SOLUTION BRIEF

SUPERMICRO MAXIMIZES AI POWER AT THE EDGE WITH NAMUTECH'S COCKTAIL CLOUD

Simplify Edge AI Management With Supermicro Servers

Supermicro Single Socket Server

Supermicro Edge Server

TABLE OF CONTENTS

Executive Summary

AI workloads often involve large amounts of data, but internet connections at edge locations can be unreliable or slow. Even a few hundred milliseconds of delay in real-time interactions can significantly impact user experience. Hyperscalers charge for data ingestion, movement between availability zones, and extraction. These costs can accumulate when dealing with millions of AI interactions. Many AI workloads involve sensitive and regulated data. Processing this data in the cloud raises concerns about data privacy and security.

Solution Description

Together, Supermicro and NAMUTECH provide a platform for your AI inference at the edge with Kubernetes management and modern application deployment. Our solution offers a rich suite of capabilities to address specific requirements throughout the lifecycle of edge infrastructure and AI software stacks:

- It cuts through the complexity of deploying and managing AI stacks at the edge, using your preferred OS and Kubernetes distributions. Popular AI model engines like Kubeflow, OLLAMA, OpenAI, and Claude are supported.

- Locks down edge infrastructure, intellectual property, and model data with uncompromising configurations and ironclad role-based access control, strictly adhering to policies.

- Strips away inefficiencies in distributed inferencing, compelling organizations to leverage multiple edge nodes for simultaneous execution, slashing model latency.

- Demands federated training, accelerating model improvement through relentless on-device learning and stringent local data control.

Supermicro is a global technology leader committed to delivering first-to-market innovation for Enterprise, Cloud, AI, Metaverse, and 5G Telco/Edge IT Infrastructure.

NAMUTECH is a global provider of enterprise-level cloud-native software solutions specializing in hybrid cloud, modern virtualization, container, and Kubernetes technologies.

Maximize GPU Power at the Edge

Cocktail Cloud offers dedicated GPU monitoring, providing complete visibility into the resource utilization of whichever GPU hardware you're running at the Edge.

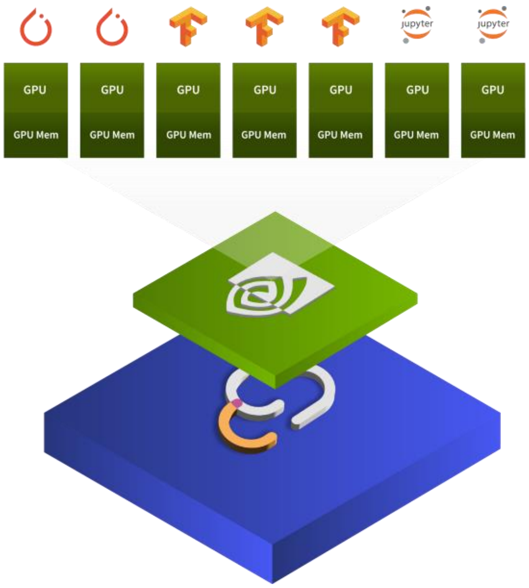

Figure 1 – Embedded NVIDIA GPU Sharing Capabilities

Embedded NVIDIA GPU Sharing Capabilities

The embedded NVIDIA GPU sharing capabilities provide even more flexibility, allowing both burst workloads and smaller, simultaneous workloads to be efficiently managed. Currently, there are two NVIDIA GPU sharing capabilities available compatible with the following GPU hardware models – A100, H100, and A30:

- Time Slicing Sharing: Allows the sharing of physical GPUs, enabling workloads that require bursting to utilize GPU resources without waste.

- Multi-Instance GPU (MIG): Allows the division of one physical GPU into a maximum of seven instances, suitable for cases that require a small amount of dedicated GPUs.

Use Cases

With Supermicro's Superedge servers and Cocktail Cloud, this solution provides a centralized platform to manage applications and AI workloads running at the edge and across a data center and multiple public clouds. It effortlessly automates DevOps and Day 1 & 2 operations, reduces resource requirements, and ultimately saves time and costs.

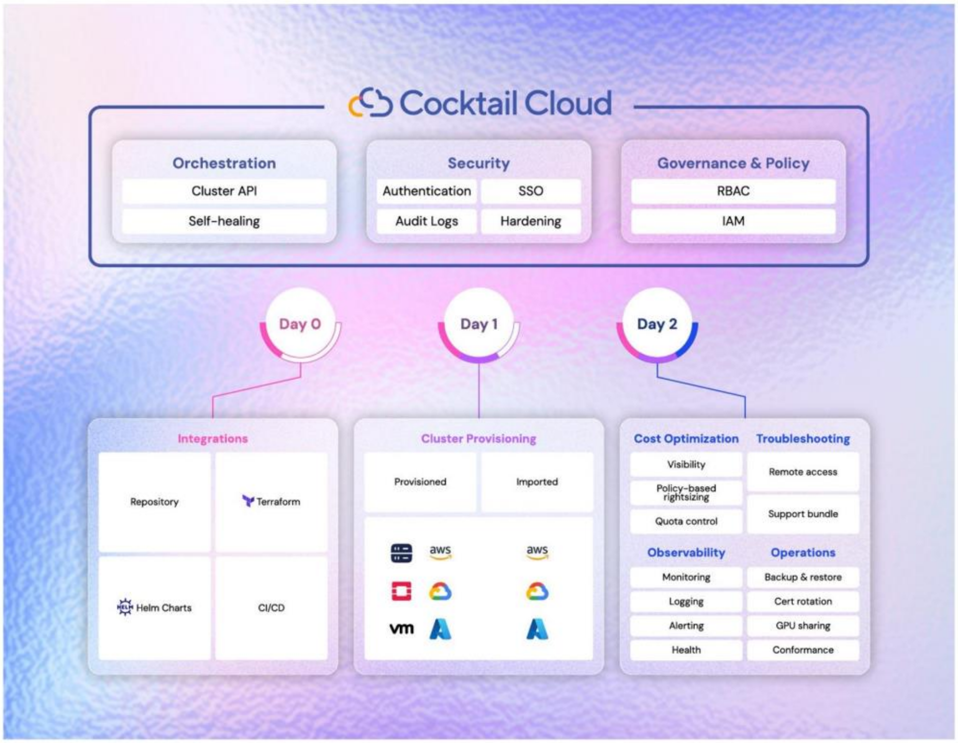

Figure 2 - Cocktail Cloud automates DevOps and Day 1 &2 operations

- **Mining** – Edge AI can analyze sensor data from mining equipment to prevent unexpected breakdowns, reduce downtime, and optimize maintenance schedules. Edge AI can also analyze geological data, monitoring miners' vital signs, movements, and location in real time to ensure their safety.

- **Robotics** – The dynamic world of robotics demands instant decision-making, and Edge AI is the key to unlocking this potential by bringing data processing closer to the source. Edge computing dramatically reduces latency and allows robots to respond to their environment in real time. This shift is crucial, especially in applications where every millisecond counts, like autonomous driving or precision manufacturing.

- **Hospital** – Edge AI enables the monitoring of continuous and real-time patient vital signs using wearable devices, sensors, and IoT devices. AI algorithms deployed at the edge can process and analyze the data locally, allowing healthcare providers to receive immediate insights and alerts in critical situations. Edge computing can significantly accelerate the interpretation of medical imaging data. AI models deployed on edge devices can process images locally, allowing for rapid analysis and diagnosis without the need for data transfer to a remote server.

- **Retail** – Edge AI analyzes customer data and predicts future purchasing patterns, leading to optimized product offerings and improved sales. Personalization through tailored marketing campaigns and personalized product recommendations can boost customer loyalty and repeat business.

- **Inventory Management** – AI empowers retailers to create more resilient supply chains that respond quickly to changing consumer demand and effectively manage inventory distribution. Edge AI can track Point of Sales (POS) and online shopping transaction inventory levels in real time, identify discrepancies, and prevent stockouts, leading to cost savings and increased revenue. The technology can be applied in smart shelves and robotics for inventory management and stocking shelves, resulting in increased efficiency and reduced labor costs.

Cocktail Cloud (PaaS)

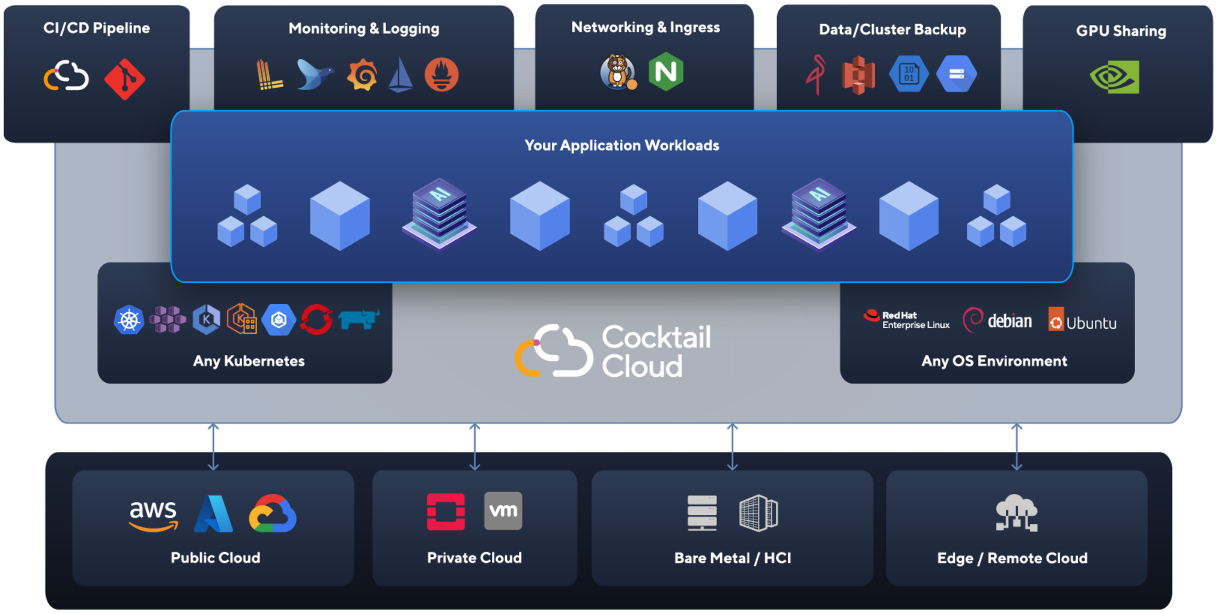

Figure 3 - Cocktail Cloud Platform Components

The Cocktail Cloud platform provides six core foundational components:

1. Application Deployment – Easily deploy and manage container applications across multiple infrastructure environments using the embedded or any other CI/CD engines.

2. Observability – Embedded centralized monitoring and logging capabilities with Cocktail Cloud and, integrated tools like Grafana and Prometheus that provide world-class monitoring capabilities, and integrated OpenSearch to provide users with best-in-class logging capabilities.

3. Integrated access to AI modeling through Application Catalog— Integrated access to third-party applications, including CI/CD pipeline tools or AI/ML modeling frameworks such as Kubeflow, TensorFlow, and PyTorch, simplifying deployment and version control.

4. Backup/Restore – Protect the data to any cloud by creating backups of Kubernetes Cluster to AWS S3, Azure Blob, Google Cloud Storage, and MinIO.

5. NVIDIA GPU Sharing capabilities—Enhance performance and reduce cost by allowing multiple applications to share a single GPU and efficiently use GPU resources.

6. Security – Ensures that all container-based applications and data are protected with robust security features such as role-based access control (RBAC), network policies, and encryption at rest and in transit.

Cocktail Cloud is beyond just an edge AI management platform. Cocktail Cloud is both cloud agnostic and supports x86 on-premises hardware for end-customer deployment. Its control plane runs in client data centers, at 3rd party colocation providers, and on major public cloud providers like AWS, Google Cloud, and Microsoft Azure.

Platform as a Service ("PaaS") – combines the Cocktail Cloud Platform edge solution with x86 physical edge-capable hardware, tailored to specific requirements, with managed services.

Solution Architecture

This solution demonstrated that Supermicro servers operate from the data center (CloudDC, Hyper, Big Twin) to the edge by deploying Kubernetes clusters on a single socket SuperServer SYS-111C-NR. Primary nodes can also be hosted on Supermicro servers within the data center for added flexibility.

Kubernetes cluster was deployed on UP SuperServer SYS-111C-NR to demonstrate that this AI edge solution can run from the data center CloudDC, Hyper, and Big Twin to the edge. The master nodes can also be run on Supermicro servers in your data center, for example:

- 3 control plane nodes: UP SuperServer SYS-111C-NR
- 1 image registry node: UP SuperServer SYS-111C-NR
- 2 worker nodes: 1 worker in the data center: UP SuperServer SYS-111C-NR, and 1 worker in an edge location: SYS-E403-12P-FN2T

Edge Node

- Hardware: SYS-E403-12P-FN2T | Box PC | SuperServer | Products | Supermicro
- GPU: Nvidia A100
- Edge node software: Cocktail Cloud, Stable diffusion, OLLAMA, and Open WebUI

S2S VPN was used to connect the edge node to the Supermicro Cluster. The MIG mode of the Nvidia GPU was activated on the edge node. Cocktail Cloud's built-in CI/CD tools deployed Stable diffusion and Ollama + Open WebUI on the edge node.

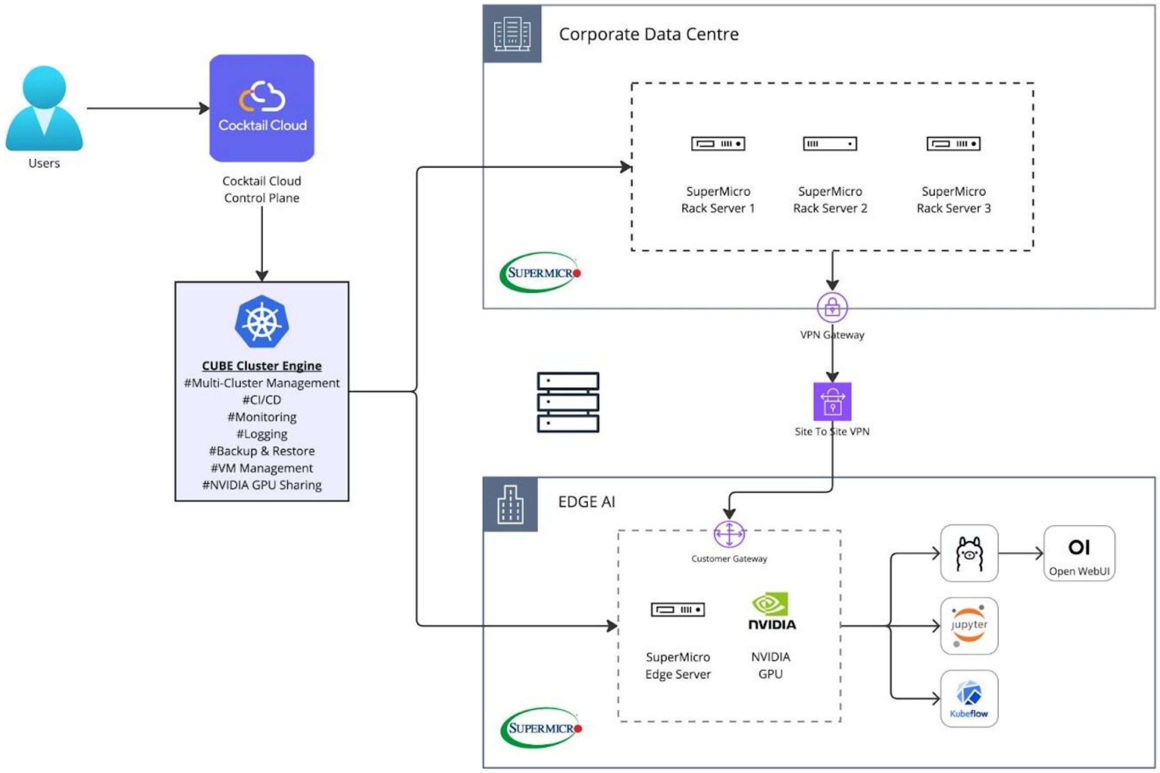

Figure 4 – Network Connections

Performance Evaluation on AI Inferencing at the Edge

Overview

The solution leverages Cocktail Cloud’s integrated NVIDIA GPU sharing capabilities with an NVIDIA A100 GPU, demonstrating the robustness and efficiency of this comprehensive Edge AI solution. It underscores the effectiveness of integrating Cocktail Cloud on Supermicro's hardware for Edge AI applications. Despite some marginal differences in raw performance metrics compared to NVIDIA's industry benchmarks, the advanced configuration capabilities of Multi-Instance GPU (MIG) via Cocktail Cloud provide substantial additional value.

Cocktail Cloud not only simplifies the configuration of MIG without requiring extensive setup on NVIDIA's CUDA platform but also offers a suite of DevOps tools for managing application deployment. This makes it easier for enterprises to deploy, monitor, and scale their AI applications seamlessly. Combining these features ensures that enterprises can achieve optimal performance and operational efficiency with minimal effort.

Calculating GPU Usage for Model Deployment

When determining the GPU usage required to deploy a model, our primary consideration is the model’s parameter size. The Llama2-70B model, for instance, is a large language model with 70 billion parameters. Utilizing 8-bit quantization, each parameter is stored using 8 bits. Consequently, for 70 billion parameters, the required GPU memory is calculated as follows:

GPU Memory= 70 × 109 × 8 bits = 70GB

This calculation aligns with the observed usage in the table mentioned above, demonstrating the model’s efficiency and the effective memory utilization facilitated by Cocktail Cloud. By understanding and planning for these memory requirements, enterprises can better allocate their GPU resources, ensuring that even large models like Llama2-70B can be deployed effectively and efficiently.

MIG 1g.20gb:

- Typical Usage: Suitable for smaller AI inference tasks with less memory and compute power.
- MIG Configuration: MIG (Multi-Instance GPU) allows a single physical GPU to be split into multiple isolated instances. In this case, "1g.20gb" means that one instance (1g) is allocated out of the full GPU, with 20 GB of dedicated GPU memory.
- Performance: Low power usage and temperature, designed for light tasks that don't need the full power of the GPU.

Llama3:70b-instruct-q8_0 under MIG 1g.20gb

Cocktail Cloud could maximize memory utilization, there was no issue in scaling, and the performance was not up to the mark due to the constrained GPU. The stats after the test execution looked something like this:

GPU Memory Utilization	24 %
GPU Memory usage	19.2 GB
GPU Temperature	38-40 degrees Celsius
GPU Power usage	Varying between 61w - 72w
Time taken to execute the task	20 minutes
Time taken (%)	100%

MIG 3g.40gb:

- Typical Usage: Mid-sized tasks that require more memory and compute power, such as training smaller models or running moderate inference tasks.

- MIG Configuration: The "3g.40gb" configuration allocates 3 GPU slices with a total of 40 GB of memory. This provides a larger portion of the GPU's compute power and memory, making it suitable for medium-sized AI tasks where more resources are needed than the 1g.20gb configuration.

- Performance: Medium power usage, moderate temperature, balancing performance and efficiency.

Llama3:70b-instruct-q8_0 under MIG 3g.40gb

The GPU usage spiked to a bit more than 39, and the task was executed much faster in 11 minutes.

GPU Memory Utilization	49%
GPU Memory usage	39.5 GB
GPU Temperature	44 - 45 degrees Celsius
GPU Power usage	118W and then spiking to 178
Time taken to execute the task	11 minutes
Time taken (%)	Reduction by almost 45%

MIG 7g.80gb:

- Typical Usage: Large-scale AI training and inference tasks require maximum memory and compute power.

- MIG Configuration: "7g.80gb" is the highest configuration available on this GPU, allocating 7 GPU slices and providing the full 80 GB of memory. This setup is optimized for the most demanding AI tasks, including large model training or high-performance inference.

- Performance: High power usage, higher temperatures, optimized for maximum performance and speed.

Llama3:70b-instruct-q8_0 under MIG 7g.80gb

The same task was executed in a mere 1.5 minutes. The execution was fast, and there was no lag, no hitches.

GPU Memory Utilization	89%
GPU Memory usage	72.2 GB
GPU Temperature	70 degrees Celsius
GPU Power usage	295w - 300w
Time taken to execute the task	1.5 minutes
Time taken (%)	Reduction by almost 93%

Comparative Analysis Table

Based on the NVIDIA documentation and performance metrics provided in their technical blog, here is a comparative analysis:

Metric	Our Results (MIG 1g.20GB)	Industry Standard (MIG 1g.20GB)	Our Results (MIG 3g.40GB)	Industry Standard (MIG 3g.40GB)	Our Results (MIG 7g.80GB)	Industry Standard (MIG 7g.80GB)
GPU Memory Utilization	24%	~25%	49%	~50%	89%	~90%
GPU Memory Usage (GB)	19.2	20	39.5	40	72.2	80
GPU Temperature (°C)	38-40	35-40	44-45	45-50	70	70-75
GPU Power Usage (W)	61-72	60-70	118-178	120-180	295-300	290-300
Time Taken (Minutes)	20	18-22	11	10-13	1.5	1-2

Key Insights

1. Enhanced Resource Allocation via Cocktail Cloud:

   a. Scalability and Flexibility: Cocktail Cloud’ s MIG configuration allows for dynamic resource allocation, enabling enterprises to run multiple tasks simultaneously while maintaining high efficiency. This flexibility ensures that workloads are optimally matched to available resources, maximizing hardware utilization.

   b. Adaptability: The ability to adjust GPU resources dynamically allows enterprises to swiftly adapt to changing workloads without extensive reconfiguration, saving both time and operational costs.

2. Operational Efficiency:

   a. Cost Management: Efficient management of GPU resources through Cocktail Cloud helps reduce operational costs. Enterprises can allocate the right amount of GPU power needed for specific tasks, preventing wastage and optimizing performance.

   b. Energy Efficiency: Optimized power usage and temperature control lead to reduced energy consumption, which is both cost-effective and environmentally friendly.

3. Reliability and Consistency:

    a. Stable Performance: Despite slightly lower raw performance metrics, the stability and predictability of the solution across different configurations ensure reliable operations. This consistency is crucial for mission-critical applications where performance reliability is more important than peak performance.

    b. Fault Isolation: The MIG configuration in Cocktail Cloud provides fault isolation between instances, ensuring that issues in one instance do not impact others. This enhances the system's overall reliability and uptime.

4. Comprehensive Management and Monitoring:

    a. User-Friendly Interface: Cocktail Cloud offers an intuitive interface for managing and monitoring GPU resources, making it easier for IT teams to oversee operations without deep technical expertise.

    b. Real-Time Insights: Advanced monitoring capabilities provide real-time insights into GPU performance, allowing for proactive management and optimization of resources.

5. Day 2 Operations:

    a. CI/CD Integration: Enterprises can easily implement continuous integration and continuous deployment (CI/CD) pipelines, streamlining the development and deployment of AI models.

    b. Logging and Monitoring: Comprehensive logging and monitoring capabilities ensure that any issues can be quickly identified and resolved, maintaining smooth operations.

    c. Backup & Restore: Reliable backup and restore functionalities protect critical data and models, ensuring business continuity.

    d. Multi-Tenancy: Support for multi-tenancy allows multiple users or departments to share the same infrastructure securely, optimizing resource utilization and cost-efficiency.

Conclusion

This solution benefits AI data inference and processes at the edge. All those industries where edge devices might have to be deployed in challenging and extreme environments like mining, heavy-duty industrial factories, remote locations with patchy network conditions, retail, hospitals, and many more would benefit from a solution like this.

The solution’s raw performance metrics are closely aligned with industry standards. The value lies in the advanced configuration capabilities and additional benefits Cocktail Cloud provides. These include enhanced resource allocation, operational efficiency, reliability, comprehensive management and monitoring features, and robust Day 2 operations such as CI/CD, logging, monitoring, backup & restore, and multi-tenancy. These attributes ensure that enterprises can rely on the complete solution of SuperMicro and Cocktail Cloud for Edge AI solutions to deliver robust, scalable, and efficient performance in real-world applications, making it a competitive choice for edge AI deployments.

Summary

The combination of Supermicro's NVIDIA GPU-powered Edge servers with Cocktail Cloud is tailored for various edge AI applications, offering adaptable and powerful solutions suitable for industries requiring immediate, high-performance AI processing and workloads and scaling modern application deployment and operations across any infrastructure. These attributes make it a highly competitive choice for enterprises seeking to deploy robust and adaptable AI solutions at the edge.

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.

For more information: www.supermicro.com

NAMUTECH

NAMUTECH is a cloud solutions company that helps customers with their digital transformation and effective operation, from virtualization to cloud,infrastructure, big data, and AI. Since its establishment in November 2001, NAMUTECH has continuously grown and has served over 300 clients, including renowned enterprises such as Samsung, LG, and Hyundai as well as several government agencies. For more information visit https://www.namutech.io