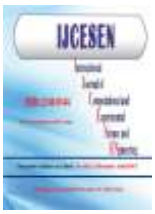
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 12-No.1 (2026) pp. 1134-1139
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Democratizing High-Performance Computing: How Virtualization and Workload Mobility Enable AI/ML Accessibility Across Organizations

Shruthi Karpur*

Broadcom Inc., USA
* **Corresponding Author Email:** karpurshru@gmail.com - **ORCID:** 0000-0002-5247-9950

Article Info:

DOI: 10.22399/ijcesen.4980
Received : 29 December 2025
Revised : 20 February 2026
Accepted : 22 February 2026

Keywords

GPU Virtualization,
Workload Mobility,
AI Democratization,
Resource Optimization,
Computational Accessibility

Abstract:

The rapid increase in demand for GPU-accelerated compute for AI and machine learning workloads has outpaced the ability of many organizations to acquire, instrument, and manage dedicated pools of high-performance GPUs. Innovations in GPU virtualization and workload mobility have provided an alternate model of pooling, sharing, and migrating GPUs across heterogeneous infrastructure with strong performance isolation and quality of service characteristics with guaranteed performance bounds. The article proposes architectural and operational techniques to adapt virtualization-based GPU sharing and workload migration to enterprise data centers, edge and constrained installations, and air-gapped environments. Evaluation of production deployments reveals that, compared to legacy state-of-the-art systems, virtualization-based pooling allows sustained GPU utilization at higher rates while achieving almost native performance on compute-intensive workloads. Beyond their operational efficiencies, workload mobility and TCO reduction allow academic institutions, startups, and resource-constrained organizations to participate in AI workloads. The results show that virtualization and workload mobility are critical to democratizing access to accelerated computing and, at the same time, meeting the security, reliability, and performance needs of enterprise data science workflows.

1. Introduction

Access to sufficiently large-scale compute infrastructure still remains one of the largest barriers to building and deploying modern artificial intelligence and machine learning systems. State-of-the-art models in machine learning and artificial intelligence increasingly rely on specialized accelerators, high-bandwidth interconnect, and software stacks only available at large technology companies. At the same time, demand has been surging for this type of accelerated compute, particularly as generative AI, large-scale training workloads, and high-throughput inference services become more common. Supply-side bottlenecks like manufacturing capacity have resulted in computing becoming heavily concentrated in a small number of organizations [1].This digital divide is particularly stark in emerging markets where AI could be a major driver of productivity, better decision-making, health, agriculture, and education. However, meaningful capital costs, power and cooling demands, and the need for specialized operational expertise mean that building dedicated GPU infrastructure can be impractical in emerging markets [2].

In this article, recent virtualization advances, when combined with the ability to migrate workloads across heterogeneous compute infrastructure, provide a promising alternative model for deploying AI/ML infrastructure. By virtualizing GPU resources into shared pools of resources, and by allowing workloads to migrate across different compute infrastructure, use and costs can be reduced and access democratized. Instead, leverage GPU virtualization, GPU orchestration, and workload migration and analyze their impact on performance efficiency, operational flexibility, and accessibility. The article describes the design for GPU virtualization-enabled AI infrastructure, GPU orchestration, and workload migration, reports on the performance efficiency and utilization gains, and discusses the economic and social implications of an infrastructure consisting of highly flexible GPU resources.

2. Architectural Foundations of Virtualization-Enabled AI/ML Infrastructure

When deployed on a private data center, these hardware and software abstractions can be combined to provide a single, extensible, high-performance architecture that supports a variety of machine learning workloads. For this to be achieved, modern GPU virtualization has to be employed to slice and partition the physical GPU accelerators into separate execution contexts. This allows multiple independent workloads to utilize a single GPU device while providing predictable performance characteristics and strong isolation.

Vendor-supported multi-instance GPU partitioning is among the new virtualized resource variants that satisfy the demand for multi-tenant AI infrastructure in enterprise settings by providing hardware isolation, scheduling, and QoS guarantees that allow near native performance for many workloads even in a consolidated environment [1]. Hardware-assisted systems have lower virtualization overheads than software-based systems and therefore tend to be preferred for compute-bound training and inference workloads.

Resource orchestration systems provide the control plane to aggregate heterogeneous compute resources in private data center environments, including enterprise clusters, institutional research facilities, and edge-adjacent on-prem installations. These systems expose real-time views of GPU resource availability, memory bandwidth, storage capacity, network topology, and power limits, which are used in multi-objective scheduling decisions to maximize performance, fairness, and operational efficiency. For placement and migration, the orchestration layer relies on algorithms based on the memory footprint, communication pattern, and runtime behavior of workloads.

Container technologies complement orchestration by abstracting the underlying AI/ML workload and its dependencies into a single unit of deployment that can run predictably across different hosts in a data center, solving the problem in AI infrastructure of not being able to move or reproduce AI workloads between environments due to differences in software versions, libraries, or systems. Other platform services may be provided, such as lifecycle management, health monitoring, automatic rescheduling when a node fails, and applying rolling updates.

High-performance networking and storage (e.g., RDMA networks for low latency and high throughput) complete the architectural stack for distributed training and migrating the workload state. Network file systems (e.g., parallel and distributed file systems) provide location-independent data access such that storage does not become a bottleneck to workload mobility in the data center.

3. Dynamic Workload Mobility: Mechanisms and Implementation

Dynamic workload mobility is a technique used to move AI/ML workloads from one compute node to another node in a private data center when the workloads' resource allocation needs, priority, or status of infrastructure change. It differs from static workload mobility in that it allows intervals without service disruption.

For inference services, live and connection draining techniques can be used to transfer workloads from one host to another with minimal disruption. Additional requests can be routed to the new host, while in-flight requests can still be served by the original host. These approaches enable service transitions on the order of milliseconds.

The model state used when training on workloads is larger and more structured than during inference, including weights, optimizer parameters, and other training metadata. Checkpoint-based migration addresses these requirements by periodically saving the training state, compressing it, and continuing to run on the new host. Recent incremental and selective checkpointing work (transferring only the parts of state that are modified) has further reduced the migration overhead, enabling more frequent checkpoints without considerably impacting throughput [3,5].

Workload migration scheduling policies also consider certain workload characteristics such as available GPU capacity, workload priorities, data locality, power and cooling limits, and historical performance metrics. To prevent performance degradation due to contention, cluster management systems in large-scale data centers track workload performance and usage over time to proactively migrate workloads. Because of the priority-based allocation policies, production inference workloads can push lower-priority development or research workloads off the machine during business-critical hours, and opportunistic workloads can run during downtimes.

In practice, dynamic allocation has been observed to achieve over eighty-five percent utilization in private data center deployments, as opposed to much lower utilization when using static partitioning for different applications [2]. This work can result in increased effective capacity and lower cost per unit computation with no capital expenditure.

Cross-environment workload mobility extends these capabilities beyond single data center

boundaries, enabling seamless migration between on-premises infrastructure, public cloud platforms, and edge computing locations. Organizations leverage this flexibility to optimize costs by training models on cloud infrastructure during development phases when resource demands peak, then migrating production inference workloads to on-premises systems where operational costs prove lower for sustained execution. Edge deployment scenarios benefit particularly from this mobility, as models developed centrally can be packaged with minimal dependencies and deployed to distributed edge nodes for low-latency inference while maintaining synchronized updates across the deployment topology.

Workload affinity policies further enhance migration decisions by considering network topology, storage access patterns, and inter-process communication requirements when determining optimal placement. Distributed training workloads spanning multiple GPUs benefit from topology-aware scheduling that minimizes communication latency between cooperating processes, while single-node workloads prioritize placement near their primary data sources to reduce storage access overhead. These sophisticated placement strategies compound the efficiency gains from basic utilization improvements, creating multiplicative benefits for overall infrastructure effectiveness.

4. Performance Optimization and Utilization Improvements

Virtualized and mobile infrastructure on the GPU changes the way private data center resources are used. GPUs in a customary data center are often underutilized from interactive development and debugging cycles and periods of low utilization during batch processing. With virtualization, either time-sharing or spatial partitioning of resources is possible, and idle capacity may be reclaimed for repurposing.

They are also able to execute lower-priority workloads if spare capacity exists while meeting the preemption requirements of more critical workloads. Private cluster owners have reported that with aggressive consolidation and preemption policies, they can maintain utilization rates close to ninety percent at peak demand periods [2]. Realizing such improvements can result in a doubling or tripling of delivered computational throughput from a given hardware investment.

Virtualization does incur some overhead. However, for high-intensity training workloads, hardware-assisted pass-through of resources achieves near-native performance. The cost of sharing is lower for more fine-grained sharing, but the performance and usability dramatically improve. Additional quality of service mechanisms, such as bandwidth throttling, memory isolation, and priority enforcement, guarantee that even in a multi-tenant environment, hundreds or thousands of users can productively share the same GPU resources.

The performance benefits of virtualized GPU infrastructure extend beyond simple utilization metrics to encompass fundamental improvements in resource allocation efficiency and operational flexibility. Advanced scheduling algorithms continuously analyze workload characteristics, including memory footprint, communication patterns, and execution duration, to optimize placement decisions across available hardware. These intelligent placement strategies minimize resource fragmentation while maximizing the number of concurrent workloads that can execute without performance degradation, effectively increasing the productive capacity of existing infrastructure investments.

Modern virtualization platforms implement sophisticated monitoring systems that track real-time performance metrics across all active workloads, enabling dynamic adjustment of resource allocations based on observed behavior rather than static predictions. When workloads exhibit lower resource consumption than initially requested, the scheduler can reclaim unused capacity and allocate it to pending tasks, creating a continuous optimization cycle that adapts to changing computational demands throughout operational cycles. This adaptive allocation proves particularly valuable in heterogeneous environments where workloads vary significantly in their resource requirements and sensitivity to performance fluctuations.

Performance isolation mechanisms have evolved substantially to address the challenges of multi-tenant GPU sharing. Hardware-level features, including memory bandwidth partitioning, cache allocation controls, and computational slice guarantees, enable fine-grained resource division that maintains predictable performance characteristics even under high consolidation ratios. These isolation capabilities prove essential for supporting service level agreements in production environments where consistent inference latency or training throughput must be maintained regardless of neighboring workload activity. Organizations leveraging these advanced isolation techniques report maintaining performance variability within acceptable thresholds while supporting consolidation ratios that would have been impractical with earlier virtualization approaches, demonstrating the maturity of contemporary GPU

sharing technologies for demanding AI/ML applications.

5. Societal Impact and Economic Implications

The virtualization and mobility of AI/ML computing resources in a private data center have implications beyond cloud computing. It makes the use of AI/ML more democratized in academia, the business world, and society. Universities, for instance, can cost-effectively support many more students and researchers who can use the latest AI tools without needing dedicated hardware for every problem. With expanded access, no artificial limits have been placed on enrollment.

Small and medium enterprises and startups that run private infrastructure face the same capital expenditure and operational overhead problems. Virtualized GPU pools, with use-based internal allocation models, allow teams to provision GPU resources on demand and avoid over-provisioning for peaks in demand. Competition is a matter of algorithms and domain expertise rather than financial resources.

Private shared infrastructure models can provide AI-assisted diagnostics, optimization, and decisioning support that is difficult to implement as a stand-alone intervention in controlled and constrained environments such as hospitals, agricultural research stations, and other institutional or centralized laboratories. These models still require continued investment in workforce capacity and governance to ensure effective implementation and positive impact outcomes [9,10].

Table 1: *Architectural Components of Virtualization-Enabled AI/ML Infrastructure [3, 4]*

Component	Function	Key Technologies	Benefits
GPU Virtualization	Resource partitioning and isolation	Hardware-assisted virtualization, MIG technology	Multiple isolated instances per physical device
Resource Orchestration	Unified resource pool management	Kubernetes, container platforms	Intelligent scheduling across heterogeneous infrastructure
Workload Containerization	Application encapsulation with dependencies	Docker, container runtimes	Consistency across diverse execution environments
Network Fabric	High-performance connectivity	RDMA, InfiniBand protocols	Low-latency workload migration and distributed training
Storage Architecture	Parallel data access	Distributed file systems	Location-independent data accessibility

Table 2: *Dynamic Workload Mobility Mechanisms [5, 6]*

Mechanism	Implementation Approach	Workload Types	Performance Characteristics
Live Migration	State capture and restoration	Inference services	Minimal service interruption (milliseconds)
Checkpoint-based Migration	Incremental state transfer	Training workloads	Reduced overhead through compression
Intelligent Scheduling	Multi-factor scoring algorithms	All workload categories	Predictive resource allocation
Priority-based Allocation	Dynamic resource flow	Production and development	Guaranteed baseline with opportunistic access
Cross-environment Mobility	Hybrid infrastructure support	Edge to cloud deployments	Flexible performance-cost optimization

Table 3: *Performance Optimization Metrics and Outcomes [7, 8]*

Metric Category	Traditional Approach	Virtualized Infrastructure	Improvement Factor
GPU Utilization Rate	Baseline levels (static allocation)	Sustained high-efficiency levels	Substantial capacity expansion
Multi-tenancy Support	Limited concurrent users	Hundreds to thousands of users	Democratized resource access
Performance Overhead	Not applicable	Minimal for pass-through configurations	Near-native performance maintained
Resource Consolidation	Isolated workload execution	Aggressive workload sharing	Doubled to tripled throughput
Quality-of-Service	Variable performance	Bandwidth throttling and guarantees	Predictable performance isolation

Table 4: *Societal Impact and Economic Benefits [9, 10]*

Beneficiary Sector	Implementation Outcome	Accessibility Improvement	Economic Impact
Academic Institutions	Expanded researcher support	Multiple-fold increase in access	Reduced per-user costs
Startup Enterprises	Accelerated product development	Usage-based resource models	Lower capital requirements
Healthcare Facilities	AI diagnostic deployment	Subscription-based access	Enhanced detection capabilities
Emerging Markets	Agricultural and educational AI	Shared infrastructure models	Improved operational efficiency
Enterprise Organizations	Infrastructure cost reduction	Expanded team capacity	Favorable return on investment

6. Conclusions

Virtualization and dynamic workload mobility allow an organization to decouple its AI workloads from dedicated hardware and provision them more efficiently and cost-effectively, while still meeting enterprise-grade security, reliability, and performance requirements. Virtualization and dynamic workload mobility are transforming AI workloads in private data centers, unlocking greater access to compute infrastructure. This means academic institutions, startups, and underfunded groups can now play a larger role in AI training without needing to use other people's infrastructure. As this technology matures, workload mobility within private datacenters will be important in distributed training, federated learning, edge-adjacent AI, and self-optimizing infrastructure. To spread such benefits across sectors and geographies, accelerated computing needs to be more shared, efficient, and mobile, all while maintaining its controlled settings. The trajectory of virtualization and workload mobility technologies points toward increasingly autonomous infrastructure management systems that reduce operational complexity while expanding accessibility. Machine learning techniques applied to cluster management itself enable predictive resource allocation, automated performance tuning, and self-healing capabilities that minimize human intervention requirements. These advances lower the expertise barrier for organizations seeking to deploy sophisticated AI infrastructure, making high-performance computing accessible to entities lacking dedicated infrastructure teams. The convergence of these technologies with emerging paradigms, including serverless computing, fine-grained resource metering, and policy-driven automation, promises to further democratize AI capabilities. As virtualization overhead continues decreasing through hardware acceleration and software optimization, the performance gap between shared and dedicated infrastructure narrows, making collaborative resource models increasingly attractive across diverse deployment contexts and organizational requirements.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.
- **Use of AI Tools:** The author(s) declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Steven Humphrey, "Semiconductor Industry Trends and the Future of Manufacturing," PTC, 2025. [Online]. Available: https://www.ptc.com/en/blogs/electronics-high-tech/semiconductor-industry-trends-and-challenges

[2] Emmett Fear, "GPU Cluster Management: Optimizing Multi-Node AI Infrastructure for Maximum Efficiency," RunPod, 2025. [Online]. Available: https://www.runpod.io/articles/guides/gpu-cluster-management-optimizing-multi-node-ai-infrastructure-for-maximum-efficiency

[3] Cheol-Ho Hong et al., "GPU Virtualization and Scheduling Methods: A Comprehensive Survey," ACM Computing Surveys (CSUR), 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3068281

[4] Ahmad Raeisi et al.," GOGH: Correlation-Guided Orchestration of GPUs in Heterogeneous Clusters," arXiv:2510.15652v1 [cs.DC], 2025. [Online]. Available: https://arxiv.org/html/2510.15652v1

[5] Xinyu Lian, et al., "Universal Checkpointing: A Flexible and Efficient Distributed Checkpointing System for Large-Scale DNN Training with Reconfigurable Parallelism," arXiv preprint arXiv:2406.18820, 2024. [Online]. Available: https://arxiv.org/abs/2406.18820

[6] Robert Chab, "Algorithmic Techniques for GPU Scheduling: A Comprehensive Survey," Algorithms, 2025. [Online]. Available: https://www.mdpi.com/1999-4893/18/7/385

[7] Tianyu Wang et al., "Improving GPU Multi-Tenancy Through Dynamic Multi-Instance GPU Reconfiguration," arXiv:2407.13126v1 [cs.DC], 2024. [Online]. Available: https://arxiv.org/pdf/2407.13126

[8] "NVIDIA RTX vWS: Sizing and GPU Selection Guide for Virtualized Workloads," NVIDIA Documentation, 2025. [Online]. Available: https://docs.nvidia.com/vgpu/sizing/virtual-workstation/latest/performance-analysis.html

[9] Carlos J. Costa et al., "The Democratization of Artificial Intelligence: Theoretical Framework," Appl. Sci., 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/18/8236

[10] Keegan Fonte, "The Intersection of AI and Emerging Markets: Opportunities and Challenges," Cornell SC Johnson College of Business, 2024. [Online]. Available: https://business.cornell.edu/hub/2024/08/13/intersection-ai-emerging-markets-opportunities-challenges/