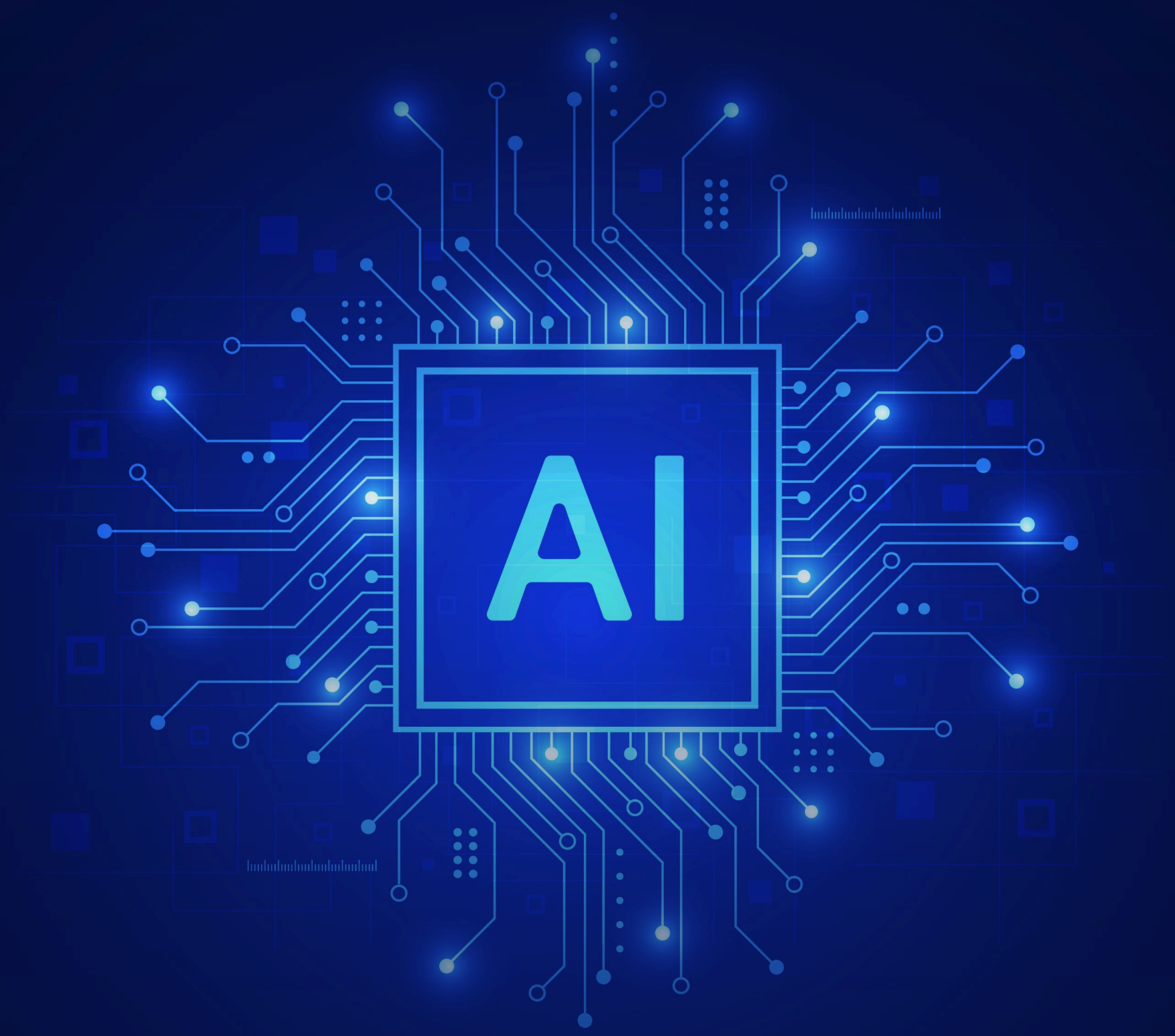
LARSEN & TOUBRO
VYOMA



Cost Optimization for A I Workloads: GPU-as-a-Service vs Owning GPUs

Executive Summary

Artificial intelligence workloads are driving unprecedented demand for high-performance GPU infrastructure. Enterprises are increasingly investing in compute environments to support model training, inference, and large-scale data processing. However, the cost structure of AI infrastructure particularly GPU acquisition and utilization has become a critical concern.

Organizations must balance performance requirements with financial efficiency. Traditional approaches involving outright GPU ownership often result in high upfront capital expenditure (CapEx), underutilized resources, and limited flexibility in scaling.

Larsen & Toubro-Vyoma addresses these challenges through flexible infrastructure models, including GPU-as-a-Service (GPUaaS), GPU cloud, and bare-metal GPU deployments. These offerings enable enterprises to optimize cost structures, improve utilization, and scale AI workloads dynamically without overcommitting capital.

The Cost Challenge in AI Infrastructure

AI and machine learning workloads require specialized hardware, particularly GPUs capable of parallel processing and high-memory operations. As organizations adopt generative AI, HPC, and advanced analytics, infrastructure demand becomes both intensive and variable.

Unlike traditional IT workloads, AI demand is:

Non-linear, with spikes during training cycles
Resource-intensive, requiring high-end GPUs
Rapidly evolving, with newer architectures emerging frequently

Owning GPU infrastructure in such an environment introduces inefficiencies, especially when capacity planning does not align with actual usage patterns. As a result, cost optimization is emerging as a central consideration in AI infrastructure strategy.

Key Customer Challenges

Organizations building AI capabilities commonly face the following cost and infrastructure challenges:

1. Infrastructure Constraints

Setting up GPU infrastructure requires significant planning across compute, networking, cooling, and power, making deployments complex and time-intensive.

2. High Capital Expenditure (CapEx)

Upfront investment in GPU hardware- especially high-performance units creates financial pressure and long payback periods.

3. Underutilized and Idle GPU Resources

AI workloads are often intermittent. During non-training periods, expensive GPU resources remain idle, leading to inefficient utilization.

4. Unpredictable Workload Demand

Fluctuating demand across AI use cases makes it difficult to accurately size infrastructure, resulting in either overprovisioning or performance bottlenecks.

4. Scalability Limitations

Expanding owned infrastructure requires additional procurement cycles, delaying the ability to respond to changing workload requirements.

Larsen & Toubro-Vyoma's Approach to Cost Optimization

Larsen & Toubro-Vyoma enables organizations to optimize AI infrastructure costs by offering multiple GPU deployment models tailored to different workload needs.

1. GPU-as-a-Service (GPUaaS)

Provides on-demand access to GPU resources, allowing organizations to pay based on usage rather than ownership.

2. GPU Cloud

Offers scalable, virtualized GPU environments for AI/ML workloads, supporting dynamic scaling and rapid deployment.

3. GPU Bare Metal

Enables dedicated GPU infrastructure for performance-intensive workloads requiring full hardware control.

4. Advanced GPU Infrastructure

Deployment of high-performance GPUs, including NVIDIA H200 systems, supports next-generation AI workloads with improved efficiency and compute density.

These models allow organizations to align infrastructure strategy with workload patterns, balancing performance, control, and cost efficiency.

Cost Efficiency and Utilization Optimization

Optimizing GPU infrastructure costs requires aligning resource consumption with actual workload demand.

1. CapEx vs OpEx Flexibility

- ***Traditional ownership models require large upfront investment***
- ***GPUaaS and cloud-based models convert infrastructure costs into operational expenditure (OpEx)***
- ***This improves cash flow and reduces financial risk associated with technology obsolescence***

2. Improved Resource Utilization

- ***On-demand provisioning reduces idle GPU time***
- ***Shared infrastructure models improve overall utilization rates***
- ***Workloads can scale up during peak demand and scale down during idle periods***

3. Elastic Infrastructure Scaling

- ***Dynamic allocation ensures resources are available when needed***
- ***Eliminates overprovisioning while maintaining performance***

4. Workload-Specific Optimization

- ***GenAI, ML, and HPC workloads can be mapped to appropriate infrastructure models***
- ***High-performance tasks can leverage dedicated GPUs, while variable workloads use shared environments***

These capabilities enable organizations to achieve more efficient cost structures without compromising performance.

Operational Impact and Cost Outcomes

Organizations adopting flexible GPU infrastructure models can achieve measurable improvements:

1) Reduction in idle GPU costs, through on-demand provisioning and shared usage models
2) Improved utilization rates, aligning infrastructure consumption with actual workload demand
3) Lower total cost of ownership (TCO) compared to fully owned GPU environments
4) Predictable cost structures, supporting better financial planning
5) Faster deployment cycles, reducing delays associated with hardware procurement

Real-world benchmarks across GenAI, machine learning, and HPC workloads indicate that flexible infrastructure models significantly improve cost efficiency when compared to static, owned GPU deployments.

Future Roadmap

Larsen & Toubro-Vyoma continues to invest in next-generation infrastructure to support evolving AI workloads.

1. Next-Generation GPU Readiness

Infrastructure is being designed to support upcoming architectures, including NVIDIA Vera Rubin and GB300 systems.

2. Mahape Data Center Expansion

The development of advanced infrastructure in Mahape reflects a focus on supporting high-density GPU environments with scalability and efficiency.

These initiatives ensure long-term readiness for increasingly compute-intensive AI workloads.

Conclusion

AI workloads are fundamentally changing how enterprises approach infrastructure investment. Traditional GPU ownership models, while offering control, often result in inefficiencies due to high upfront costs and underutilized resources.

By adopting flexible models such as GPU-as-a-Service, GPU cloud, and bare-metal deployments, organizations can align infrastructure costs with actual usage, improve utilization, and respond more effectively to changing workload demands.

Larsen & Toubro-Vyoma's approach enables enterprises to build cost-efficient, scalable AI infrastructure that supports both current requirements and future growth, without the constraints of rigid ownership models.

LARSEN & TOUBRO

vyoma

Start Your Journey With Us

Contact Us