning environments without the need to construct environments for the wide variety of diverse, complex, and frequently updated deep learning frameworks.

3.3 Operational monitoring platform supporting stable operation of large-scale systems

In the development of generative AI, simultaneously using a large number of GPUs for distributed training necessitates high availability in the AI supercomputer. However, due to the high utilization of multiple GPUs, the failure rate is higher than in typical systems, making operational monitoring important.

Therefore, NEC has developed an operational monitoring platform that conducts monitoring of the heartbeat, logs, resources, and performance—which are specific to the AI supercomputer—in an integrated manner to realize high availability. Specifically, in addition to general monitoring items, various types of monitoring are conducted in different layers, such as the metrics on Layer 1 of the optical fibers connecting GPU servers that are specific to the AI supercomputer, the communication metrics specific to distributed training such as RoCE v2, and errors in the hardware/software of each GPU. This enables the integrated detection of faults and failures on each layer, immediately removing abnormal nodes from Kubernetes and ensuring high stability.

The operation monitoring platform uses Prometheus, Grafana, Fluent Bit, and Elasticsearch. These software tools allow all monitoring functions to be extended using the Go language. By ensuring that the extensions can be developed in the same Go language as the execution platform, engineering scalability is enhanced, enabling efficient development of operation monitoring functions for the entire complex, advanced AI supercomputer.

4. Conclusion

In the future, generative AI will evolve into multimodal generative AI that encompasses language, images, and audio and it will evolve into a fundamental technology supporting advanced decision-making by humans. Even foundation models for language alone require a large amount of computation, so the development of multimodal generative AI technology requires immense computational power.

Therefore, NEC will leverage the largest-scale AI supercomputer—which serves as a source of competitiveness—in the industry in Japan to accelerate the research and development of foundation models for multimodal generative AI, using NEC's strengths in language, image, and speech AI technologies. Through collaboration with customers and partners, NEC aims to realize a center of excellence for AI research that generates innovative social value.

References

1) NEC Press Release: NEC begins to build the largest supercomputer in Japan to advance the study of AI, May 2022 (Japanese)
https://jpn.nec.com/press/202205/20220517_02.html
2) Accelerating Social Value Creation: NEC's AI Supercomputer.
https://www.nec.com/en/global/rd/aisupercomputer/index.html

Author's Profile

KITANO Takatoshi
Director
Global Innovation Strategy Department

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.17 No.2 Special Issue on Revolutionizing Business Practices with Generative AI
– Advancing the Societal Adoption of AI with the Support of Generative AI Technologies

Remarks for Special Issue on Revolutionizing Business Practices with Generative AI
Approaches to Generative AI Technology: From Foundational Technologies to Application Development and Guideline Creation

Vol.17 No.2
June 2024

Special Issue TOP

Papers for Special Issue

Market Application of Rapidly Spreading Generative AI
NEC Innovation Day 2023: NEC's Generative AI Initiatives
Streamlining Doctors' Work by Assisting with Medical Recording and Documentation
Using Video Recognition AI x LLM to Automate the Creation of Reports
Understanding of Behaviors in Real World through Video Analysis and Generative AI
Automated Generation of Cyber Threat Intelligence
NEC Generative AI Service (NGS) Promoting Internal Use of Generative AI
Utilization of Generative AI for Software and System Development
LLMs and MI Bring Innovation to Material Development Platforms
Disaster Damage Assessment Using LLMs and Image Analysis

Fundamental Technologies that Enhance the Potential of Generative AI
NEC's LLM with Superior Japanese Language Proficiency
NEC's AI Supercomputer: One of the Largest in Japan to Support Generative AI
Towards Safer Large Language Models (LLMs)
Federated Learning Technology that Enables Collaboration While Keeping Data Confidential and its Applicability to LLMs
Large Language Models (LLMs) Enable Few-Shot Clustering
Knowledge-enhanced Prompt Learning for Open-domain Commonsense Reasoning
Foundational Vision-LLM for AI Linkage and Orchestration
Optimizing LLM API usage costs with novel query-aware reduction of relevant enterprise data

For AI Technology to Penetrate Society
Movements in AI Standardization and Rule Making and NEC Initiatives
NEC's Initiatives on AI Governance toward Respecting Human Rights
Case Study of Human Resources Development for AI Risk Management Using RCModel

NEC Information

2023 C&C Prize Ceremony