SOLUTION BRIEF

SUPERMICRO REFERENCE ARCHITECTURE END-TO-END AI SOLUTION WITH CNVRG.IO

The Next Generation GPU systems with 3rd Gen Intel® Xeon® Scalable Processors

TABLE OF CONTENTS

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Executive Summary

Machine Learning (ML) has seen extraordinary growth in the last two years. While 70% of enterprise AI practitioners are exploring AI use cases, only 7% are able to deploy and scale across business units.[1] Over the next two years, it's predicted by Gartner that 50% of IT leaders will struggle to move their AI predictive projects past proof of concept to a production level of maturity.[2] The top barrier enterprises face to scaling AI is the complexity of integrating the solution within existing enterprise applications and infrastructure.

Supermicro provides state-of-the-art servers for computing, storage, and networking and works with partners to provide integrated, differentiated, and affordable solutions for various use-cases.

cnvrg.io delivers an operating system for machine learning that provides Data Science and IT practitioners a unified control plane that manages diverse ML and DL workloads and enables ML in production at scale.

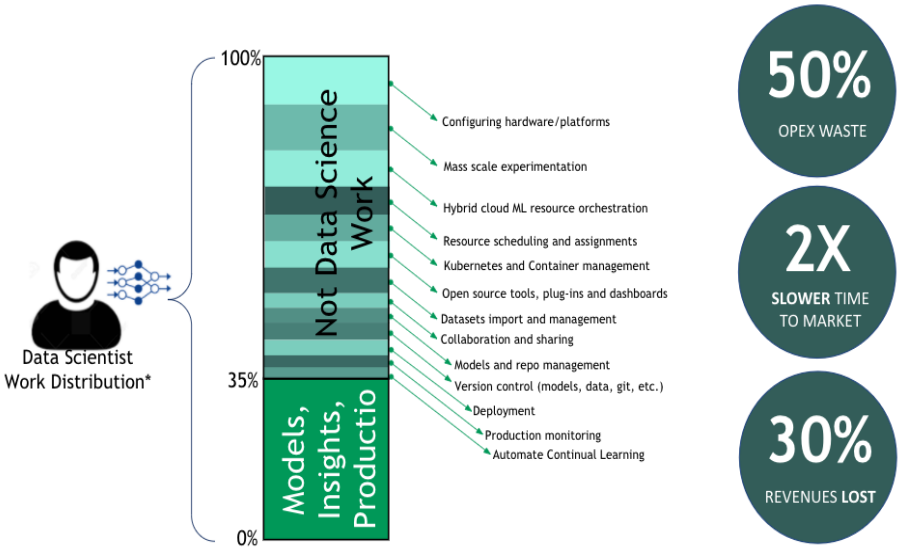
[1] Shimmin, Bradley. Omdia, 2020, *Fundamentals of MLOps*, omdia.tech.informa.com/OM012245/Fundamentals-of-MLOps.

[2] Brethenoux, Erick. Gartner, 2020, *Top Strategic Technology Trends for 2021: AI Engineering*, www.gartner.com/document/3994947?ref=solrAll&refval=282401476.

Supermicro and cnvrg.io – An End-to-End AI Solution

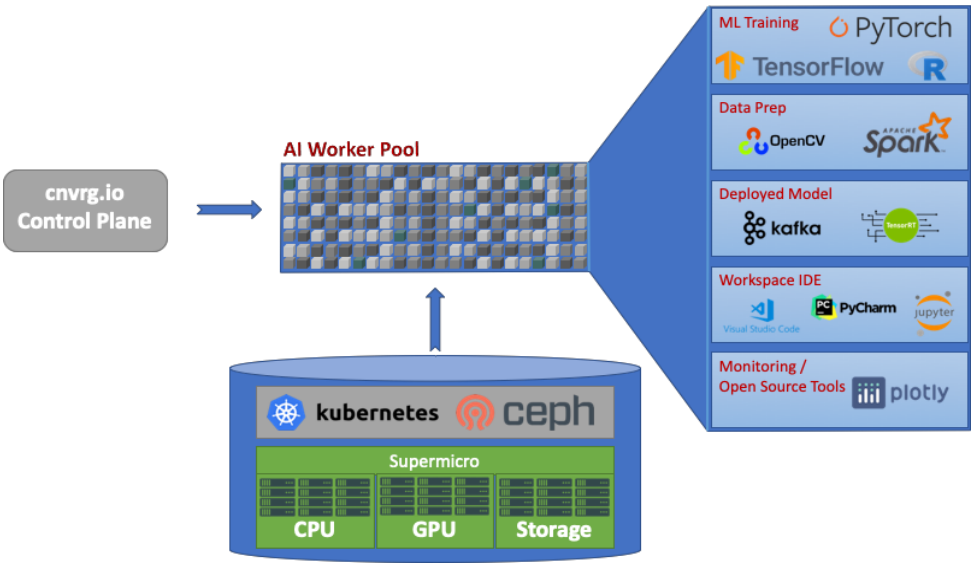
Industry Problem

65% of data scientists' time is being spent on heavy engineering tasks like data verification, monitoring, configuration, compute resource management, serving infrastructure, and feature extraction.[3] This is causing "Technical Debt" that reduces AI applications' ROI and a loss in revenue. On the other hand, data scientists work in silos. IT managers cannot provide scalability and implement required compliance over the operations performed by the data scientists.



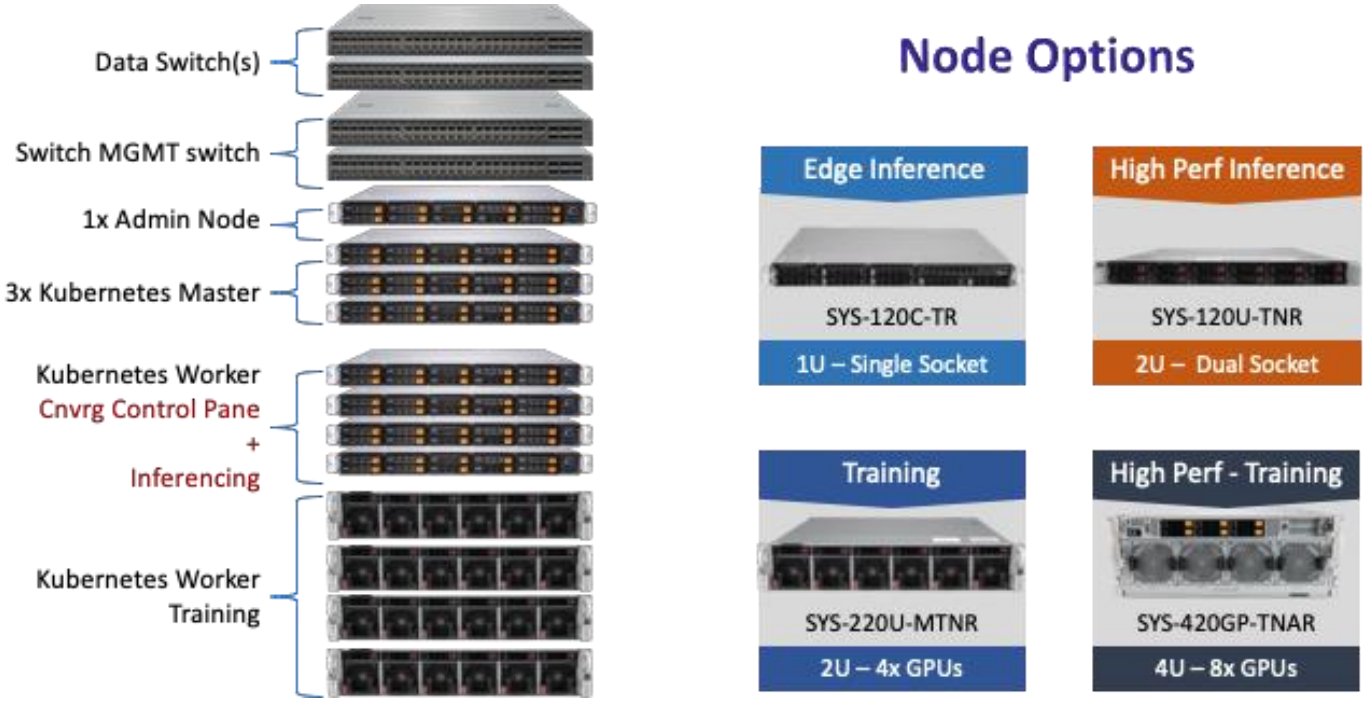
Unified AI Operating System Solution

Supermicro worked with cnvrg.io to provide an end-to-end solution for this problem. The foundation of this solution is built on top of a scalable storage platform using Supermicro storage servers. Then the control and compute layers were built using Supermicro's state-of-the-art CloudDC and GPU systems. Ceph storage and Kubernetes are installed on these control, compute, and storage components which together provide a scalable infrastructure as a service (IAAS) layer.



[3] *cnvrg.io internal survey (https://cnvrg.io/wp-content/uploads/2020/10/cnvrg-data-sheet.pdf)*

Below is the rack diagram and the sample configuration for the cnvrg.io POC cluster.



cnvrg.io is installed on top of this Supermicro IAAS layer. cnvrg.io serves the extreme needs of DevOps and IT Ops by providing:

- Sophisticated meta-scheduling that accomplish high infrastructure utilization (>90%)
- Infrastructure utilization dashboards with different views (per data scientist, project, server, etc.) providingability to segment available compute infrastructure per team or as one pool.
- It is quick and easy onboarding, all based on containers and packaged for simple installation and stand-up (helm, Kubernetes, etc.).
- Integration with one container catalog optimized to increase the performance of Supermicro servers, featuring 3rd Gen Intel® Xeon® Scalable Processors for accelerated AI workloads.
- Accelerated workloads, with the ability to launch any ML or DL framework in one click with direct integration to optimized containers from NVIDIA NGC
- Compatible with existing infrastructure, hence no silos need to be created. Scaling your Supermicro AI servers with cnvrg.io is simple. With one click, you can add more installed hardware to the managed infrastructure.

Launch Any ML or DL Framework with One-Click

Accelerated workloads, with the ability to launch any ML or DL frameworks in just one click with direct integration to optimized containers from both Intel and NVIDIA, engineered to increase the performance of Supermicro servers featuring 3rd Gen Intel® Xeon® Processors and NVIDIA GPUs for AI workloads.

Accelerate Time to Value with Frictionless MLOps Workflows

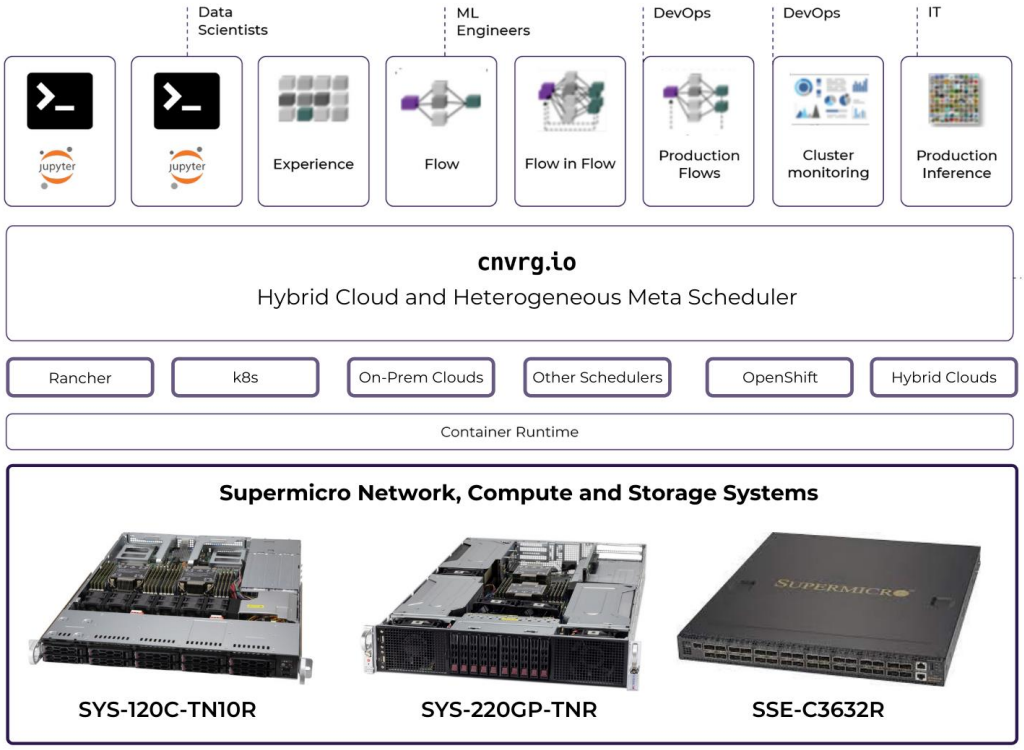
The cnvrg.io platform, running on Supermicro AI-ready servers, provides data scientists with accelerated project execution from research, to training, to production. cnvrg.io enables enterprises to manage and scale ML in any environment quickly. Data scientists can run ML pipelines on diverse workloads to achieve maximum performance and accelerate time to production.

Increase Data Science Productivity

Data scientists can easily manage, experiment, track, version, and deploy models in one click in one unified solution. Its container-based framework is designed to be agnostic and portable. It solves key MLOps challenges to help data scientists deliver more models to production fast.

Maximize Server Utilization with AI Workload Visibility

This solution allows real-time visibility of your AI workloads. The MLOps Dashboard summarizes granular perspectives of your entire workflows to identify bottlenecks so you can maximize Supermicro server utilization. cnvrg.io MLOps helps streamline AI application delivery, so data science teams and IT can more effectively manage users, workloads, models, datasets, experiments, and more while speeding continuous application delivery.



Focus on Speed and Best in Class Experience

- cnvrg.io running on Supermicro AI-ready servers delivers data scientists and AI practitioners one unified control plane to build, deploy and manage machine learning workloads. With cnvrg.io, data scientists can bring high-impact models to production faster and monitor models on top of the Supermicro AI purpose-built solutions.
- Supermicro and cnvrg.io integrated solutions improve collaboration and governance, providing a single place for IT engineers, DevOps engineers, ML engineers, data scientists, and researchers to share and achieve AI-driven results.
- cnvrg.io offers data scientists execution of a complete pipeline on Supermicro servers without understanding the infrastructure or configurations. With cnvrg.io and Supermicro, data scientists can build dynamic end-to-end ML/DL solutions on heterogeneous compute by running different tasks on maximized CPU and GPUs.
- Data scientists designed cnvrg.io for data scientists. Through research, experimentation, and deployment, all pipeline stages from datasets versioning and management are either Python SDK, CLI, REST APIs, or GUI.
- cnvrg.io offers a quick onboarding for data scientists and improves accessibility to Supermicro servers for any ML or DL workload. Industry-known tools and GUI are used for each stage of the pipeline, optimizing the data scientists' experience and eliminating the learning curve for new tools/GUI.
- cnvrg.io improves the scalability of your AI infrastructure. cnvrg.io makes the coordination of resources easier to manage and effectively pools your infrastructure into a more consumable manner.
- While cnvrg.io offers one-click development and deployment of ML pipelines to Supermicro Servers, it can also auto-detect the lack of resources and burst your workload to the cloud of your choice.
- Unlike other platforms, each stage of the pipeline can be attached to a different resource, including the support of multi-cluster and heterogeneous compute pipelines.
- cnvrg.io will support the environment of your choice, be it Kubernetes, bare-metal, or any hypervisor (e.g., KVM. All can co-exist within your Supermicro clusters.

Key Features:

Heterogeneous Compute Pipelines:

Launch and manage end-to-end heterogeneous ML pipelines where each component or stage (in a single pipeline) can run on a different compute architecture optimized for the specific use-case: preprocessing and/or inferencing on CPU, deep learning training on GPUs, and inference in the edge.

MLOps Utilization Dashboard:

Improve visibility across all ML runs. cnvrg.io dashboards have been proven to increase infrastructure utilization by up to 80% with advanced resource management. cnvrg.io gives data engineers and IT the tools to monitor utilization, properly size the compute components and visualize who uses what, with extensive cluster utilization visualization.

Open Platform:

cnvrg.io was designed to be an open and flexible platform. As a code-first platform, users are free to build in any environment, in the language of their choice, and with the tools they love. cnvrg.io is container-based making it easy to use any image or framework easily.

High-performance Infrastructure:

cnvrg.io will help increase utilization of your infrastructure while Supermicro delivers the highest performance for both your AI training, inferencing, and deployment workloads.

Sample Configurations:

For use cases that require acceleration for frequent training or high throughput inferencing of deep models.

2U 6-GPU System Configuration

Type	Description	Per System	Cluster Config
System	SYS-220GP-TNR 2U 6-GPU SuperServer	1	6
CPU	3rd Gen Intel® Xeon® Gold 6330N Processor (28C, 165W, 2.2GHz)	2	12
GPU	NVIDIA A100 40GB HBM2 PCIe 4.0	6	36
Memory	64GB DDR4-3200 2Rx4 LP (16Gb) ECC RDIMM	16	6TB
AOC M.2	LP, PCIe3 x8, dual-port NVMe M.2 carrier	1	6
OS M.2	Intel DC P4511 2T NVMe M.2 22x110mm up to 1DWPD	2	12TB
Storage	Intel D3-S4510 3.84T SATA 6Gb/s 3D TLC 2.5" 7mm <2DWPD Rev.2	4	92TB
AOC	[NR]Mellanox MCX651105A-EDAT PCIe Single-port 100Gb/s IB-HDR QSFP56	1	6
AIOM	AIOM dual-port 10GBase-T, Intel X550	2	12

Use cases targeting, general or uniform infrastructure, traditional ML-algorithms, DL inferencing, and training in both single instance/node and distributed execution patterns.

CloudDC System Configuration

Type	Description	Per System	Cluster Config
System	SYS-120C-TN10R CloudDC Server	1	3
CPU	3rd Gen Intel® Xeon® Platinum 8368 (38C, 2.40G, 270W)	2	6
Memory	64GB DDR4-3200 2Rx4 LP (16Gb) ECC RDIMM	16	3TB
OS M.2	Kioxia XG6-P 2TB NVMe M.2 22x80mm 0.3DWPD	2	6TB
Storage	Kioxia CD6 7.68TB NVMe PCIe 4x4 BiCS4 2.5"15mm SIE 1DWPD	10	230TB
AOC	Standard low-profile card based on Broadcom BCM57504. Two QSFP28 100Gbps Ethernet port PCIe 4.0 x 16	1	3
AIOM	AIOM dual-port 10GBase-T, Intel X550	2	6
Fans	Counter-rotating, 23.3K-20.3K RPM	6	18

Conclusion

With a wide variety of Supermicro systems, cnvrg.io could maximize its AI OS potential, significantly reduce the time-to-market value, clearly visualized ML pipeline workflow, and dramatically improve end-user experiences. For on-prem private cloud, the 1U CloudDC fits in the role of infrastructure and ML inferencing nodes. The 2U GPU server, on the other hand, is equipped with six NVIDIA A100 PCIe 40GB GPUs to provide ample resources for cnvrg.io AI OS to train large and extensive models. Combining new 3rd Gen Intel Xeon Scalable processor-powered Supermicro systems with the cnvrg.io platform is one of the best AI solutions on the market.

About cnvrg.io

cnvrg.io is an AI OS, transforming the way enterprises manage, scale, and accelerate AI and data science development from research to production. The code-first platform is built by data scientists for data scientists and offers unrivaled flexibility to run on-premise or cloud. From advanced MLOps to continual learning, cnvrg.io brings top-of-the-line technology to data science teams so they can spend less time on DevOps and focus on the real magic - algorithms. Since using cnvrg.io, teams across industries have gotten more models to production resulting in increased business value.