International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| AECTIC2025-Conference Proceeding
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.AECTIC-104

Original Article

GPU Fleet FinOps: Scheduling, Right-Sizing, and Cost Governance for DGX, MIG, and Preemptible Capacity

Santosh Pashikanti
Lead Cloud Architect, Independent Researcher, USA.

Abstract - *As organizations increasingly scale Generative AI (GenAI) and Large Language Model (LLM) workloads, GPU-accelerated computing has become the dominant line item in cloud expenditure [1],.[2] This paper presents a "GPU Fleet FinOps" blueprint, a unified operating model for the financial and operational optimization of large-scale GPU fleets, including NVIDIA DGX systems, MIG-partitioned GPUs, and preemptible (spot) capacity. We identify critical, unaddressed challenges: chronic low utilization of premium hardware [3], [4], [5], "capacity island" fragmentation from Multi-Instance GPU (MIG) [6], [7], [8], [9], the high failure rate of workloads on preemptible instances [10], [11], [12], and a lack of financial accountability. We propose an integrated solution built on three pillars: a FinOps-aware GPU scheduling layer [13], policy-driven right-sizing with quota management [14], [15], and robust, interruption-aware job design [16],.[17] This framework connects low-level scheduling and governance decisions to business-centric unit economics, such as "cost per training run" and "cost per 1k inferences" [18], [19], providing a practical architecture for aligning high-performance GPU investments with measurable business value.*

Keywords - *GPU Fleet Finops, Cost Governance, GPU Scheduling, Right-Sizing, NVIDIA DGX, Multi-Instance GPU (MIG), Preemptible Capacity, Kubernetes, Unit Economics.*

1. Introduction

The proliferation of Generative AI (GenAI) and Large Language Models (LLMs) has transitioned GPU- accelerated computing from a specialized tool to a primary driver of business value. With this shift, GPU costs have begun to dominate AI spend, often becoming the single largest line item in cloud and IT budgets [1],.[2] As organizations scale these workloads, they face a critical realization: without a systematic framework for financial and operational governance, this level of expenditure quickly becomes unsustainable.

This topic is critical because it addresses a fundamental misalignment in the industry. While enterprises want predictable price-performance and clear unit economics (e.g., cost per training run), technical operations remain focused on raw performance.[18]

This disconnect is exacerbated by several key challenges:

- Capacity Scarcity and Fragmentation: GPU capacity is scarce. Technologies like NVIDIA's Multi-Instance GPU (MIG), designed to help, often create "capacity islands" that lead to stranded resources and over-provisioning when managed improperly [6], [7], [8],.[9]
- Lack of Governance: Most organizations lack clear policies for who can use premium GPUs, when to use spot versus on-demand capacity, or how to enforce budgets.
- The Business Need: Ultimately, making GPU fleets more efficient, governed, and affordable is what enables wider AI adoption, allowing more teams to experiment and move workloads to production.

This paper provides a blueprint for GPU FinOps, arguing that GPU management must be treated as a unified architecture, scheduling, and financial governance problem. We will detail the key challenges observed in the field and propose an integrated operating model that connects low-level infrastructure decisions to high-level business KPIs.

2. Key Challenges in GPU Fleet Management

We have observed a set of recurring problems that prevent organizations from achieving a positive return on their significant GPU investments.

2.1. Low and Uneven GPU Utilization

The most significant issue is the chronic under-utilization of expensive, premium hardware. In a typical scenario, a global bank invested in several NVIDIA DGX pods for fraud detection and GenAI pilots [20],.[21] Project teams requested full, dedicated GPUs "just in case," fearing resource contention. As a result, most jobs used only 20-30% of the GPU's memory or Streaming Multiprocessors (SMs) while blocking the entire card. Monthly utilization reports showed an average usage below 35% [3], [4], [5], yet there was no mechanism to enforce right-sizing or safely share these premium assets.

2.2. MIG-Induced Fragmentation

NVIDIA MIG is a powerful technology that partitions a single GPU into multiple, isolated instances, which is powerful for multi-tenancy.[9] However, in practice, it creates "capacity islands" and significant resource fragmentation [6], [7],.[8] We observed an enterprise AI platform that enabled MIG on its A100s to serve small inference workloads. Teams selected arbitrary MIG profiles (e.g., 1g.5gb, 2g.10gb), and the scheduler had no packing policy. This resulted in clusters frequently showing 60-70% GPU memory allocated but only a few of the seven MIG slices usable for new jobs, forcing the company to buy more GPUs even though real utilization was low.

2.3. Unpredictable Preemptible (Spot) GPUs

Teams are attracted to the lower price of preemptible/spot GPUs but lack the patterns to handle interruptions.[10] A startup, seeking to cut costs, moved training workloads to spot capacity but did not implement standardized checkpointing or job resubmission policies. When the cloud provider reclaimed the capacity, long-running training jobs failed near completion, wasting hours of expensive GPU time [11],.[12] After several such incidents, engineers reverted to expensive on-demand GPUs, negating all potential savings.

2.4. Lack of Visibility and Accountability

It is common to find limited end-to-end visibility, making it impossible to trace GPU usage from a cluster down to a specific team, project, or model. This prevents show back or chargeback [22], [23] and makes it impossible to explain which workloads are driving spend or waste.

2.5. Organizational Misalignment

These issues stem from a fundamental misalignment between engineering and FinOps. Platform and ML teams optimize for performance, while finance focuses on budgets [24],.[25] There is no shared set of unit- economics metrics (e.g., "cost per training run") [18], [19] to align them, creating conflict and inefficiency.

3. An Integrated Blueprint for GPU Fleet FinOps

To solve these interconnected problems, we propose an integrated operating model that treats GPU management as a unified FinOps and architecture challenge. This blueprint is built on a foundation of visibility and enforced through policy-driven scheduling, right-sizing, and governance.

3.1. Pillar 1: Foundational Visibility and Unit Economics

The first best practice is to "start with visibility, not tools." This involves tagging every single GPU workload with its owner (team, project, model) and piping telemetry to a centralized observability stack [26], [27],.[28]

- Tools: This stack typically consists of the NVIDIA Data Center GPU Manager (DCGM) for granular telemetry (SM utilization, memory, power), Prometheus for metrics collection, and Grafana for visualization [26], [27],.[28] Cloud-native cost tools like Kubecost can also integrate this data for detailed GPU cost monitoring [22], [23],.[29]
- The Goal (Unit Economics): The output is not just a hardware-monitoring dashboard but a "unit- economics dashboard." This dashboard, shared by Finance, Platform, and ML teams, must expose the business-centric metrics that bridge the organizational divide: "cost per training run," "cost per 1k inferences," and "utilization by business unit" [18],.[19]

3.2. Pillar 2: FinOps-Aware Scheduling and Orchestration

With visibility established, the scheduler becomes the primary tool for optimization. A FinOps-aware scheduling layer must integrate GPU metrics and cost data to make intelligent workload placement decisions.

- Custom Schedulers: Default Kubernetes schedulers are insufficient.[30] Custom schedulers like Kueue for job queueing [13], Volcano for high-performance batch scheduling [13], or Slurm integrations [31], [32] are required to pack jobs based on MIG profile, priority, and price class (on- demand vs. spot).
- Gang Scheduling: For distributed training, gang scheduling is a critical capability. It ensures that all pods for a job are scheduled simultaneously, preventing a "deadlock" scenario where a partial job reserves expensive GPUs while waiting for resources, wasting money and blocking other work [33],.[34]
- Workflow Management: These schedulers are integrated with MLOps tools like Kubeflow [35], Ray [36], or Argo Workflows [37], [38] to manage the end-to-end lifecycle of distributed training and inference pipelines.

3.3. Pillar 3: Policy-Driven Right-Sizing and Governance

Governance must be automated and encoded as policy, not managed through manual approval tickets. This is the only scalable way to enforce right-sizing and control costs.

- Standardize GPU Shapes: A critical best practice is to offer a small, standardized catalog of "S/M/L" training and inference profiles. This directly combats the MIG fragmentation caused by arbitrary user requests.[6]
- Policy as Code: These rules are enforced using Kubernetes-native admission controllers like Open Policy Agent

(OPA) [15], [39] or Kyverno [14],.[40] These tools can block or mutate workloads that violate governance (e.g., "Error: 'premium-dgx' tier requires 'project-x' finance code").

- Proactive Right-Sizing: This model forces "right-sizing from day one." New models must undergo a "profiling run" to understand their true resource needs, which are then locked in as the smallest viable GPU profile.

3.4. Pillar 4: Interruption-Aware Job Design

To safely unlock the massive savings of preemptible capacity, the platform must "design for preemption by default." This means making spot instances "safe and boring" for engineers to use [16],.[17] This is achieved by standardizing and automating interruption-aware job design. The platform must provide robust, built-in checkpointing frequency, idempotent job definitions, and automated retry policies as a default, "batteries-included" feature of the MLOps pipelines [16], [36], [41],.

4. Measuring Success: Metrics and Case Studies

The efficacy of this integrated model is not theoretical. It can be measured by a specific set of metrics and has been validated by real-world enterprise success stories.

4.1. Criteria for Evaluation

A successful GPU FinOps program must be evaluated on a mix of technical, financial, and operational metrics:

- GPU Utilization & Efficiency: Average/P95 GPU utilization (SM and memory) [26]; MIG packing efficiency and fragmentation index.
- Cost & Unit Economics: Total GPU spend (trended); Cost per training run/experiment; Cost per 1k inferences/tokens [18], [19]; percent of workloads on spot/preemptible and realized savings.
- Reliability & Performance: Job success rate (especially on spot GPUs) [10]; impact on queue wait times and time-to-train.
- Governance & Fairness: Accuracy of chargeback/showback by team and model [22], [23]; policy compliance rates.

4.2. Real-World Outcomes

This integrated approach yields tangible results, as seen in the following cases:

- Success (DGX Utilization): A global bank implemented standard GPU profiles and MIG-based packing. They increased average DGX utilization from ~35% to ~70% and reduced GPU spend per model by ~40%, all while decreasing experiment wait times [3], [4], [5],.
- Success (Spot-First Training): A large retailer refactored its LLM training jobs with robust checkpointing and gang scheduling. This allowed them to safely shift 60-70% of their training pipelines to spot/preemptible GPUs, achieving a ~50% cost reduction with no material impact on time-to-train [10],.[16]
- Failure (Lack of Governance): An LLM lab allowed its GenAI projects to add DGX nodes without any governance. Within a year, GPU costs became the largest single IT line item, yet average utilization remained below 30% [3], [4], [5],. Leadership eventually froze all GPU purchases, mandating a painful, reactive optimization program. This proves governance must start early.

5. Gaps and Future Directions

While this blueprint addresses today's core problems, the field is evolving. We see several gaps in current practice and future trends that will reshape this domain.

5.1. Gaps in Research and Practice

- Immature FinOps Models: End-to-end "GPU FinOps" models that unify architecture, pricing, and governance are still immature [24],.[25]
- Limited Unit Economics: Most platforms still report on aggregate spend, not the business-centric unit economics that truly measure value [18],.[19]
- MIG Economics: The economics of MIG, including how to design profiles and avoid fragmentation, are under-explored in practice.[6]
- Robust Spot Patterns: Concrete, reusable patterns (e.g., SDKs, controllers) for interruption-tolerant AI are still emerging, leaving implementation ad-hoc [10], [11], [12],.

5.2. Future Trends

Over the next five years, GPU FinOps will move from an advanced optimization to a standard pillar of AI platform design.

- Heterogeneous Accelerators: Fleets will become more complex, mixing next-gen NVIDIA GPUs (H100/B100s) with TPUs and custom ASICs. FinOps-aware schedulers must learn to optimize workloads across these different price-performance profiles [42],.[43]
- Smaller, Distilled Models: A shift toward domain-specific models will increase demand for right-

sized MIG slices, making over-provisioning less tolerable.

- AI-Driven Scheduling: Reinforcement learning (RL) and AI-based schedulers will learn optimal placement, right-sizing, and spot/on-demand mix from telemetry—creating "self-optimizing" fleets[44, 45, 46].
- Policy-as-Code: Governance will become fully programmable. Declarative policies (e.g., via OPA [15, 39] or Kyverno [14, 40]) will encode business rules like carbon budgets [47], max cost per run, and auto-reclamation of idle capacity.

Ultimately, unit economics like "cost per 1k tokens" will become board-level metrics, and carbon-aware constraints will be integrated into scheduling, making economic and environmental optimization two sides of the same problem.[47]

6. Conclusion

This paper has contributed a practitioner's blueprint that combines multi-cloud architecture, GPU platforms, and FinOps into one operating model. Based on lessons from real-world deployments, this framework provides concrete patterns, guardrails, and a reference architecture rather than purely theoretical optimizations.

For industry, this paper serves as a playbook for CTOs, platform teams, and FinOps leaders to run DGX, MIG, and preemptible fleets efficiently, turning GPU spending from an opaque liability into a controllable, value-linked investment. For the academic community, it frames new questions around fleet efficiency, unit economics, and policy-aware, autonomous schedulers.

The key takeaway is that AI doesn't become strategic just because it runs on GPUs; it becomes strategic when every GPU hour is tied to business value. With GPU FinOps, we can treat capacity as a governed, measurable product—not a black-box expense to achieve faster models, lower costs, and more sustainable AI at the same time.

References

[1] Federated Learning-based Personalized Recommendation Systems: An Overview on Security and Privacy Challenges - CyberSecDome, accessed October 29, 2025, https://cybersecdome.eu/wp-content/uploads/2024/01/IEEE-Transactions-on-Consumer- Electronics-Federated-Learning-based-Personalized-Recommendati-2023.pdf

[2] (PDF) E-commerce Personalized Recommendations: a Deep Neural Collaborative Filtering Approach - ResearchGate, accessed October 29, 2025, https://www.researchgate.net/publication/377081826_E-commerce_Personalized_Recommendations_a_Deep_Neural_Collaborative_Filtering_Ap proach

[3] (PDF) Federated Learning on Recommender Systems - ResearchGate, accessed October 29, 2025, https://www.researchgate.net/publication/388088244_Federated_Learning_on_Recommen der_Systems

[4] Federated Learning on Recommender Systems - IEEE Computer Society, accessed October 29, 2025, https://www.computer.org/csdl/proceedings- article/bigdata/2024/10825895/23yjUjOpcOY

[5] (PDF) A Survey on Federated Recommendation Systems, accessed October 29, 2025, https://www.researchgate.net/publication/366821201_A_Survey_on_Federated_Recomme ndation_Systems

[6] Recommendation Systems Using Federated Learning - Meegle, accessed October 29, 2025, https://www.meegle.com/en_us/topics/recommendation- algorithms/recommendation-systems-using-federated-learning

[7] Analysis of Privacy Preservation Enhancements in Federated Learning Frameworks - Shaping the Future of IoT with Edge Intelligence - NCBI, accessed October 29, 2025, https://www.ncbi.nlm.nih.gov/books/NBK602365/

[8] Digital Markets Act Summary: EU DMA Law Explained - Usercentrics, accessed October 29, 2025, https://usercentrics.com/knowledge-hub/digital-markets-act-dma-impacts-user- privacy-and-consent-management/

[9] The Digital Markets Act: Shaping Fair Competition in the Digital Age, accessed October 29, 2025, https://business.trustedshops.com/blog/digital-markets-act

[10] Federated Learning: The Decentralized Revolution Transforming AI While Preserving Privacy | by Nicolasseverino | Oct, 2025 | Medium, accessed October 29, 2025, https://medium.com/@nicolasseverino/federated-learning-the-decentralized-revolution- transforming-ai-while-preserving-privacy-2e0a0122d8b8

[11] How is federated learning used in personalized recommendations?, accessed October 29, 2025, https://milvus.io/ai-quick-reference/how-is-federated-learning-used-in-personalized- recommendations

[12] Federated Learning: A Privacy-Preserving Approach to ... - Netguru, accessed October 29, 2025, https://www.netguru.com/blog/federated-learning

[13] Low-Latency Collaborative Predictive Maintenance: Over-the-Air Federated Learning in Noisy Industrial Environments - MDPI, accessed October 29, 2025, https://www.mdpi.com/1424-8220/23/18/7840

[14] LoLaFL: Low-Latency Federated Learning via Forward-only ..., accessed October 29, 2025, https://arxiv.org/abs/2412.14668

[15] Privacy-Preserving Federated Learning - Hasso-Plattner-Institut, accessed October 29, 2025, https://hpi.de/arnrich/research-areas/privacy-preserving-federated-learning.html

[16] (PDF) Federated Learning Architectures for Privacy-Preserving Artificial Intelligence Applications on Edge Devices - ResearchGate, accessed October 29, 2025,

https://www.researchgate.net/publication/392749199_Federated_Learning_Architectures_f or_Privacy-Preserving_Artificial_Intelligence_Applications_on_Edge_Devices
[17] Federated Learning for Cybersecurity: A Privacy-Preserving Approach, accessed October 29, 2025, https://www.mdpi.com/2076-3417/15/12/6878
[18] State of Cloud Costs | Datadog, accessed November 18, 2025, https://www.datadoghq.com/state-of-cloud-costs/
[19] Optimizing GenAI Usage: A FinOps Perspective on Cost ..., accessed November 18, 2025, https://www.finops.org/wg/optimizing-genai-usage/
[20] How one company went from 28% GPU utilization to 73% with Run:ai, accessed November 18, 2025, https://pages.run.ai/hubfs/PDFs/Case-Study-from-28-to-73-percent- GPU-Utilization.pdf
[21] (PDF) Flex-MIG: Enabling Distributed Execution on MIG, accessed November 18, 2025, https://www.researchgate.net/publication/397556277_Flex- MIG_Enabling_Distributed_Execution_on_MIG
[22] Maximize GPU Efficiency: Smarter Fixes for Checkpointing Challenges - Clockwork.io, accessed November 18, 2025, https://clockwork.io/blog/maximize-gpu-efficiency-smarter- fixes-for-checkpointing-challenges/
[23] Parcae: Proactive, Liveput-Optimized DNN Training on Preemptible Instances - arXiv, accessed November 18, 2025, https://arxiv.org/html/2403.14097v1
[24] Batch Scheduling on Kubernetes: Comparing Apache YuniKorn ..., accessed November 18, 2025, https://www.infracloud.io/blogs/batch-scheduling-on-kubernetes/
[25] Compare Custom Schedulers for Kubernetes - Rafay Product ..., accessed November 18, 2025, https://docs.rafay.co/blog/2024/10/11/compare-custom-schedulers-for-kubernetes/
[26] A Slurm on Kubernetes Implementation for HPC and ... - CoreWeave, accessed November 18, 2025, https://www.coreweave.com/blog/sunk-slurm-on-kubernetes-implementations
[27] [PDF] Parcae: Proactive, Liveput-Optimized DNN Training on Preemptible Instances, accessed November 18, 2025, https://www.semanticscholar.org/paper/01f6de9e8670b613f4eccf0a699acb45673c19c5
[28] How Nvidia DGX Cloud Uses Kyverno to Enforce Kubernetes Pod Security Standards, accessed November 18, 2025, https://nirmata.com/2024/12/15/how-nvidia-dgx-cloud-uses- kyverno/
[29] Cloud Cost Management for AI/ML Workloads - emma, accessed November 18, 2025, https://www.emma.ms/cloud-solutions/ai-cost-management
[30] FinOps for AI: A Guide To Managing AI Cloud Costs - ProsperOps, accessed November 18, 2025, https://www.prosperops.com/blog/finops-for-ai/
[31] AI and ML perspective: Cost optimization | Cloud Architecture Center ..., accessed November 18, 2025, https://docs.cloud.google.com/architecture/framework/perspectives/ai-ml/cost- optimization
[32] Role of AI in cloud cost optimization and FinOps (Financial Operations) - | World Journal of Advanced Engineering Technology and Sciences, accessed November 18, 2025, https://journalwjaets.com/sites/default/files/fulltext_pdf/WJAETS-2025-0218.pdf
[33] FinOps For AI: How Crawl, Walk, Run Works For Managing AI Costs - CloudZero, accessed November 18, 2025, https://www.cloudzero.com/blog/finops-for-ai/
[34] DGX Platform: Built for Enterprise AI - NVIDIA, accessed November 18, 2025, https://www.nvidia.com/en-us/data-center/dgx-platform/
[35] Measure and Improve AI Workload Performance with NVIDIA DGX Cloud Benchmarking, accessed November 18, 2025, https://developer.nvidia.com/blog/measure- and-improve-ai-workload-performance-with-nvidia-dgx-cloud-benchmarking/
[36] Case Study: NVIDIA Boosts BMW Group's Production Efficiency with AI, accessed November 18, 2025, https://www.nvidia.com/en-us/customer-stories/bmw-optimizes- production-with-ai-and-dgx-systems/
[37] Scalable AI Infrastructure Accelerates Autonomous Vehicle Development, accessed November 18, 2025, http://images.nvidia.cn/content/dgx/dgx-1-zenuity-case-study-us- 675763-r7-web.pdf
[38] NVIDIA A100 Tensor Core GPU Architecture, accessed November 18, 2025, https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere- architecture-whitepaper.pdf.
[39] [2502.01909] A Multi-Objective Framework for Optimizing GPU-Enabled VM Placement in Cloud Data Centers with Multi-Instance GPU Technology - arXiv, accessed November 18, 2025, https://arxiv.org/abs/2502.01909
[40] More-efficient recovery from failures during large-ML-model training - Amazon Science, accessed November 18, 2025, https://www.amazon.science/blog/more-efficient-recovery- from-failures-during-large-ml-model-training
[41] Generative AI Cost Optimization Strategies | AWS Cloud Enterprise Strategy Blog, accessed November 18, 2025, https://aws.amazon.com/blogs/enterprise- strategy/generative-ai-cost-optimization-strategies/
[42] Monitoring GPU workloads on Amazon EKS using AWS managed open-source services, accessed November 18, 2025, https://aws.amazon.com/blogs/mt/monitoring-gpu- workloads-on-amazon-eks-using-aws-managed-open-source-services/
[43] Seamless Nvidia GPU Observability! | by Nitee Shah - Medium, accessed November 18, 2025, https://niteeshah95.medium.com/seamless-nvidia-gpu-observability-b8291e4fa2d1
[44] Best Practices for GPU Observability in Modern AI Infrastructure - Techstrong.ai, accessed November 18, 2025, https://techstrong.ai/social-facebook/best-practices-for-gpu- observability-in-modern-ai-infrastructure/
[45] Integrating observability stack into your Kubernetes cluster - Crusoe Cookbook, accessed November 18, 2025,

https://cookbook.crusoe.ai/observability-kubernetes
[46] GPU observability in Azure Kubernetes Service (AKS) - Microsoft Learn, accessed November 18, 2025, https://learn.microsoft.com/en-us/azure/aks/monitor-gpu-metrics
[47] The FinOps playbook for AI: Optimizing costs and performance - Flexera, accessed November 18, 2025, https://www.flexera.com/blog/finops/the-finops-playbook-for-ai- optimizing-costs-and-performance/