NEC ADVANCES AI RESEARCH WITH ADVANCED GPU SYSTEMS FROM SUPERMICRO

NEC uses Supermicro GPU servers with NVIDIA® A100s for Building a Supercomputer for AI Research

INDUSTRY

AI Research

CHALLENGES

- Requirement for 580 Petaflops of GPU+CPU performance.
- Trusted partner with the ability to deliver high-end servers with GPUs.

Introduction

NEC has been a leader in designing and distributing a wide range of systems for many years. In addition, NEC has a track record of half a century in AI technology R&D and created many of the world's highest levels of AI in various AI fields, such as image and video recognition, language and semantic understanding, data analysis, predictive detection, and optimal planning and control.

NEC is the world's No.1 in face authentication and iris authentication, and the world's No.8 company in terms of the number of acceptances at top international conferences on machine learning. NEC conducts AI research at all its laboratories, and the number of AI researchers is in the hundreds. In this way, NEC has one of the world's leading AI research and development capabilities.

NEC chose Supermicro GPU and Ultra systems to strengthen the technological competitiveness of AI and to maintain and strengthen business competitiveness in the field of AI.

Challenges

NEC recognized that its challenge is accelerating the creation of advanced AI and speeding up social value creation. In the actual development of AI, while changing data, algorithms, etc., trial and error will be repeated thousands of times. However, due to the increase in the required amount of computation of deep learning, there are some cases

where learning consumes up to several thousand hours per training run, and in some cases, even more.

For example, GPT-3, a huge language model, took up to 355 years of GPU time to learn with a single NVIDIA Tesla V100 GPU. The time needed for this type of training run meant that results were not usable for many industries, where training needed to be completed in hours, not years. This real-world example shows that an AI supercomputer is necessary for many training models, which requires in-house expertise or a trusted supplier to design and implement such a large cluster.

Therefore, by using AI supercomputers, it is expected that advanced AI algorithms can be researched and developed without the restriction of limited computing resources. Additionally, an extensive AI system needs to have the capacity so that hundreds of researchers have simultaneous access to the system to advance their specific needs.

Solution

NEC decided to work with Supermicro to design and deliver the NEC AI supercomputer. There were two main reasons that NEC chose Supermicro as a strategic partner.

1) NEC was able to freely customize and configure the internal configuration on the GPU server according to the architecture for the AI supercomputer required by NEC. However, to optimize for AI learning workloads and deep learning, NEC needed to change or reconfigure many things, including the server's physical hardware configuration, BIOS settings, and fan settings. Therefore, the ability to work with Supermicro to modify these settings was critical to selecting the Supermicro GPU servers.

2) NEC realized that as the technology moves forward, the hardware that would be used needed to be extensible and able to house future generations of AI acceleration. In addition, as Deep Learning evolves dramatically from year to year, the freedom and expandability of hardware that could freely change and extend the physical configuration were very attractive to NEC.

Since NVIDIA's GPUs are the de facto standard for many applications in deep learning frameworks such as PyTorch and TensorFlow, there are a lot of challenges or difficulties with using non-NVIDIA GPUs, and it is difficult to conduct research on deep learning at this time if their underlying infrastructure was to be changed. Therefore, NEC definitely believes that their AI researchers would best be able to perform their research on NVIDIA GPUs.

NEC choose to use the NVIDIA A100 Tensor Core GPUs. The main reason is that the total calculation speed for AI operations using the NVIDIA A100 is exceptional. In addition, the GPU memory bandwidth of the A100 is 2TB/s, but in deep learning, it is basically a memory bandwidth limit. So, with the TF32 used in the A100, the range is the same as the FP32, but the precision can be treated as FP16, and although the accuracy is not compromised, the memory bandwidth bottleneck can be alleviated. In other words, it is easier to take advantage of the calculation performance of the A100. In particular, many groups in NEC are researching and developing AI using images, such as biometric authentication, image recognition, and video recognition, to be more memory-limiting for GPUs. Therefore, NEC has adopted the A100 with the TF32 supported.

SOLUTION

Supermicro GPU and Ultra Servers

SYS-420GP-TNAR

- Dual 3rd Gen Intel® Xeon® Platinum 8358 Processor
- 1 TB Memory/Node
- 8 x NVIDIA A100 80GB Tensor Core GPU SXM
- NVIDIA HGX™ platform with NVIDIA NVLink™ and NVSwitch™

SYS-120U-TNR

- Dual 3rd Gen Intel Xeon Gold 6342 Processors
- 256GB Memory/Node

BENEFITS

- Faster and More Complete AI Workloads
- Future ready for next generation GPUs

The second reason that NEC adopted the HGX A100 is that it is equipped with an NVIDIA A100 Tensor core GPUs, and it uses the third generation NVIDIA NVSwitch™ to communicate between GPUs at 600GB/sec. This internal switch is required because data communication between GPUs can be performed at high speed in the server, so it can learn at a very high rate.

Additionally, NEC decided that the GPUs needed to communicate with each other for training applications. With the GPU-to-GPU communication within, multiple GPUs can communicate at the high speed of 600 GB/s, NEC was confident that they could fully utilize the computing performance of the GPU with distributed learning processing, so the AI R&D efficiency will be improved significantly.

Another reason that NEC chose NVIDIA's products for network switches and NICs is the ConnextX-6 solution, which supports end-to-end RoCEv2 communication from the server to the switch, and the actual performance can be close to the limit of 200GbE. In distributed deep learning processing, it is necessary to process each iteration of AllReduce to exchange parameters, and a significant amount of communication occurs within the clusters. Therefore, in recent large-scale AI models, the expected processing time will not be achieved if the communication speed is lower due to the narrow bandwidth and higher latency. As a result, the distributed efficiency isn't met, no matter how many servers are being used. To bypass the CPU and reduce latency, it is important to be able to perform NVIDIA GPUDirect® RDMA using RoCE v2. Supermicro GPU server and NVIDIA technologies can easily connect all switches and servers end-to-end, and NEC has adopted those products and solutions.

SYS-420G-TNAR

SYS-120U-TNR

NEC

NEC established itself as a leader in the integration of IT and network technologies while promoting the brand statement of “Orchestrating a brighter world.” NEC enables businesses and communities to adapt to rapid changes taking place in both society and the market as it provides for the social values of safety, security, fairness and efficiency to promote a more sustainable world where everyone has the chance to reach their full potential. For more information, visit NEC at https://www.nec.com

Solution Specifics:

Quantity	Supermicro Server	CPU Per Node	Memory Per Node	GPU Node
116 nodes with a total of 928 GPUs	SYS-420GP-TNAR	2 x 3rd Gen Intel Xeon Platinum 8358 Processors 32 cores, 2.6GHz	1,024GB	8 x NVIDIA A100 80GB NVLink, NVSwitch
119 nodes	SYS-120U-TNR	2 x 3rd Gen Intel Xeon Gold 6342 Processors 24 cores, 2.8GHz	256GB	

Benefits

After running many tests, NEC determined that with the new AI supercomputer, compared with one Tesla V100, the deep learning performance will be up to 4,600 times faster, NEC decided that the AI development time could be greatly reduced. As a result, For example, GPT-3, a huge language model, took up to 355 years of GPU time to learn NEC researchers will now be able to achieve significantly more research into many AI domains with the new AI Supercomputer from Supermicro.

"Computational capability is a source of competitiveness in the age of AI. We are very pleased to be able to significantly accelerate AI research by providing researchers with a distributed deep learning environment with amazing performance with GPU servers by NVIDIA and Intel technologies. NEC continues to work closely with Supermicro to accelerate the creation of social value by enhancing our AI research." - Takatoshi Kitano, Senior AI Platform Architect at NEC

For more information, please visit:

https://www.supermicro.com/en/products/gpu
https://www.nec.com/en/global/rd/aisupercomputer/