



Activating Untapped Tier 0 Storage Within Your GPU Servers

*Accelerate AI Workloads with the Highest
Performing NVMe Storage Available Today...
Which You Already Own.*

By Floyd Christofferson

V. 1.2 - Feb 2026

www.HAMMERSPACE.com

Executive Summary:

Key Ideas Introduced in This Paper

- **Every modern GPU (and many CPU) servers already contain ultra-fast NVMe storage that is largely underutilized.**

This local storage is typically used only as cache or scratch space, leaving enormous performance potential stranded inside the servers themselves.

- **The primary limiter for AI and other high-performance workloads is no longer storage media, but the network path to it.**

Even the fastest external NVMe systems remain constrained by network latency, congestion, cost, and power requirements.

- **Hammerspace's Tier 0 architecture creates a new shared storage tier by activating local NVMe inside compute servers.**

By aggregating server-local NVMe across a cluster, Tier 0 delivers performance that exceeds any external storage while remaining shared, protected, and usable at scale.

- **This approach changes the economics of performance.**

Tier 0 uses capacity that organizations have already paid for within their servers, avoids the need for extreme networking upgrades, and dramatically improves GPU utilization and ROI.

- **Tier 0 is not a new silo or a special-purpose cache.**

Data written to Tier 0 remains part of a unified global namespace and can be orchestrated automatically for protection, reuse, and downstream workflows.

- **The same architecture applies beyond GPUs**

CPU-based and ARM servers with local NVMe can benefit from the same ultra-low latency and high-throughput Tier 0 capabilities for a wide range of data-intensive workloads.

- **This paper explains how Tier 0 works, why it matters now, and how it fits into modern AI and high-performance environments.**

It shows how organizations can unlock unprecedented performance using infrastructure they already own, without introducing proprietary clients or disruptive architectural changes.



Why Tier 0 Matters Now

AI and other high-performance workloads are exposing a structural inefficiency in modern infrastructure. While GPU servers have become dramatically more powerful, the data paths feeding them have not kept pace.

External storage systems, regardless of media speed, remain constrained by network latency, congestion, cost, and power. At the same time, nearly every GPU server already includes multiple high-performance NVMe drives that are largely underutilized because they cannot be safely shared or integrated at scale.

Hammerspace's Tier 0 architecture emerges at this moment as a fundamentally different approach: activating this stranded, server-local NVMe as a shared, protected, ultra-high-performance storage tier. Rather than adding new infrastructure, Tier 0 rethinks how existing resources are used, delivering performance, efficiency, and economics that traditional architectures cannot match.

Overview:

The industry response to AI has largely focused on building faster storage systems as the ideal solution to performance and scale challenges for AI workloads. But there is a fundamental architectural constraint that is still unaddressed: the network bottleneck between compute and storage.

This is not simply about which storage vendor delivers the greatest performance to accelerate data throughput to GPU servers. As AI models grow more complex and inference sessions handle longer context lengths, workloads increasingly saturate GPU memory and main system RAM. This forces greater data movement to and from external storage, resulting in increased I/O wait times and growing network congestion. That network pressure directly impacts GPU utilization and increases costs in both on-premises and cloud-based data centers.

Here's the rub: Even with vendors promoting faster storage systems or requiring proprietary client software to be installed on GPU servers to enable caching, token optimization, or other acceleration techniques, the structural issue remains. Data moving between external storage systems and GPU servers must still traverse the network, and that pathway is inherently constrained by bandwidth, regardless of how fast the underlying storage system may be.

Adding intelligence to prioritize which data is routed to GPUs can improve performance and reduce unnecessary traffic. But even with these optimizations, an ever-increasing volume of data must still traverse the network from external storage to feed the GPUs. That is the crux of the problem.

The good news is that a far better solution already exists. It is hidden in plain sight within nearly every GPU server, on-premises and in the cloud: the underutilized local NVMe storage built directly into those systems, activated via the standard Linux capabilities already present on the server.



This paper provides an overview of how organizations of any size can unlock cost savings and performance gains by activating this idle capacity as a seamless, protected, ultra-fast shared storage tier to dramatically accelerate GPU workloads and improve GPU utilization.

Hammerspace Tier 0 is the most performant shared storage type currently available for AI workloads, with benchmarks demonstrating **it can support 16.5x more GPUs than Lustre with an equivalent number of clients.** When this capacity is aggregated across multiple GPU servers in a cluster, organizations can drive significantly greater output from their GPU investments at lower cost, both on-premises and in the cloud.

Critically, this innovation works with existing on-premises and cloud-based GPU or CPU/ARM servers from any vendor by activating the local NVMe capacity already installed in those systems.

This bears repeating: **This ultra-fast local NVMe capacity is largely underutilized and represents a sunk cost** already included in the price of GPU servers, both on-premises and in the cloud. It adds no additional power burden, which is a meaningful advantage for data centers operating near their power limits.

With Hammerspace, this local NVMe is activated as Tier 0 shared storage using capabilities built directly into the Linux kernel and included in standard distributions. **No proprietary client software is required, and no alterations to existing servers are necessary.** As outlined below, Tier 0 delivers immediate value across environments of all sizes, from small clusters with a handful of GPU servers to hyperscale deployments spanning many thousands of GPUs.

Utilizing local Tier 0 capacity reduces or can even eliminate the need for additional high-performance networking and storage infrastructure, which are often deployed as costly attempts to overcome the network bottleneck and keep GPUs fully utilized.

With Hammerspace, Tier 0 delivers these benefits within an open, standards-based platform that provides global, multi-protocol file and object access through a unified namespace. Data orchestration and protection are automated across storage from any vendor, on-premises or in the cloud.

The result is that Hammerspace enables customers to unlock greater performance and efficiency from the AI infrastructure they already own, including activating underutilized storage within existing compute resources. This reduces the cost and friction of scaling AI and other high-performance data initiatives while accelerating time to value.

This ultra-fast local NVMe capacity – which is largely underutilized – is a sunk cost, since it is already included within the price of GPU servers in both on-premises and cloud-based deployments.



Activating Tier 0: Local NVMe Storage within GPU Servers

Lurking within every GPU server is an untapped resource that can dramatically reshape GPU-computing workloads. Virtually all GPU servers—whether on-premises or in the cloud—run standard distributions of Linux that already include NFS and pNFS v4.2 performance capabilities, which Hammerspace has contributed to the open source community over the last eight-plus years. In addition, these GPU servers include at least two, and usually eight high-performance local NVMe drives. Some vendors offer servers that support up to 16 local NVMe drives.

Different manufacturers ship varying amounts and capacities of NVMe drives in their servers. Some allow the drive slots to be only partially populated, while others—including cloud-based GPU servers—include these drives by default.

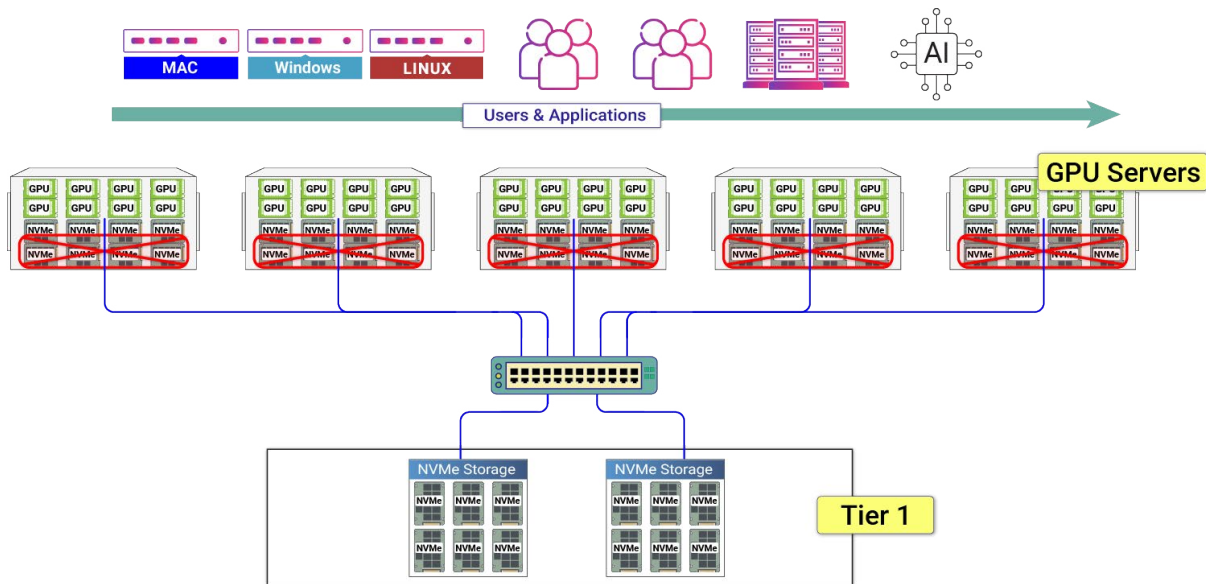


Fig. 1: Stranded NVMe capacity in GPU servers is mostly underutilized in GPU clusters. Data needed by the GPUs must go across the network to/from external storage, which slows down workloads, and decreases GPU utilization.

The problem is that while some applications can partially leverage this local capacity as a high-performance cache or temporary scratch space, particularly in a single server or with very small training datasets, this NVMe storage often becomes stranded and siloed in clusters of two or more GPU servers or when working with larger datasets. The reasons for this are as follows:

- **Capacity is siloed:** Local NVMe drives are isolated within each server. GPU servers cannot share this capacity across the cluster.
- **Lack of shared access:** Without some way to bridge these drives to create a shared storage environment across servers, any data on local storage must be manually copied, adding overhead, latency, and operational complexity.
- **Limited data protection:** Because NVMe storage on GPU servers lacks built-in redundancy across a cluster, there's a high risk of data loss or disruption if a server crashes, or is disrupted in other ways.



Large SSD drives and larger GPU server clusters change the equation

As SSD drive densities increase and the number of GPU servers in a cluster grows, the landscape changes drastically. As of early 2026, GPU server vendors ship systems capable of supporting large local NVMe capacities, commonly in the tens of terabytes per server when configured with multiple SSDs—for example, $8 \times \sim 3.8\text{TB}$ NVMe in leading AI servers today.

Drive densities continue to increase, with 30TB and 60TB enterprise SSDs now broadly available. Higher-capacity NVMe drives, such as 122-TB-class devices and beyond, are being introduced as manufacturers push density limits further. However, in the current supply-constrained environment, availability and lead times for high-capacity drives remain uneven across vendors and regions.

This means that even with the smaller drive capacities typically deployed in servers today, an eight-server cluster can contain nearly a Petabyte (PB) of ultra-fast NVMe storage in default configurations. For configurable GPU servers, upgrading the same eight-server cluster to 30TB drives would increase the available capacity to 3.8PB.

What is game-changing about the Tier 0 innovation is that this extreme performance is available within the existing GPU servers and power budget, without the need to add external high-performance storage systems with associated switching, rack space, power overhead, and other costs. In a period of constrained SSD supply and extended lead times, this also reduces dependence on acquiring additional flash capacity.

On a cost-per-terabyte basis, expanding NVMe capacity within existing GPU servers—such as populating unused drive bays—is materially less expensive than adding equivalent capacity through external storage arrays¹. Even when purchasing new enterprise SSDs at current market prices, internal server-based NVMe typically delivers lower total cost than external flash platforms due to the elimination of additional enclosure, switching, power, and infrastructure overhead.

But price per terabyte alone does not tell the full story. The more meaningful comparison is price/performance. When evaluating the effective cost of delivering data to GPUs, external Tier 1 NVMe storage carries a significantly higher fully loaded cost per GB per second than internal Tier 0 storage, even when assuming a 200Gb/s network. That fully loaded cost includes not only SSD media, but also enclosures, power and cooling, rack space, and high-speed networking infrastructure required to move data between storage and compute.

Rather than adding new external flash arrays, organizations can increase performance using server-local NVMe they already own.

¹ Enterprise-class SSD pricing and availability vary significantly based on capacity, volume, manufacturer, and prevailing market conditions. In 2026, flash pricing and lead times are highly volatile due to structural supply constraints. The comparisons in this paper reflect a fully loaded cost model that includes SSD media, enclosures, power and cooling, rack space, and networking infrastructure required to deliver performance to GPUs, rather than relying on static price-per-terabyte assumptions.



In the current supply-constrained environment, flash pricing and availability are highly volatile. Under these conditions, the advantage of activating NVMe capacity already installed within GPU servers—or expanding capacity within those same servers—becomes even more pronounced. Rather than depending on additional flash arrays, organizations can increase performance using storage they already own or can deploy directly within existing compute infrastructure.

Increasing network speeds to 400Gb/s or 800Gb/s may narrow performance gaps in specific scenarios, but doing so materially increases infrastructure cost and power consumption. The structural advantage of Tier 0 remains: local NVMe avoids the network bottleneck entirely, delivering higher effective throughput at lower total cost.

As will be detailed below, the implications of activating such massive volumes of high-performance capacity, combined with the growing numbers of GPU servers being deployed within AI clusters, are prompting organizations to rethink their IT architectures and their investments in external high-performance storage and networking infrastructure.

The problem is the network bottleneck

The key point here is that any external storage, however fast, is only as fast as the speed of the network connection between the servers and storage. This is not news, and has always been the case in all IT environments.

But the implications of this with GPU-heavy workloads for AI training and inference are far more serious than in traditional enterprise IT use cases. If faster network links are required to keep GPUs fully utilized, infrastructure costs increase proportionally. And the bottleneck problem only gets worse as data volumes explode with larger GPU clusters, and with the added traffic caused by more complex models and larger context lengths that saturate server system memory.

Even using expensive 400Gb/s InfiniBand switches and the fastest external storage available, **writing data across the network takes ~2.5 times longer** than writing directly to local NVMe Tier 0 storage. This delay means valuable GPU cycles are wasted waiting for data to arrive. And that's true even when caching or other token optimization techniques are in use.

If the network is 100Gb/s, it takes nearly **ten times longer** to write the same amount of data to external Tier 1 storage as it does when writing to Tier 0 within the servers.

There are variables inherent in these numbers that can optimize the data flows, such as protection methods and affinization strategies, which intelligently places frequently accessed data as close as possible to the GPU that needs it. But the point remains: **any data that needs to move across the network is going to be slower than utilizing the local NVMe** already built into the servers, as shown in the tables below.

In cloud-based instances the difference is even more dramatic: In AWS, 500GB checkpoints in a cloud-based eight-GPU H100 instance will write to local NVMe in ~4.5 seconds. If it is written to high-performance EBS (io2) storage in AWS, the same checkpoint takes ~139 seconds (~2.3 minutes), or **more than 30× longer**. Writing checkpoints via the more affordable AWS EBS (gp3) storage will take ~9 minutes, **or nearly 124 times longer**.



Microsoft Azure offers a very fast Ultra disk, which has speeds equivalent to an 80Gb/s on-premises network, which is very fast for cloud offerings. **But this is still 11 times slower than leveraging Tier 0 within an Azure cluster.**

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)
	200 Gb/s Network	25 GB/s	22.5 GB/s	~22 sec (~1 min)
	400 Gb/s Network	50 GB/s	45 GB/s	~11.1 sec

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

Checkpointing use cases are particularly telling, since they are needed to save the model's current state to protect long-running jobs against possible node failure. The problem is that in many existing training frameworks the GPUs must sit idle while waiting for the checkpoint to be written to storage. In cloud-based GPU workloads this idle time is particularly painful, since you're paying for the GPUs whether they are being used or not.

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)
	AWS EBS (io2)	4 GB/s	3.6 GB/s	~138.9 sec (~2.3 min)
	AWS S3 (via high-speed network)	1 GB/s	0.9 GB/s	~555.6 sec (~9.3 min)
	Microsoft Azure Ultra Disk	10 GB/s	0.9 GB/s	~55.6 sec
	Microsoft Azure Premium SSD v2	1.2 GB/s	1.08 GB/s	~462 sec (~7.7 min)
	Oracle - OCI Block Volume	1 GB/s	0.9 GB/s	~555.6 sec (~9.3 min)

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

Newer training frameworks address this issue with asynchronous checkpointing to minimize GPU downtime. But the performance differential between Tier 0 and external storage is not only about checkpointing. The same performance differences seen in the examples above also affect all other data traffic between the GPUs and external storage. And the volume of data can be far greater than a 500GB checkpoint.



Let's be clear here: purchasing the most expensive external storage and 800Gb/s networking options can approach the throughput of Tier 0 NVMe storage for on-premises environments, albeit at the cost of several thousand dollars per network port, not to mention the added power requirements and other overheads.

But with Tier 0 this performance can be achieved with NVMe capacity that organizations have already paid for within their GPU servers, and using the network infrastructure they already own. Even with existing 100Gb/s networks that list for about \$200 per port, your workloads will still move faster by utilizing Tier 0.

In large environments with hundreds of GPUs, the cost penalty of adding that many more expensive network ports and external storage units explodes to extreme scales.

Simply put, adding the substantial expense of external networking and storage to an AI project—especially when you already own higher-performing local NVMe storage in your GPU servers—will directly reduce the ROI of your AI investment. And in cloud-based or hybrid on-premises/cloud deployments, the impact is even greater.

With Tier 0, and especially with larger GPU clusters and drive capacities, you can get the best of both worlds: The highest performing storage available on the market to keep your GPUs fed and your workloads moving, while using the NVMe capacity and network infrastructure you already own.

With Tier 0 this performance can be achieved with NVMe capacity that organizations have already paid for within their GPU servers, and using the network infrastructure they already own.

How to Activate Tier 0 Capacity within Your Server Clusters

When considering Tier 0, the obvious first question is this: If activating local NVMe storage within CPU and GPU server clusters provides such a dramatic performance and cost benefit, then why isn't everyone already taking advantage of it? Why is the storage industry focused primarily on the performance of external storage systems, or requiring customers to install proprietary client software on their application servers simply to optimize which data they feed to the GPUs?

The simple answer is that **it requires capabilities uniquely provided by the Hammerspace Data Platform to safely and effectively activate local NVMe capacity as ultra-fast Tier 0** shared storage across a cluster.

Hammerspace has created a data platform that simplifies adapting existing infrastructure to become AI-ready by eliminating storage silos and automating cross-platform data orchestration. Without requiring customers to install any client software, or alter their existing infrastructure, Hammerspace creates a high-performance parallel file system to provide unified file/object access to data on storage from any vendor, including local NVMe within servers, remote sites, and the cloud.



For the scale and performance needed for even hyperscale AI workloads, Hammerspace uniquely leverages the parallelism of Linux-standard pNFS v4.2 with Flex Files to automate data orchestration to the GPUs at extreme performance levels. As a standard that Hammerspace contributed to the Linux kernel years ago, the pNFSv4.2 client is already installed in every data center on the planet, and has been included in virtually all Linux distributions for over five years.

This has enabled Hammerspace to achieve the linear scalability needed by Meta, for example, to power its Llama 2, 3 and now Llama 4 large language models (LLMs) across many thousands of GPUs both in on-premises and cloud-based data centers.

Hammerspace achieves this performance without needing to route data through proprietary controller nodes, or requiring customers to install proprietary client software on existing user computers and application servers, and without the need to alter existing servers and storage from any vendor².

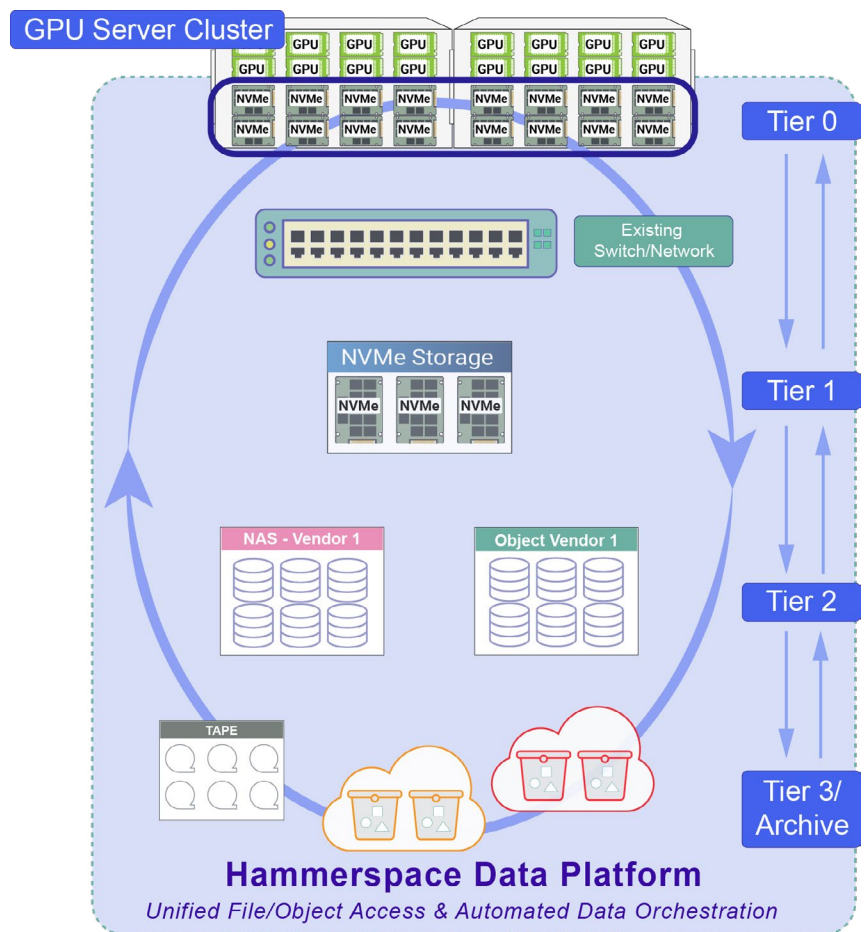


Fig. 2: Hammerspace activates the stranded capacity within the GPU servers, so that it is a seamless Tier 0 within a multi-tier, multi-vendor storage environment. Hammerspace data orchestration automates data placement and data protection, to maximize GPU efficiency.

The Keys to Unlocking Tier 0

1. Data protection & orchestration

To activate and protect local NVMe storage across a cluster of GPU servers, Hammerspace aggregates this capacity into a seamless global namespace that can comprise multiple tiers of any storage type from any vendor, including across multiple sites and clouds.

But spanning all the storage types in a global file system is not enough. The key Hammerspace superpower that makes this possible is the ability to automate non-disruptive file-granular data orchestration between Tier 0 and any other class of storage, both on-premises and remote.

²For more information on the industry-leading capabilities of pNFSv4.2 with Flex Files, please refer to the white paper "How to Adapt Legacy Infrastructure to Power GPU-Based AI/DL Workloads Using Parallel NFS v4.2"



Unlike the simple policies used by HSMs (hierarchical storage management) or other point solutions that are typically one-dimensional commands with limited options, Hammerspace enables administrators to create comprehensive service-level Objectives that can be finely-tuned based upon business logic, such as data durability, data placement, workflow automation, and other data services.

Multiple Objectives can be customized to accommodate a virtually unlimited number of business rules, use cases, and workflows, and may apply to all or intelligently-selected subsets of data across any storage type. The Objectives may be applied conditionally to determine with file/object granularity how data, directories, and/or buckets should be managed across any storage type from any vendor, including the cloud.

In the example above where a checkpoint file is written to local Tier 0, a service-level Objective would be set in Hammerspace to define how that checkpoint is to be protected, the level of data durability that is required, where it will be placed, etc.

For example, as soon as the file is written to Tier 0 Hammerspace would asynchronously copy or move that file downstream to an object store or any other target. Or in the case of clusters with larger Tier 0 capacities, the file could also be replicated to other nodes in the cluster for rapid recovery if needed. As many instantiations of the file as may be required for performance and/or durability can be determined with simple declarative Objectives.

With Hammerspace, these are not forked file copies that need to be wrangled later. Rather, they are instantiations of the same file, with shared file metadata that is globally accessible to users/applications and managed by Hammerspace across the entire unified namespace.

2. Data orchestration activates Tier 0 for all workloads

But Hammerspace data orchestration is not only used to protect checkpoint files. Leveraging the extreme performance Hammerspace delivers with pNFSv4.2 with Flex Files, Hammerspace Objectives can also automate any data movement into and out of Tier 0 to parallelize workloads across the GPU cluster.

Remember that the “east-west” inter-node networking speeds between the servers in a cluster is much faster than the “north-south” networking between the servers and external storage. This means that Hammerspace’s ability to stage data forward into Tier 0 close to the compute nodes, plus its ability to leverage pNFSv4.2 to further accelerate I/O, together result in multiple ways that activating Tier 0 provides a net benefit to any high-performance workload.

Additionally, since networks are bi-directional, Hammerspace can simultaneously move checkpoint files out of a Tier 0 (GPU node sending) to external storage or to another Tier 0 node in the background without impacting the bandwidth needed to feed the GPUs reading training data (GPU node receiving) from the same Tier 0.

But spanning all storage types in a global file system is not enough. The key superpower is the ability to automate non-disruptive file-granular data orchestration between storage types and locations.



3. Where are the GPUs? Enabling burst workloads to cloud-based GPU clusters

A significant problem for organizations trying to launch AI initiatives is the availability of on-premises GPUs. In addition, even if GPUs are available to purchase, the planned usage or available data center power envelope may not make owning the GPUs the best choice. Which is why many organizations are using third-party providers offering GPUs-as-a-service, or renting cloud-based GPUs. But both of these options are complicated, and require manually pushing files/objects to and from remote resources, and managing these workloads in yet another silo.

Hammerspace eliminates these issues, by extending its global file system and data orchestration between on-premises data centers and one or more clouds. It only takes a few minutes to spin up another Hammerspace cluster in any of the major public cloud providers, which can be scaled to whatever size is needed for the job.

Once a new Hammerspace instance is activated in the cloud, within minutes it automatically synchronizes file system metadata with the existing on-premises Hammerspace cluster. From then on, any cloud-based storage, including Tier 0 storage within cloud-based GPU clusters, becomes a seamless extension of the on-premises Hammerspace global file system.

Hammerspace Objectives can now orchestrate data to and from the cloud automatically, including Tier 0 within the cloud-based GPUs just as is done between any other storage class in one or more on-premises data centers. This multi-site capability can bridge as many as 16 on-premises and cloud-based Hammerspace clusters to create a single unified, high-performance namespace that can span existing or new storage from any vendor in a truly global data environment.

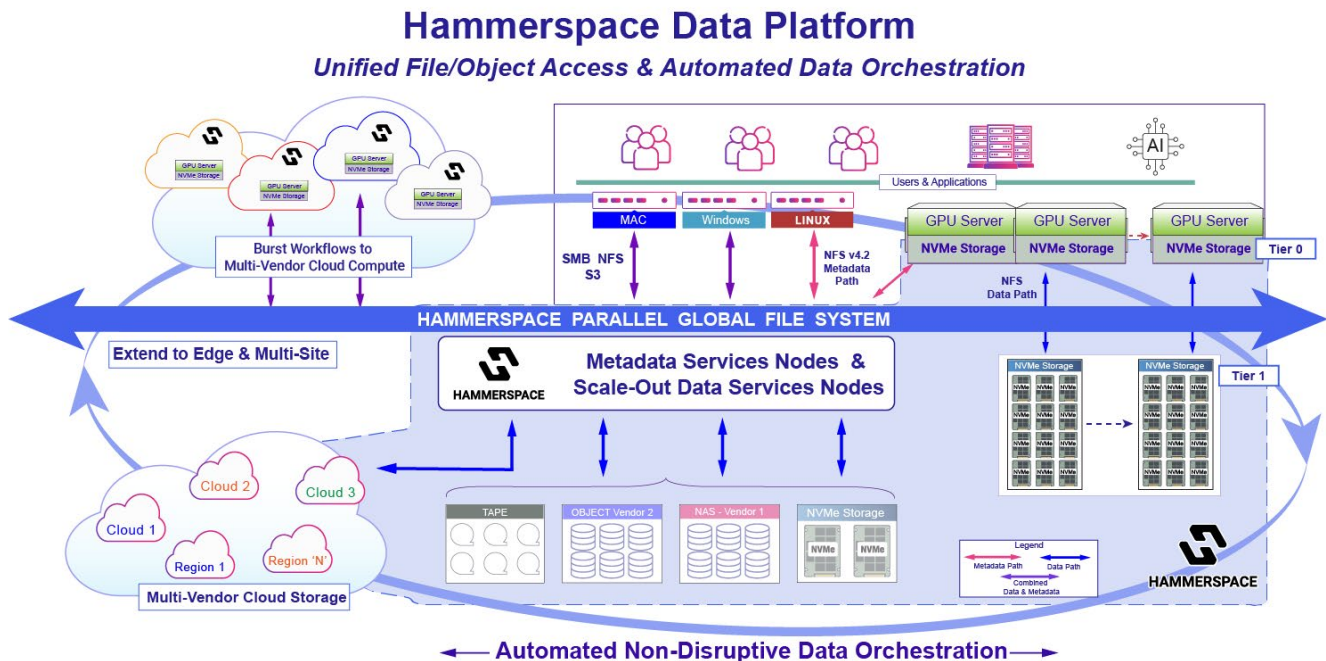


Fig. 3: The Hammerspace Data Platform activates Tier 0 within GPU servers, with unified file/object access and data orchestration globally across on-prem and cloud-based resources.



Users and applications see the file system via standard SMB, NFS, and S3 mount points as before, without needing to install client software or to alter their workflows. Hammerspace handles data placement and other data services, plus automates workflows in the background across all silos, all without interruption to users or applications.

And to maximize cost savings, when workloads are finished the cloud-based Hammerspace cluster can be immediately decommissioned to reduce cloud compute fees. Orchestration of data will have already happened transparently as a background operation, ensuring the output data generated by the GPUs is automatically protected and placed wherever it is needed, in any storage type or location.

4. Shared standards-based file/object access

As mentioned above, Hammerspace requires no proprietary client software to be installed on users' computers or application servers, and no alterations to existing storage. Users may not even be aware that Hammerspace is there. Both end-users and applications simply see shares in the global file system via standard NFS, SMB and/or S3 protocols exactly as they did before, except now they have unified access to the global namespace of all data, across all storage types, silos, and locations.

Only the data that users and applications have permissions to see is visible to them via shares across the entire global file system. Applications that expect to see S3 buckets continue to see S3 buckets, including data from any storage type. Applications that expect to see a file hierarchy via NFS or SMB shares see that same data presented in a standard file/folder structure.

The storage type the data happens to reside on at any given moment is governed by the Hammerspace service-level Objectives based on business rules, and is completely transparent to users and applications. In fact, Hammerspace incorporates a key feature called Live Data Mobility that it contributed into the Linux kernel as part of pNFSv4.2 with Flex Files so that even files that are open and actively being read or written can be orchestrated across storage tiers transparently as a background operation. There is no interruption to application or user workflows during this process.

Hammerspace's ability to provide universal access via standard protocols across any storage, combined with its ability to orchestrate data transparently at file- and object-level granularity, is critical for AI workloads. This is particularly the case for organizations with large quantities of unstructured data that they want to activate.

Rather than migrating entire volumes to new repositories, Hammerspace orchestrates only the specific files or objects required for the job, again without creating forked copies or interrupting other users or applications.

Gone are the days of having to perform costly migrations and pushing data copies into net-new repositories, which themselves become new isolated silos.

Applications that expect to see S3 buckets, see S3 buckets including any or all data on any storage. Applications that expect to see a file hierarchy via NFS or SMB shares see that same data presented in a standard file/folder structure.



What about HCI and CPU-based servers?

Although this paper focuses on GPU-based workloads for AI use cases, CPU-based servers often ship with the same ultra-fast local NVMe drives that can be activated as Tier 0. This is also true for low-power ARM CPUs.

These server types are often confused with Hyperconverged Infrastructure (HCI), which typically combine compute and storage in a single unit. HCI systems are predominantly used for structured data, VMs, and other block-oriented I/O patterns that are generally not optimized for CPU-intensive unstructured data workloads. And while HCI systems can use distributed storage architectures and other tiering solutions, their focus on block-oriented use cases makes them unsuitable for processing unstructured data.

But structured data only makes up about 20% of enterprise datasets, and according to industry analysts, is growing at only about 20% per year. On the other hand, unstructured file/object data makes up about 80% of all enterprise data and is growing annually at an estimated 65%.

There are numerous CPU-intensive use cases that need large volumes of unstructured data, and that can take advantage of the same performance and cost improvements of Tier 0 as the GPU-computing workloads outlined above. Among them are other AI use cases, which include CPU-based AI inference, predictive analytics, or machine-learning pipelines dealing directly with unstructured datasets.

However, beyond AI use cases there are numerous other CPU-based application workflows that demand high performance and that can directly benefit from the ultra-low latency and enhanced throughput provided by activating Tier 0 on local NVMe drives within CPU servers.

Some of these use cases include:

Real-time analytics and data streaming	Rapid processing of log data, sensor inputs, or clickstream data to generate insights instantly.
High-performance search and indexing	Fast indexing and search across large-scale document repositories or datasets (e.g., Elasticsearch, Splunk).
Video and media processing	Low-latency editing, transcoding, rendering, and content delivery requiring quick, frequent data access.
Financial trading and modeling	Real-time risk analysis, algorithmic trading, and financial modeling needing immediate access to large datasets.
Genomics and bioinformatics	Rapid processing of genome sequencing data, imaging data, or bioinformatics workloads demanding near-instantaneous



The benefits for CPU-based Tier 0 include the same advantages highlighted in the GPU-based examples above:

- Lower cost by activating ultra-high-performance local NVMe within servers rather than using expensive network paths to external arrays.
- Significant performance gains and lower latency compared to external storage, directly benefiting these workloads.

Additional Hammerspace performance advantages for CPU-based workloads

In addition to leveraging the extreme performance made possible with pNFSv4.2 with Flex Files and activating Tier 0 within CPU-based servers, Hammerspace takes advantage of its tight integration with standard Linux to provide another direct performance advantage for CPU-based workloads.

HCI systems and some other vendor solutions use a polling-based methodology to process I/O. This user-space approach keeps the CPU constantly polling for incoming data. This polling occurs constantly, even when there is no data present, and consumes a significant amount of CPU resources. The result of this excessive load on the CPUs is increased latency, and degraded performance.

Hammerspace takes advantage of standard NFS, which is Linux kernel-based, to accelerate I/O with an IRQ approach (IRQ=interrupt requests) that significantly lowers CPU overhead and reduces latency.

Unlike the constant CPU load needed by polling-based approaches, the kernel-based IRQ approach used by Hammerspace means the CPU is alerted only when the data actually arrives, enabling far more efficient event-driven processing.

The result is faster, lower-latency I/O and more efficient CPU usage compared to user-space solutions that rely on CPU-intensive polling loops.

On the horizon: Coming soon to a Linux distribution near you

Another key innovation that Hammerspace has recently contributed upstream into Linux—and which will directly improve performance of local NVMe in Tier 0 even more—is an NFS Protocol Bypass in the Linux kernel.

Known in the Linux community as LOCALIO, this enhancement adds intelligence to the NFS clients and servers to detect when they are running on the same host. This eliminates unnecessary latency of going through the NFS protocol stack in the kernel to loop back to the local file system and storage. The feature was implemented in a robust way that works even if the NFS client and server are running in containers.

LOCALIO was initially released in Linux kernel 6.12 in November 2024, with additional enhancements added in kernel 6.14 in March 2025. It was subsequently incorporated into Red Hat Enterprise Linux (RHEL) 10 in May 2025, marking its availability in a mainstream enterprise distribution.



In benchmark testing, preliminary results show that bypassing the NFS protocol has a significant impact on reducing CPU utilization within the server. In addition, testing has so far shown a marked boost in read performance within Tier 0 when using LOCALIO. Early testing also indicates we'll see write performance improvements, although how much is still being quantified. This testing is ongoing based upon the latest kernel updates, and will be documented when results are ready.

Combined with other upcoming enhancements to increase affinization of data placement across the GPU cluster, the LOCALIO feature has already shown that it will provide significant performance benefits even over and above what Tier 0 delivers today.

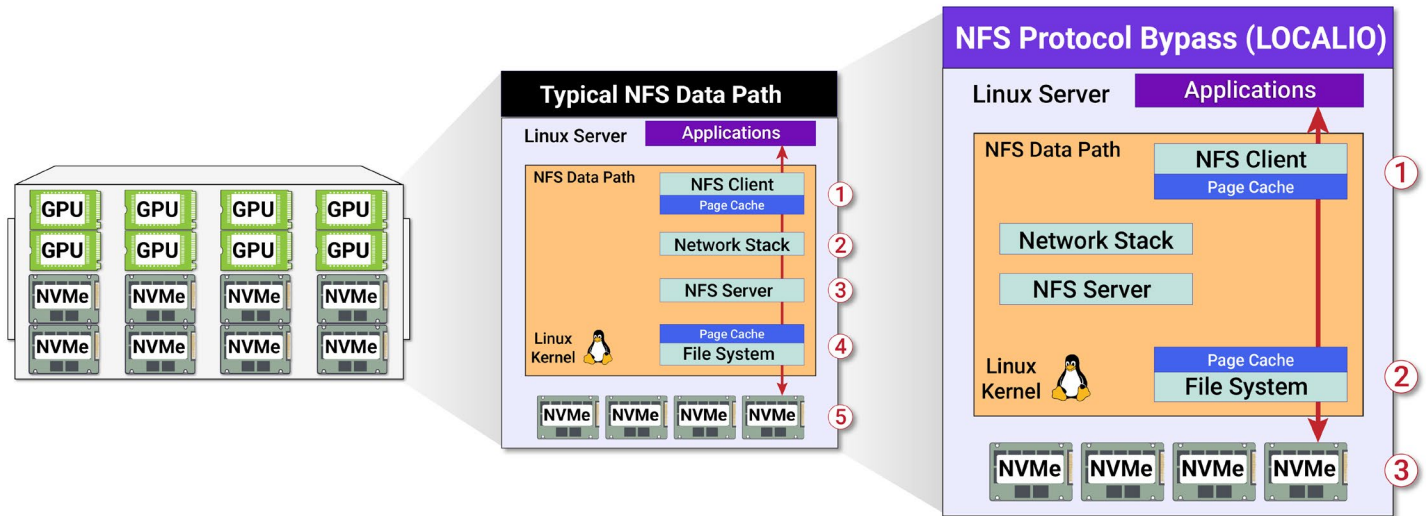


Fig. 4: Hammerspace engineering developed LOCALIO and contributed it into the Linux kernel to further increase performance of local NVMe in GPU and CPU-based servers.

Although LOCALIO was included in Linux 6.12, which was released as the Long-Term Support (LTS) version in November 2024, it takes some time for enterprise Linux distributions to adopt new releases. But the demand for this feature is significant, and the most recent release of Red Hat Enterprise Linux (RHEL 10) in May 2025 was based upon kernel 6.12, and also included the kernel 6.14 enhancements to LOCALIO.

It will take some time before organizations adopt to RHEL10 and other distributions as they upgrade to the latest Linux kernel, but when they do so the improvements Hammerspace has contributed will already be there waiting for them to take advantage of. This will further improve the dramatic performance and efficiency improvements available now with Tier 0.

Standards are the key

Hammerspace has a long history of contributing key performance enhancements into Linux, with over 2,400 features and patches contributed into the open source kernel in recent years. Trond Myklebust, the Hammerspace CTO, has been the Linux NFS client kernel maintainer for more than 20 years, and in that role has ensured that Hammerspace's work adheres to the standards maintained in the community.



Hammerspace Senior Principal Software Engineer Mike Schnitzer is also a Linux kernel maintainer, responsible for the upstream Linux kernel's Device Mapper (DM) subsystem. And Tom Haynes, another Hammerspace software engineer, is a co-author of RFC 8435, which defined the pNFS Flexible File Layout in 2018, the key feature of pNFS that was adopted into the v4.2 spec in 2019.

This focus on open standards aligns with Hammerspace's vision to overcome the limitations of proprietary vendor silos and vendor lock-in for enterprise data environments, while maintaining compatibility with existing infrastructure. On the user/application side, customers never have to install proprietary software on their application servers, or alter their user's computers. And on the storage side, Hammerspace is compatible with any storage, from any vendor.

As importantly, the extreme performance improvements Hammerspace has engineered into standard Linux—and which are part of all major distributions—are also tightly integrated within Hammerspace software, which is uniquely positioned to take advantage of them.

Summary

Organizations across industries, both public and private, are grappling with how to retool their IT infrastructures to accommodate the explosive promise of AI. Gone are the days where enterprise IT was a mostly linear proposition, with a one-to-one ratio between applications and data, where unstructured data would rapidly age out into an archive.

While organizations generally realize they need an AI strategy, the return on investment in AI initiatives is far less certain, particularly for those who are not among the hyperscale elite. But a quick scan of the industry trade press often would lead you to think that the only way to launch an AI project is by creating a second AI-focused IT architecture alongside the infrastructure they already own.

Such solutions promote purchasing net-new high-performance storage repositories and expensive high-speed networking, or modifying customer's application servers with proprietary client software, plus multiple other expensive and vendor-locked solutions.

All of these offerings perpetuate the problems of the past, where proprietary data silos fragment access and hinder efficient utilization of data. Such an approach leads to data gravity and substantially higher costs in both CAPEX and OPEX.

Hammerspace has already proven that its standards-based approach enables organizations to unlock greater performance and efficiency from their existing IT infrastructure. For organizations embarking on or expanding their AI journey, this means AI-focused resources can now be additive and tightly integrated with existing IT infrastructure, rather than needing to be an isolated net-new silo.

For organizations this means AI-focused resources can now be additive and tightly integrated with existing IT infrastructure, rather than an isolated net-new silo.



The Hammerspace Data Platform has demonstrated this vision in real-world environments that span on-premises and multi-cloud by enabling unified file/object access, and high-performance data orchestration in organizations of all sizes, scaling to extreme volumes and performance levels across on-premises and hybrid cloud environments, in all verticals including HPC and hyperscale AI environments.

These capabilities and this real-world experience are what make Tier 0 a solution that is having a game-changing impact on high-performance workloads for both GPU- and CPU-based computing.

Hammerspace enables organizations of all sizes in any industry to take advantage of the unprecedented performance boost Tier 0 achieves, surpassing the fastest external storage types available on the market. And this is matched only by the incredible cost savings possible by activating this underutilized extreme-performance storage tier.

The fact that this activates a performance tier that you may already own is icing on the cake.

This is the Hammerspace innovation.

