MINISFORUM

ENTERPRISE AI MINI WORKSTATION

BUYER'S GUIDE

Global Edge Computing Solution Brand & System Integrator

Table of Contents

ENTERPRISE AI MINI WORKSTATION BUYER'S GUIDE

Compact Workstation Architecture for Local AI and Edge Computing

1 Overview

The Breakthrough Path for Next-Generation Enterprise Computing Architecture

For modern enterprises, where computing power is deployed determines both the speed and the boundaries of business operations.

With the rapid rise of Large Language Models (LLMs), the widespread adoption of Industrial IoT (IIoT), and the growing importance of data privacy, traditional cloud-centric centralized IT architectures are facing three major challenges: network latency, escalating bandwidth costs, and increasing data security risks. Enterprise IT infrastructure is undergoing an irreversible transition toward distributed architectures and edge computing.

Against this backdrop, MINISFORUM's AI Mini Workstation Series introduces a new form factor for enterprise computing. Breaking free from the size and power limitations of traditional tower workstations and rack servers, these systems deliver uncompromising enterprise-grade computing and AI inference performance in a chassis as small as 1.7 liters.

Its core business value includes:

- **Low-Cost Local AI Deployment**

For enterprise knowledge base assistants, AI Agents, and automated customer service applications, traditional deployment methods often rely on extremely expensive multi-GPU servers, such as systems equipped with four RTX 4090 graphics cards.
Leveraging the Unified Memory Architecture (UMA) of the MINISFORUM MS-S1 MAX, a single unit can allocate up to 96GB of VRAM, enabling smooth execution of large-scale AI models. Furthermore, by clustering just four systems together, enterprises can privately deploy full-scale models such as DeepSeek-R1 671B Q4 locally, achieving both computing freedom and data security at a significantly lower cost.

- **Redefining the Total Cost of Ownership (TCO)**

Compared with traditional rack servers that often consume between 500W and 1500W and require dedicated server rooms and precision cooling

systems, MINISFORUM AI Mini Workstations typically operate within a power range of only 65W to 300W.

Combined with flexible deployment options—including desktops, compact server cabinets, and production-line environments—organizations can significantly reduce electricity costs, space requirements, and cooling expenses throughout a five-year lifecycle.

- **Enterprise-Class Reliability and Manageability**

The product lineup is deeply integrated with enterprise IT requirements. For example, the MS-02 Ultra incorporates dual 25GbE high-speed network ports, ECC memory support, and hardware RAID capabilities within a compact 4.8-liter chassis. Combined with Intel vPro® platform technology, IT administrators can perform low-level remote diagnostics, repairs, and system recovery even when a system is unresponsive, dramatically reducing the operational costs associated with geographically distributed edge computing nodes.

MINISFORUM AI Mini Workstations are not merely smaller PCs. They serve as powerful computational neurons connecting cloud infrastructure with edge endpoints, becoming critical IT assets for enterprises seeking greater efficiency, lower operational costs, and accelerated digital transformation in the AI era.

Advancing Toward High-Density, High-Efficiency Computing

Driven by digital transformation and AI adoption, enterprises no longer need to invest heavily in oversized and expensive traditional servers for non-core data center workloads.

With exceptional performance, ultra-compact form factors, flexible expandability, and low total lifecycle costs, MINISFORUM AI Mini Workstations demonstrate that compact size and powerful computing can coexist. They provide an ideal foundation for software development, edge cybersecurity, local AI knowledge base deployment, and high-end commercial applications.

2 Trends in Enterprise Computing Architecture

Enterprise IT infrastructure is increasingly transitioning toward distributed computing architectures.

Table 1: Evolution of Enterprise Computing Architecture

Stage	Computing Location	Key Characteristics
Traditional IT	On-Premises Enterprise Servers	Centralized Computing
Cloud Computing	Cloud Data Centers	Outsourced Computing
Edge Computing	Near the Data Source	Distributed Computing

In traditional architectures, most workloads are processed within centralized data center servers. In modern enterprise environments, however, computing tasks are increasingly distributed across multiple nodes, including development servers, AI inference nodes, and edge computing devices.

This transformation is primarily driven by three major trends:

- Growth in data volume
- Widespread adoption of AI technologies
- Increasing demand for edge computing

Table 2: Edge Computing Architecture Overview

Architecture Layer	Typical Devices / Components	Primary Functions	Data Flow
Endpoint Layer	Surveillance cameras, IoT sensors, industrial equipment, PCs, mobile devices	Generate raw data such as video streams, sensor readings, and user requests	Data is generated and sent to the edge computing layer
Edge Computing Layer	Edge servers, AI Mini Workstations, edge AI nodes	Local data processing, AI inference, data filtering, real-time analytics	Receive device data → process locally → send summaries or results to the cloud
Cloud & Data Center Layer	Cloud servers, data center servers, centralized storage systems	Large-scale data storage, model training, long-term analytics, centralized management	Receive processed data from edge nodes → perform deep analysis or model training

3 Architectural Characteristics of AI Mini Workstations

AI Mini Workstations are designed to deliver enterprise-grade computing capabilities in a compact and space-efficient form factor.

Compared with conventional PCs, these systems typically offer the following characteristics:

- Compact form factor
- Excellent power efficiency
- Enterprise-grade networking interfaces

- Customizable hardware configurations
- Easy maintenance and hardware upgrades
- Stable and reliable performance

Table 3: AI Mini Workstation Internal Architecture Overview

Architecture Module	Key Components	Description	Relationship to the CPU
CPU	Processor (CPU)	The core computing unit of the system, responsible for executing programs, processing data, and coordinating hardware resources.	Central component of the system; all modules connect to the CPU through the motherboard.
Memory	DDR4 / DDR5 System Memory	Provides high-speed data caching to support application execution and multitasking.	Directly connected to the CPU through the memory controller.
Storage	PCIe M.2 NVMe SSD	Stores the operating system, applications, and data while delivering high read/write performance.	Connected to the CPU via the PCIe bus.
PCIe Expansion	PCIe x16 / x8 / x4 Slots	Supports the expansion of high-performance hardware such as GPUs, network adapters, and storage controllers.	Communicates with the CPU through PCIe lanes.
Network Ports	2.5GbE / 5GbE / 10GbE / 25GbE Ethernet Ports	Provides high-speed network connectivity for data transfer, server communication, and cluster deployment.	Connected to the CPU through PCIe interfaces or onboard controllers.

4 MINISFORUM AI Mini Workstation Portfolio

When selecting an AI Mini Workstation, enterprises typically evaluate the following key criteria:

- CPU performance
- AI inference capability
- Network bandwidth
- Memory and storage performance
- Expandability

These critical factors directly influence the selection of an AI Mini Workstation for enterprise deployment.

- CPU performance determines the system's computing speed and processing capability, directly affecting task execution efficiency. It is particularly important for AI inference workloads that require substantial computational resources.
- AI inference capability determines the size of models the system can support and the speed at which inference can be performed. For local deployment of Large Language Models (LLMs), memory allocation, available VRAM, and compute resources are especially critical.

- Network bandwidth affects data transfer speed. In distributed computing environments, workstations equipped with higher-bandwidth networking can synchronize data more efficiently and reduce latency.
- Memory and storage performance determine the efficiency of data access and processing. When working with large-scale datasets, both memory capacity and storage throughput become increasingly important.
- Finally, expandability determines the platform's long-term flexibility. Support for additional expansion cards, storage devices, networking interfaces, and other hardware upgrades enables the workstation to accommodate future business and infrastructure requirements.

By considering these factors collectively, enterprises can select the workstation configuration that best aligns with their computing requirements and budget.

The current MINISFORUM AI Mini Workstation lineup includes:

- MS-01
- MS-A1
- MS-A2
- MS-S1 MAX
- MS-02 Ultra

Let us first compare their hardware specifications:

Table 4: Hardware Specifications Comparison

Model	MS-01	MS-A1	MS-A2	MS-S1 MAX	MS-02 Ultra
					
CPU	Intel Core i9-13900H	AMD Ryzen 7 8700G	AMD Ryzen 9 9955HX	AMD Ryzen AI Max+ 395	Intel Core Ultra 9 285HX
CPU Spec	14C/20T, Max 5.4GHz	8C/16T, Max 5.1GHz	16C/32T, Max 5.4GHz	16C/32T, Max 5.1GHz	24C/24T, Max 5.5GHz
GPU	Intel Iris Xe Graphics eligible	AMD Radeon 780M	AMD Radeon 610M	AMD Radeon 8060S Graphics	Intel Graphics
Memory	2×DDR5 SO-DIMM, up to 5600MT/s & 96GB	2×DDR5 SO-DIMM, up to 5600MT/s & 96GB	2×DDR5 SO-DIMM, up to 5600MT/s & 96GB	LPDDR5x UMA memory, 8000MHz & 128GB	4×DDR5 SO-DIMM, up to 4800MT/s & 256GB, ECC supported
Storage	3×M.2 NVMe SSD, up to 23TB	4×M.2 NVMe SSD, up to 16TB	3×M.2 NVMe SSD, up to 23TB	2×M.2 NVMe SSD, up to 16TB	4×M.2 NVMe SSD, up to 32TB
Wireless	Wi-Fi 6E, BlueTooth 5.2	Wi-Fi 6E, BlueTooth 5.3	Wi-Fi 6E, BlueTooth 5.2	Wi-Fi 7, BlueTooth 5.4	Wi-Fi 7, BlueTooth 5.4
I/O	1×USB 3.2 Gen1 Type-A 2×USB 2.0 Type-A 1×3.5mm Audio Combo Jack 1×USB4 1×HDMI™ 2.0 2×USB3.2 Gen2 Type-A 2×2.5GbE LAN 2×10GbE SFP+ LAN 1×DC IN 1×PCIe x16 (PCIe 4.0 x8)	1×DC IN 1×3.5mm Audio Combo Jack 2×2.5GbE LAN 1×USB-C 1×DP 1×HDMI™ 1×USB 3.2 Gen2 2×USB 2.0 1×USB 3.0 Gen1 1×OCuLink	1×3.5mm Audio Combo Jack 2×USB 3.2 Gen1 1×USB 2.0 2×USB 3.2 Gen2 Type-C 1×HDMI™ 2.1 2×2.5GbE LAN 2×10GbE LAN 1×USB 3.2 Gen2 1×USB 3.2 Gen1	1×AC IN 2×10GbE LAN 2×USB 3.2 Gen2 2×USB 2.0 2×USB4 V2 1×HDMI™ 2.1 FRL 2×USB4 1×USB 3.2 Gen2 1×3.5mm Audio Combo Jack 2×DMIC 1×PCIe x16 (PCIe 4.0 x4)	2×25GbE LAN 1×10GbE LAN 1×2.5GbE LAN 1×HDMI™ 2.1 FRL 1×USB4 4×USB 3.2 Gen 2 2×USB4 V2 1×3.5mm Audio Combo Jack 1×PCIe 5.0 x16 1×PCIe 4.0 x4 1×PCIe 4.0 x16 (Physically populated)
Network	2×2.5GbE LAN 2×10GbE SFP+ LAN	2×2.5GbE LAN Wi-Fi 6E	2×2.5GbE LAN 2×10GbE LAN Wi-Fi 6E	2×10GbE LAN Wi-Fi 7	2×25GbE LAN 1×10GbE LAN 1×2.5GbE LAN Wi-Fi 7
Dimensions	196×189×48mm	189.5×186×48mm	196×189×48mm	222.1×206.3×77.1mm	221.5×97×225mm

Model	MS-01	MS-A1	MS-A2	MS-S1 MAX	MS-02 Ultra
Weight	1.39kg	1.8kg	1.4kg	2.8kg	3.45kg
OS	Windows 11	Windows 11	Windows 11	Windows 11	Windows 11
Features	• Ultra-compact 1.7L chassis • Slide-out motherboard design • PCIe expansion slot • Supports low-profile graphics cards • Rich high-speed networking connectivity • Supports Intel vPro® technology • Supports enterprise-grade U.2 SSDs	• Ultra-compact 1.7L chassis • Slide-out motherboard design • OCuLink interface for external GPU expansion • Supports user-upgradeable AM5 desktop processors (up to AMD Ryzen™ 9 9950X)	• Ultra-compact 1.7L chassis • Slide-out motherboard design • PCIe expansion slot • Supports low-profile graphics cards • Rich high-speed networking connectivity • Supports enterprise-grade U.2 SSDs	• Up to 126 TOPS of AI compute performance • High-performance Radeon™ 8060S integrated graphics • 128GB high-speed unified memory architecture (UMA) • Up to 96GB memory allocation for GPU workloads • Slide-out motherboard design • PCIe expansion slot • Dual USB4 ports (40Gbps) • Dual USB4 Version 2 ports (80Gbps) • Peak power consumption of 160W, sustained at 130W • Built-in 320W power supply unit	• High-performance desktop-class processor • Multiple PCIe expansion slots • Supports dual-slot discrete graphics cards • Supports Intel vPro® technology • Supports ECC memory • Enterprise-grade 25GbE high-speed networking • Supports memory expansion up to 256GB • Built-in 350W power supply unit
Typical application scenarios	• Remote IT Management • Virtualizatio	• IT Operations • Virtualization	• IT Operations • Virtualization	• AI Development • IT Operations • Professional Workstation for Media Production, Music, Architecture, and Healthcare	• Remote IT Management • Professional Workstation for Media Production, Music, and Architecture

Note: All specifications listed in the table are based on the highest-configuration version of each model. Product specifications may be adjusted due to product revisions or optimization updates. Please refer to the latest information available on the MINISFORUM official website.

As shown above, each model offers its own unique positioning and strengths. The following sections provide a closer look at the core features and recommended deployment scenarios for each system.

MINISFORUM MS-01

The MS-01 is a mini workstation designed with a strong focus on networking capabilities and edge deployment.

Its key features include:

- Intel platform processor with Intel vPro® support
- Dual 10GbE SFP+ high-speed network ports
- PCIe x16 expansion slot
- Support for low-profile GPUs (such as the NVIDIA RTX A2000)
- Support for multiple NVMe SSDs

These features make it particularly suitable for:

- Virtualization nodes (VMware / Proxmox)
- Network security appliances
- Edge computing servers
- High-speed NAS controller nodes

For IT environments that require high-bandwidth network connectivity, the MS-01 offers exceptional deployment flexibility. Support for Intel vPro® enables IT administrators to remotely power on, recover, repair, reinstall operating systems, and modify BIOS settings through a wired network connection, even when the system is powered off, unresponsive, or experiencing operating system failures.

For organizations operating multiple distributed edge nodes, Intel vPro® can significantly reduce the need for on-site maintenance visits, lowering operational expenses while improving management efficiency and hardware-level security.

MINISFORUM MS-A1 and MS-A2

The MS-A1 and MS-A2 are both built on the AMD Ryzen™ platform and are positioned as high-performance productivity-focused mini workstations. The primary differences between the two models are: The MS-A1 places greater emphasis on expandability, including OCuLink eGPU support; The MS-A2 offers higher network bandwidth and stronger CPU performance.

These systems are well suited for:

- Software development teams
- Creative design studios

- Data analysis workloads
- Small-scale server deployments

With pricing comparable to conventional PCs while offering superior performance and expandability, both models deliver excellent value for small and medium-sized businesses.

MINISFORUM MS-S1 MAX

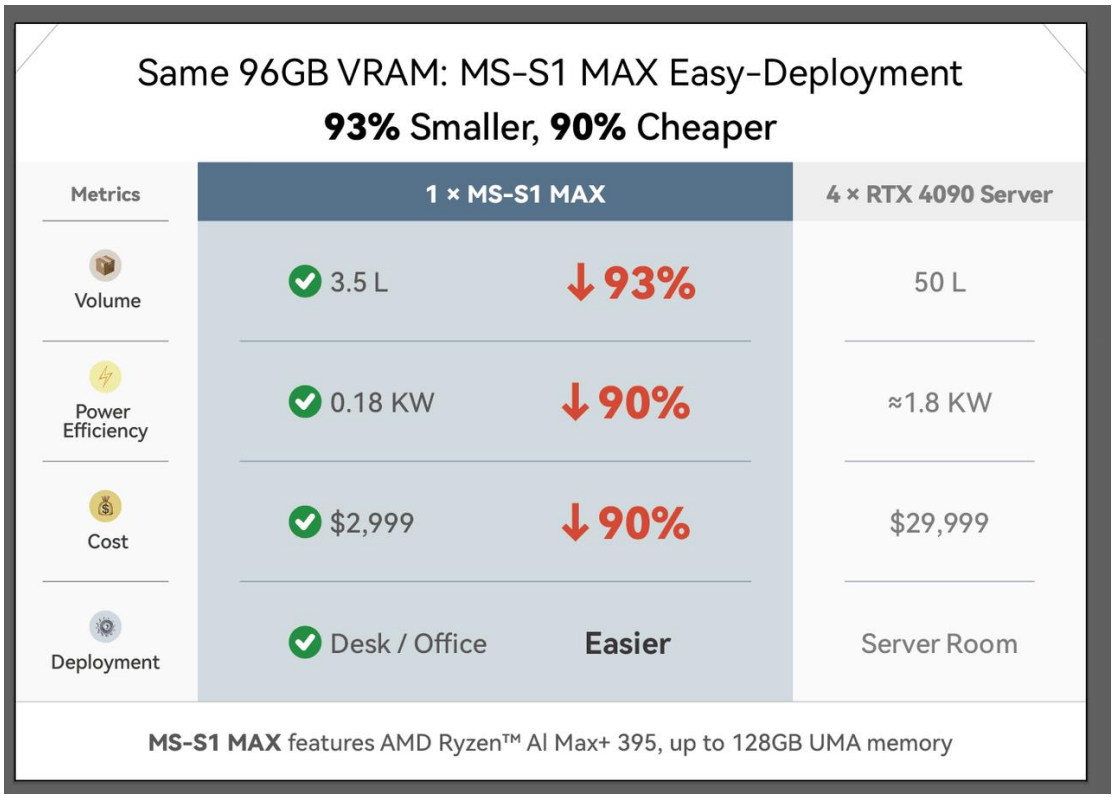
The MS-S1 MAX is a flagship AI Mini Workstation specifically designed for enterprise local AI inference.

Its most distinctive feature is its 128GB Unified Memory Architecture (UMA), of which up to 96GB can be allocated to GPU workloads. Since memory capacity is often the primary constraint when running large language models, this architecture enables support for significantly larger AI models.

The system delivers up to 126 TOPS of AI compute performance. Multiple units can also be combined into an AI cluster. For example, a cluster of four systems can locally deploy the full DeepSeek-R1 671B Q4 model. Here, Q4 refers to the model's quantization level, which is currently one of the most widely adopted options for edge deployment, providing a practical balance between model accuracy and resource requirements.

In many scenarios, this architecture can achieve AI inference performance approaching that of multi-GPU server configurations, such as systems equipped with four RTX 4090 graphics cards. Even a single MS-S1 MAX offers significant advantages in terms of physical footprint, power consumption, and deployment cost compared with traditional multi-GPU server solutions.

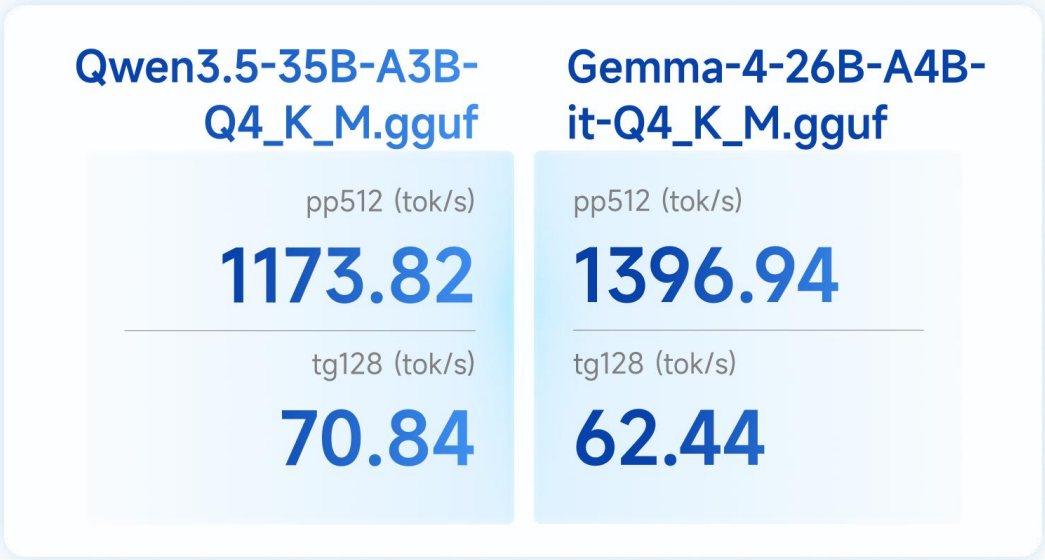
Chart 5: MS-S1 MAX vs. Quad RTX 4090 Comparison



The MS-S1 MAX also demonstrates strong local AI benchmark performance. In tests using the 35B-parameter Qwen model and the 26B-parameter Gemma model, the system achieved excellent results across both the PP512 and TG128 benchmark categories. In real-world

deployments, this translates into a smooth and responsive user experience.

Table 6: MS-S1 MAX Local AI Benchmark Results



Note: Testing was conducted on a single MS-S1 MAX configured with 128GB of memory using llama-b8783-win-vulkan-x64

Typical applications for the MS-S1 MAX include:

- Enterprise AI knowledge bases
- AI Agent
- Data collection and analysis
- AI-generated content creation

MINISFORUM MS-02 Ultra

The MS-02 Ultra is a flagship AI Mini Workstation designed for enterprise high-performance computing. Despite its compact 4.8L chassis, it offers

expansion capabilities approaching those of a traditional tower workstation.

Key features include:

- Intel® Core™ Ultra processor
- ECC memory support
- Intel vPro® support
- RAID 0 / 1 / 5 / 10 support
- Dual 25GbE ultra-high-speed network ports
- Three PCIe expansion slots
- Support for dual-slot GPUs

On selected configurations, one PCIe slot is pre-installed with a 25GbE network adapter.

The MS-02 Ultra is available with multiple processor options:

- Intel® Core™ Ultra 9 285HX
- Intel® Core™ Ultra 9 275HX
- Intel® Core™ Ultra 5 235HX

Different processor configurations may also include variations in other hardware features. Details are provided in the following table.

Table 7: MS-02 Ultra Configuration Comparison

Based on the five key evaluation criteria discussed earlier, we can visualize the positioning of each model through radar charts, providing a more intuitive comparison of their strengths and intended use cases.

Figure 8: Product Positioning Radar Charts

Organizations can select the most suitable model according to their specific performance requirements, deployment scenarios, and budget considerations.

5 Core Application Scenarios and Deployment Guidelines

AI Mini Workstations serve a wide range of enterprise applications. They are capable of meeting the requirements of diverse industries and workloads, including virtualization environments, local AI inference, large-scale data processing, and multitasking operations. This section explores how AI Mini Workstations can be efficiently deployed across these scenarios to maximize performance and provide stable, high-efficiency computing resources for enterprise workloads.

Virtualization Platform Support

MINISFORUM AI Mini Workstations provide robust support for virtualization platforms, enabling organizations to easily deploy and manage a wide variety of virtualized applications and environments. Common virtualization platforms include:

- **VMware**

VMware is one of the world's leading virtualization platforms and is widely used in data centers and cloud computing environments. With VMware, MINISFORUM AI Mini Workstations can run multiple virtual machines simultaneously while providing efficient resource allocation and management, ensuring stable performance under parallel workloads.

- **Proxmox**

Proxmox is an open-source virtualization platform known for its intuitive web-based management interface and flexible deployment options. Combined with MINISFORUM AI Mini Workstations, Proxmox enables organizations to build and manage efficient virtualized environments while keeping infrastructure and operational costs under control.

- **KVM**

KVM (Kernel-based Virtual Machine) is an open-source virtualization technology built into the Linux kernel. It delivers near bare-metal performance and is well suited for demanding virtualization workloads. With their strong hardware capabilities and system stability, MINISFORUM AI Mini Workstations provide an ideal platform for KVM deployments, meeting enterprise requirements for resource isolation and high-concurrency workloads.

All of the virtualization platforms above can fully leverage the hardware advantages of MINISFORUM AI Mini Workstations, improving the flexibility and reliability of enterprise IT infrastructure while supporting organizations of all sizes.

Local AI Deployment

In the past, the value of local AI inference was often debated. Many argued that users could simply access large language models through a web browser whenever an internet connection was available, making local deployment unnecessary except in situations involving privacy or security concerns.

Using the MS-S1 MAX as an example, some have questioned the necessity of a high-performance AI Mini Workstation specifically designed for local AI inference.

However, the growing adoption of AI agents such as OpenClaw has made the value of local AI deployment increasingly clear.

Many modern AI applications, particularly LLM-based AI agents, consume a substantial number of tokens during operation. In cloud-based

environments, token usage can become extremely expensive, especially when AI inference is performed at scale. As request volumes increase, enterprises often face continuously rising operational costs.

Under these circumstances, AI Mini Workstations such as the MS-S1 MAX become especially valuable. Equipped with a powerful APU and 128GB of Unified Memory Architecture (UMA), the system enables organizations to deploy AI inference locally. By running OpenClaw together with locally hosted large language models through platforms such as LM Studio, enterprises can eliminate the ongoing token costs associated with cloud-hosted models.

Compared with traditional cloud-based AI services, local deployment can significantly reduce operational expenses while providing a more efficient and cost-effective AI experience.

Compared with earlier generations of AI Mini Workstations, the local AI deployment capabilities of the MS-S1 MAX represent a major advancement. Previous systems often lacked the resources required to run advanced AI models locally, forcing users to rely on cloud computing services and incur token-related costs for every AI agent interaction.

With its powerful hardware architecture, the MS-S1 MAX can not only handle large-scale AI inference workloads but also significantly reduce cloud computing expenses through local deployment, providing enterprises with a practical path toward AI cost optimization.

AI Model Scale and Memory Requirements

Large language models are available in a wide range of parameter sizes. In general, larger models tend to provide stronger reasoning and language capabilities, although they also require greater computational resources.

As AI technology continues to advance, memory capacity has become one of the primary constraints when deploying large-scale AI models. This is particularly true for models containing tens or even hundreds of billions of parameters.

Taking the MS-S1 MAX as an example, the platform supports four-node clustering and is capable of running extremely large models such as DeepSeek-R1 671B at Q4 quantization. The following sections outline several common model sizes and their typical memory requirements.

- **7B Models**

Models with approximately 7 billion parameters typically require relatively modest memory resources. In most cases, 16GB to 24GB of available GPU memory is sufficient for inference.

MINISFORUM AI Mini Workstations can efficiently support models of this scale, making them ideal for everyday AI tasks. Smaller models often deliver faster response times and higher operating efficiency.

- **13B Models**

Models with approximately 13 billion parameters generally require at least 24GB of available GPU memory.

Typical applications include natural language processing, image recognition, and document analysis.

- **70B Models**

As model size increases, memory requirements rise significantly. Models with approximately 70 billion parameters typically require 48GB or more of available GPU memory.

The MS-S1 MAX can support models of this scale, making it suitable for computationally intensive workloads such as data analytics, speech recognition, and machine translation.

- **671B Models**

Models with approximately 671 billion parameters represent some of the largest AI models currently available for local deployment. Their memory requirements are extremely demanding and often require hundreds of gigabytes of available GPU memory.

Traditionally, deploying models at this scale required highly expensive systems equipped with four RTX 4090 GPUs or similar multi-GPU configurations.

The MS-S1 MAX introduces a new deployment approach through four-node clustering. By combining four MS-S1 MAX units, organizations can deploy DeepSeek-R1 671B at Q4 quantization locally.

This architecture overcomes the limitations of single-system memory capacity while leveraging the ability of each MS-S1 MAX to allocate up to 96GB of its 128GB UMA memory pool to GPU workloads. The result is improved memory utilization and more efficient large-scale AI inference.

Deployment Scenarios

MINISFORUM AI Mini Workstations are suitable not only for traditional computing workloads but also for a wide range of enterprise deployment scenarios.

- **High-Efficiency Office Environments**

For organizations requiring efficient collaboration and virtual desktop infrastructure, MINISFORUM AI Mini Workstations provide a powerful platform capable of running multiple virtual machines while maintaining stable resource allocation and performance. This helps improve productivity and user experience across the organization.

- **Data Centers and Server Clusters**

For data-intensive workloads, MINISFORUM AI Mini Workstations can serve as compute nodes within data centers and server clusters.

They are capable of supporting high-performance computing tasks such as big data analytics, rendering workloads, and distributed processing. Their compact form factor and energy-efficient design also help reduce space requirements and power consumption within the data center.

- **AI Applications**

For AI development teams, MINISFORUM AI Mini Workstations can serve as local AI inference platforms for model development, testing, and deployment.

When combined with solutions such as LM Studio and OpenClaw, these systems enable organizations to build efficient AI-powered applications, including digital employees capable of operating around the clock and other emerging AI-driven business solutions.

- **Remote Work and Flexible Work Environments**

As remote work continues to grow in popularity, MINISFORUM AI Mini Workstations provide an effective platform for flexible work arrangements.

Their compact design, rapid deployment capabilities, and support for virtual desktop infrastructure allow remote employees to access enterprise computing resources with a user experience comparable to working in a physical office, while maintaining collaboration efficiency and information security.

Through support for these diverse deployment scenarios, MINISFORUM AI Mini Workstations help organizations improve resource utilization,

streamline workload management, and enhance operational efficiency across a wide range of industries and applications.

6 AI Mini Workstations vs. Traditional Servers

As efficient, compact, and powerful computing platforms, AI Mini Workstations are well suited to a wide range of enterprise requirements. Compared with traditional tower workstations and rack servers, AI Mini Workstations offer distinct advantages in many deployment scenarios, particularly in terms of cost, space utilization, and power efficiency.

Reduced Space Requirements and Lower Costs

The compact design of AI Mini Workstations enables deployment in environments where space is limited, making them especially suitable for offices, laboratories, and other constrained workspaces.

Thanks to their small footprint, AI Mini Workstations can be placed directly on desks or integrated into existing work environments without requiring

dedicated server rooms. This significantly improves space utilization and reduces infrastructure requirements.

In addition, their lower power consumption helps organizations reduce electricity costs while also lowering expenses associated with cooling systems, air conditioning, and facility maintenance. Together, these factors contribute to a lower overall cost of operation.

High Performance with Flexible Deployment

Despite their compact size, AI Mini Workstations deliver substantial computing performance capable of supporting everyday productivity workloads, edge computing applications, and a range of AI inference tasks. Their energy-efficient design provides an effective balance between performance and power consumption.

At the same time, hardware expansion options such as M.2 and U.2 NVMe SSD slots, PCIe expansion slots, and additional networking capabilities allow organizations to scale system resources as business requirements evolve.

Selected MINISFORUM AI Mini Workstation models are also available in barebone configurations. This allows organizations to reuse existing memory modules and SSDs, helping reduce procurement costs while providing greater flexibility when configuring storage and memory resources in response to rising hardware costs.

Support for Diverse Application Scenarios

AI Mini Workstations are suitable not only for everyday office workloads and light computing tasks but also for virtualization deployments, local AI inference, and computing workloads commonly found in small and medium-sized enterprises.

For organizations that require capable computing resources without the high costs and deployment complexity associated with traditional servers, AI Mini Workstations provide an attractive alternative.

For example, in edge computing environments, AI Mini Workstations can perform low-latency local data processing, enabling real-time responsiveness without relying heavily on cloud infrastructure.

Reliability and Ease of Maintenance

AI Mini Workstations are efficient, compact, and powerful computing platforms that are well suited to a wide range of enterprise requirements. Compared with traditional tower workstations and rack servers, AI Mini Workstations offer distinct advantages in many deployment scenarios, particularly in terms of cost, space utilization, and power consumption.

With their strong computing performance, low power consumption, compact design, and flexible expandability, AI Mini Workstations provide enterprises with a computing platform that is both efficient and cost-effective. They are particularly suitable for small businesses and teams that need to balance performance requirements with space constraints. These advantages enable organizations to maximize the utilization of existing resources while maintaining efficient operations without compromising performance.

Table 9: AI Mini Workstations vs. Traditional Servers

Comparison Category	AI Mini Workstation	Tower Workstation	Rack Server
Power Consumption	Low to moderate power consumption, typically around 60W–150W, making it suitable for long-term operation in edge computing and office environments.	Moderate power consumption, typically 200W–800W, depending on GPU and expansion configurations.	High power consumption, typically 300W–2000W or higher, designed for data center-scale workloads.
Physical Footprint	Extremely compact, typically around 1L–5L in volume.	Traditional tower chassis, typically ranging from 20L–50L in volume.	Standard 19-inch rack-mounted equipment, typically ranging from 1U–4U in height.
Deployment Model	Can be deployed directly on desks, in network cabinets, or at edge locations, offering excellent deployment flexibility.	Typically deployed in offices or workstation environments and requires more physical space.	Deployed in data centers and server rooms, requiring racks, dedicated power infrastructure, and cooling systems.
Overall Cost	Lower acquisition and operating costs, making it suitable for small- to medium-scale computing nodes.	Moderate cost, typically used for high-performance personal workstations.	Highest acquisition and operational costs, requiring dedicated data center facilities and professional maintenance.
Expandability	Provides a reasonable level of expansion, including NVMe SSDs, PCIe slots, and networking upgrades.	Strong expandability, supporting multiple GPUs, storage devices, and expansion cards.	Highest level of expandability, supporting large-scale storage, networking, and compute resources.

7 Cost Model

When selecting an appropriate computing platform, understanding the Total Cost of Ownership (TCO) is essential. This section compares the cost models of AI Mini Workstations, Tower Workstations, and Rack Servers, covering factors such as hardware acquisition costs, power consumption,

space requirements, and maintenance expenses. Through a detailed analysis of these cost factors, organizations can make more informed decisions and select the platform that best aligns with their operational requirements.

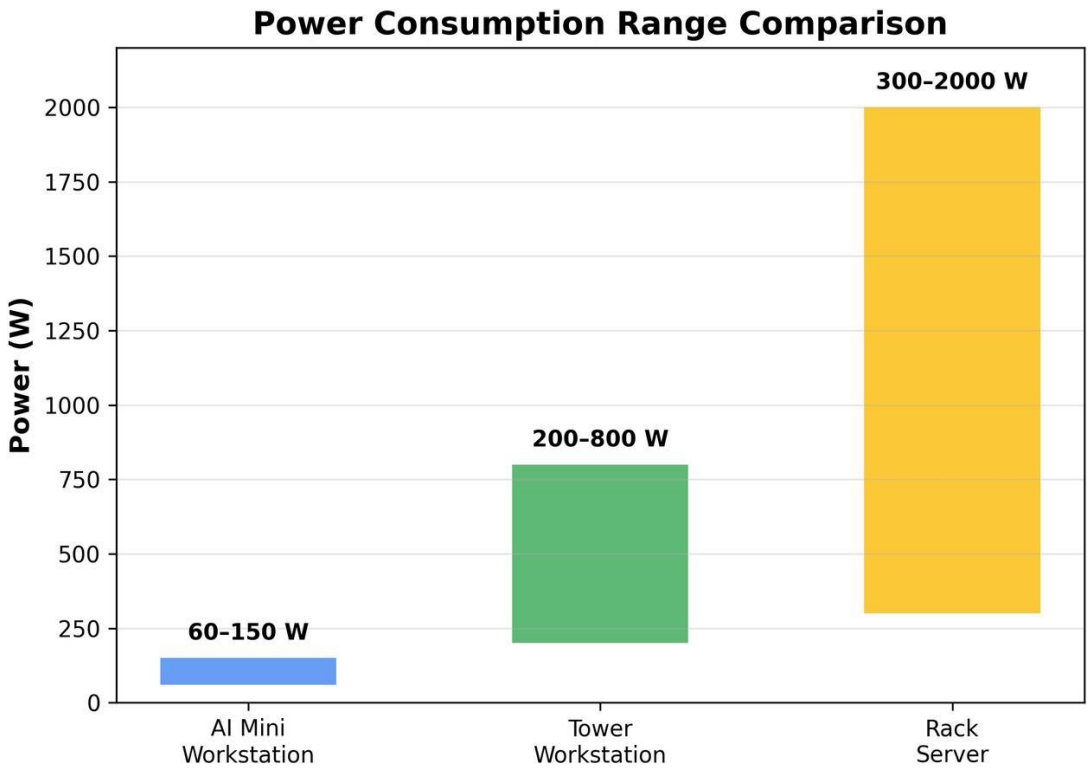
Power Consumption and Cost Analysis

Power consumption is one of the most important factors affecting long-term operating costs.

Figure 10 illustrates the power consumption comparison among AI Mini Workstations, Tower Workstations, and Rack Servers across different power ranges. As shown by the chart, AI Mini Workstations typically operate with significantly lower power consumption, making them particularly suitable for applications such as office productivity, edge computing, and other energy-efficient deployment scenarios.

By comparison, Tower Workstations commonly consume several hundred watts of power, while Rack Servers often operate at substantially higher power levels, in some cases exceeding one kilowatt.

Figure 10: Power Consumption Comparison



Note: The values shown are typical power consumption ranges intended to illustrate the relative power characteristics of different computing platforms. They do not represent actual measured power consumption of any specific product.

Deployment Costs

Figure 11 compares the deployment costs of AI Mini Workstations, Tower Workstations, and Rack Servers across several dimensions, including hardware, power consumption, space requirements, and maintenance.

- **Hardware Cost**

AI Mini Workstations generally offer lower hardware acquisition costs, making them suitable for small businesses and teams with limited budgets.

Tower Workstations typically fall within the mid-range cost category and are often selected by organizations requiring stronger computing capabilities.

Rack Servers usually involve the highest hardware investment and are best suited for organizations that require large-scale computing resources and high-performance infrastructure.

- **Power Consumption**

As power consumption increases, electricity costs rise accordingly.

AI Mini Workstations consume relatively little power and can significantly reduce energy expenses over time. Rack Servers, by contrast, typically incur higher long-term energy costs due to their greater processing capabilities and higher power requirements.

- **Space Requirements**

AI Mini Workstations occupy the smallest physical footprint, making them ideal for environments where space is limited.

Tower Workstations require more space and are typically deployed in offices, studios, or dedicated workstation environments.

Rack Servers require dedicated server rooms or data center facilities and must also accommodate racks, power distribution systems, and cooling infrastructure.

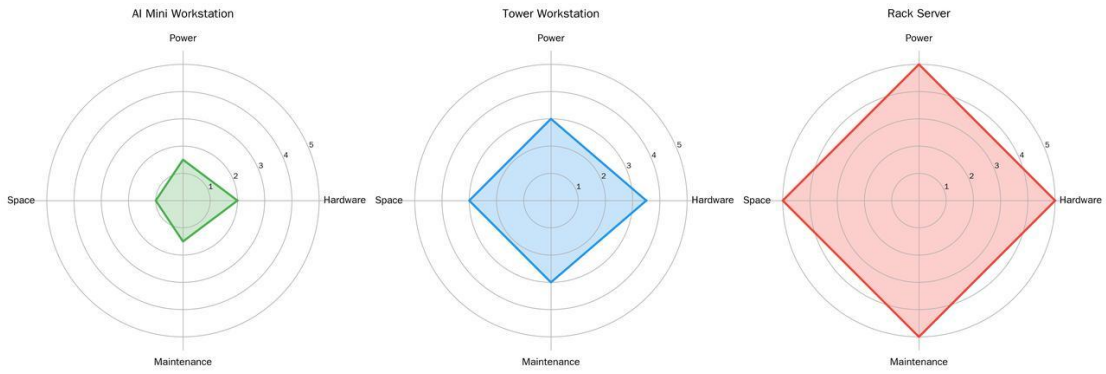
- **Maintenance Costs**

The relatively simple hardware architecture of AI Mini Workstations contributes to lower maintenance costs.

Tower Workstations generally involve moderate maintenance requirements, primarily related to routine hardware inspections and upgrades.

Rack Servers typically have the highest maintenance costs and often require specialized personnel for ongoing administration, monitoring, and troubleshooting.

Figure 11: Deployment Cost Comparison



Note: The values shown are relative deployment cost indicators intended to illustrate the characteristics and comparative deployment costs of different computing platforms. They do not represent actual measured costs of any specific product.

Five-Year TCO (Total Cost of Ownership)

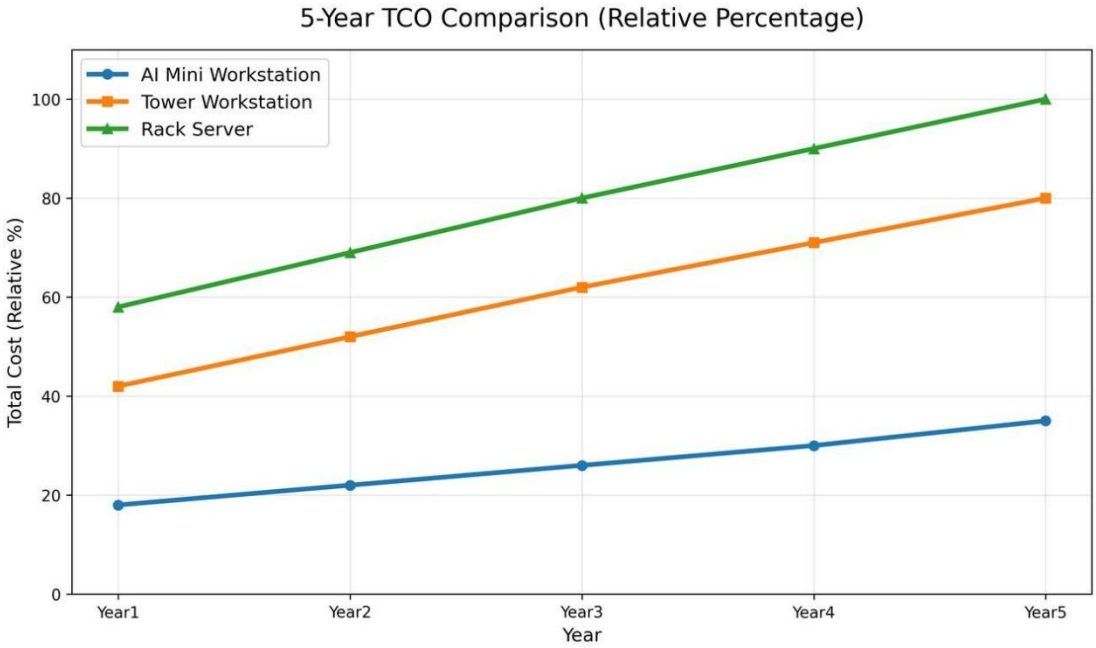
Figure 12 compares the estimated five-year Total Cost of Ownership (TCO) of different computing platforms.

The horizontal axis represents the deployment timeline, while the vertical axis represents total accumulated cost.

As illustrated by the chart, AI Mini Workstations generally require a lower initial investment and maintain a relatively stable cost profile throughout the deployment lifecycle. Their lower power consumption further reduces operating expenses over time, making them particularly attractive for small and medium-sized businesses (SMBs) and small studios.

In comparison, Tower Workstations and Rack Servers typically require higher upfront investment and are generally more suitable for organizations with sustained high-performance computing requirements and large-scale workloads.

Figure 12: Five-Year TCO Comparison



Note: The values shown are relative TCO estimates based on typical industry deployment assumptions. They are intended to illustrate general cost trends and comparative differences among computing platforms and do not represent the actual costs of any specific product, configuration, or deployment environment.

8 Procurement Guide

To help your organization rapidly deploy a cost-effective computing infrastructure, we recommend that your IT and procurement teams follow the evaluation and purchasing process outlined below.

Step 1: Identify Core Workloads and Business Challenges

Begin by assessing your organization's primary computing requirements and operational pain points.

- Networking and Edge Virtualization
  - Do you require high-bandwidth networking and concurrent workloads?
  - Are you planning to deploy software-defined networking solutions, virtual routers, or lightweight virtualization clusters such as VMware or Proxmox?
- Local LLM Deployment and AI Agents
  - Do you need to process large volumes of sensitive or proprietary business data locally?
  - Has the high cost of professional GPUs become a barrier to AI adoption within your organization?
- Mission-Critical Development and Compute Workloads
  - Do your workloads require exceptional single-threaded performance?
  - Are ECC memory protection and 25GbE-class networking throughput important for reliability and productivity?

Step 2: Match Your Requirements with the Ideal MINISFORUM Solution

Select the platform that best aligns with your business objectives.

- MS-01: Optimized for Networking and Edge Deployments

Equipped with dual 10GbE networking and PCIe expansion capabilities, the MS-01 is an ideal platform for virtualization environments, network infrastructure applications, and compact NAS controller deployments.

- MS-A1 / MS-A2: Cost-Effective Productivity Platforms

Designed for software development teams and creative studios, the MS-A1 and MS-A2 deliver performance and flexibility at a cost comparable to traditional PCs. The MS-A1 additionally offers OCuLink-based external GPU expansion for enhanced scalability.

- MS-S1 MAX: Private AI Computing Powerhouse

Designed for large language model (LLM) inference, the MS-S1 MAX combines up to 126 TOPS of AI compute performance with a Unified Memory Architecture capable of allocating up to 96GB of memory to GPU workloads, making it a compelling platform for edge AI clusters and local AI deployment.

- MS-02 Ultra: Enterprise Computing Flagship

For organizations seeking maximum performance, reliability, and expandability, the MS-02 Ultra delivers a powerful solution featuring up to

a 24-core processor, dual 25GbE networking, and Intel vPro® remote management capabilities for mission-critical workloads.

Step 3: Evaluate Emerging Subscription and Hardware-as-a-Service Models

In addition to traditional capital expenditure (CapEx) purchasing models, organizations may consider hardware leasing or Device-as-a-Service (DaaS) offerings when planning AI infrastructure investments.

These models can convert large upfront capital expenditures into predictable operational expenditures (OpEx), reducing financial pressure while helping organizations stay current with rapidly evolving AI hardware technologies.

Step 4: Start with a Proof of Concept

Nothing validates infrastructure decisions better than real-world testing.

The true value of hardware lies in how effectively it supports your applications, workflows, and business objectives. Before committing to large-scale deployment, organizations should conduct a proof-of-concept (PoC) evaluation using representative workloads to verify performance, compatibility, and return on investment.

👉Purchase intent form: https://s.minisforum.com/4vQ1Gyk

🛒 MINISFORUM Official Store: https://store.minisforum.com/

🛒 Amazon Store (US): https://www.amazon.com/minisforum

🛒 Amazon Store (DE): https://bit.ly/44rLY0Z

📧B2B Contact: opportunity@minisforum.com

👉US Leasing Service: https://reurl.cc/4lAMKV