eastern
computer exchange

SUCCESS STORY

AI Infrastructure Deployment

INTRODUCTION

A Health Care Life Sciences (HCLS) customer sought to establish a robust Artificial Intelligence (AI) infrastructure to support its various lines of business. This initiative required a comprehensive technical solution encompassing hardware, software, networking, and security infrastucture to ensure scalability, reliability, and performance. To address these needs, a pre-engineered and pre-built Supermicro (SMCI) liquid-cooled Rack Scale solution was proposed, providing a turn-key AI infrastrcture ready for deployment.

CHALLENGES AND REQUIREMENTS

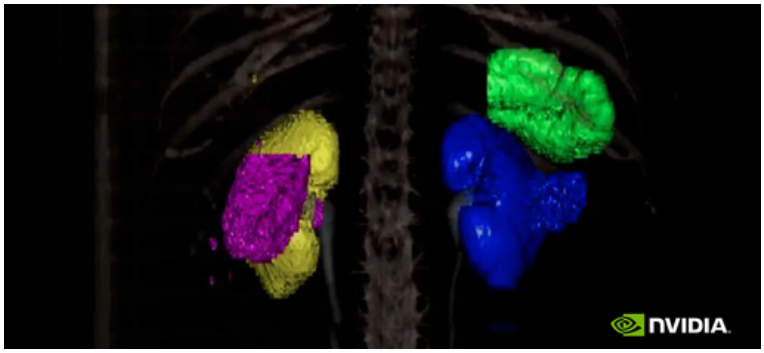
The HCLS customer faced several challenges in deploying an AI infrastructure, including the need for high-performance computing with GPU acceleration, scalable and efficient storage solutions to handle large datasets, high-speed networking for low-latency communication between AI components, robust security measures to meet compliance and regulatory requirements, and an energy-efficient infrastructure optimized for power and liquid-cooling solutions.

SOLUTION OVERVIEW

The proposed Supermicro (SMCI) Rack Scale solution offered a fully integrated and scalable platform that met the customer's AI infrastructure needs. The compute infrastructure included high-density servers with support for NVIDIA H200 GPUs, a modular architecture for seamless scalability, and 8 H200 GPUs deployed across GPU worker nodes. Storage solutions featured all-flash NVMe storage with 200TB of usable capacity, optimized for AI workloads and real-time data processing. Additionally, the solution incorporated Starfish Storage, delivering comprehensive data management, data protection, and archival capabilities. Starfish Storage ensures seamless lifecycle management, enabling intelligent data tiering and automated archiving for cost-effective long-term retention. With built-in redundancy and advanced security protocols, it safeguards critical AI datasets while maintaining accessibility and performance.

Networking infrastructure provided high-speed 400Gbps InfiniBand connectivity for GPU worker nodes, ensuring low-latency communication to support large-scale AI workloads. Energy efficiency and cooling mechanisms incorporated liquid-cooled infrastructure to manage GPU thermal output, along with high-efficiency power supplies and airflow optimization for sustainability. Security and compliance measures included hardware-based security features such as secure boot and encryption, ensuring adherence to industry regulations and data protection standards.

eastern@ecei.com www.ecei.com 203.877.4334
HQ: 61 Wilton Road, Westport, CT 06880



SOFTWARE INTEGRATION WITH NIVIDIA AI ENTERPRISE (NVAIE)

To maximize the AI infrastructure's potential, the solution incorporated NVIDIA AI Enterprise (NVAIE), a suite of software tools and frameworks designed for enterprise AI applications. This included NVIDIA RAPIDS for accelerated data science and analytics, NVIDIA TensorRT for deep learning inference optimization, and NVIDIA Triton Inference Server for scalable AI model deployment. Deep learning frameworks such as TensorFlow, PyTorch, and the NGC Catalog enabled seamless AI development. Kubernetes and NVIDIA Inference Microservices (NIM) facilitated AI workload orchestration and scalability. Additionally, NVIDIA BioNeMo provided foundation models for protein structure prediction and drug discovery, NVIDIA MONAI supported medical imaging AI applications, and NVIDIA Clara Discovery enabled AI-driven drug discovery, accelerating molecular simulations and generative chemistry.

BUSINESS IMPACT

The deployment of Supermicro Rack Scale and NVIDIA AI Enterprise delivered significant benefits to the HCLS customer. Optimized compute, storage, and networking ensured seamless AI model training and deployment, significantly increasing AI performance. The modular design provided scalability for future growth as AI workloads evolved. Enhanced security and compliance measures safeguard sensitive data while ensuring regulatory adherence. A pre-engineered and pre-built solution reduced time-to-deployment, minimizing integration challenges and accelerating implementation. Furthermore, advanced energy management improved operational efficiency by reducing power, cooling consumption and lowering operational costs.

CONCLUSION

This successful sales campaign demonstrated the value of a fully integrated AI infrastructure solution tailored to enterprise needs. By leveraging Supermicro's Rack Scale technology and NVIDIA AI Enterprise, the customer achieved a cutting-edge AI environment that supports ongoing innovation and business objectives. This case serves as a model for other organizations seeking to implement high-performance AI solutions efficiently and effectively.

eastern@ecei.com www.ecei.com 203.877.4334
HQ: 61 Wilton Road, Westport, CT 06880