MemVerge

Memory Machine™ for AI

GPU-as-a-Service
For the Enterprise

Product Brief

Memory Machine™ for AI

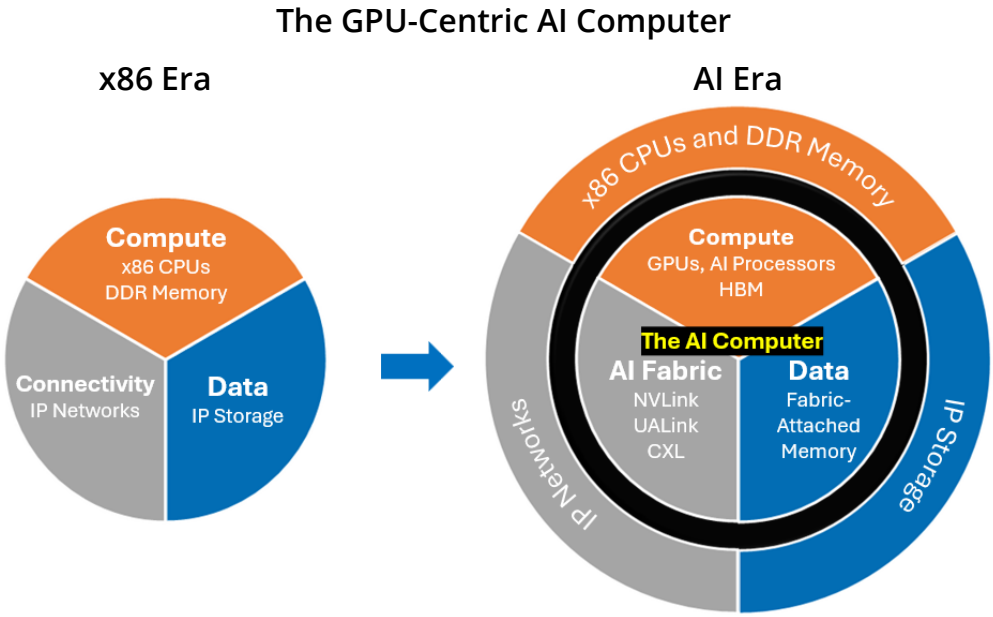
GPU-as-a-Service for the Enterprise

In today's competitive landscape, Enterprises are increasingly building their own GPU cluster infrastructures to minimize costs while ensuring data security and sovereignty. However, these high-value resources must be shared across multiple departments and users, leading to optimal resource allocation and management challenges. Efficient utilization of these GPU clusters is crucial for maintaining operational efficiency and maximizing return on investment.

MemVerge Memory Machine for AI addresses these challenges head-on. Designed specifically for AI training, inference, batch, and interactive workloads, the advanced software allows your workloads to surf GPU resources for continuous optimization. Serving GPU on-demand, Memory Machine ensures your clusters are fully utilized, delivering GPU-as-a-Service for superior performance, security, user experience, and cost savings.

Challenges MemVerge Solves for MLOps Administrators

Managing on-premises GPU clusters for multiple departments, projects, and users requires complex resource

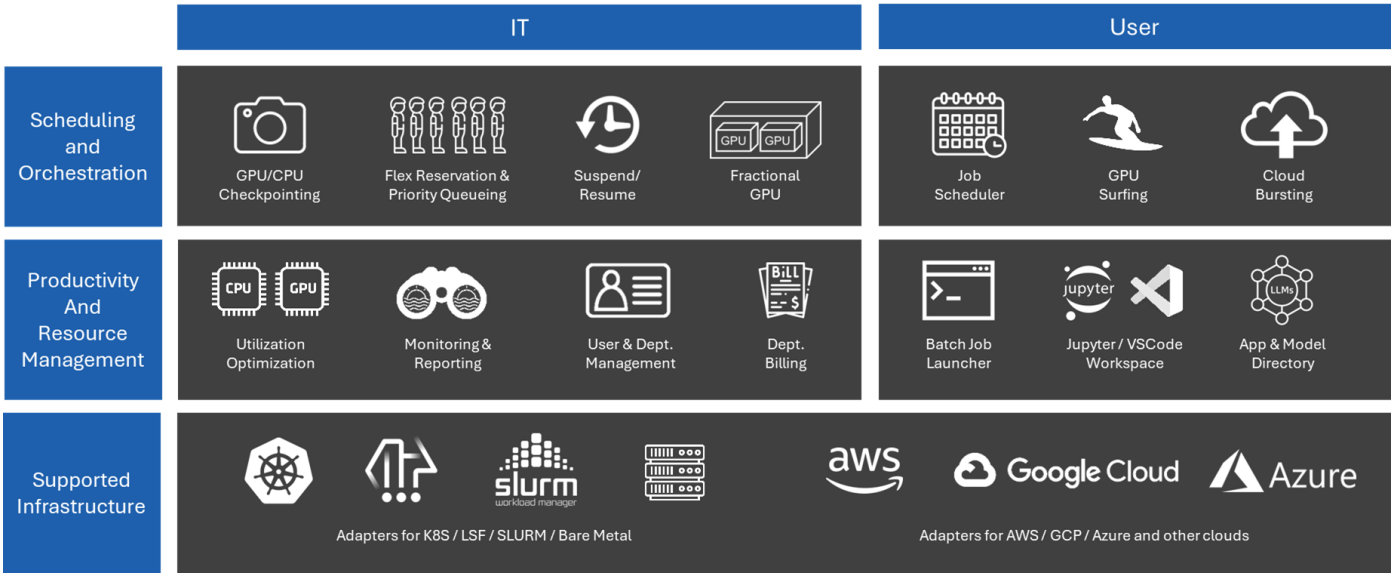


allocation and job scheduling to achieve high utilization. The Memory Machine for AI excels in providing solutions for:

- **GPU Availability and Cost Management:** Strategic resource allocation and shared GPU usage policies to ensure critical projects get necessary computational power while minimizing idle time.
- **Flexible GPU Sharing:** Allows GPUs to be dynamically allocated to different user projects when idle and seamlessly returned to their original users when needed, optimizing resource utilization and efficiency.
- **Real-time Observability and Optimization:** Tools that provide detailed insights into GPU usage, allowing dynamic workload adjustments for optimal resource utilization.
- **Priority Management:** Intelligent scheduling algorithms to balance project deadlines, resource needs, and user demands, ensuring fair access and minimal disruptions.
- **Cloud Bursting Management:** Enables organizations to extend their on-premises infrastructure to the public cloud, providing additional capacity during peak demand.

Introducing Memory Machine for AI

The MemVerge Memory Machine for AI empowers Enterprise organizations to harness Artificial Intelligence (AI) and Machine Learning (ML) while maintaining control and sovereignty of your data and infrastructure.



Key Features Empowering MLOps Administrators and End Users

- **GPU Surfing:** Ensure uninterrupted job execution by transparently migrating user jobs to available hardware resources when the original GPUs become unavailable, maintaining continuous operation and maximizing resource efficiency.
- **Automatically Suspend and Resume Jobs Transparently:** Seamlessly move user jobs across the AI Platform, safeguard against out-of-memory conditions, and prioritize critical tasks by automatically suspending and resuming lower-priority jobs, ensuring uninterrupted and efficient resource management.
- **Optimal GPU Utilization:** Intelligent GPU sharing algorithms eliminate idle resources and maximize utilization.
- **Intuitive User Experience:** Easy-to-use UI, CLI, and API for seamless workload management. The user interface provides proactive monitoring and optimization.
- **Intelligent Job Queueing & Scheduling:** Optimizes user jobs for the available hardware by employing a variety of advanced scheduling policies and algorithms to ensure maximum efficiency and performance.
- **Flexible GPU Allocations:** Optimize resource utilization by dynamically reallocating idle GPUs from other projects, ensuring efficient use of available hardware and minimizing downtime.
- **Optimized for NVIDIA GPUs:** Leverage advanced NVidia GPU capabilities for superior performance and efficiency with tailored optimizations that maximize the potential of NVidia hardware in AI/ML workloads.
- **Granular Resource Assignment:** Partition your infrastructure into specific departments and projects to ensure precise and optimal allocation of resources, maximizing efficiency and reducing waste.
- **Comprehensive Workload Support:** Accommodate diverse user workloads, including Training, Inference, Interactive, and Distributed tasks, with an integrated application and model directory that stores existing and customized Docker images in a secure private repository for streamlined deployment and management.
- **Extensive Infrastructure Support:** Seamlessly integrate with diverse infrastructure environments, including Kubernetes, LSF, SLURM, bare metal, and public cloud platforms such as AWS, GCP, Azure, and more, providing unparalleled flexibility and scalability for AI/ML workloads.
- **Cloud Bursting:** Enable seamless scheduling of user jobs on public cloud resources, ensuring continuous operation and scalability without compromising performance.

How Memory Machine for AI Boosts Productivity for the Enterprise

1. **Enhanced Resource Efficiency:** With features like Optimal GPU Utilization and Flexible GPU Allocations, the platform ensures that GPU resources are used to their fullest potential, minimizing idle times and maximizing throughput across various projects and departments.
2. **Uninterrupted Operations:** The GPU Surfing and Automatic Job, Suspend, and Resume features guarantee continuous job execution, even when underlying hardware resources become unavailable or high-priority jobs need to take precedence. This dynamic resource management ensures minimal disruptions and maximized productivity.
3. **Streamlined Workload Management:** The Intuitive User Experience and Intelligent Job Queueing & Scheduling provide MLOps administrators and end-users with an easy-to-use interface and advanced scheduling algorithms. This leads to efficient workload management and optimized performance for AI/ML tasks.
4. **Comprehensive Infrastructure Support:** The platform offers unparalleled flexibility and scalability by supporting a wide range of infrastructure environments, including Kubernetes, LSF, SLURM, bare metal, and public cloud platforms like AWS, GCP, and Azure. Enterprises can seamlessly integrate their existing on-prem systems and scale their AI/ML operations as needed.
5. **Data Security and Sovereignty:** The platform's design prioritizes data security and sovereignty, which is crucial for Enterprise customers. By enabling on-premises GPU cluster management and secure private repositories for Docker images, the platform ensures that sensitive data remains protected while complying with industry regulations and standards.

Key Use Cases

- **AI/ML Model Training and Inference:** Enterprises can leverage the platform's advanced GPU capabilities and comprehensive workload support to train and deploy machine learning models efficiently. The seamless resource management ensures optimal performance during intensive training and inference tasks.
- **Data-Driven Research and Development:** Researchers can utilize the platform for large-scale data analysis and experimentation. With features like GPU Surfing and Automatic Job Suspend and Resume, research workflows remain uninterrupted, allowing for continuous discovery and innovation.
- **Multi-Department Resource Allocation:** Organizations with multiple departments can benefit from Granular Resource Assignment and Flexible GPU Allocations. These features enable precise and efficient distribution of GPU resources across various projects, ensuring each department has the computational power needed without wasting resources.
- **Real-Time Data Processing and Analytics:** The platform's Intelligent Job Queueing & Scheduling and Cloud Bursting capabilities make it ideal for real-time data processing and analytics. Enterprises can dynamically scale their operations to handle large data streams and compute-intensive tasks, ensuring timely and accurate insights.
- **Enterprise-Scale AI Deployment:** The platform is perfect for deploying AI solutions at an enterprise scale with extensive infrastructure support and optimized performance for NVIDIA GPUs. Businesses can seamlessly integrate the platform into their existing infrastructure, utilize public cloud resources when needed, and maintain data security and sovereignty while scaling their AI initiatives.

Get Ready to Transform Your AI/ML Operations

Memory Machine for AI is a revolutionary solution for enterprises to unlock the full potential of AI/ML initiatives while addressing critical challenges. With its powerful features and industry-leading capabilities, MemVerge is the go-to platform for accelerating your AI/ML journey. Empower your IT teams to deliver cost-effective resource management of GPUs and provide an exceptional user experience to AI practitioners.

Request a Demo

About MemVerge

MemVerge is at the forefront of data center infrastructure optimization with its innovative Memory Machine software. Memory Machine for AI allows workloads to surf GPUs in real time to maximize their utilization; Memory Machine for Cloud lets workloads surf cloud computing resources during job runtime to slash costs; and Memory Machine for CXL enables workloads to surf resources across a memory fabric.

Hang loose and take control of your data center infrastructure with Memory Machine software from MemVerge. Learn more at www.memverge.com.

Memory Machine™