SUPERMICRO AND STATUSCORE OFFER GENERATIVE AI AND GENERATIVE DESIGN WORKFLOW SOLUTION

Innovative Solution Powered by Supermicro Workstation, Ravel Orchestrate, and AMD

Supermicro Superworksation AS -5014A-TT

Executive Summary

This solution brief aims to offer prospective customers, resellers, and integrators a comprehensive turnkey solution for effectively handling Generative AI and Generative Design workflows and workloads. As Generative AI and Generative Design generate vast amounts of data, managing additional machines and workflows becomes increasingly complex, necessitating an environment management approach. This management entails four key areas:

1. Ensuring efficient management of local and federated infrastructure and networking.
2. Streamlining software and AI engine imaging for swift deployment across the infrastructure.
3. Securely allocating infrastructure resources to teams.
4. Intelligently orchestrating all these components through a unified interface.

Table of Contents

Introduction

Generative AI and Generative Design have emerged as critical considerations in various industries such as Media & Entertainment, Product Design and Manufacturing, Gaming, and AEC. Efficiently building and managing secure turnkey environments for these industries is paramount. To achieve a successful Generative AI and Generative Design environment, three primary challenges need to be addressed:

1. Adequate computing power: A powerful machine, workstation, or server is essential to handle complex computational requirements.
2. Data and workflow management: Effectively managing the data and image workflows associated with content generation is crucial.
3. Orchestration layer: An orchestration layer is needed to enable seamless communication and coordination between the infrastructure and the workflow, yielding desired results.

Supermicro and RAVEL have collaborated to develop the industry's first turnkey SuperWorkstation, which was designed specifically for Generative AI and Generative Design using AMD AI accelerators. This SuperWorkstation caters to the needs of DevOps administrators, managers, and other end users, empowering them to manage, control, and scale Generative AI and Design workloads for both local and remote content creation teams.

Solution Description

The first iteration of this collaboration debuted at NAB 2023 in Las Vegas, where it was awarded Product of the Year. Including RAVEL's no-code RAVEL Orchestrate™ GenAI software further enhances the practical utilization of our AI SuperWorkstation within the content creation industries, allowing teams to quickly and efficiently commence their work. By managing Supermicro's AI SuperWorkstation's technical resources and other ecosystem components, this solution provides customers with a comprehensive and integrated approach.

This solution enables complex workflows and heavy workloads to be efficiently executed within a desk-sized space or easily integrated into existing infrastructures. Its small power footprint ensures a lower total cost of operation and ownership, making it a clear advantage for all stakeholders.

Today's DevOps customers, who require swift and efficient initiation of complex workflows with substantial workloads for their content creation teams, will find this solution to be highly beneficial. While the SuperWorkstation excels as a dedicated AI compute machine, it also offers the versatility of being transformed into a virtualized command center through RAVEL Orchestrate GenAI. This allows administrators to deploy virtualized technical resources to both local and remote content creation team members, showcasing unparalleled range and flexibility.

In conclusion, this solution effectively addresses the needs of modern content creation workflows by offering quick start capabilities, managing large workloads, and providing extensive utility through its SuperWorkstation and virtualized command center functionalities facilitated by RAVEL Orchestrate GenAI.

A deeper look at Supermicro's SuperWorkstation integrated with RAVEL Orchestrate GenAI's SMART Assembly™.

Creators and intellectual property (IP) owners are increasingly interested in harnessing the potential of Generative AI and Generative Design technologies. They have a range of options at their disposal, including AMD optimized open source technologies such as Amuse, KoboldCPP, Open AI, and Nod-ai SHARK, as well as Adobe Firefly, which offers a content authenticity model. Recognizing users' diverse needs and preferences, our partnership with the SuperWorkstation solution ensures flexibility to support exploration and IP protection.

This collaborative solution allows creators and IP owners to freely explore the open source technologies mentioned, empowering them to experiment and innovate with Generative AI and Generative Design. Simultaneously, it caters to those who prioritize safeguarding their intellectual property by integrating Adobe Firefly, which offers robust content authenticity measures. By providing support for both options, our SuperWorkstation solution ensures that users can choose the best approach to their objectives, whether exploration or protection of their valuable IP assets.

May 2024
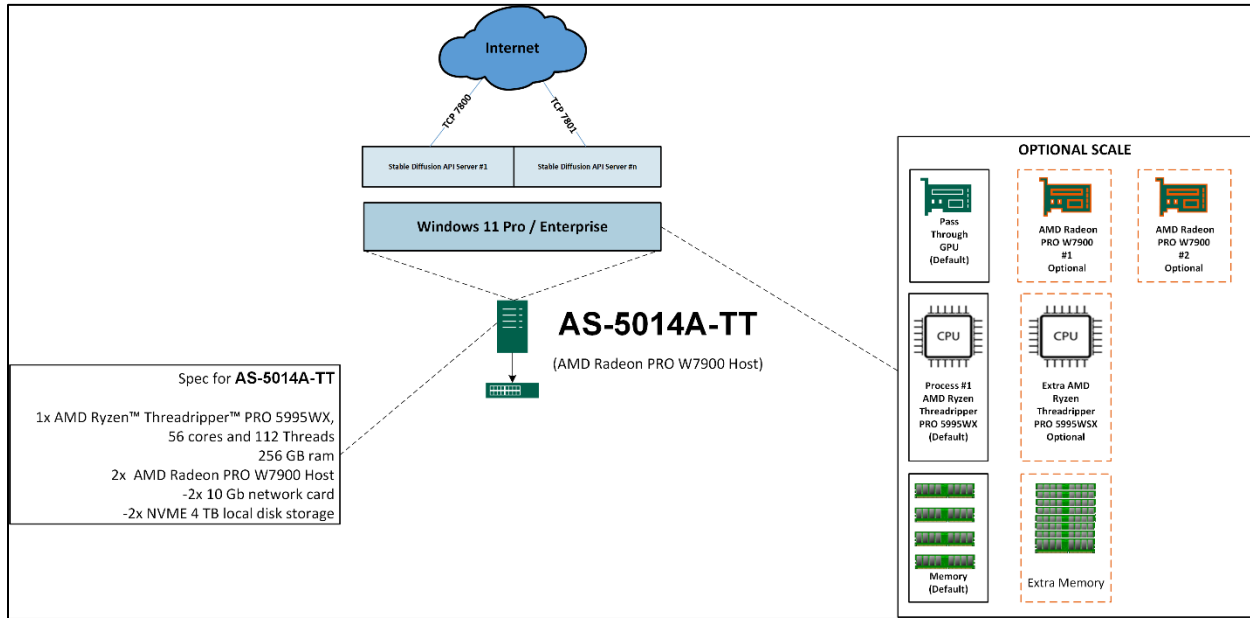
Use Case #1

Use Case #1 Description: The Generative AI and Generative Design Compute Working Server (GPU Super Server)

In this specific use case, an existing environment is already in place, whether on-premise, in the cloud, or in a hybrid setup. The objective is to integrate the Super Server seamlessly with the existing systems. To accomplish this, two primary requirements need to be addressed:

Validate the equipment and architecture outlined above as the optimized and approved schematic that will maximize the utilization of the Super Server. Deploy RAVEL Orchestrate GenAI to manage the associated infrastructure and environment efficiently.

By implementing RAVEL Orchestrate GenAI, the Generative AI and Generative Design workflows can be effectively managed in conjunction with virtual workstations located either on-premise or in a data center. It ensures seamless integration with the associated infrastructure components, including license servers, storage area networks (SANs), network-attached storage (NAS), and more. Additionally, the flexibility of RAVEL Orchestrate GenAI extends to managing and enabling virtual environments for various content creation workflows.

This configuration provides comprehensive support for AI engine training, inference, 3D modeling, rendering, high-powered virtual workstations, as well as persistent servers and workstations. Moreover, it enables remote users to access and utilize the system efficiently, regardless of their location.
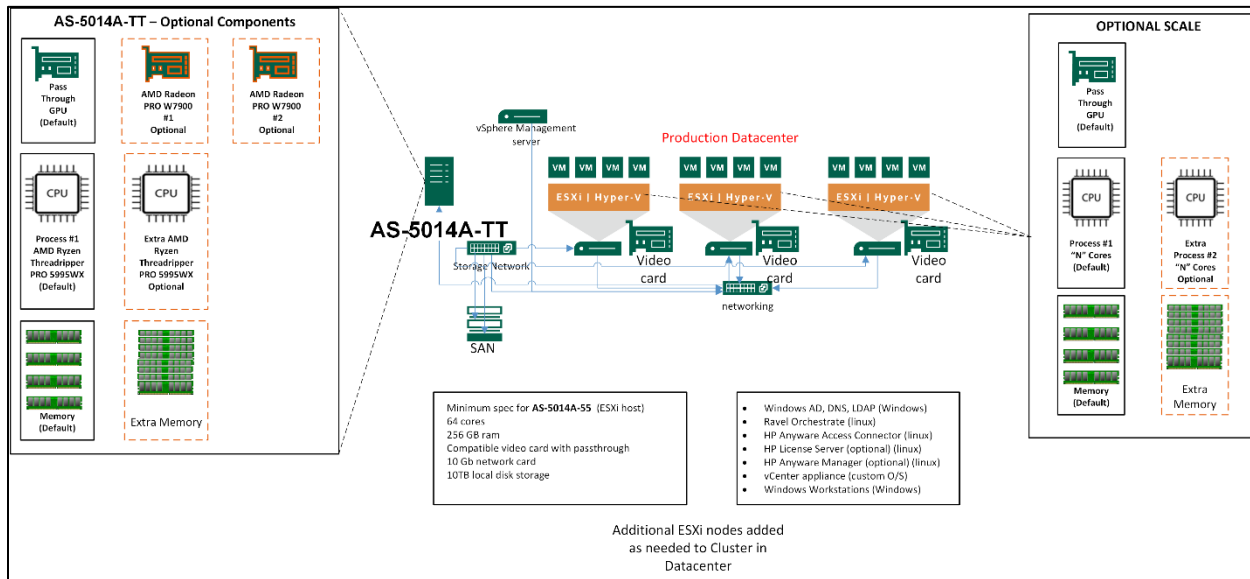
Use Case #2
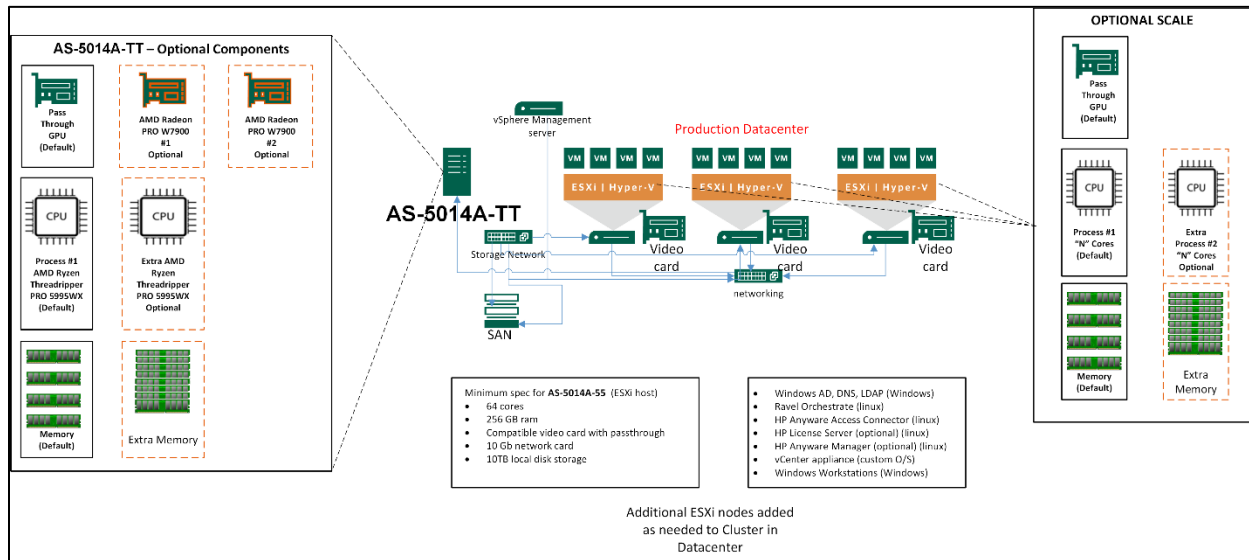
Use Case #2 Description: A clustered environment for Generative AI and Generative Design for SMBs or smaller installations with the ability to scale and collaborate with other team members.

This configuration provides a streamlined infrastructure solution in an established environment, whether on-premise, in the cloud, or a hybrid setup. It focuses on deploying a compact and scalable system that efficiently supports GPU and AI/ML generative content production. This setup is particularly suitable for demonstration purposes, allowing for easy node expansion to enhance disaster recovery and ensure high availability in production scenarios.

RAVEL Orchestrate GenAI is crucial in managing Generative AI and Generative Design workflows, seamlessly integrating with virtual workstations on-premise or in a data center. It effectively handles associated infrastructure components such as license servers, storage area networks (SANs), network-attached storage (NAS), and more. Furthermore, RAVEL Orchestrate GenAI exhibits the necessary flexibility to manage and enable virtual environments for any content creation workflow.

This configuration provides robust support for various tasks, including AI engine training, inference, 3D modeling, rendering, high-powered virtual workstations, and persistent servers and workstations. Moreover, it offers the advantage of facilitating remote user access, allowing individuals to utilize the system effectively regardless of physical location.

Use Case #3

Use Case #3 Description: A core cluster that is ready to scale immediately for medium and larger organizations and team members.

This configuration is intended for teams that have already established their environment, whether on-premise, in the cloud, or in a hybrid setup, and are prepared to leverage the power of Generative AI and Generative Design for their operations.

Designed to optimize medium to large-scale infrastructures, this configuration is adaptable to various established environments, including on-premise, cloud, and hybrid systems. Utilizing virtualized servers enables efficient GPU and AI/ML-driven content generation. This solution is particularly well-suited for production environments where high-availability and disaster recovery mechanisms are already in place. It provides an optimal approach for deploying GPU-enabled workstations on a medium-to-large scale, regardless of whether the operating environment is cloud-based, hybrid, or on-premise.

RAVEL Orchestrate GenAI is the management tool for overseeing Generative AI and Generative Design workflows, seamlessly integrating with virtual workstations on-premise or in a data center. It effectively handles the associated infrastructure components, including license servers, storage area networks (SANs), network-attached storage (NAS), and more. Furthermore, RAVEL Orchestrate GenAI is versatile enough to manage and enable virtual environments for any content creation workflow.

This configuration offers comprehensive support for a wide range of tasks, such as AI engine training, inference, 3D modeling, rendering, high-powered virtual workstations, and persistent servers and workstations. Additionally, it facilitates remote access, allowing users to operate the system efficiently from anywhere.

Supermicro, in connection with RAVEL Orchestrate GenAI, supports the following Generative AI and Design workflows:

AutoDesk, Inc	Adobe, Inc	AVID, Inc.	AI Generative Software*	Knowledge Generative*	Game Development
Media & Entertainment Collection including Maya, 3DS Max, Motion Builder,	**Creative Cloud** Including Photoshop, AfterEffects, Premiere Pro	NEXIS Storage	Amuse Nod-ai SHARK	KoboldCPP for ROCm - Large Language Models (LLM)	Unity Unreal
Architecture, Engineering, and Construction: Collection including AutoCAD, Revit, Civil 3D	**Generative AI & Design:** Amuse Nod-ai SHARK	File Transfer Service Media Central - Production Engine Media Index Production Services Cloud UX Ingest Asset Management	KoboldCPP for ROCm - Large Language Models (LLM)	KoboldCPP for ROCm - Large Language Models (LLM)	
Product Design & Manufacturing: Collection including Inventor, AutoCAD, Fusion360					

*Other use cases, including Life Sciences and AI & Deep Learning, are also applicable

Supermicro Hardware Configurations:

Product Type	Super Micro SKU	Processor	Memory	OS	Expansion	Network	Storage	Display	Power
AMD system for Gen AI/Design	AS-5014A-TT	AMD Ryzen Threadripper PRO 5995WX, 64 cores	64GB – up to 1TB DDR4 RAM	Windows 11 Pro	2x AMD Radeon PRO W7900 with 48GB GDDR6	1x 1Gb and 1x 10Gb Base T Lan ports	1 x 2TB NVME for OS and & 1 - 3x 4TB NVMe for storage	2x AMD Radeon PRO W7900 with 48GB GDDR6	2000W PS2 Multi-output 80PLUS Platinum Level
ESXi Host	SYS-751A-I	Single or Dual Xeon LGA4677 SPR 8480+ - up to 112 Cores	256 GB DDR5 RAM - Up to 1 TB RAM	VMWare ESXi v8u3 with NVIDIA Drivers		2x 10Gb Base T Lan ports	2x 6.7TB 2.5" Magnetic (OS and App installation) RAID Cluster	2x Nvidia RTX 4000	2000W PS2 Multi-output 80PLUS Platinum Level
ESXi Host - Storage for vSphere	SSG-5029P-E1CTR12L	-	-	-	-	-	Storage SAN - iSCSI or FC connection. RAID 5 or better	-	-
ESXi - Virtual workstations	-	16 Core - Up to 96 per compute resource	96 GB of RAM	Windows 10 Pro, 11 Pro, Server 2019, or Server 2022	-	-	-	GPU Pass-through required for graphic acceleration	-

Learn more about the Supermicro AMD Workstation at:
https://www.supermicro.com/en/products/system/workstation/5u/as-5014a-tt.

The Trial and "Verify" Opportunity

The SuperWorkstation powered by RAVEL Orchestrate GenAI is now available for trial. The trial program is called Verify. Verify showcases a sophisticated Generative AI and Generative Design workflow seamlessly executed on Supermicro's AI SuperWorkstation and integrated with RAVEL Orchestrate GenAI. This demonstration leverages a combination of cutting-edge technologies, including:

1. Open source latent text-to-image and image-to-image diffusion models, Amuze and Nod-ai SHARK, to generate high-quality content.
2. Adobe's creative tools, such as Firefly, Photoshop, and Premiere Pro, enable creative enhancements and content refinement.
3. Autodesk Maya and 3DS Max, providing robust 3D modeling and animation capabilities.
4. Virtualization technology for efficient resource allocation and management.
5. Secure access control and identity management to ensure data integrity and user authentication.

Sign up for our Verify Turnkey Trial

RAVEL and Supermicro have collaboratively developed a comprehensive turnkey solution, allowing end users to trial and experience a Generative AI and Generative Design Workflow. This solution incorporates the powerful capabilities of RAVEL Orchestrate GenAI, a no-code SMART Assembly product, ensuring seamless workflow management.

As part of the Verify trial, users can customize software images for commonly used creative applications. Select customers and partners are granted virtualized access to the SuperWorkstation trial, hosted at Supermicro's state-of-the-art lab in San Jose. Please note that availability for the trial is limited and provided on a first-come, first-serve basis. For further information and to sign up for this valuable opportunity to trial the solution, please visit the following link: https://ravelinc.com/orchestrate-verify-supermicro/

For Integrators & Resellers

Interested in partnering to host your own Generative AI and Generative Design trial? Please contact us at
https://ravelinc.com/partner/

Conclusion

Supermicro's SuperWorkstations are optimized for applications requiring powerful compute and graphics capabilities. RAVEL Orchestrate GenAI's SMART Assembly capabilities facilitate rapid deployments, scheduling, and management of various software images on a SuperWorkstations virtual machines. Together, they deliver a robust, multi-model Generative AI and Generative Design workflow solution that offers IT and DevOps Administrators and their content creation teams an all

encompassing and secure environment, meticulously crafted for the heavy compute workloads of complex Generative AI and content creation workflows.

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.

Learn more at www.supermicro.com

RAVEL

RAVEL (formerly StratusCore) is at the forefront of digital transformation, providing innovative no-code orchestration software designed to streamline the complexities faced by DevOps and IT teams. With a singular focus on reducing friction in the creation of secure, virtual content creation environments, RAVEL's flagship product RAVEL Orchestrate™ empowers organizations to achieve rapid assembly of software applications, infrastructure (cloud, on-premise, and remote hardware), and networking—all seamlessly managed through a single-pane interface.

Learn more at: www.ravelinc.com