HEROZ SELECTS SUPERMICRO SUPERBLADE TO BOOST AI INFERENCE PERFORMANCE

HEROZ

HEROZ Selects Supermicro SuperBlade to Provide Consistent Support for AI, Which Plays a Central Role in DX in Various Industries, from Conceptualization to Implementation and Operation

INDUSTRY

AI Gaming, AI Research

CHALLENGES

- High operating costs for GPU servers
- Power consumption
- Expansion expenses

Introduction

HEROZ focuses on "Core Operations," which are the source of value creation in various industries and implements the high-value Real-world AI technology in Core business. AI engineers in HEROZ developed the famed AI-Shogi, which defeated a professional Shogi player. They continue to work daily on developing other AI tools, including machine learning, which has culminated in developing games such as *Shogi Wars*, *CHESS HEROZ*, and *BackgammonAce.* HEROZ has attended the World Computer Shogi Championship for three years. HEROZ had won the championship and was runner-up multiple times. HEROZ is also developing a new AI system as a member of the Japan Deep Learning Association (JDLA). HEROZ is also a supporting member of the Japanese Society for Artificial Intelligence, and they keep abreast of cutting-edge trends in AI. Besides brain games, HEROZ-developed AI plays a key role in many other industries, including major financial institutions.

Challenges

The previous high-performance GPU server systems, NVIDIA DGX-1 V100 GPU server with many optional licenses, were too expensive to operate, and the power consumption was too high. In addition, the GPU server had performance bottlenecks on specific AI workloads, such as matrix multiplications, data preprocessing, and inference tasks. Thus, HEROZ required a new generation of servers that could handle the computing demands of HEROZ customers.

SOLUTION

Supermicro SuperBlade®

- 20 Blades & 1 InfiniBand Switch per Enclosure
- Total of 120 Blades

SBI-4129P-T3N Blade

- Intel® Xeon® Silver 4210R Processor
- Intel® AVX and Intel® AVX-512
- 512GB Memory
- InfiniBand EDR 100G HCA Card

SBI-421E-1T3N Blade

- Intel® Xeon® Silver 4416+ Processor
- Intel® AVX and Intel® AVX-512
- 512GB Memory
- InfiniBand HDR 200G HCA Card

SYS-1029P-N32R

- Intel Xeon Silver 4114
- 245TB NVMe storage

Solution

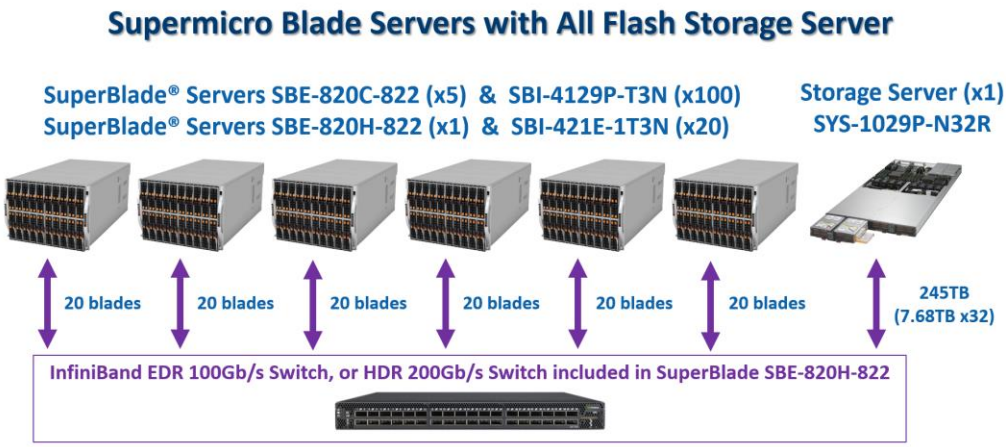
While GPUs generally offer higher raw performance for parallel processing tasks, the Intel Xeon Scalable Processors with built-in AI accelerators provided competitive performance for HEROZ's highly specialized AI workloads. The significant performance boost offered by the 8U SuperBlade system with Intel Xeon Scalable Processors allows for more complex and responsive AI in our gaming applications.

HEROZ uses the NVIDIA GPU server to train deep-learning models and the Supermicro SuperBlade server to implement a high-density CPU node for inference purposes based on the on-premises environment. HEROZ purchased Supermicro SuperBlade SBI-4129P-T3N of 100 nodes in 2019 and an additional 20 nodes of Supermicro SuperBlade SBI-421E-1T3N, which demonstrated a significant performance improvement by using dual 2nd Gen Intel® Xeon® 4210R Scalable Processors 10C/20T 2.4G, a total of 120 blades, and dual 4th Gen Intel® Xeon® 4416+ 20C of 40 blades on Supermicro SuperBlade servers.

To improve scalability, the high-density design of the SuperBlade system enables HEROZ to scale its infrastructure efficiently, allowing HEROZ to expand its computational resources to meet growing demands quickly.

Supermicro SuperBlade is designed to share cooling fans, power supplies, and networking in the SuperBlade system to enhance energy efficiency and reduce TCO.

HEROZ provides technologies for AI tools related to space comfort control, structural design support, business diagnosis, stock portfolio proposal, electricity market price prediction, image generation, brain games, etc. HEROZ will use the Supermicro SuperBlade to host and serve their application.



Solution Specifics:

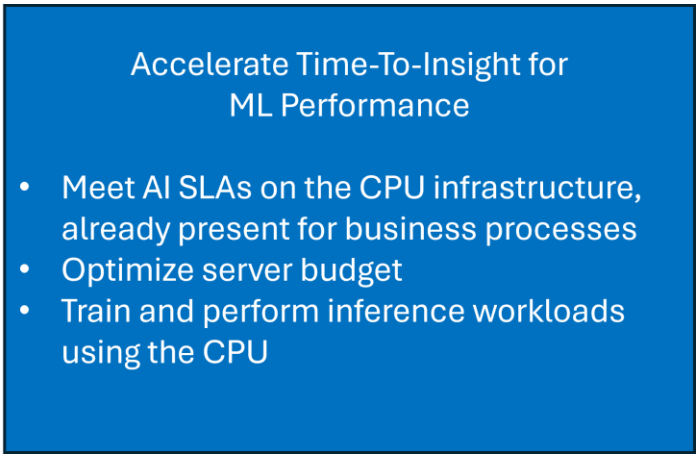
Quantity	Supermicro Server	CPU Per Blade (Node)	Memory Per Node
5x Blade Enclosures 100 Blades (Nodes)	SBE-820C-822 8U Enclosure with IB Switch	-	-
	SBI-4129P-T3N	Dual, 2nd Gen Intel® Xeon® Silver 4210R Processor 10C/20T, 2.4GHz Intel® AVX, Intel® AVX2, Intel® AVX-512	512GB
1x Blade Enclosure 20 Blades (Nodes)	SBE-820H-822 8U Enclosure with IB Switch	-	-
	SBI-421E-1T3N	Dual, 4th Gen Intel® Xeon® Silver 4416+ Processor 20C/40T, 2.0GHz Intel® AVX, Intel® AVX2, Intel® AVX-512	512GB
1x 245TB Storage Server	SYS-1029P-N32R	2 x Intel Xeon Silver 4114 Processor 20 cores, 2.2GHz,	512GB

Benefits

Intel® Xeon® Scalable Processors come with Intel Advanced Matrix Extensions (AMX), which significantly boost AI performance, especially for deep-learning tasks like natural language processing and recommendation systems. Intel Advanced Matrix Extensions (AMX) helps HEROZ improve the performance of its AI models, making its games more responsive and challenging. In addition, Intel Advanced Vector Extensions (AVX) enables HEROZ to perform multiple operations simultaneously, improving its AI algorithms' overall efficiency and speed.

BENEFITS

- Higher Performance
- Lower Power consumption
- Lower CAPEX and OPEX



The high-density and scalable 8U Supermicro SuperBlade system supports 20 blades in a single enclosure to meet increasing computational demands. Supermicro SuperBlade's design also enables optimal energy efficiency by sharing fans, power supplies, and networking throughout the whole SuperBlade system. This enhances energy efficiency, which is crucial for reducing OPEX. Hot-swappable blades and storage drives ensure high availability, flexibility, and easy maintenance. In addition, SuperBlade with embedded 200G InfiniBand switches can save up to 90% of external cables, and the future-proof enclosure can be reused for several generations, helping HEROZ minimize the expense of upgrading the system. HEROZ has chosen Supermicro SuperBlade with Intel Xeon Scalable Processors supported by Intel AMX or AVX-512 AI capabilities for their AI workloads due to these key factors.

"We are very pleased to be able to significantly accelerate AI technology by using deep learning and inference in a well-balanced on-premises solution. We continue to work closely with Supermicro to accelerate the target for the new Generative AI segment." -- Keiichi Iguchi, Chief Technology Officer, HEROZ, Inc.

SUPERMICRO

Supermicro is a global leader in high performance, green computing server technology and innovation. We provide our global customers with application-optimized servers and workstations customized with blade, storage, and GPU solutions. Our products offer proven reliability, superior design, and one of the industry's broadest array of product configurations, to fit all computational need.

Visit https://www.supermicro.com

HEROZ, INC.

HEROZ, founded by amateur Grandmaster of Japanese chess, possesses a unique core technology of artificial intelligence (AI), named as "HEROZ Kishin. "We have accumulated such technology as a machine learning and deep learning through the development of AI for Japanese Chess, Chess, Backgammon and Go. Now we are applying its cutting-edge AI technology into various industries to solve difficult issues. For more information, visit HEROZ at https://heroz.co.jp/en/