Federator.ai GPU Booster®

GPU Performance Maximization with AI-Enhanced Dynamic Allocation for LLMs

- ***Intelligently managing MultiTenant AI and ML tasks by minimizing latency resulting from resource contention***
- ***Efficient allocation of GPU resources through patented multi-layer correlation analytics in Kubernetes***
- ***Significantly improving job processing time and maximizing the total throughput of diverse training and inferencing tasks***
- ***Enhancing GPU resource utilization up to 90% by eliminating idle resources and optimizing placements in a MultiCloud environment***

Challenges

The fast-paced evolution of artificial intelligence demands an unparalleled level of GPU performance optimization, particularly for training Large Language Models (LLMs). However, organizations today are up against considerable challenges in maximizing the effectiveness of these critical resources:

- **Resource Contention and Latency:** MultiTenant AI and ML tasks vie for limited GPU resources, and the resultant latency impacts the speed and efficiency of computational operations, slowing down the pace of innovation.
- **Inefficient Resource Allocation:** Contemporary systems frequently grapple with dynamically allocating GPU resources. In MultiTenant environments, improper scheduling and GPU resource allocation can lead to fragmented and idle compute resources, resulting in significant inefficiency and wastage.
- **Suboptimal Throughput:** The lack of agile infrastructure that responds to the volatile and diverse resource demands of AI workloads results in extended job processing times, which in turn impact overall throughput.

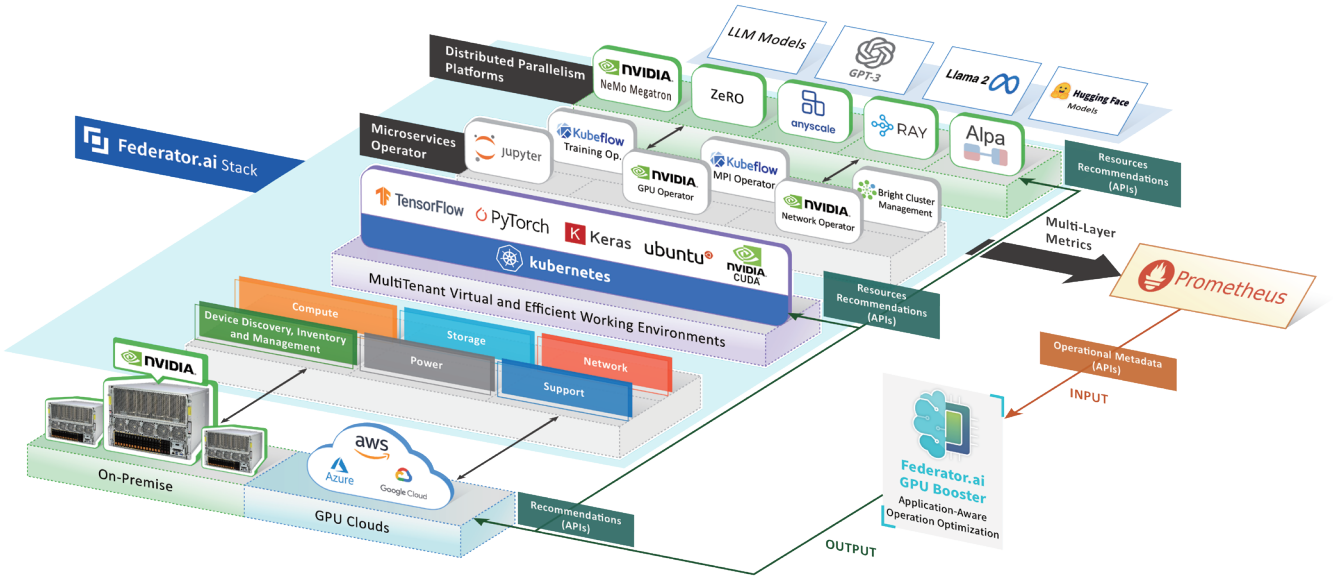
The landscape is further complicated by the pressing issues of supply constraints against the backdrop of an ever-increasing market growth demand for AI compute. In conjunction with escalating prices and operational costs, these factors heighten the urgency for a more strategic approach to GPU utilization. Additionally, Environmental, Social, and Governance (ESG) concerns place additional pressure on organizations to optimize their computational efficiency while adhering to sustainability standards. Amidst these market dynamics, an intelligent solution that can navigate these complexities becomes not just a value-add, but a necessity.

The Solution - Federator.ai GPU Booster

Federator.ai GPU Booster is designed as an innovative solution to optimize GPU performance for LLM training. At its core, it intelligently manages the configuration and allocation of GPU resources, ensuring that computational power aligns seamlessly with the workload demands. This system of smart resource management effectively reduces latency and maximizes the utilization rates of available GPU resources.

Harnessing the power of AI, Federator.ai GPU Booster anticipates and dynamically adapts to fluctuating demands, allowing for predictive scaling and proactive resource adjustments. The result is a streamlined operation that not only elevates performance but also mitigates the common pitfalls associated with the static allocation of GPUs.

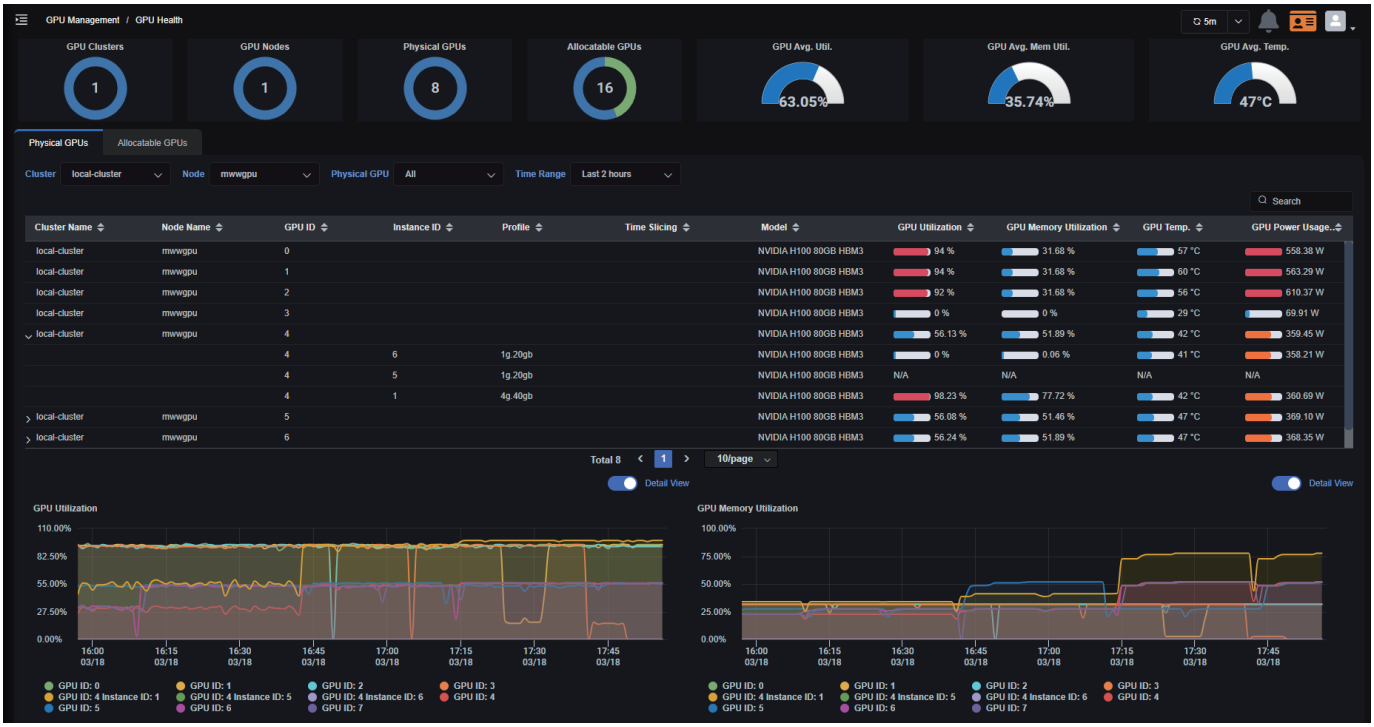
Beyond operational excellence, Federator.ai GPU Booster contributes to cost efficiency, navigating the delicate balance between resource allocation and budget constraints. It ensures that organizations can leverage their AI and ML capabilities to the fullest, propelling them forward in an increasingly competitive landscape while also upholding their commitment to sustainability and environmental responsibility.



Features

- **Optimal GPU Resource Configuration:** Federator.ai GPU Booster provides recommendations for the most appropriate GPU resources and enhanced configurations for GPU servers, ensuring they operate at peak efficiency. This feature guarantees superior performance by managing multiple parallel tasks on GPU servers without compromise.
- **Predictive Analytics for Resource Scaling:** Leveraging AI-driven predictive analytics, Federator.ai GPU Booster forecasts workload requirements, enabling proactive scaling of resources. Such forward-looking resource management effectively reduces contention and maximizes computational throughput.
- **Patented Multi-Layer Correlation Analytics:** Utilizing a sophisticated analytics engine, Federator.ai GPU Booster performs multi-layer correlation analysis. This unique approach facilitates a deeper understanding of resource interactions across the Kubernetes environment, leading to more informed decision-making and resource distribution.
- **Intelligent Task Management in MultiTenant Environments:** In MultiTenant settings, where various AI and ML tasks compete for resources, Federator.ai GPU Booster ensures that computational resources are allocated effectively. It guarantees that tasks are processed efficiently, receiving the computational power they require precisely when it's needed, thus minimizing idle times and enhancing resource utilization.
- **ESG/Green IT Compliant:** Emphasizing sustainability, Federator.ai GPU Booster supports organizations in their Green IT initiatives. Through efficient resource use and optimization, it contributes to reducing the carbon footprint of GPU-powered operations, aligning with Environmental, Social, and Governance (ESG) goals.
- **Seamless Integration with Kubernetes:** Designed to work within the Kubernetes ecosystem, Federator.ai GPU Booster integrates smoothly, enhancing existing workflows without the need for extensive system reconfiguration. This compatibility ensures that organizations can leverage the solution's benefits with minimal setup time and effort.

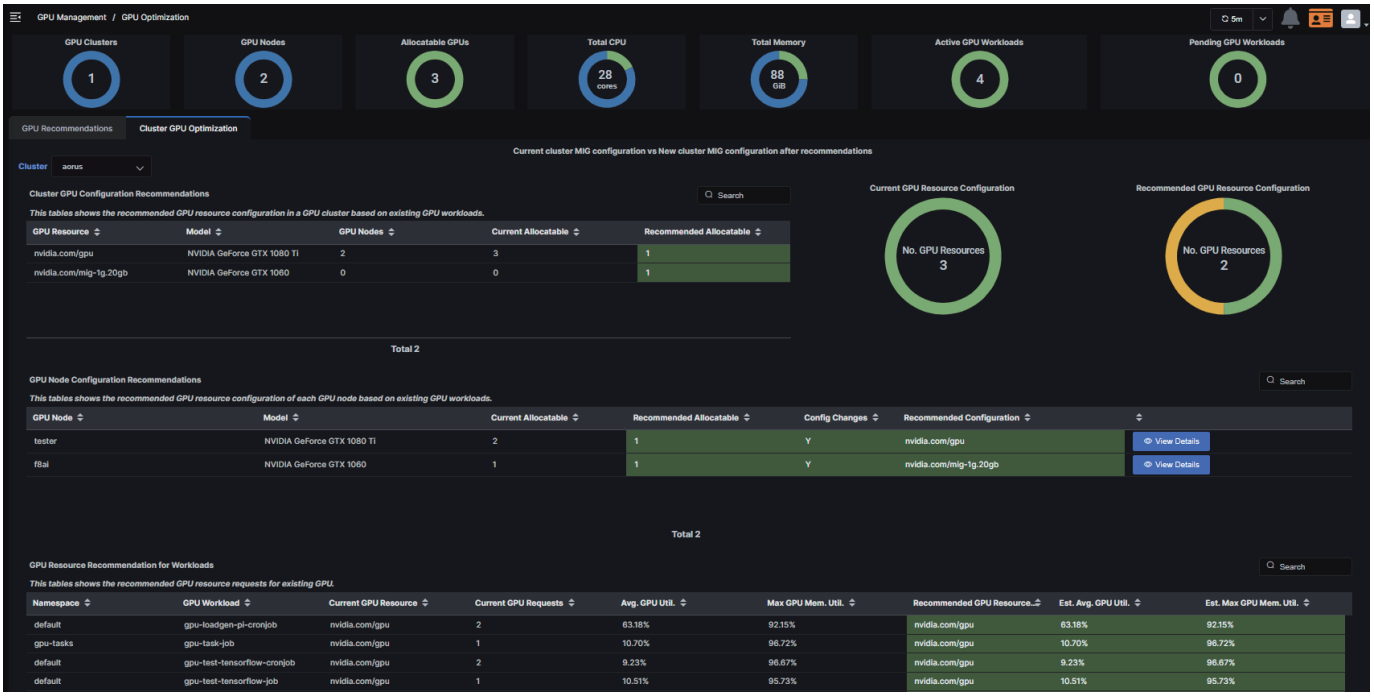
Each feature of Federator.ai GPU Booster converges to create a solution that not only addresses the immediate needs of GPU optimization for LLM training but also paves the way for sustainable, efficient, and cost-effective AI operations.



▲ Overview of GPU and GPU memory utilization

▲ GPU workload management

▲ GPU workload management



▲ GPU recommendations & optimization

Benefits

- **Maximized GPU Utilization:** Federator.ai GPU Booster drastically improves the efficiency of GPU usage, pushing utilization rates up to near-total capacity. This ensures that the vast computational power of GPUs is not left idle but is instead fully leveraged for intensive LLM training tasks, maximizing the return on investment in GPU technology.
- **Reduced Computational Latency:** By intelligently placing resources for AI and ML tasks and optimizing resource utilization, the GPU Booster significantly reduces latency in processing. This leads to faster job completion times, enabling more rapid development and deployment of AI models.
- **Enhanced Cost Efficiency:** The solution's ability to dynamically allocate resources and optimize placements in multi-cloud environments translates into substantial cost savings. Organizations can avoid over-provisioning and underutilization, ensuring that financial resources are allocated as efficiently as computational ones.
- **Support for ESG and Sustainability Goals:** Federator.ai GPU Booster aligns with the growing emphasis on sustainability within the tech industry. By optimizing the power usage of GPU resources, it aids organizations in minimizing their environmental impact, contributing to broader ESG objectives without compromising performance.
- **Streamlined Operations in MultiTenant Environments:** The solution excels in managing the complexities of MultiTenant AI & ML workloads, ensuring equitable and efficient distribution of resources. This streamlines operations, minimizing the administrative overhead associated with managing competing tasks and workloads.
- **Agility in Scaling AI Workloads:** With predictive analytics for resource scaling, Federator.ai GPU Booster provides organizations with the agility needed to scale AI workloads in response to evolving demands. This capability ensures that AI and ML initiatives can grow seamlessly alongside organizational needs.
- **Simplified Kubernetes Integration:** Federator.ai GPU Booster is designed for easy integration with existing Kubernetes infrastructures, allowing organizations to deploy and benefit from its features without significant system overhauls. This ease of integration ensures a smooth adoption process and quick realization of benefits.

In sum, Federator.ai GPU Booster offers a comprehensive suite of benefits, from operational efficiency and cost savings to environmental sustainability and scalability. These advantages make it an indispensable tool for organizations looking to optimize their AI and ML operations in today's dynamic technological landscape.

Feature Details and Specifications

Holistic visibility of GPU utilization	• Detailed view of GPU/ GPU memory utilization of both physical and allocatable GPUs • Track GPU resource usage by all GPU workloads, identify workloads with pending GPU requests • Analyze and track usage and allocation of different GPU resources
AI-based multi-layer workload predictions	• Continuous workload predictions for multi-layer Kubernetes resources: clusters, nodes, namespaces, applications, and controllers • Generate daily, weekly, and monthly workload predictions • Obtain near-immediate prediction results when historical metric data is available
Intelligent recommendations for GPU resource planning	• Provide GPU/ GPU memory resource recommendations for different sizes of MIG, or multi-instance GPU, within a GPU, reducing the wastage resulting from idle and fragmented resources • Recommend and optimize GPU resource configuration of each GPU node based on existing GPU workloads
Intelligent autoscaling for CPU/ memory	• Other than MIG recommendations, the efficiency of CPU and memory in a GPU server can be optimized and autoscaling based on AI-based workload predictions
Intelligent cost management	• Machine-learning-based cost optimization and recommendations for clusters, nodes, namespaces, and applications • Predictive analytics for cost trends of clusters, cluster nodes, namespaces, and applications based on expected workload • Analysis of cost efficiency for better planning and optimization
Alert Management	• Custom-defined monitors on resource shortage or overage based on resource predictions • Auto-alert cancellation when the condition recovers • Automatic email notification of new alerts
Auto-discovery of GPU workloads	• AI/ML workloads utilizing GPU resources can be auto-discovered, eliminating the need for individual setups and accelerating the adoption and training process
Config DB backup and restore	• Auto backup of configuration during the upgrade procedure, and auto restore backup config during downgrade • Backup/restore the configuration DB to/from an external system • Backup DB is encrypted and password protected
Open REST API	• Open REST API for resource predictions and recommendations • Open REST API for cost management and cluster configuration recommendations
One-step installation	• Easy installation with a complete software stack for all necessary Nvidia drivers and tools. • Support installation by Helm charts

Easy-to-use UI	・Support monitoring for multiple AI/ ML workloads ・Visualization of resource usages and predictions for multi-layer of Kubernetes resources ・Support monitoring for multiple clusters
Integration with third-party monitoring services	・Operational data collection through APIs from open-source services, such as Prometheus ・Security through operational metrics and metadata collection alone
Software requirements	・Kubernetes v1.24 or above ・NVIDIA Driver 535.1540 or above ・CUDA Toolkit 12.2 or above ・CPU Operator v23.9.1 ・Data source: Prometheus
Supported Platforms	・Kubernetes ・Red Hat OpenShift ・SUSE/Rancher

ProphetStor Data Services, Inc.

Headquarters
830 Hillview Court, Suite 100
Milpitas, CA 95035
+1 408 508 6255
www.prophetstor.com

Paris Office
2 place de Touraine
78000 Versailles
France
+33 1 7029 0866

Tokyo Office
7F, Wakamatsu Bldg. , 3-3-6
Nihonbashihoncho, Chuo-Ku
Tokyo 103-0023, Japan
+81 3 3249 6378

Taipei Office
16F, No. 182, Sec. 2, Dunhua S. Rd.
Da'an Dist. , Taipei City
Taiwan 10669
+886 2 8219 2814

Taichung Office
13F, No. 219, Minquan Rd.
West Dist. , Taichung City
Taiwan 40341
+886 4 2305 1816

ProphetStor

Visit us at ***www.prophetstor.com*** to find out more,email us at ***info@prophetstor.com*** or contact your local ProphetStor office.