INVERGENT

Datasheet

7U DenseMax Appliance

Professional AI for the enterprise

Bring AI In-House with Confidence

The DenseMax appliance is a fully integrated, plug-and-play hardware platform engineered for enterprises that demand control over cost, performance, and privacy from their AI deployments.

It simplifies the journey from experimentation to deploying enterprise AI applications by providing enthusiasts, developers, and AI builders with pre-optimized models and industry-standard APIs for building powerful AI agents, co-pilots, chatbots, and assistants.

Instant Deployment, No Complexity

- Pre-configured 8-GPU system optimized for high-throughput LLM inference.
- Drop-in setup: power it up, connect it to your network, and start serving models within minutes.

Complete Data Sovereignty

- Keep sensitive data on-premise — no cloud dependency or data egress risk.
- Role-based access control, audit logging, and optional air-gapped operation.

Next-Gen Inference Performance

- Custom software components built for the Blackwell architecture deliver sub-second latency and massive throughput even under multi-user load.
- Includes the latest carefully chosen open-weight LLMs

Key Features

- Complete Plug & Play AI solution for enterprises
- Simple integration into AI applications, development frameworks, and workflows.
- Integrated security and guardrails.
- Audit, monitoring and observability.
- Optimized inferencing and fine-tuning engine.
- Carefully chosen, up-to-date models.
- Up to **8x** GPU cards:
  - NVIDIA RTX 5090 32GB
- Up to **256GB** GDDR7 VRAM

Fine-Tuning

- Fine-tune models directly on the appliance with a user-friendly UI — no need for external tools or code.
- Supports parameter-efficient tuning methods (LoRA, QLoRA) to maximize speed and minimize resource usage.

Enterprise-Grade Monitoring & Observability

- Real-time dashboards for GPU usage, memory, token throughput, and latency.
- Integrated alerts and logging for proactive maintenance and issue detection.

Secure by Design

- Encrypted storage, secure boot, and containerized model isolation.
- TLS-enabled APIs and full compliance with enterprise IT policies.

Enterprise Reliability

Designed for professionals who demand the best, the 7U DenseMax appliance delivers unparalleled performance, reliability, and support. Every component is rigorously tested for a wide range of design, engineering, and AI workflows and is continually optimized.
With enterprise-grade support, the 7U DenseMax appliance is engineered to facilitate seamless AI inferencing for the latest AI foundation models and is the trusted choice for enterprise and mission-critical work.

Get AI Infrastructure That Works for You

Whether you're modernizing operations, building internal copilots, or enhancing security posture, our appliance gives you speed, flexibility, and full control — all in one box.

Use Cases

🧠 Agentic Workflows

Deploy AI agents that take actions across internal systems, apps, and APIs — ideal for process automation, research, and task orchestration.

🔐Use & Protect Sensitive Data

Run inference and fine-tuning on data that cannot leave your infrastructure.

🧰Customize LLMs

Tailor models to domain-specific language, tone, and behavior using internal datasets — all managed through a low-code UI.

🧩Internal Tools and Workflows

Connect models to CRMs, ERPs, ticketing systems, document stores, or proprietary UIs for AI-native productivity.

🌐Multi-Tenant AI Across Teams

Serve different departments (e.g., legal, HR, marketing) from the same appliance with isolated, parallel model deployments.

🔁 Continuous Learning Loops

Use internal feedback and usage data to fine-tune and improve models regularly — keeping performance aligned with evolving needs.

🧪 Experiment & Iterate Locally

Rapidly test prompts, tune configurations, and evaluate model behavior without reliance on cloud costs or vendor limitations.

💸Predictable, Contained Costs

Avoid unpredictable usage-based cloud pricing. Run unlimited inference and fine-tuning workloads on a fixed-cost platform, eliminating API costs and reducing TCO over time.

Let's bring AI home.

Technical Specifications

Hardware

- 2x AMD Epyc 9005, 9004 or 97x4 processors
- 24x DDR5 ECC RDIMM slots
- 8x PCIe5 x16 GPU slots
- up to 8x NVIDIA RTX 5090
- up to 256GB GDDR7 VRAM, 1.8 TB/s memory bandwidth
- 5x 2000W industrial-grade PSU
- 2x Intenal PCIe3.0x4 M.2 NVMe Slots
- 4x U.2 PCIe5.0x4 NVMe slots
- 2x PCIe5.0x8 slots for networking, can be merged into 1x PCIe5.0x16
- Front I/O with IPMI port and 2x 1GbE RJ45
- Operation temperature: 10°C - 35°C / Non operation temperature: -40°C ~70°C
- Non operation humidity: 20% - 90% (Non condensing)
- 1x DB15 VGA port
- 4 Type-A (USB3.2 Gen1)

Software

- Automatic Prefix Caching
- Disaggregated Prefilling
- Fine-tune and serve multi-LoRA adapters
- Reinforcement Fine-Tuning (RFT)
- Speculative Decoding
- Prefill and Decoding (PD) Disaggregation
- Parallel Inference with multi-GPU, multi-Node sharding
- Structured Outputs
- Tool and Function Calling
- Weight Quantization: AutoAWQ, GPTQ, BitsAndBytes, INT4 W4A16, INT8 W8A8, FP8 W8A8
- KV Cache Quantization: FP8

Ready to Get Started?

To learn more, visit: invergent.ai

Let's bring AI home.