SOLUTION BRIEF

SUPERMICRO'S AI CLUSTERS GET SUPERCHARGED BY MIRANTIS K0RDENT AI

Sovereign AI and Hybrid GPU Clouds Benefit from an Integrated Solution

MIRANTIS

AMD

TABLE OF CONTENTS

Executive Summary

In the high-stakes arena of sovereign AI and hybrid GPU clouds, organizations are choosing full-stack AI infrastructure solutions to accelerate GPU operationalization, ensure efficient utilization, and enforce security and compliance at scale. Mirantis k0rdent AI is a turnkey, production-ready "super control plane" for managing these complex environments, that automates provisioning, lifecycle management, and orchestration of infrastructure and core services, from Metal-to-Model™.

This solution brief details the successful validation of Supermicro's modular server architecture with the k0rdent AI ecosystem. For this verification, the models GPU A+ Server AS -8126GS-TNMR and BigTwin A+ Server AS -2124BT-HNTR were used. By unifying bare-metal provisioning, AMD-powered acceleration, and modern virtualization, this stack provides a blueprint for the next generation of AI data centers.

The Foundation: k0rdent Bare Metal Operator

Validation begins at the physical layer. The k0rdent Bare-metal Operator (utilizing Metal3 and Ironic) serves as the bridge between the declarative Kubernetes API and Supermicro servers.

- Bare Metal Integration: k0rdent communicates directly with the Supermicro Baseboard Management Controller (BMC) through the Redfish protocol to configure all nodes from the metal up. This includes automated BIOS configuration, firmware updates, and RAID orchestration without human intervention.
- Provisioning Workflow: During validation, Supermicro nodes were discovered as BareMetalHost objects. k0rdent successfully automated the PXE-less booting process, deploying a hardened host OS and the k0s Kubernetes distribution across the fleet.

AI at Scale: AMD GPU & Network Operators

To meet the demands of Generative AI, the validation environment used AMD Instinct MI-325X accelerators. The integration was managed via two critical operators:

AMD GPU Operator

The AMD GPU Operator was deployed via the k0rdent catalog, which provides ready-to-deploy Service Templates.

- Automated ROCm™ Stack: The operator automatically injected the ROCm 6.3 or 6.4 driver stack and GPU resource allocation.
- Validation Result: Utilizing the ROCm Validation Suite (rvs), the system confirmed peer-to-peer (P2P) bandwidth and memory throughput, ensuring the 8-GPU "fully-meshed" Infinity Fabric™ was operating at peak efficiency.

AMD Network Operator

High-performance networking enables efficient GPU-to-GPU communication in an AI cluster. The AMD Network Operator (v1.0.0) was validated for its ability to:

- Configure AI NICs: Automate the driver installation of AMD Pensando™ Pollara 400 NICs and backend NICs resource allocation.

Validation Reference Stack

Component	Version / Model
Control Plane	Mirantis k0rdent Enterprise v1.2.2
Hardware	Supermicro AS -8126GS-TNMR (AMD EPYC™ 9xx5 Series)
GPU	8x AMD Instinct™ MI325X
Network	AMD Pensando™ Pollara 400 (400GbE)
Virtualization	KubeVirt (HCO) v1.7.0

Hardware Test Bench

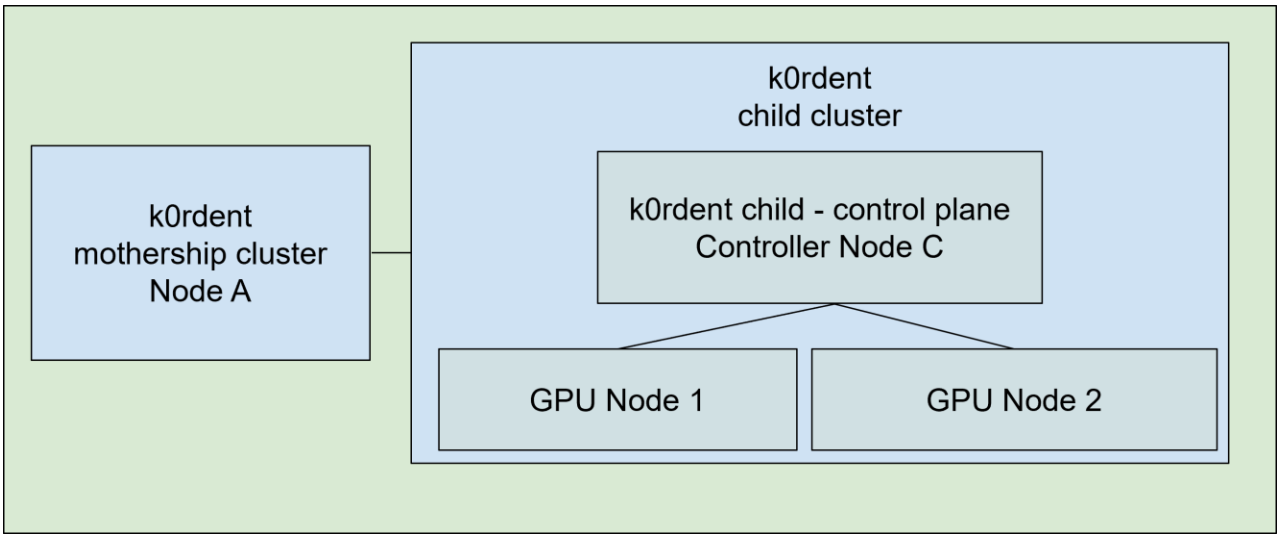

Figure 1 – System Architecture

Name	Model	CPU	GPU
k0rdent Node A	AS -2124BT-HNTR	AMD EPYC 7352 (24-Core Processor)	-
Controller node C	AS -2124BT-HNTR	AMD EPYC 7352 (24-Core Processor)	-
GPU node 1	AS -8126GS-TNMR	2 x AMD EPYC 9965 (192-Core Processor)	8 x AMD MI325x
GPU node 2	AS -8126GS-TNMR	2 x AMD EPYC 9965 (192-Core Processor)	8 x AMD MI325x

Validation Results Summary

The validation process concluded with a suite of tests to ensure "Day 1" production readiness:

- Provisioning: Bare metal provisioning with k0rdent

The validation successfully demonstrated the automated provisioning of production-grade Kubernetes clusters on Supermicro bare-metal hardware using k0rdent's declarative orchestration engine and the Bare Metal Operator (BMO).

By leveraging the integration of Metal3 and Ironic, k0rdent managed the entire lifecycle of the Supermicro nodes—from out-of-band discovery via BMC/IPMI and hardware introspection to automated OS imaging and Kubernetes bootstrapping. This process eliminated manual configuration and hypervisor overhead, providing a high-performance, consistent, and repeatable deployment model that adheres to Cluster API (CAPI) standards.

The validation confirms that k0rdent effectively bridges the gap between physical server management and cloud-native agility, making it an ideal solution for resource-intensive workloads requiring direct hardware access and deterministic performance on Supermicro infrastructure.

- GPU Validation:
    - To get the cluster up and running for high-performance AI workloads, we used k0rdent to configure the environment with the ROCm and amdgpu-dkms packages to ensure full hardware compatibility.
    - On the orchestration side, we deployed Kubernetes with k0rdent utilizing Cert-manager 1.15.1 for certificate management and the AMD GPU Operator 1.4.1 to streamline driver and resource provisioning across the fleet.
    - To wrap things up, we confirmed that all 8 GPUs per node were correctly registered and available within the cluster. This was verified using custom scripts built on top of the official rocm/vllm container image, leveraging a suite of tools, including rocm-smi, Python, and PyTorch, to ensure the hardware wasn't only visible but also fully functional for AI models.
- GPU Performance Evaluation:

We conducted performance benchmarks in accordance with the official AMD documentation. The evaluation environment utilized the rocm/vllm Docker image, focusing on a single GPU configuration as per the official testing framework.

Test Configuration

- Model: amd/Llama-3.1-8B-Instruct-FP8-KV
- Environment: ROCm-optimized vLLM container
- Scope: Single GPU unit performance

Validation Methodology

To validate the performance metrics, we utilized a custom PyTorch script to measure raw compute throughput across different precisions. Specifically, we evaluated the TFLOPs for the following:

	FP64 (Tensor/Matrix) TFLOPs	FP64 (Vector) TFLOPs
Actual Results	**160.6**	**82.4**
Expected Results	163.4	81.7

Conclusion

Validating Supermicro hardware with Mirantis k0rdent AI represents a shift from "building" clusters to "composing" them.

By leveraging the automated operators for AMD hardware and the flexibility of k0rdent AI and KubeVirt, enterprise customers can run their entire portfolio—from legacy apps to cutting-edge LLMs—on a single, unified, bare-metal platform with automatic deployment and comprehensive platform management from the bare metal up.

The shift from manual to automated deployment and management reduces the workload on the cluster management teams and eliminates human error and inconsistencies from the deployment and management process.

More Information:

Supermicro AMD Servers: www.supermicro.com/aplus
Mirantis Software: www.mirantis.com

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. Visit www.supermicro.com

AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics and visualization technologies. Billions of people, leading Fortune 500 businesses and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible.

Learn more at www.amd.com

MIRANTIS

Mirantis helps organizations achieve digital self-determination by giving them complete control over their strategic infrastructure. The company combines intelligent automation and cloud-native expertise to manage and operate virtual machines, containers, Kubernetes, and cloud environments. Mirantis lets platform teams deliver a public cloud experience on any infrastructure, from the data center to the edge, with a single, cohesive cloud experience for complete application and operations portability, a single pane of glass, and automated full-stack lifecycle management, all based on open source and open standard APIs.

Visit www.mirantis.com