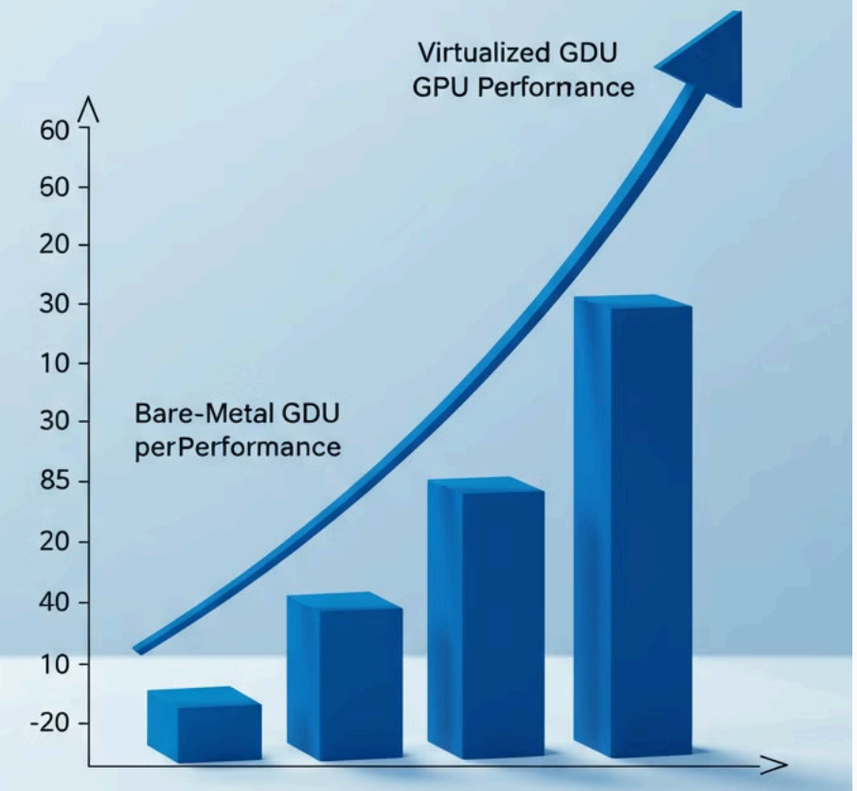
Virtualized GPU for AI in Enterprise Storage: Cost-Optimized Solutions

The integration of artificial intelligence capabilities into enterprise operations has become a strategic imperative across industries. Organizations face significant challenges related to infrastructure costs, operational complexity, and deployment efficiency.

This presentation examines how Dell Technologies, VMware, and NVIDIA have collaborated to develop virtualized GPU solutions specifically designed for enterprise storage environments, creating a more flexible and cost-effective approach to AI implementation.

By: **By: Venkat Chainuru Nivas**



Market Growth and Performance Potential

$11.83B

Data Center GPU Market

Global market value in 2023

29.3%

CAGR

Expected growth rate from 2024 to 2030

87%

Performance

Virtualized GPU environments can achieve up to 87% of bare-metal performance

The remarkable growth trajectory underscores the critical importance of optimized GPU solutions in modern enterprise infrastructure. Research demonstrates that while virtualized GPUs exhibit an average overhead of 13-18% across various workloads, the benefits of resource utilization, management flexibility, and cost optimization often outweigh these performance considerations.

The Collaborative Approach

Dell Technologies

Delivers PowerEdge servers equipped with NVIDIA A100 80GB PCIe Tensor Core GPUs, achieving 3.9x performance improvement for AI inference workloads across enterprise environments

NVIDIA

Supplies the comprehensive AI Enterprise software suite specifically optimized for virtualized environments, preserving efficiency above 92% even in fully virtualized configurations

1

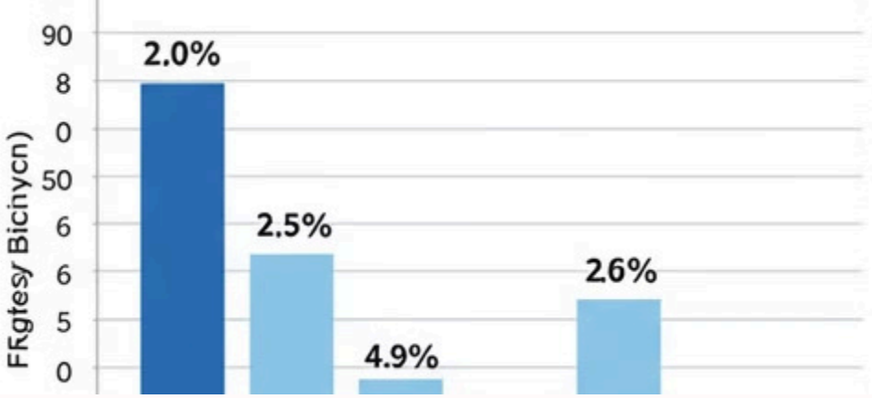
2

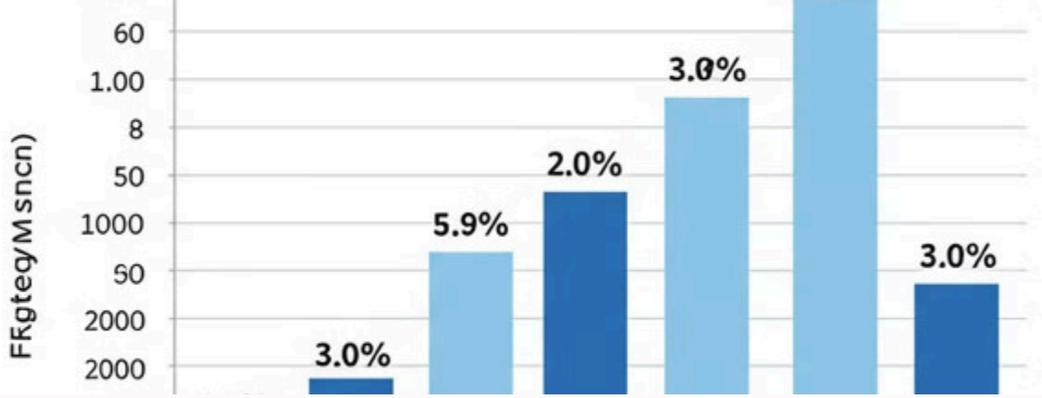
3

VMware

Provides enterprise-grade vSphere with Tanzu solution, enabling virtualization with minimal performance overhead of just 1-3% compared to bare-metal implementations

This strategic alliance has produced exceptional outcomes, demonstrating near-linear performance scaling from 1 to 8 GPUs within a single server for deep learning training workloads. Independent research confirms that state-of-the-art vGPU implementations consistently deliver 82-94% of native bare-metal performance across diverse AI workloads.

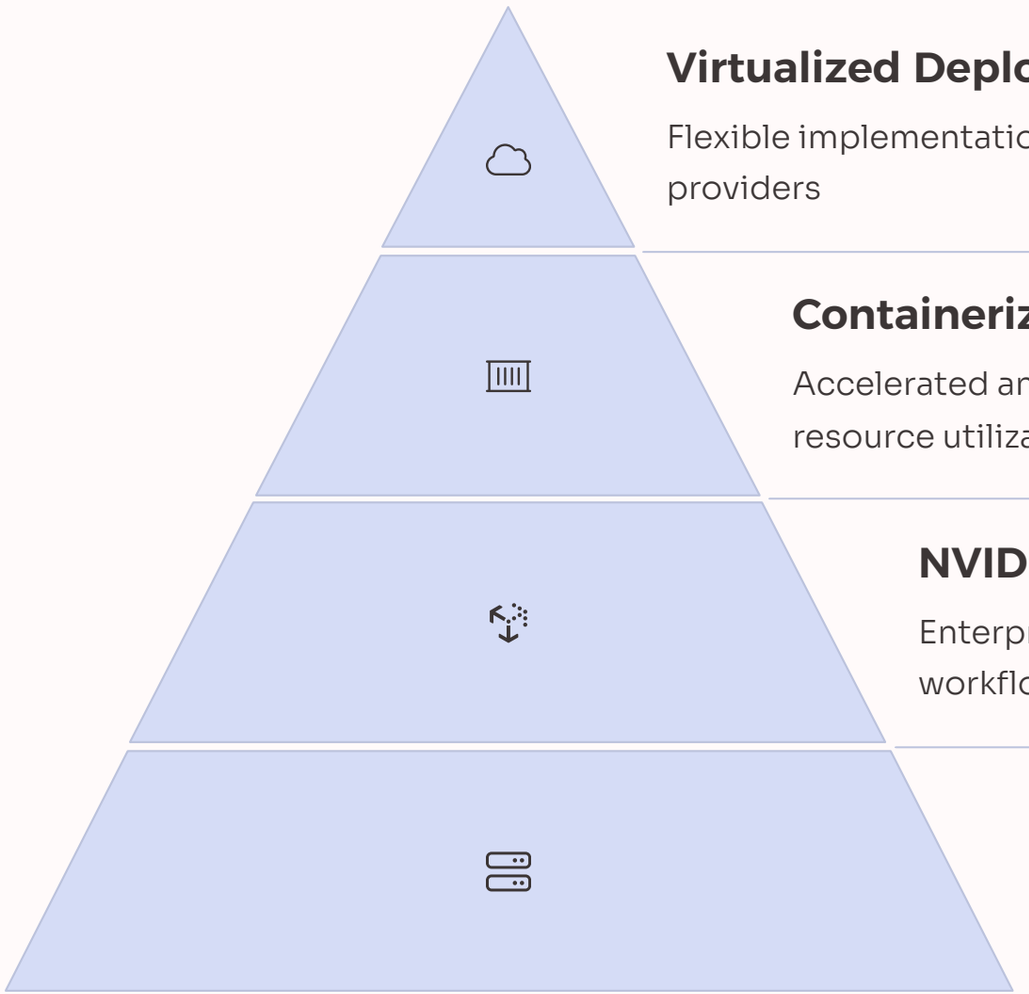




Performance Comparison Metrics

Metric	Value	Comparison Point
NVIDIA A100 80GB PCIe GPU Performance Improvement	3.9x	Compared to previous generation A30 GPUs
Virtualization Performance Degradation	1-3%	Compared to bare-metal implementations
GPU Utilization in Virtualized Environments	>92%	Efficiency in fully virtualized configurations
vGPU Performance (Average)	82-94%	Percentage of native bare-metal performance
Memory-Intensive AI Workload Penalty	7-12%	Performance penalty in virtualized environments
Compute-Intensive AI Workload Penalty	5-9%	Performance penalty in virtualized environments

Key Components of the Solution



Virtualized Deployment

Flexible implementation across on-premises infrastructure or cloud service providers

Containerization

Accelerated and simplified AI workload deployment with optimized resource utilization

NVIDIA AI Enterprise

Enterprise-grade software suite optimized for virtualized AI workflows

NVIDIA-Certified Systems

Rigorously tested hardware foundation ensuring optimal performance

This enterprise-ready validated architecture integrates NVIDIA AI Enterprise software, empowering organizations leveraging VMware vSphere to efficiently virtualize and containerize mission-critical AI workloads. Comprehensive performance testing reveals exceptional results, with deep learning training workloads achieving up to 96% of bare-metal performance efficiency. Inference workloads demonstrate even more impressive results, reaching 98% equivalence to bare-metal implementations in specific deployment scenarios.

Virtualized vs. Bare-Metal Performance

Performance analyses confirm that virtualized environments can achieve performance comparable to bare-metal implementations across various AI workloads.

Deep Learning Training

96% of bare-metal performance in virtualized environments, with minimal performance penalty

Inference Workloads

98% equivalence to bare-metal implementations, showing highest compatibility

TensorFlow Image Classification

95% performance retention, processing up to 4,500 images per second

General AI Workflows

96.5% performance efficiency across standard enterprise AI applications

Memory-Intensive Workloads

91.5% performance retention, showing only 4-6% decrease compared to non-virtualized environments

Operational Benefits

Reduced Deployment Time

Infrastructure provisioning tasks that previously took weeks can be completed in days or even hours, with up to 60% reduction in deployment time for AI workloads.

Improved Operational Efficiency

Organizations leveraging VMware expertise for AI deployments achieved 83% greater operational efficiency and required 30% less specialized training than those implementing dedicated AI infrastructure.

Enhanced Resource Utilization

Organizations implementing virtualized GPU solutions achieved 3.2x better resource utilization than dedicated AI infrastructure and reduced maintenance overhead by approximately 54%.

Lower Total Cost of Ownership

The unified approach reduces overall IT operational expenses by 33% and lowers the total cost of ownership by approximately 40% over a five-year period.

Operational Improvements from Virtualized GPU Solutions

AI Workload Deployment Time

60% reduction in implementation timelines, transforming weeks-long projects into days or even hours

Operational Efficiency

83% greater operational efficiency for organizations leveraging existing VMware expertise for AI deployments

Specialized Training Requirements

30% decrease in specialized training needs compared to implementing dedicated AI infrastructure

Resource Utilization

3.2x better hardware resource utilization with virtualized solutions, maximizing investment returns

Maintenance Overhead

54% reduction in ongoing maintenance requirements compared to traditional siloed AI infrastructure

Enterprise-Wide AI Integration

Human Resources

AI-powered talent acquisition systems reduced time-to-hire by 37% while improving quality-of-hire metrics by 28%. These applications process an average of 6,000 resumes per day while requiring only 15-20% of the computational resources needed in non-virtualized environments.

Organizations implementing these solutions through virtualized GPU infrastructure reported achieving full deployment 42% faster than those using dedicated systems.

Information Technology

AI-enhanced cybersecurity solutions demonstrated impressive results, with organizations reporting a 47% reduction in security incidents and cost savings averaging $3.1 million annually from prevented breaches.

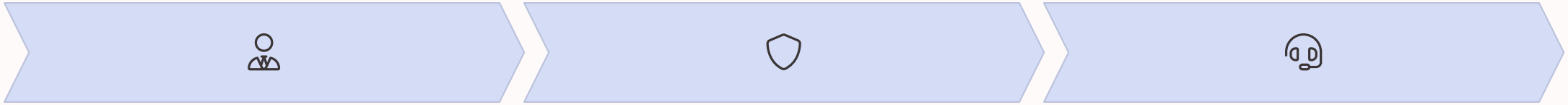
Organizations adopting virtualized GPU solutions for their AI initiatives reported an average ROI of 134% over a three-year period, with deployment costs 27-35% lower than dedicated infrastructure approaches.

Customer Service

Organizations deploying AI-powered conversation systems could handle 65% more customer interactions while reducing staffing requirements by 23%, with virtualized GPU solutions enabling these systems to be deployed 3.5 times faster than traditional approaches.

Leveraging a common infrastructure platform across multiple departments was a critical success factor, with 38% lower training costs and 41% faster implementation timelines.

Department-Specific AI Implementation Benefits



Human Resources

- Accelerated recruitment with 37% reduction in time-to-hire
- Enhanced talent acquisition with 28% improvement in quality-of-hire metrics
- Strengthened workforce stability through 24% improvement in employee retention
- Boosted operational effectiveness with 31% increase in workforce productivity

Information Technology

- Fortified security posture with 47% reduction in critical incidents
- Delivered $3.1 million in annual cost savings from prevented breaches
- Generated substantial 134% ROI over three-year implementation period
- Achieved 27-35% lower infrastructure deployment costs

Customer Service

- Dramatically expanded capacity with 65% increase in customer interactions
- Optimized resource allocation with 23% reduction in staffing requirements
- Expedited implementation with 3.5x faster deployment timelines
- Reduced operational expenses through 38% lower training costs

Conclusion: Accelerating Enterprise AI Adoption

Unified Infrastructure Approach

The collaborative solution from Dell Technologies, VMware, and NVIDIA represents a significant advancement in enterprise AI infrastructure, addressing both technological and operational challenges that have traditionally impeded widespread AI adoption.

Comparable Performance

Performance analyses confirm that virtualized GPU environments deliver results comparable to bare-metal implementations for most AI workloads, while offering substantial operational and financial benefits including faster deployment and improved resource utilization.

Enterprise-Wide Implementation

The solution's support for diverse AI requirements across enterprise departments enables organizations to implement comprehensive AI strategies without creating infrastructure silos or operational complexity.

Accelerated AI Transformation

As AI continues to transform business operations across industries, the virtualized GPU approach provides a pragmatic path for enterprises seeking to balance technological innovation with operational efficiency and cost optimization.

Thankyou