

Optimizing HPC and AI Infrastructure with Standard Linux and Advanced Storage Architectures

Molly Presley

*Head of Global Marketing – Hammerspace
& Host of Data Unchained Podcast*

08 April, 2025



Industry Trends Driving Changes to Enterprise IT

LLM Training
Load and Iterate in GenAI



Large, Decentralized
Data Sets



Multi-Site
Multi-Cloud
Remote AI Researchers



The Problem: Effectively Utilizing Distributed Data



Valuable data
trapped in silos

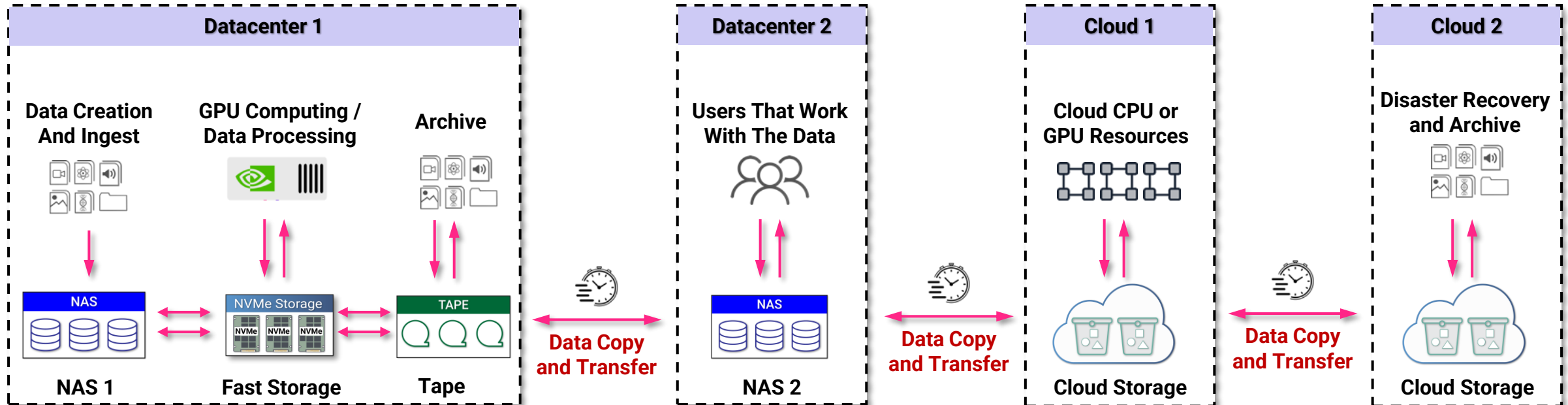
Getting data to global
users **takes too long**

Infrastructure
is **not ready for AI**

Data copy sprawl **impacts
cost, governance and security**

Lack performance to
keep GPUs utilized

Lack agility to use
elastic cloud resources



Hammerspace Unifies Incompatible Silos

Unify and automate unstructured data across any data center, any cloud, anywhere

Hammerspace Global Data Platform

Data Creation
And Ingest



NFS ↓ SMB

GPU Computing /
Data Processing



↑ pNFSv4.2
with Flex Files ↓

On-Premises
Archive



NFS ↓ SMB ↑

Users That Work
With The Data



NFS ↓ S3 ↑

Cloud CPU or
GPU Resources



↓ S3 ↑

Remote
DR & Archive



↓ S3 ↑



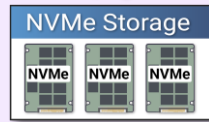
Unified Multi-protocol File/Object Access

Hammerspace Parallel Global File System

Automated Non-disruptive Data Orchestration



NAS 1



Fast Storage



Tape



NAS 2



Cloud Storage

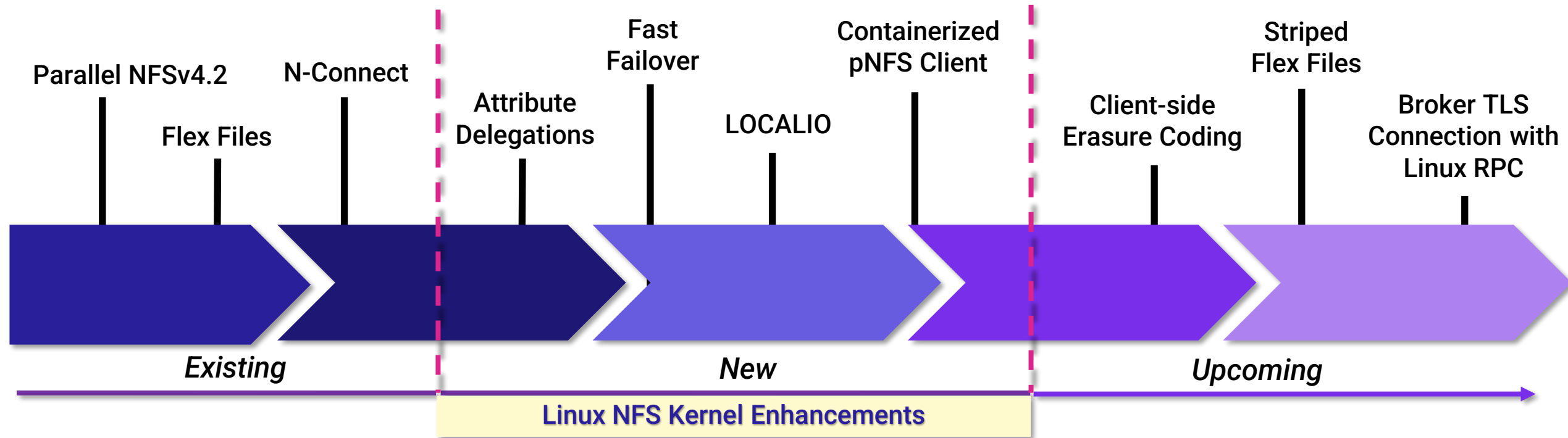


Cloud Storage



Hammerspace Commitment to Linux Standards

Hammerspace Contributions to Linux for High-Performance Storage & Distributed Data Environments



This Means Customers Never Have to Install Proprietary Client Software or Alter Existing Storage

Hammerspace is Engineered to Activate These Capabilities Which are Already Included in All Major Linux Distributions

Meta Picked Hammerspace to Power AI

"What Hammerspace does is pure magic."

-Paul Saab, Principal Engineer Meta

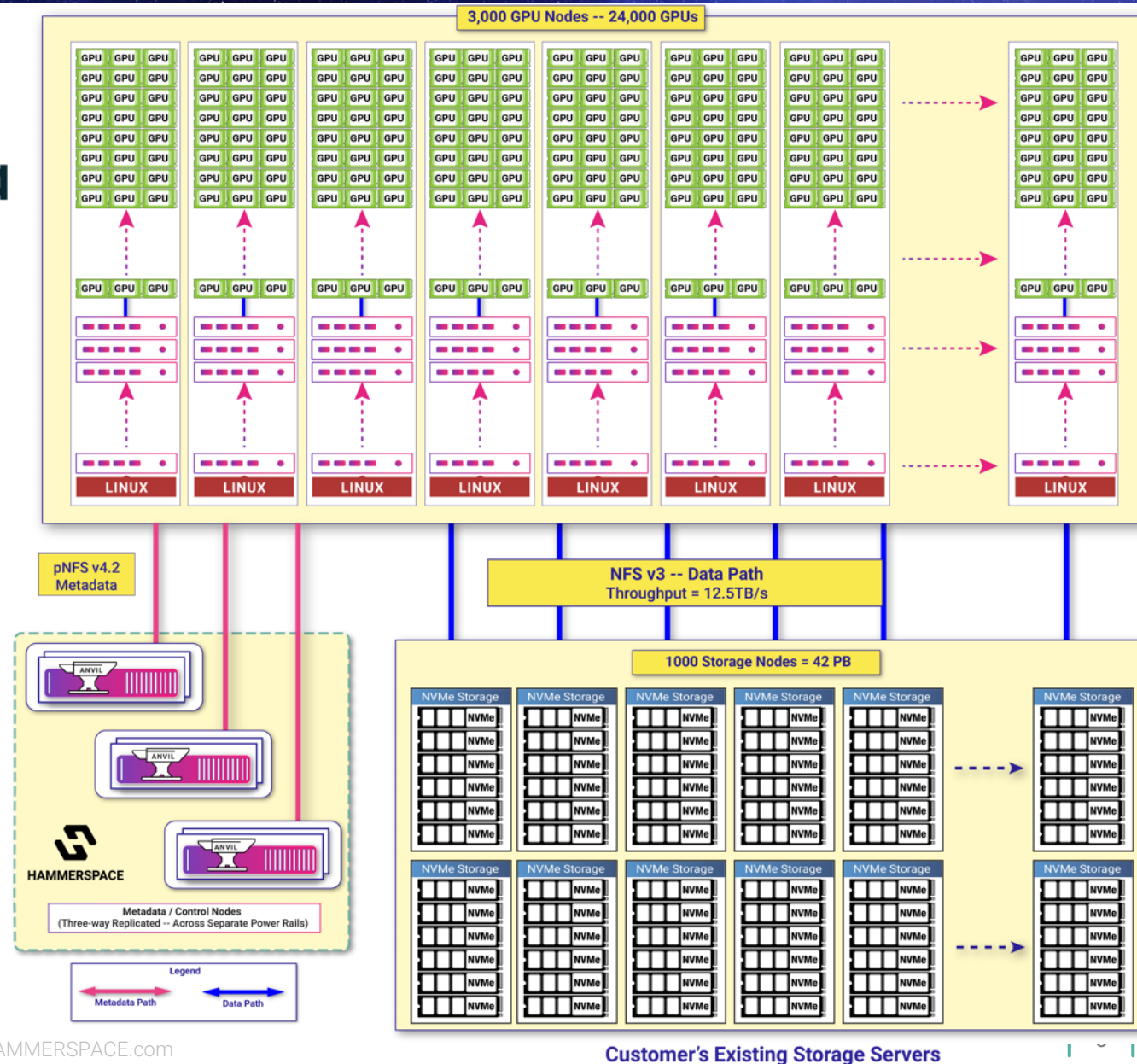


Hammerspace Solution for Llama 2 & 3

- No vendor even came close to Hammerspace's capabilities
- Aggregate performance of 12.5TB/sec (100Tb/sec)
- Everything is standards-based and plug-n-play
- Customer was able to use existing OCP storage servers

Meta Extended Llama 4 to the Cloud

- Hammerspace flexibility made the shift seamless
- Meta able to utilize existing cloud-based GPUs

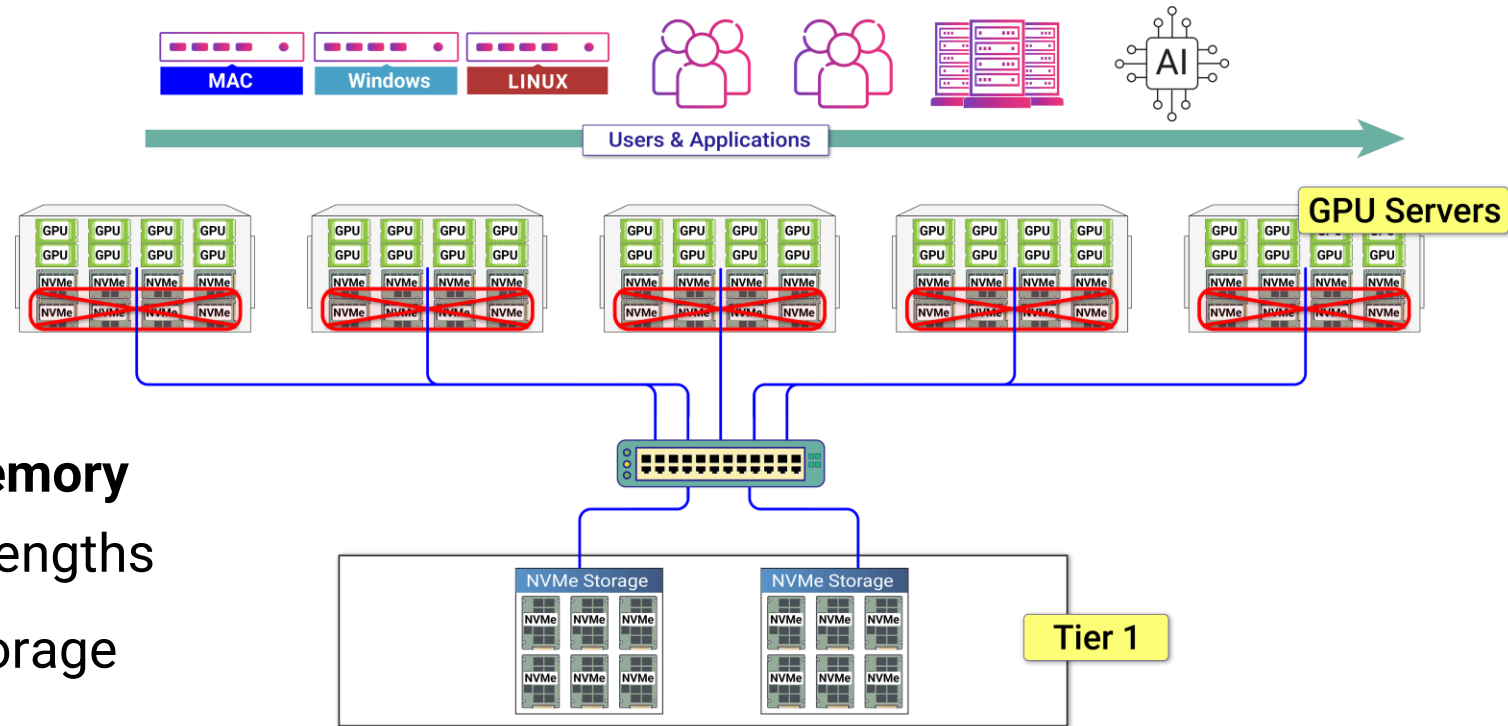


The Problem of Enough Storage Performance for GPUs

How To Activate Underutilized Local NVMe on GPU Servers

GPU Server Local NVMe Storage is Stranded, Unused Capacity

- Siloed, and largely unprotected
- Not shared across the cluster
- As NVMe densities grow, this equals large volume of stranded, under-used storage capacity



AI Workloads Surpassing GPU Server Memory

- More complex models, longer context lengths
- Forcing I/O over network to external storage
- Network bottleneck adds latency

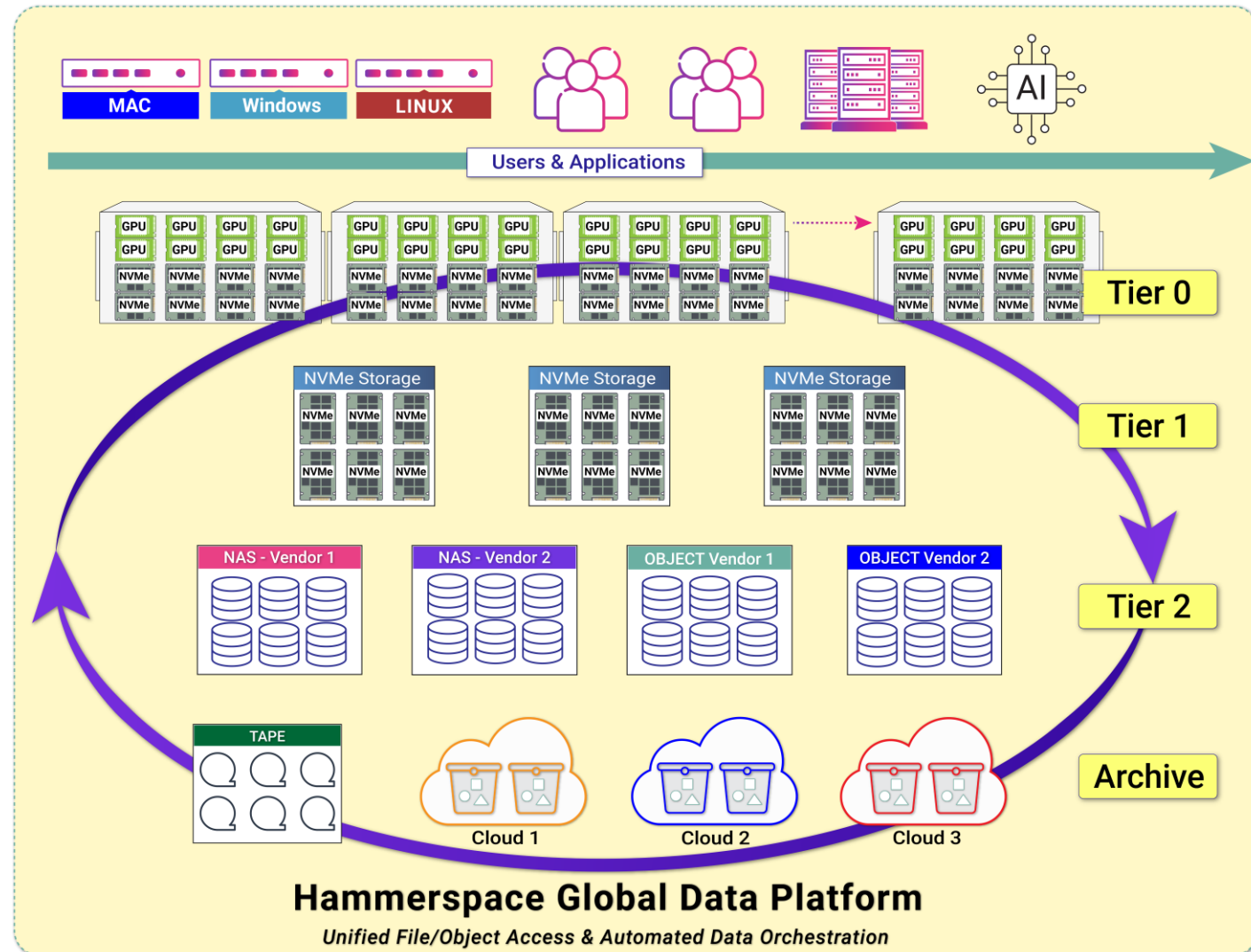


Hammerspace Data Orchestration Activates Tier 0

The Key is Putting Data in Motion Across All Tiers

Hammerspace Data Orchestration

- Unifies all storage tiers
- Eliminates vendor silos
- Automates non-disruptive data orchestration
- Ensures all data is protected
- Automates AI workloads & policy-based data placement
- Spans multi-site & cloud



Tier 0 Performance Faster than Any External Storage

On premises:

- Tier 0 is faster than any networked external storage
- Reduces need for expensive networks
- Reduces dependency on expensive Tier 1

Cloud

- Tier 0 is dramatically faster
- Increases GPU output
- Reduces downtime
- Reduces Tier 1 costs

1. Tier 0 vs. Tier 1 - On Premises (90% Network Efficiency)

(Assuming Checkpoint Size of 500GB)

	Storage Type	Optimal Throughput	Effective Throughput	Time to Write 500GB	
Tier 0	Local NVMe (8-drive GPU server)	112 GB/s	112 GB/s ✓	~4.5 sec ✓	
Tier 1	100 Gb/s Network	12.5 GB/s	11.25 GB/s	~44 sec (~2.1 min)	10x Slower
	200 Gb/s Network	25 GB/s	22.5 GB/s	~22 sec (~1 min)	5x Slower
	400 Gb/s Network	50 GB/s	45 GB/s	~11.1 sec	2.4x Slower

Note: External Tier 1 storage throughput adjusted assuming 90% practical network efficiency.

2. Tier 0 vs. Tier 1 - Cloud Storage (90% Network Efficiency)

(Assuming Checkpoint Size of 500GB)

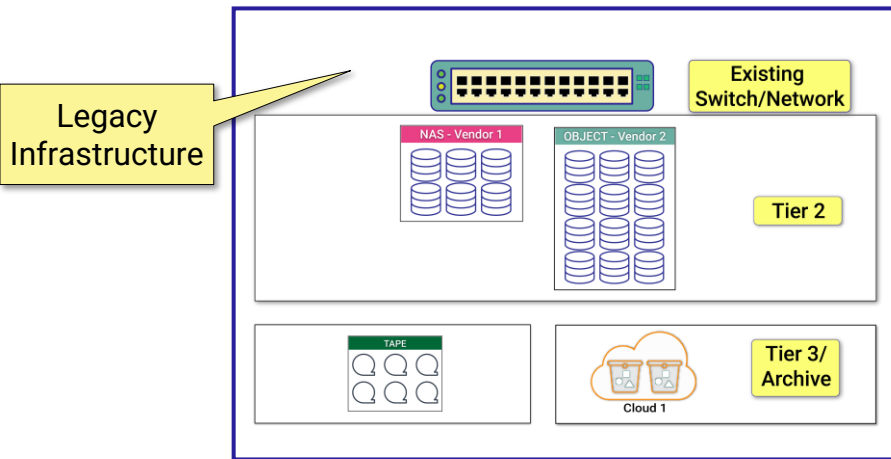
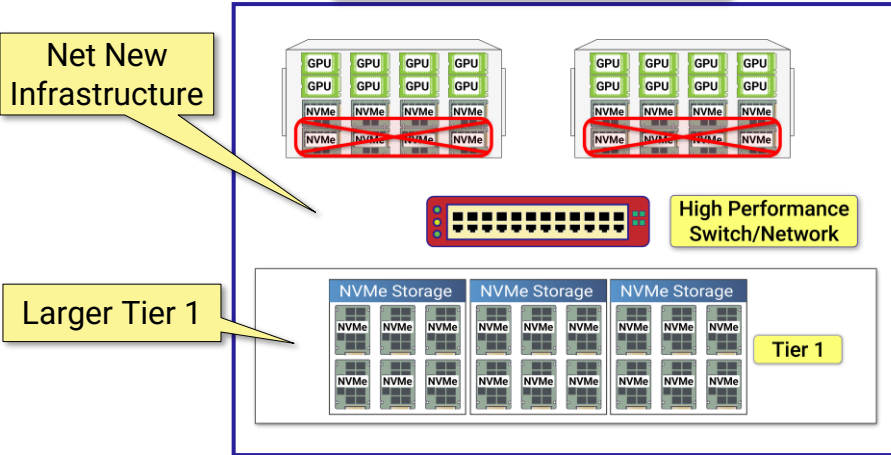
	Storage Type	Optimal Throughput	Effective Throughput	Time to Write 500GB	
Tier 0	Local NVMe (8-drive GPU server)	112 GB/s	112 GB/s ✓	~4.5 sec ✓	
Tier 1	AWS EBS (gp3)	1 GB/s	0.9 GB/s	~555.6 sec (~9.3 min)	123x Slower
	AWS EBS (io2)	4 GB/s	3.6 GB/s	~138.9 sec (~2.3 min)	30x Slower
	AWS S3 (via high-speed network)	1 GB/s	0.9 GB/s	~555.6 sec (~9.3 min)	123x Slower

Note: External Tier 1 storage throughput adjusted assuming 90% practical network efficiency.



Tier 0 = Increased ROI → Eliminate Duplicate Costs

Without Tier 0



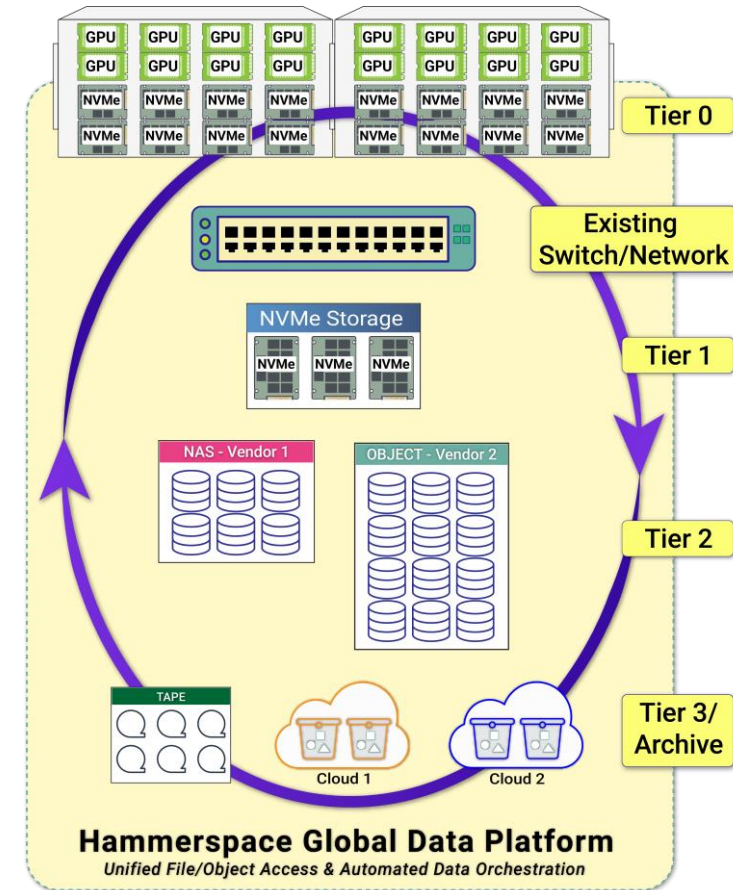
Eliminate Unnecessary Infrastructure

- Reduce net-new Tier 1 storage
- Utilize existing network
- Utilize lower-cost existing storage

Utilize Sunk Cost of Local NVMe

- It's already paid for with the servers
- Existing Tier 0 capacity is faster than any external storage from any vendor

With Tier 0

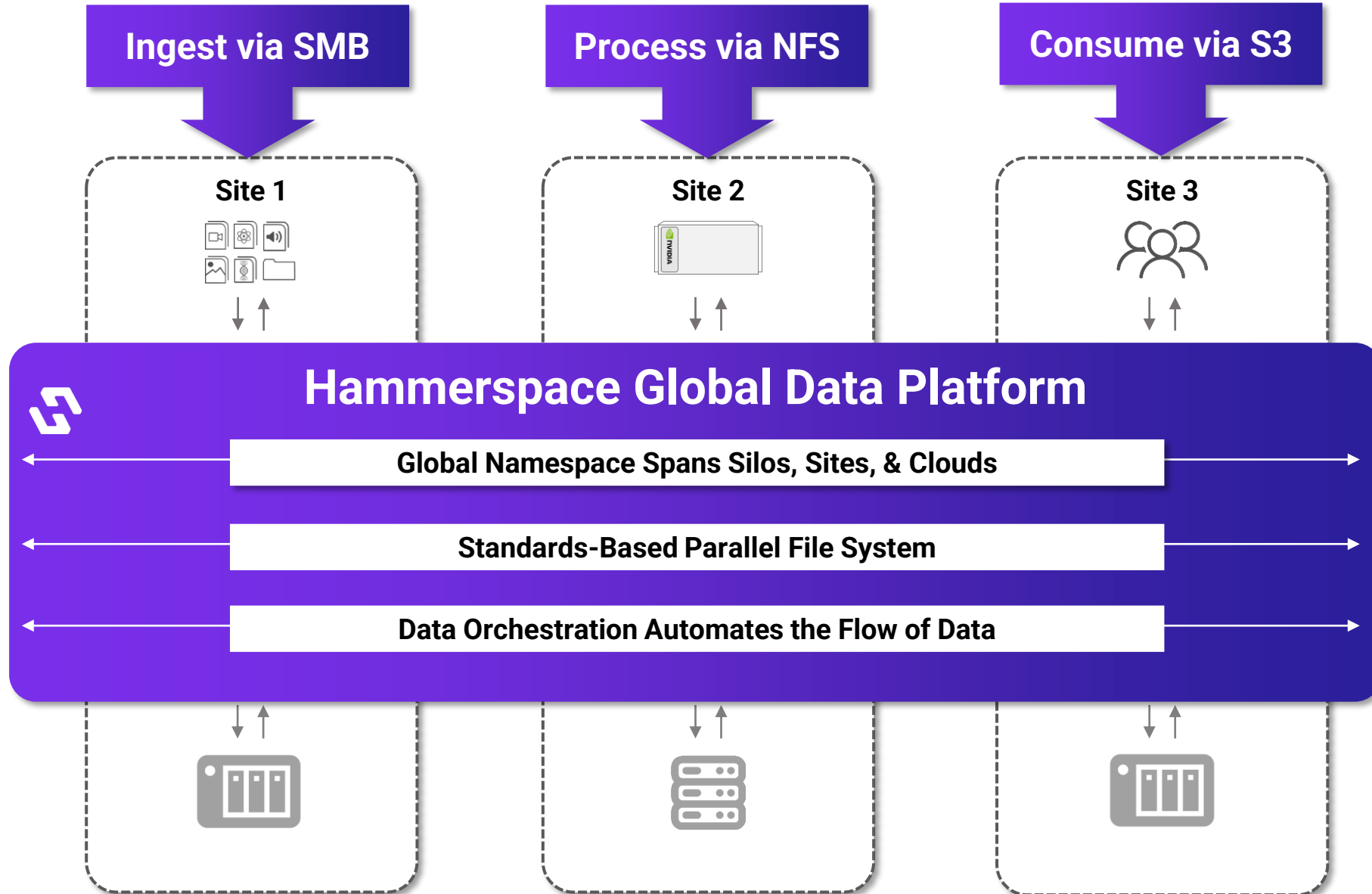


Cost Comparison between Tier 1 & Tier 0

Metric	Tier 1 External Storage (Array)	Tier 0 Internal NVMe Storage
Cost per TB	2x (More expensive)	1x (Baseline)
Cost per GB/s (200Gb/s network)	10x (More expensive)	1x (Baseline)



Unify, Orchestrate, Accelerate Distributed Data



Hammerspace Global Data Platform

Thank You

www.Hammerspace.com

