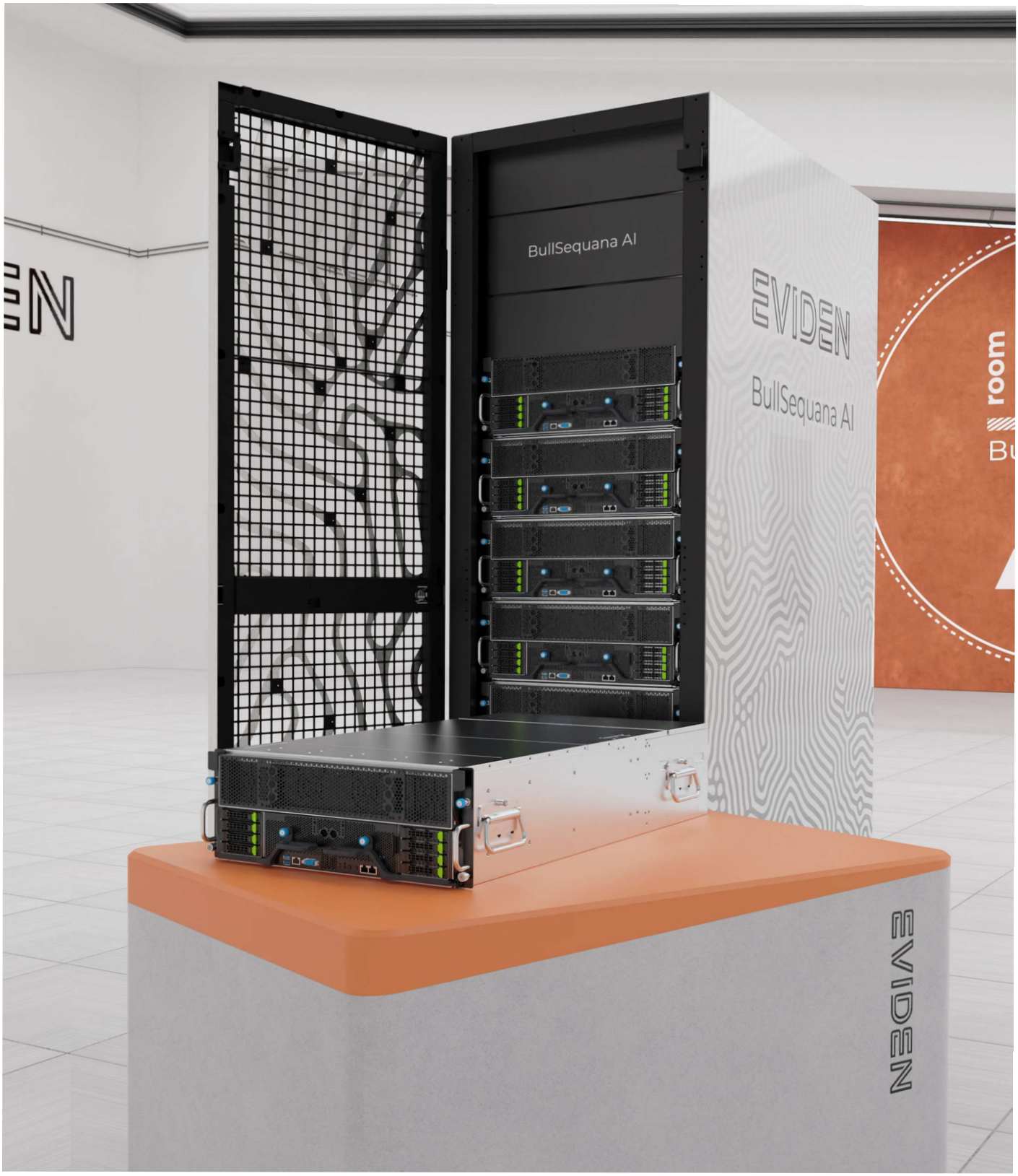
EVIDEN | AMD

BullSequana AI642-H03

A high-performance AI accelerator designed for training AI models and performing inference with large models



BullSequana AI 642-H03 Overview

The Eviden BullSequana AI 642-H03 server is an AI-optimized compute node designed to empower industries, national HPC centers, government bodies, academia, and startups to execute accelerated, low-latency AI inference and precision fine-tuning for mid-sized language models (~100M–10B parameters) and multimodal AI workloads.

The BullSequana AI 642-H03 offers superior speed and performance, with up to 48 AMD Instinct™ MI355X GPUs per rack across six servers. The system delivers 64 TB of global GPU bandwidth, enabling high-speed communication pathways and direct memory exchange without routing through the CPU. The AMD Instinct™ MI355X GPU architecture balances GPU-CPU performance for AI workloads, supports a wide range of precision formats for training and inference—including low-precision inference for mid-sized language models and offers 2.3TB HBM3e with 288GB of onboard memory per GPU.

Powered by AMD's latest AI technology, the BullSequana AI 642-H03 combines AMD EPYC processors with high-performance GPUs, making it an ideal solution for building training clusters that require eight GPUs per server.

The BullSequana AI 642-H03 server is hybrid cooled for better energy efficiency and sustained performance across demanding workloads, featuring manifold and rack-mounted CDU, with up to 70% direct liquid cooling (DLC). The remaining 30% is cooled by air via fans.

Designed for Performance, Built for Functionality

The server is purposely designed, customized, sized, and optimized to ensure the most effective power distribution unit (PDU), coolant distribution unit (CDU), effectively dissipating the hot air generated within compact 4U servers housed in 42U racks. For even higher energy efficiency, an optional rear door heat exchanger can be added, targeting compute-only workloads with enhanced thermal performance.

The standardized 800mm wide rack allows for easier cable and accessory integration, and the in-rack CDU and manifold provide direct liquid cooling. Additionally, it features 80+ titanium-certified power supplies and PWM cooling fans.

With its fully pre-integrated, pre-tested, and configured design, the BullSequana AI 642-H03 server offers fast deployment, streamlined installation, and reduces space and costs. This turnkey solution provides flexibility and efficiency for both rapid brown-field rollouts and greenfield deployments.

As part of the BullSequana AI 600H server family, the BullSequana AI 642-H03 is meticulously assembled and rigorously tested in Eviden's factory, backed by robust quality assurance processes that ensure reliability and give customers complete confidence in the product's performance and durability.

Software Infrastructure for Empowering AI Teams

AI developers can utilize software infrastructure to build, train, and optimize their models effectively. By leveraging the full capabilities of the AMD ROCm 7 software stack, AI engineers can significantly enhance both usability and performance for MI355X GPUs. This enhancement leads to faster results compared to competing solutions and accelerates the model development workflow.

AI system administrators benefit from Eviden's HPC Software Suite, which provides a unified platform for running, tuning, and managing nodes and clusters at scale. It integrates orchestration, profiling, intelligent resource control, and advanced power monitoring and management to optimize performance, enhance energy efficiency, and ensure cost-effective, reliable operations. This approach aligns fully with your team's commitment to digital sobriety and sustainable infrastructure.

Proven Track Record and Expertise

Eviden holds a leading position as the only European manufacturer of supercomputers, setting it apart in the global market. We provide end-to-end solutions—designing, delivering, installing, and configuring cluster systems for leading organizations worldwide. Our factory in France plays a strategic role in this mission, enabling high-quality production that supports innovation and digital sovereignty in AI and HPC infrastructure. With decades of experience, Eviden has earned the trust of industrial, academic, and HPC sectors across the globe.

BullSequana AI642-H03
Product brief Technical Specifications

GPU Architecture
Hybrid cooled 8 GPUs AMD MI355X UBB8
Memory: 2.3TB HBM3e / (288GB per GPU)
Interconnect: 8TB/s of global GPU bandwidth (1TB/s GPU to GPU)
Tokens per second (pending for AMD)
GPU TDP: 11200W (1400W per GPU)

CPU Architecture
Dual AMD EPYC™ 4th and 5th Gen processors
12-channels DDR5 RDIMM, 3TB total size

Storage
Front: 8 x 2.5" NVMe SSDs
Internal: 2 x M.2 NVMe SSDs

I/O Options
Interconnect BXI V3 / Ethernet, InfiniBand XDR
12 x PCIe Gen5 x16
2x 1 GbE RJ45, 2x USB 3.1, 1 x VGA port and 2x 10Gb LAN ports

System Management
BMC based on Aspeed AST2600
Supported Operating Systems and Software
Operating Systems: Red Hat Enterprise Linux
Hardware Tooling: Hardware Abstraction Layer (HAL) and
monitoring exporters & dashboards
Software Management Stack:
Smart Management Center xScale
Smart Maintenance Management Suite
Smart Performance Management Suite:
Eviden DOCA-OFED & OpenMPI
Eviden SLURM
Sylabs SingularityPRO and SingularityENT
Eviden IBMS (Interconnect fabric Monitoring System)
Job Monitoring & Cluster Efficiency: BullSequana ARGOS

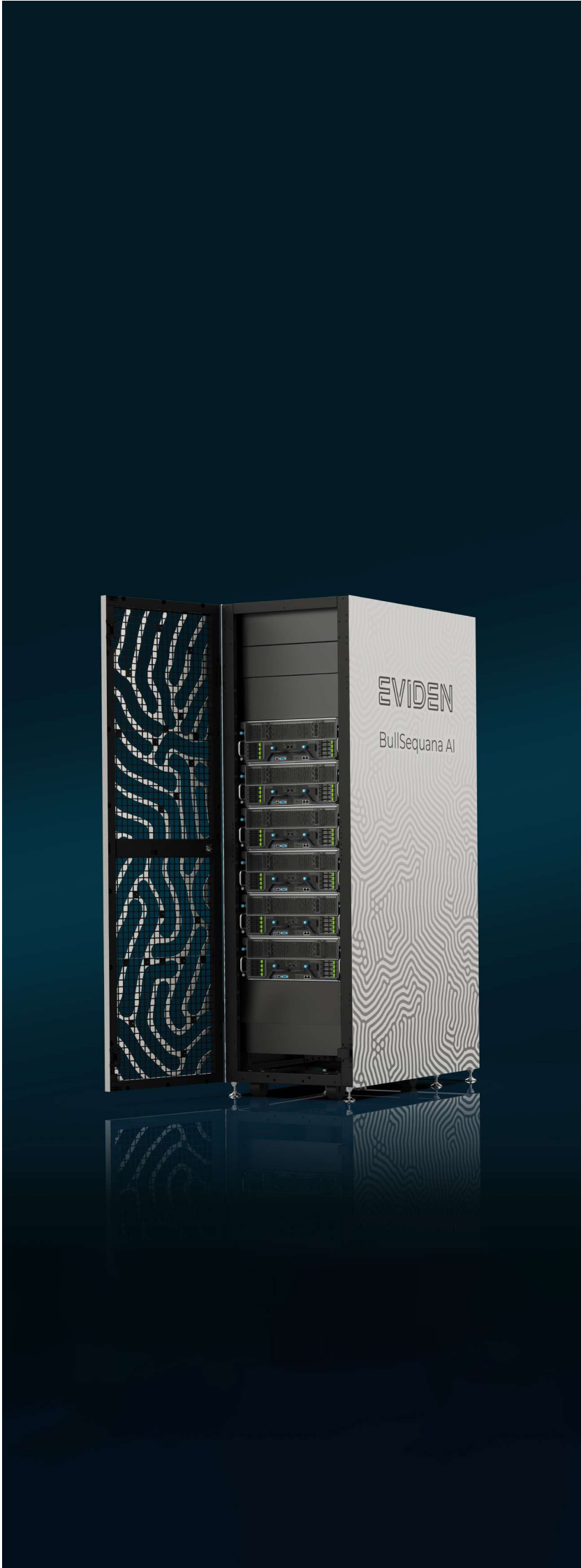
Availability and RAS Features
Smart Crises Management and Protection, Dual ROM Architecture
Hot swap devices (PSUs, PCIe blades, NVMe drives)

Security Features
Secure Boot, Secure firmware, and TPM 2.0

Power Supply Unit (PSU)
80 PLUS Titanium, Full redundancy

Physical Characteristics
4U form factor
TOTAL Weight: 110 Kg

Regulation and Safety
Compliance Global: RoHS, REACH
Country: CE, ErP Lot 9, FCC



Connect with us

/in/eviden
@EvidenLive
@evidenlive
/EvidenLive

eviden.com

ECT-250806-SB-106442 – BullSequana AI 642 Print V4