



Scaling AI Workloads Smarter:

How Avesha's Smart Scaler Delivers Up to 3x Performance Gains over Traditional HPA

The demand for high-performance AI inference and training continues to skyrocket, placing immense pressure on cloud and GPU infrastructure. AI models are getting larger, and workloads are more complex, making efficient resource utilization a critical factor in cost and performance optimization. Enter **Avesha Smart Scaler** — a **reinforcement learning-based scaling solution** that dynamically optimizes GPU/CPU resource allocation for AI workloads, delivering **unprecedented throughput gains and reduced inference latency**.

The Challenge: Scaling AI Workloads Efficiently

Traditional autoscaling mechanisms, such as **horizontal pod autoscaler (HPA)** and vertical scaling, are not optimized for the dynamic and bursty nature of AI workloads. Static provisioning leads to **resource overuse, high operational costs, and unnecessary latency**, while under-provisioning results in **bottlenecks and poor user experience**. Most AI inferencing systems today either **overcommit resources** or **struggle with unpredictable workload spikes**.

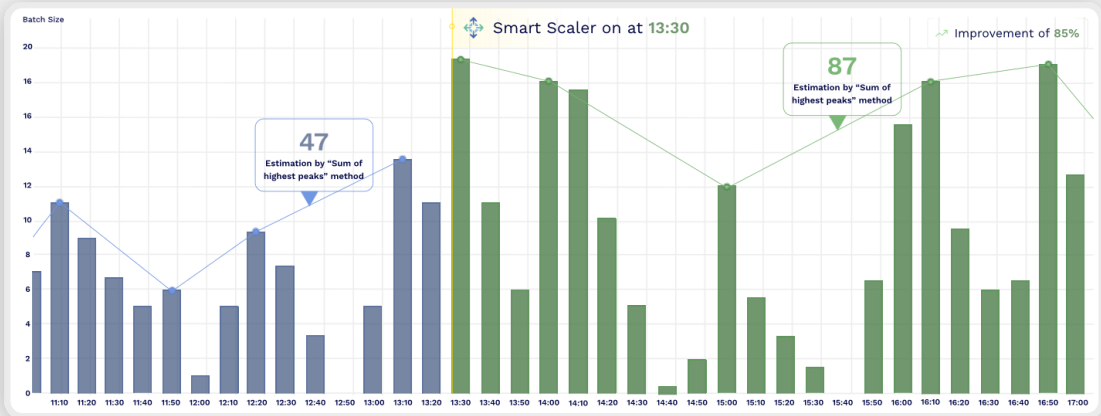
[Avesha's Smart Scaler](#) addresses this by applying reinforcement learning-based intelligence to dynamically scale GPU resources in real-time, ensuring optimal efficiency without waste.

Benchmarking Results: Smart Scaler in Action

Our latest benchmarking results demonstrate how Smart Scaler **significantly enhances AI inference performance**. Running models such as **Llama3-8B** and **DeepSeek-7B**, Smart Scaler achieved:

- **85% larger batch sizes**, increasing inference efficiency with the ability to dynamically adjust GPU resources, particularly in the context of compute-intensive workloads like deep learning, AI training, or high-performance computing (HPC).

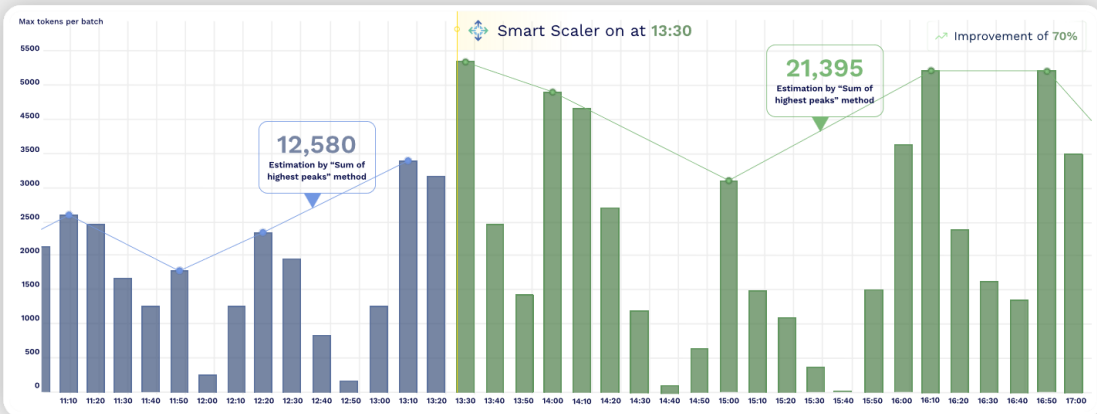
85% larger batch sizes, increasing inference efficiency



Smart Scaler turned on at 13:30pm. Batch Sizes increased to 87 from 47 (Sum of 5 highest peaks)

- **70% more tokens processed per batch**, reducing latency and improving response times with more training speed and throughput. Particularly useful for natural language processing (NLP) tasks, like training large language models (LLMs).

70% more tokens processed per batch, reducing latency and improving response times



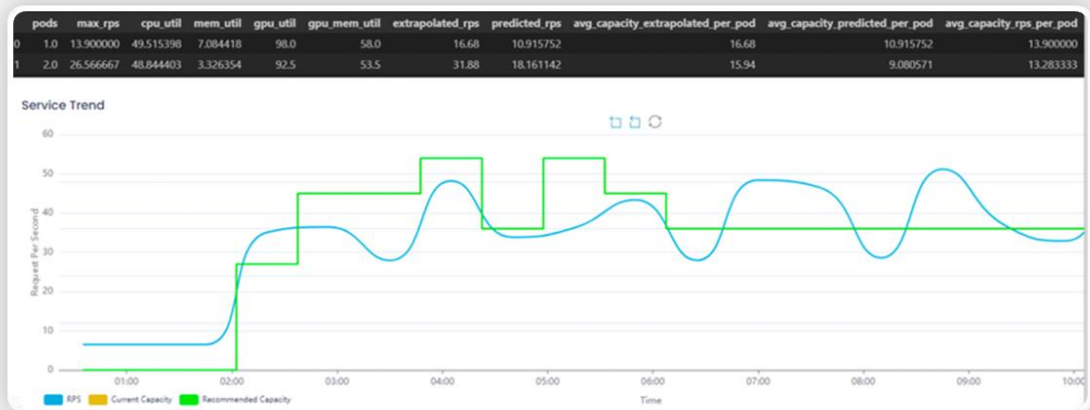
Smart Scaler turned on at 13:30pm. Number of tokens processes increased to 21,395 from 12,580 (Sum of 5 highest peaks)

- **3x higher instantaneous throughput**, during burst processing scenarios (can process three times more data per unit of time compared to a baseline)

$1.85n * 1.7m = 3.1nm$, where n and m are respectively the batch size and tokens per batch processed without Smart Scaling

- **Latency reduction from 8 seconds to 2 seconds** for AI inferencing workloads. That means 4x reduction in response time, meaning the system is now 75% faster per inference. This really matter as it transforms usability, boosts throughput, cuts costs, and can make or break a system's success—especially in real-time or high-stakes scenarios.

Latency reduction from **8 seconds** to **2 seconds** for AI inferencing workloads



This is made possible through intelligent predictive scaling, where Smart Scaler anticipates workload demand and adjusts resource allocation dynamically.

How Smart Scaler Works: Reinforcement Learning for Adaptive Scaling

Unlike traditional autoscalers, Smart Scaler is not rule-based. Instead, it uses reinforcement learning (RL) algorithms to continuously learn and optimize GPU resource utilization. It evaluates factors such as:



- Real-time workload demand:**
 Scaling up resources when an influx of AI inference requests is detected.
- Token throughput and processing efficiency:**
 Ensuring that larger batch sizes are processed efficiently without overloading the infrastructure.
- Historical workload patterns:**
 Predicting future scaling needs based on previous usage data.
- Latency and performance constraints:** Adjusting scaling decisions to minimize response times while maximizing throughput.

Smart Scaler vs. Traditional GPU Scaling Approaches

Feature	Traditional GPU Autoscaling	Avesha Smart Scaler
Scaling Mechanism	Rule-based(CPU memory thresholds)	Reinforcement learning-based adaptive scaling
Performance Gains	Limited	Up to 3x improvement
Latency Reduction	Moderate	75% reduction (8s → 2s)
Cost Efficiency	Overprovisioning wastes GPU hours	Pay-per-work-output optimization
Workload Adaptability	Static scaling logic	Dynamic, real-time optimization

Avesha Smart Scaler Advantages over Traditional GPU Scaling



The Impact: AI Scaling Without Overprovisioning

Smart Scaler eliminates **GPU underutilization** and **improves AI inference efficiency**, and thus helps businesses scale AI workloads **without incurring unnecessary compute costs**. This is particularly beneficial for:

- **LLM providers running inference at scale** (e.g., Hugging Face/TGI (Text Generation Interface), VLLM frameworks)
- **AI-powered SaaS platforms** needing **low-latency model responses**
- **Cloud GPU providers** looking to implement **pay-per-work-output pricing**

Pay-Per-Work-Output Pricing: Smarter Cost Optimization

Traditional GPU usage pricing has been based on time—charging for the number of hours GPUs are allocated, regardless of how much work is actually done. **Smart Scaler enables a shift to pay-per-work-output pricing, where users are charged based on the actual work completed, rather than just the time the GPUs are running.**

The unit of work can vary depending on the type of AI model:

- **Language models (LLMs):**
Charges can be based on **tokens processed** rather than GPU hours.
- **Drug discovery and chemical processing models:**
Workload measurement may be based on **simulations completed, molecular structures analyzed, or compounds screened.**
- **Computer vision and image processing models:**
Pricing can be based on **images or videos processed, rather than raw time usage.**

Thus, by **measuring GPU utilization in terms of actual work output**, Smart Scaler ensures more **efficient and fair pricing models**, reducing wasted GPU spend while delivering superior performance.



The Future of AI Scaling with Smart Scaler

As AI workloads continue to evolve, Smart Scaler is positioned to **redefine how enterprises manage compute resources**, delivering both cost savings and **high-performance AI inferencing**. By leveraging reinforcement learning for real-time optimization, Smart Scaler ensures that businesses can **scale seamlessly across multi-cloud and multi-cluster environments** while maintaining peak efficiency.

With up to **3x better performance and 75% lower inference latency**, Smart Scaler is not just an incremental improvement—it is a **breakthrough in AI scaling technology**.