Federator.ai GPU Booster Inference™

Achieving Zero-Downtime, High-Performance LLM Inference Through Autonomous Optimization

- **>60% Throughput Enhancement – Continuous Auto Kaizen™ optimization delivers significant higher user throughput**
- **~25% Response Latency Reduction – Auto Kaizen™ ensures consistently fast responses, reducing latency variability and improving end-user experience**
- **Up to 95%+ Memory Utilization – Memory Walking Technology ensures safe operation at peak efficiency**
- **100+ Server Scalability – Enterprise-ready federation supports seamless expansion across 100+ servers**
- **0% OOM Events – Multi-layer protection guarantees complete elimination of out-of-memory failures**

Challenges

Deploying Large Language Models (LLMs) at enterprise scale pushes GPU clusters to their absolute limits. Traditional static configurations cannot adapt to dynamic workloads, leading to instability, wasted resources, and frequent service disruptions. Key challenges include:

- **The Memory Cliff:** Running massive models like DeepSeek-R1 (671B) on 8×H20 GPUs consumes nearly all available memory, leaving less than 13% for the key-value (KV) cache, activations, and overhead. This razor-thin margin creates a constant risk of out-of-memory (OOM) crashes.
- **Unpredictable Workloads:** LLM inference demands fluctuate by language, context length, and concurrency. Static configurations fail to handle spikes such as product launches or sudden Chinese-language workloads, which require up to 2.5× more memory.
- **Hidden Cost of OOM Events:** Each OOM event triggers complete service outages of 3–5 minutes, plus cache rebuilds and manual DevOps intervention. In typical deployments, 5–10 OOM events per hour can result in up to 66% downtime.
- **Inefficient Resource Utilization:** To avoid OOM failures, enterprises often operate GPUs conservatively at 60–70% utilization, wasting expensive hardware capacity and inflating infrastructure costs.

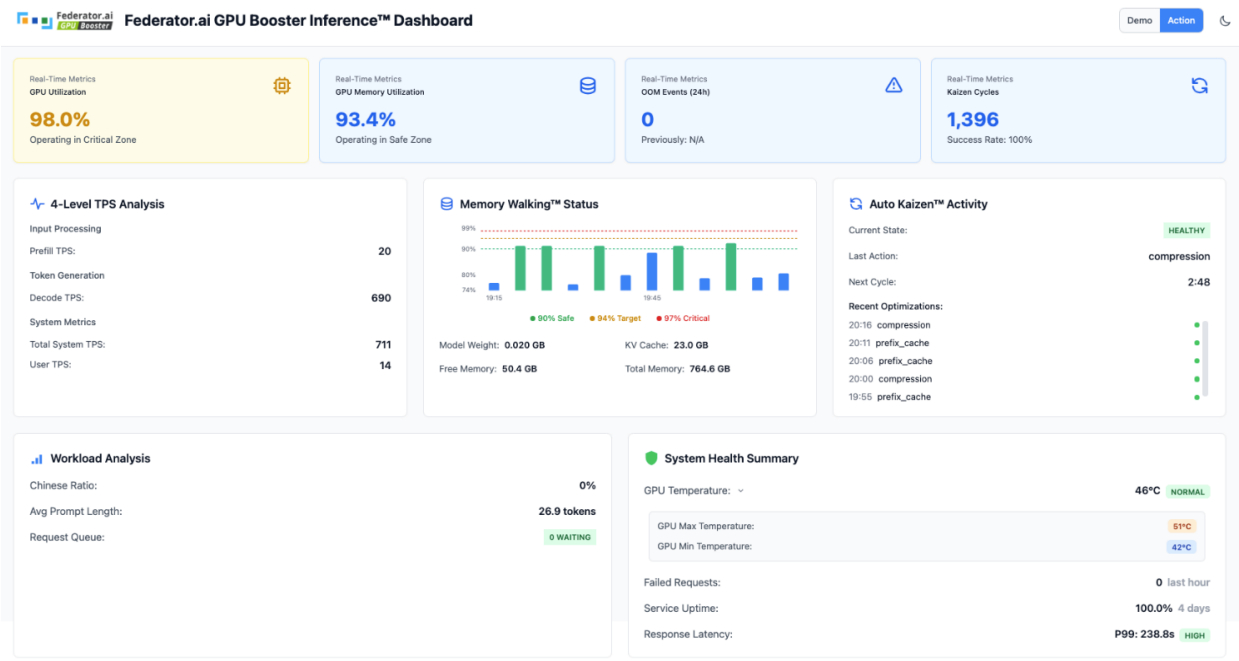
The Solution – Federator.ai GPU Booster Inference™

Federator.ai GPU Booster Inference™ eliminates the risks of static inference configurations by introducing **Auto Kaizen™**, a patent-pending continuous optimization engine. Instead of relying on fixed parameters, the platform adapts dynamically to workload changes, guaranteeing stability, efficiency, and zero downtime.

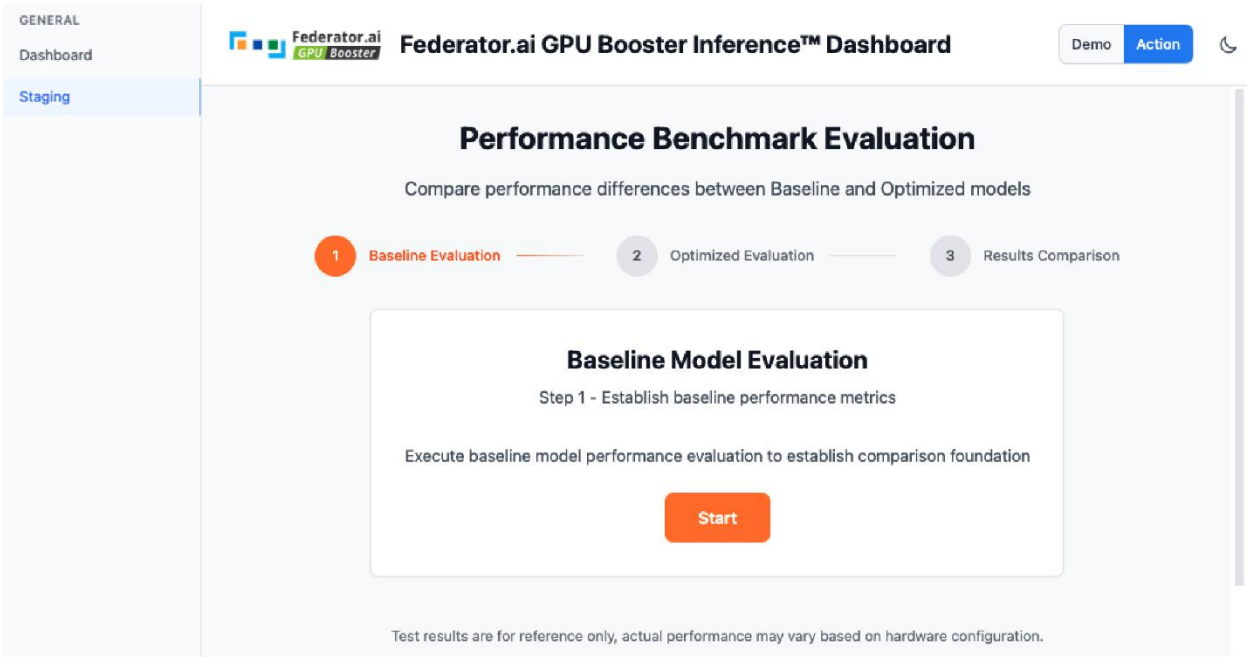
- **Auto Kaizen™ Continuous Optimization:** Continuously executes a PDCA (Plan-Do-Check-Act) cycle to adjust a substantial set of optimization dimensions—including batch size, caching, scheduling, and memory management—without human intervention.
- **Zero OOM Guarantee:** Multi-layer protection combines predictive admission control, machine-learning-based memory forecasting, token budget management, and intelligent request preemption to ensure complete elimination of OOM failures.
- **Memory Walking Technology:** Safely drives GPU utilization up to 95–96%, compared with the conservative 80–85% typical of traditional deployments.
- **Enterprise-Ready Architecture:** Docker-compatible, scalable to 100+ servers, and integrated with security, monitoring, and load balancing frameworks for high availability and seamless deployment.

Breakthrough Features

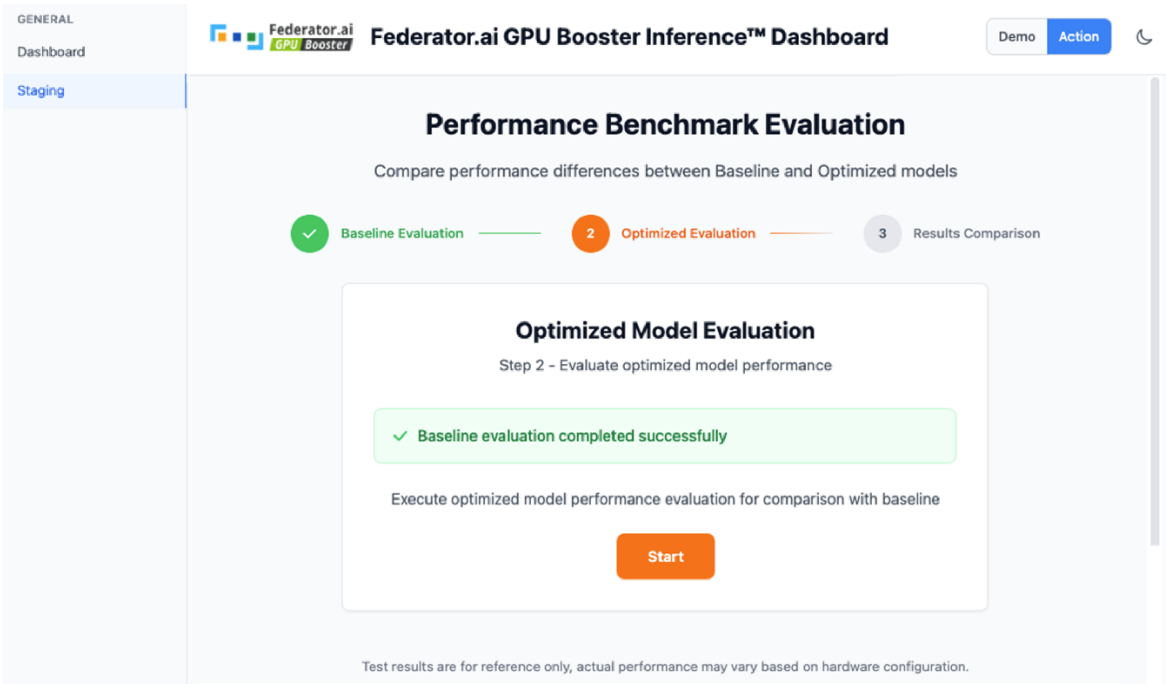
- **Auto Kaizen™ Engine:** Industry-first PDCA-based continuous optimization system that automatically tunes batch sizes, memory allocation, caching strategies, and more—improving performance with zero human intervention.
- **4-Level Observability:** Complete visibility from theoretical hardware limits to actual user experience. Track performance at hardware, model, service, and user levels to identify and eliminate bottlenecks.
- **Memory Walking Technology:** Safely utilize up to 96% through predictive modeling and instant response to pressure.



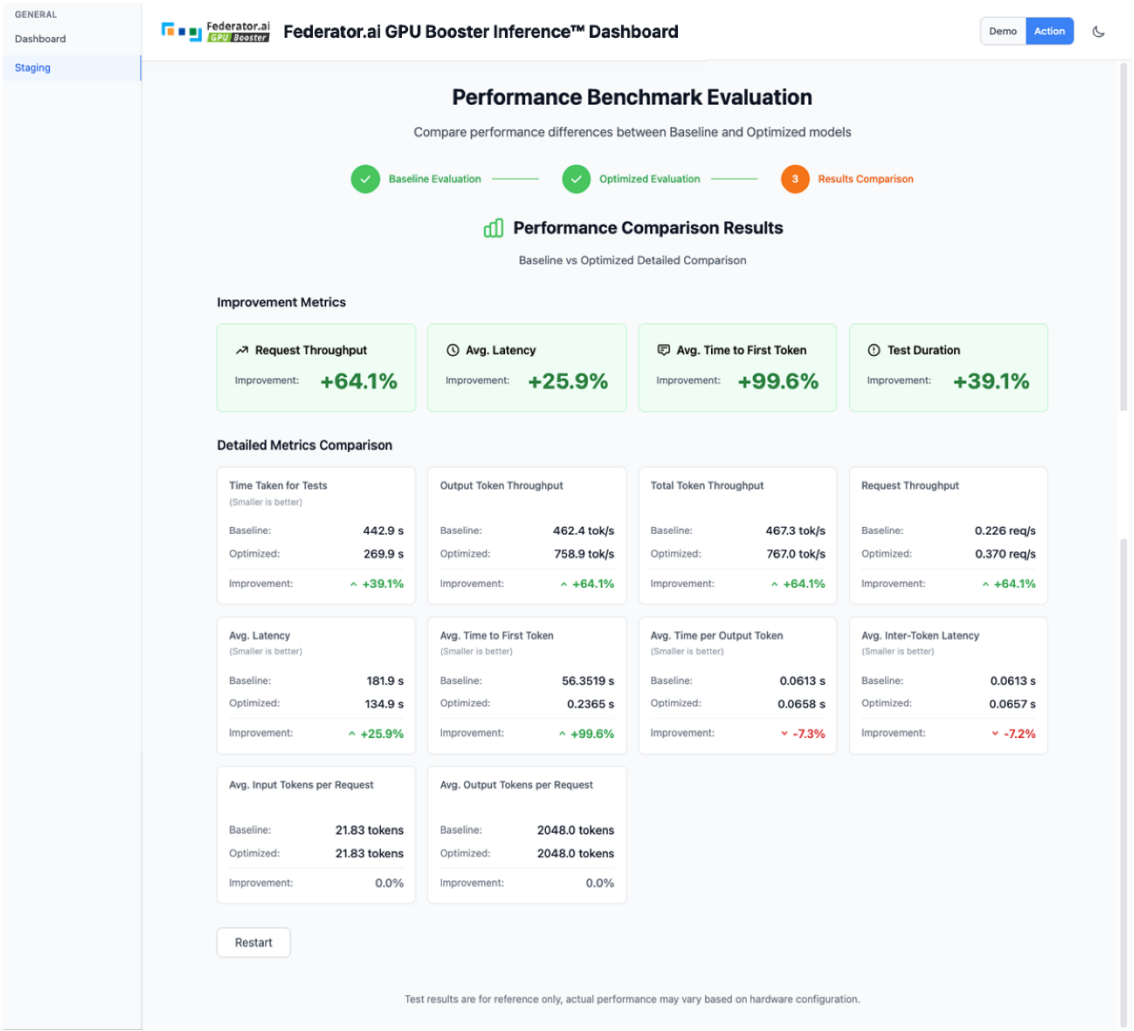
▲ Federator.ai GPU Booster Inference Dashboard



▲ Staging: Baseline Evaluation



▲ Staging: Optimized Evaluation



▲ Staging: Results Comparison

Proven Performance Gains

Metric	Traditional Deployment	With Auto Kaizen™	Improvement
User Throughput	Baseline	Significantly Higher	**+64.1%**
Response Latency	Variable	Consistently Fast	**-25.9%**
Service Disruptions (OOM)	5-10 events/hour	Zero	**Eliminated**
Memory Efficiency	Conservative (~85%)	Optimal (94-96%)	**+12%**
Manual Tuning Required	Daily	Never	**Fully Autonomous**

Enterprise Architecture

Built on industry-standard components with seamless integration:

- **Load Balancing:** Application-aware load balancing for high availability and linear scalability
- **Scalability:** Docker-ready with multi-server federation
- **Security:** TLS 1.3, API key authentication

Recommended Configuration for DeepSeek-R1

Supported Models	DeepSeek-R1 671B (and future releases)
GPU Support	NVIDIA H200, H20 (96GB)
Minimum GPUs	1 server with 8x GPUs (768GB total)
Maximum Scale	Up to 100+ servers (800+ GPUs)
Monitoring Metrics	50+ real-time metrics
API Compatibility	OpenAI-compatible REST API, just like others
Deployment Time	3 days to production

Ideal For

Enterprise AI

Deploy large language models at scale with enterprise-grade reliability

Cost Optimization

Maximize GPU ROI while reducing hardware and operational costs

High-Concurrency Applications

Customer service, financial, and e-commerce workloads requiring high throughput and low latency

Mission-Critical Services

Ensure zero-downtime delivery for applications where reliability cannot be compromised

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.