Case Study

AI in the Cloud
4th Generation Intel® Xeon® Processors
Intel® Accelerator Engines

NHN Cloud Makes Artificial Intelligence Easily Accessible to Customers With New AI Cloud Service

4th Gen Intel® Xeon® processors with built-in Intel® Accelerator Engines along with GPUs power NHN Cloud's new 88.5PF AI as a Service supercomputer (AIaaS).

Solution Summary

- Expanding services to offer Artificial Intelligence as a Service (AIaaS)
- Faster vector processing with Intel® AVX-512
- Improved deep learning (DL) performance using lower-precision without sacrificing accuracy with Intel® DL Boost

Executive Summary

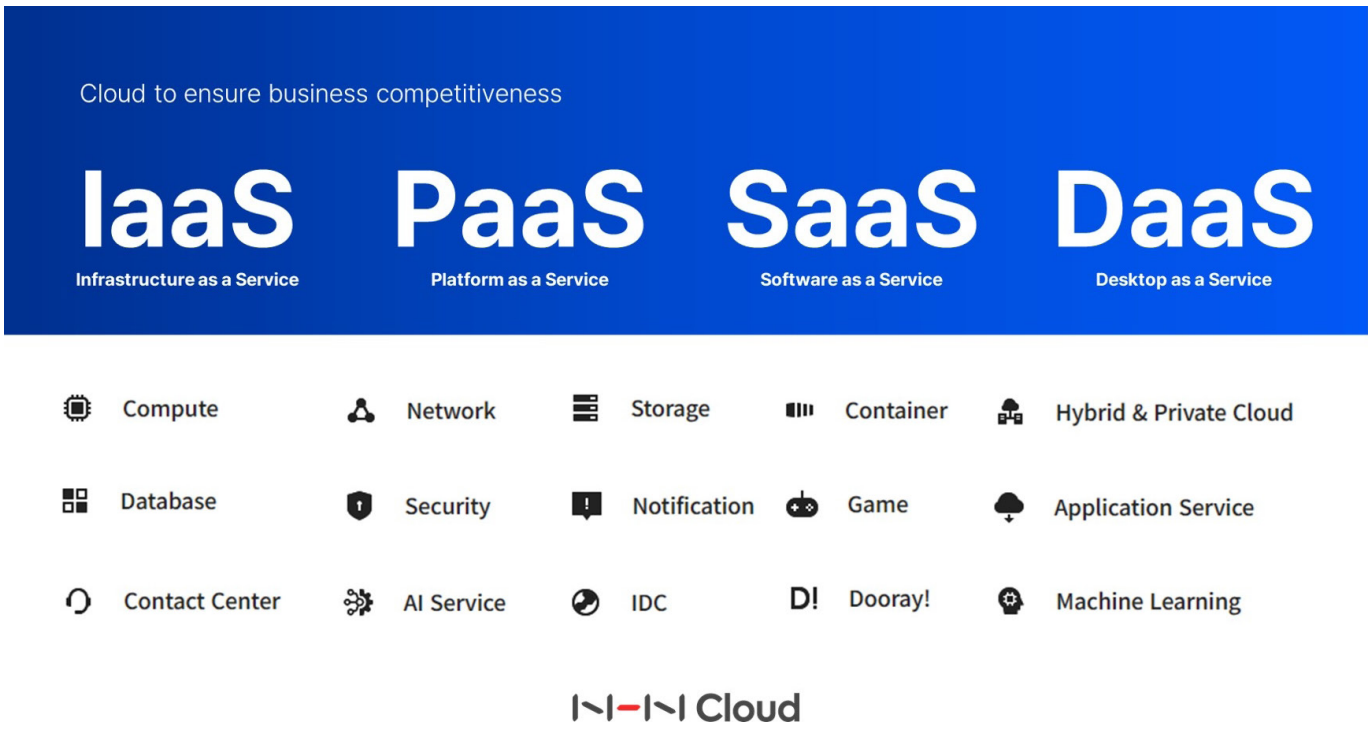
Artificial Intelligence is becoming a mainstream capability that businesses are using to stay competitive, but not all businesses can support an AI infrastructure. NHN Cloud Corp, a subsidiary of NHN, Korea's leading Cloud Service Provider (CSP), offers a range of IT services to both Korean and global customers. NHN Cloud decided to augment its existing cloud services infrastructure with Artificial Intelligence as a Service (AIaaS). To do so, NHN Cloud needed a new supercomputer designed to support cloud-based AI for a range of customers—from small businesses to enterprises and government agencies. NHN Cloud chose the 4th Gen Intel® Xeon® processor with GPUs to build their new 88.5 PetaFLOPS supercomputer for AIaaS.

Challenge

NHN Cloud delivers a full stack of Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS), including gaming, webtoon, music, advertisement, collaboration/communication, fintech, and others. NHN Cloud serves a range of customers worldwide, from small/medium to enterprise businesses and government.

Large and small companies are engaging Artificial Intelligence (AI) as a key asset to their business strategic and technical operations. Building AI infrastructure requires expertise that may not be within the core competency of many companies, but the value of AI deems it necessary for companies to access the

NHN Cloud's AI as a Service, powered by the Intel® 4th Gen Intel Xeon processor, supports a wide range of machine learning and deep learning capabilities.



NHN Cloud delivers a full stack of Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) to customers worldwide.

technology to stay competitive in their markets. This created an opportunity for NHN Cloud to support AI training and inference as a service.

With NHN Cloud's technical expertise in running IT operations, adding AIaaS is a natural step to expand its capabilities and services to customers around the world. To support an AIaaS vision, NHN Cloud built a new data center that focuses on cloud. The company is also expanding regional data centers to support local industry. But to support hyperscale AIaaS for industry and government, the company needed new computing resources built for SaaS, PaaS, IaaS, and AIaaS.

> "NHN Cloud's AI as a Service, powered by 4th Gen Intel Xeon processor architecture with its built in accelerator engines, will allow us to offer customers capabilities such as faster vector processing with Intel Advanced Vector Extensions 512, and improved deep learning performance using lower-precision without sacrificing accuracy with Intel Deep Learning Boost."
>
> *—Myung Shin Kim, NHN Cloud CTO*

Solution

NHN Cloud was able to provide a path that would enable their customers to quickly take advantage of the insights that artificial intelligence could offer. An AIaaS platform needed to support a wide range of machine learning, and deep learning capabilities. But for the company's market the infrastructure also needed to be designed to stand up instances for a range of customers' projects—from small/medium businesses to government and large enterprise projects. Such a system can offer more companies and institutions scalable AI resources.

With the cost of deploying and operating AIaaS, the system had to lead in performance, efficiency, and price/performance. To meet these diverse requirements, the company chose 4th Gen Intel Xeon processors with AI accelerator engines to build its new AI/HPC cloud infrastructure, which will be housed in their new data center. These latest generation Intel data center CPUs offer support for the latest I/O and memory, such as PCIe5 and DDR5, along with new power-saving features to enhance price/performance in the data center.

Additionally, the built-in Intel® Accelerator Engines of the 4th Gen Intel Xeon processor architecture enable NHN Cloud to offer customers processors with the most built-in accelerators of any CPU. The 4th Gen Intel Xeon processors will help speed processing of AI tasks with accelerated capabilities, such as faster vector processing with Intel® Advanced Vector Extensions 512 (Intel® AVX-512), improved deep learning performance using lower-precision (INT8) without sacrificing accuracy with Intel® Deep Learning Boost (Intel® DL Boost), and others. Additionally, with security being critical to many business operations, Intel Accelerator Engines speed up cryptography operations.

The new 88.5 PetaFLOPS system will be deployed later in 2023.

Result

With the infrastructure of the National AI Data Center based on Intel 4th Gen Xeon processors and GPUs, NHN Cloud expects to meet the high demands of their customers. The new supercomputer will provide up to 2.5 times better performance and newer features than the current system based on an earlier version of the CPU without GPUs.[1]

Compatibility with H100 GPU accelerators add significant performance and scale for the complexity of AI and HPC workloads. By launching a cloud service equipped with the latest Intel Xeon processor technology, NHN Cloud has secured capabilities that are highly competitive when compared to similar Cloud Service Providers in Korea and overseas.

Solution Summary

To deliver AIaaS to customers, NHN Cloud required advanced computing capabilities designed for AI workloads. The Korean CSP chose 4th Gen Intel Xeon processors along with GPUs to power its 88.5 pF supercomputer. Both the GPUs and the new Intel CPUs with Intel Accelerator Engines will deliver up to 2.5x more performance than its predecessor, according to NHN Cloud. The new system will be available to customers in late 2023.

Where to get more information

Learn more about NHN Cloud.

Explore the capabilities of the latest Intel® Xeon® processors with integrated Intel® Deep Learning Boost capabilities for accelerated AI inferencing.

Solution Ingredients

- Expanding services to offer Artificial Intelligence as a Service (AIaaS)
- Faster vector processing with Intel® AVX-512
- Improved deep learning (DL) performance using lower-precision without sacrificing accuracy with Intel® DL Boost.

[1] Performance benchmarks completed by NHN Cloud 11/1/2023.

102023/RJMJ/KD/PDF ♻ PleaseRecycle 356768-001US