Futurum

HAMMERSPACE

Hammerspace Tier 0: Unlocking Greater Efficiency in GPU-Driven Computing

APRIL 2025

Executive Summary

Technological advances within High Performance Computing (HPC) and AI over the past few years have led to a rapid increase in GPU-driven computing. The potential of technologies that are heavily reliant on GPU computing, such as new generative AI applications, has challenged a wide range of organizations to quickly meet new and demanding infrastructure requirements. IT organizations are tasked with standing up such environments as quickly as possible, yet face several challenges related to cost, efficiency, power, and performance.

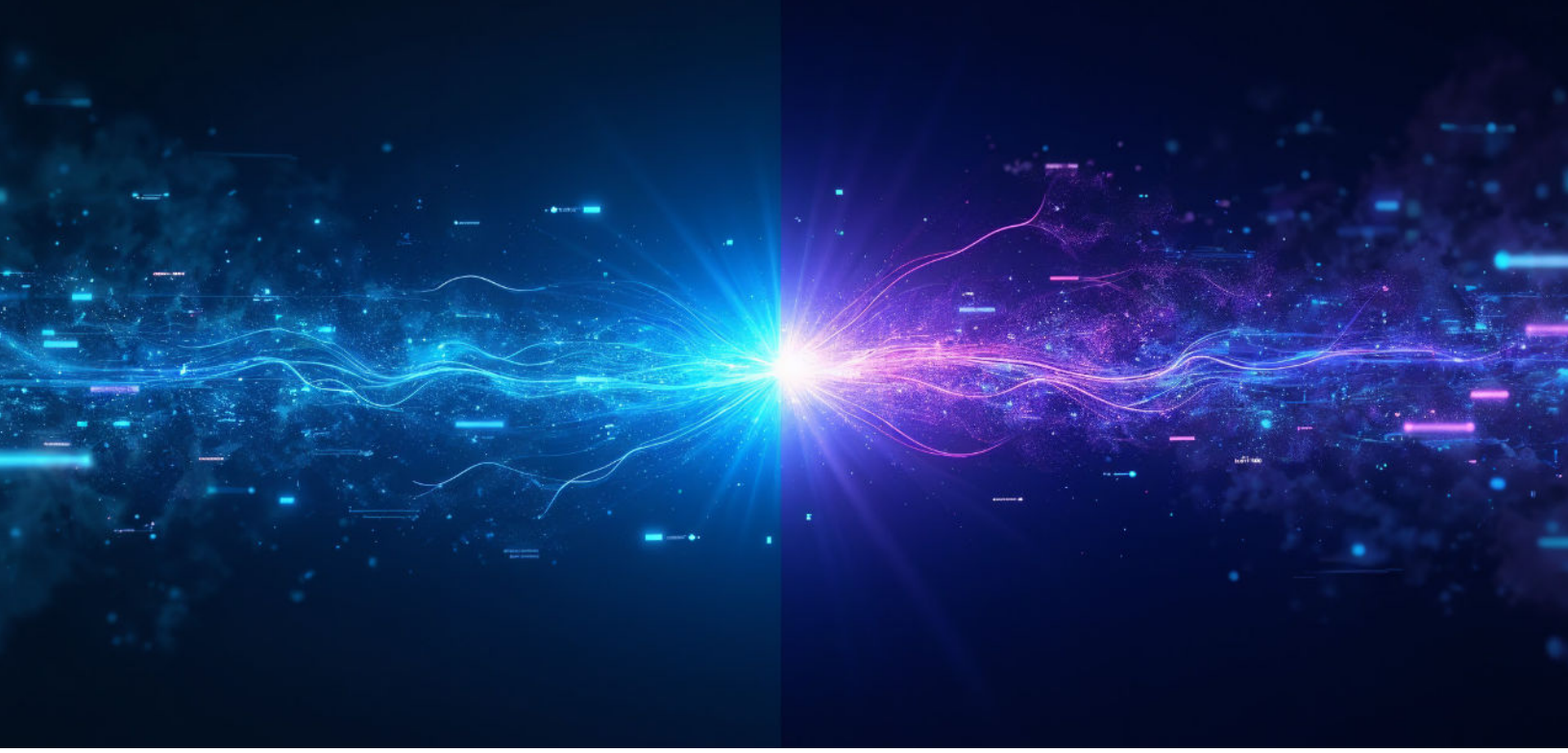
This paper explores Hammerspace Tier 0 storage that extends the Hammerspace Global Data Platform software to activate local extreme performance NVMe storage within GPU servers, unlocking previously siloed capacity and easing the challenges of GPU-based computing.

Challenges of Modern GPU-driven Data Centers

For modern data centers, large-scale utilization of GPU-based compute is critical to supporting AI and HPC workloads. Due to the recent innovation in generative AI, the requirement for GPU-based compute has grown across nearly all industries. While GPU-based compute has played a crucial role in HPC environments for some time, the widespread interest in leveraging generative AI has increasingly introduced GPU compute requirements to enterprise environments.

The GPU-driven data centers required to support innovative AI and HPC workloads present significant challenges to IT organizations. The overall infrastructure requirements are often incredibly large, including racks of GPU servers and petabyte-scale storage. Both compute and storage requirements involve significant cost, power, and space. In addition, procuring this infrastructure can be a challenging and time-consuming process, with long lead times causing significant delays and disrupting critical AI initiatives.

In order to overcome these challenges and meet their AI and HPC goals quickly, efficiently, and cost effectively, it is crucial for organizations to explore new approaches to deploying infrastructure.



Hammerspace Tier 0 – A New Tier of Data Orchestration

A key area of inefficiency within GPU-based computing use cases is data storage. The AI and HPC workloads increasingly powered by GPU server platforms typically involve large, petabyte-scale data capacities with strict performance requirements. Although modern GPU servers powering these workloads typically include significant capacities of direct-attached NVMe storage, such capacity is often left unused. Instead, organizations typically rely solely on large network-attached storage systems, adding additional cost, power, space, and complexity.

The local storage within GPU servers, while consisting of terabytes or petabytes of low-latency NVMe flash, is typically considered impractical for use in clustered environments. This local storage within each server is siloed, providing capacity only to GPUs within its own server, rather than to the entire cluster. The siloed nature of direct-attached storage additionally introduces the risk of data loss, as data is not protected and may be lost if that server goes down. Utilizing the direct-attached storage found in each GPU server alongside a shared storage environment also presents a challenge, as the process of manually moving and monitoring data becomes increasingly complex as the number of servers in the cluster grows. For large AI and HPC environments, the result is petabytes of underutilized and often wasted storage and an inefficient data architecture.

Hammerspace Tier 0 extends the Hammerspace Global Data Platform to include the captive high-performance storage capacity within GPU servers, unlocking previously stranded data and introducing a new ultra-fast tier of data storage for AI and HPC workloads. By unlocking previously stranded data capacities, Hammerspace Tier 0 can lower infrastructure costs, increase storage performance, and accelerate time to market.

Unlocking this previously stranded storage becomes increasingly notable when considered alongside the development of high-density NVMe devices. While even GPU clusters configured with relatively small drives, 4 or 8 TB drives can amount to petabytes of unutilized storage, NVMe SSDs are available at increasingly large capacities, ranging from 30-122 TB. When configured with 30TB drives, an 8-node cluster with 16 SSDs per server can support up to 3.8 PB of storage. If configured with 122 TB drives, this same cluster can support over 15 PB.

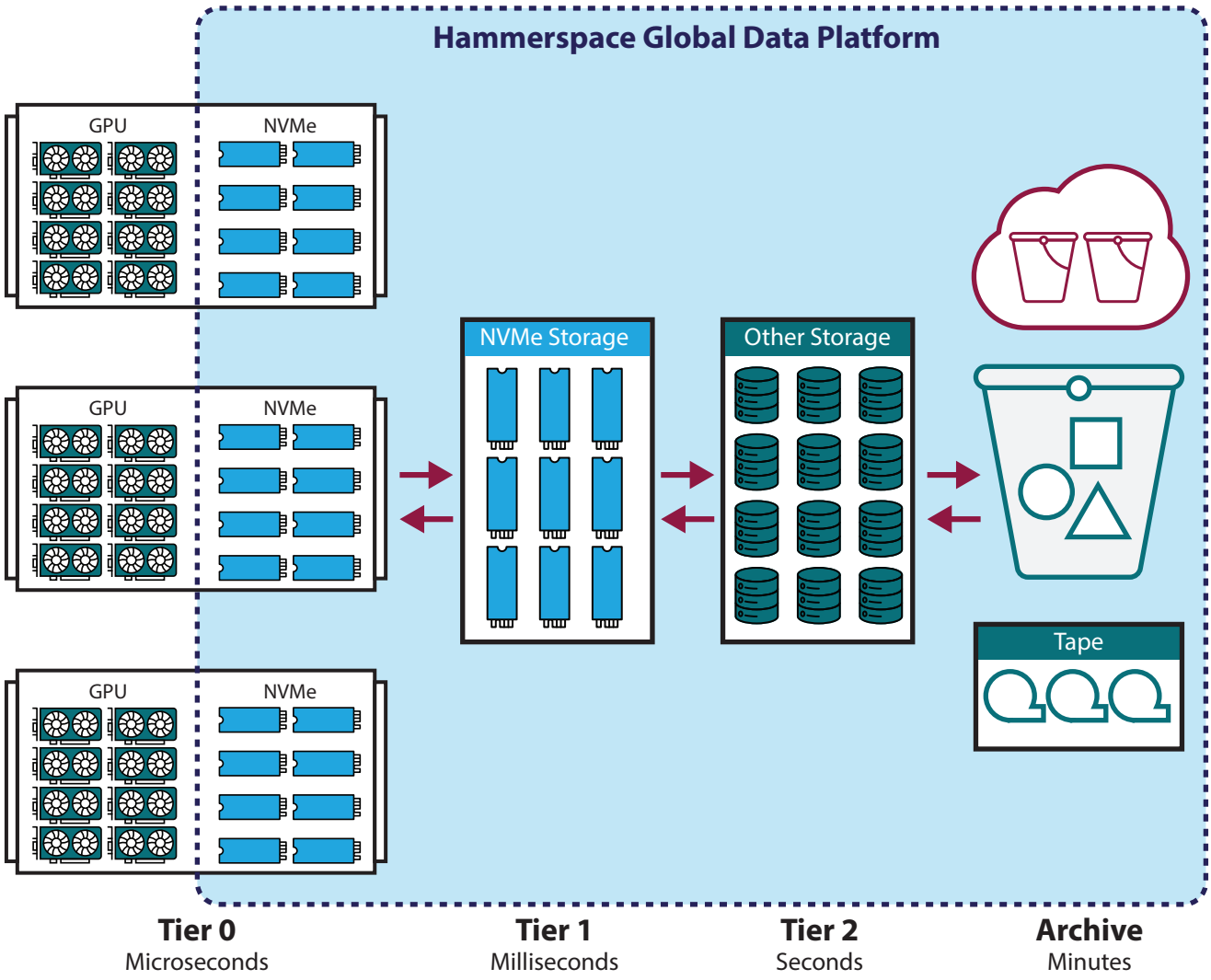


Figure 1: Hammerspace Global Data Platform

Hammerspace Tier 0 storage provides a new, flexible storage tier within the Hammerspace Global Data Platform that scales linearly with GPU servers. Tier 0 storage can be utilized either standalone, in environments in which the GPU server storage capacity meets all storage needs, or in combination with additional storage tiers. In multi-tiered environments, Tier 0 storage acts as the highest performance tier, due to the low latency of direct attached storage, and can be combined with additional tiers of various performance consisting of flash, disk, tape, or cloud-based storage. Data orchestration within the environment is performed by Hammerspace software.

Hammerspace Tier 0 Technical Overview

Hammerspace software provides a unified, global data platform capable of spanning across heterogeneous storage systems. With Hammerspace Tier 0, local storage within GPU servers is additionally made available as a protected shared storage tier within the Hammerspace global namespace. To achieve this, the local storage within each GPU server is configured as an NFS export, making it available as a volume within the Hammerspace environment. By making this storage available to the Hammerspace Global Data Platform, it is no longer siloed to each individual server.

As with all other storage volumes configured within a Hammerspace global data environment, Tier 0 storage is protected and orchestrated via Hammerspace software. This is achieved with user-defined policies known as Hammerspace Objectives. Objectives enable administrators to define flexible declarative policies to meet various needs, including data availability, tiering, archival, business workflows, and more, all down to a file granular level. For data protection, policies can be specified with objectives that define the data availability and durability requirements as a "number of 9's." Hammerspace software then intelligently places data across volumes to meet these requirements.

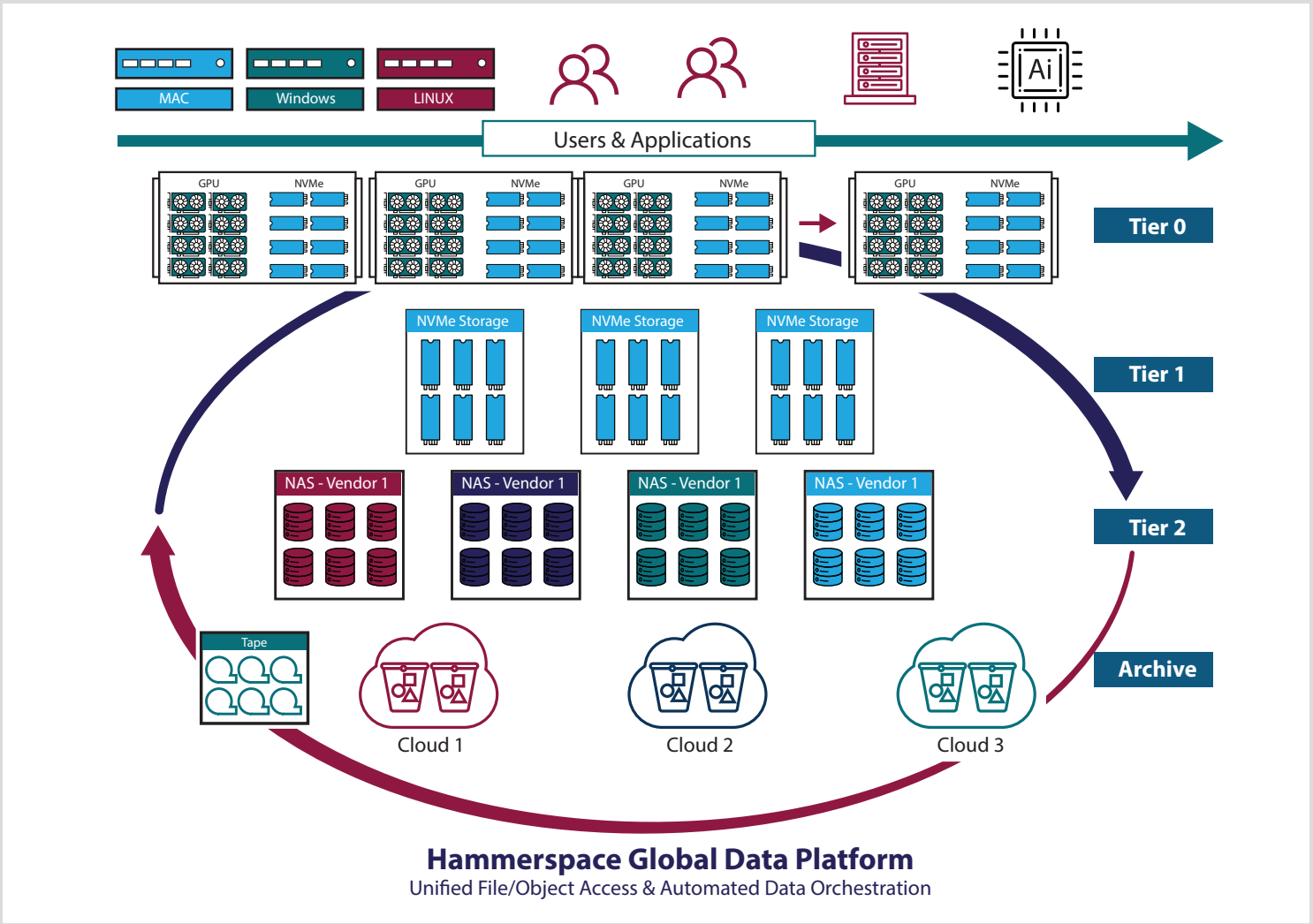


Figure 2: Hammerspace Global Data Platform

Data tiering and migration policies can similarly be specified, and Hammerspace facilitates data movement as a non-disruptive background process. By incorporating local GPU server storage into the Hammerspace platform with Tier 0, the data protection and data movement concerns typically associated with otherwise siloed local storage are easily addressed with Hammerspace Objectives. Data on Tier 0 storage is protected alongside all other data in the Hammerspace environment and automatically moved between Tier 0 and other tiers in the environment non-disruptively as needed.

Notably, Tier 0, along with the greater Hammerspace Global Data Platform, does not require the installation of any custom clients or deployment of specialized hardware. Instead, Hammerspace follows a standards-based approach, enabling users to leverage the unified global namespace via standard NFS, SMB, and S3 protocols. In large-scale AI and HPC environments, Hammerspace leverages standard pNFSv4.2, included in virtually all Linux distributions, to achieve the parallelism and scalability required of GPU-driven compute. The addition of Tier 0 seamlessly adds an additional, high-performance layer to Hammerspace's global file system.

Benefits of Hammerspace Tier 0

Hammerspace Tier 0 offers IT organizations a new approach to deploying data storage in GPU environments. The Futurum Group has identified several key areas in which Tier 0 storage provides notable benefits capable of solving critical challenges faced by IT organizations deploying AI and HPC data centers.

Improved Performance

For AI and HPC workloads, high-performance storage is critical. Due to the requirements of a large, shared storage environment, the traditional approach has been to maximize the performance of network-attached storage. While there are several approaches to achieve high performance with a networked solution, transmission over a network will always introduce additional latency.

As AI workloads are becoming more complex, the added latency introduced by network-attached storage becomes an increasing challenge. Large model sizes and complex inferencing workloads, which utilize large context lengths, can exceed the capacity of GPU memory, requiring frequent data transfer between external storage and GPU servers. Even with high-speed networks, this data transfer introduces significant latency and can hinder application performance. For many AI applications, such as those requiring real-time inference, this added latency is severely detrimental.

Direct-attached storage can offer greater performance by avoiding latency-inducing networking altogether. Hammerspace Tier 0 introduces a practical approach to utilizing direct-attached storage and its performance advantage in a shared storage environment.

Hammerspace claims data can be written to local NVMe storage approximately 2.5x faster than over a 400Gb/s Infiniband network, and nearly 10x faster than over a 100Gb/s network. In Cloud environments, the performance impact of Tier 0 can be even more dramatic. Hammerspace estimates that data can be written to Tier 0 storage in cloud-hosted GPU servers between 31x and 124x faster than to separate cloud storage instances, depending on the storage class selected. With Tier 0, this performance enhancement is achievable without adding additional infrastructure, power, cost, or networking requirements.

Time to Value

A key challenge facing many organizations deploying new AI and HPC applications is time. AI in particular has become a strategic priority, and organizations are highly motivated to deploy new AI applications as quickly as possible. In addition, the GPU servers that power these applications are expensive, and once acquired, it becomes vital for organizations to realize their value as quickly as possible. The process of purchasing and configuring large external storage systems capable of supporting such applications can introduce additional time and complexity, further delaying the total time to value.

Hammerspace Tier 0 provides a streamlined approach to deploying GPU data centers by enabling organizations to utilize the storage included in their GPU servers that they have already purchased. Due to increasingly dense SSD capacities, GPU clusters often include petabytes of high-performance NVMe storage. By unlocking this previously isolated storage capacity, Tier 0 equips organizations with a fully capable storage solution immediately upon acquiring GPU servers. Additionally, the standards-based approach of Hammerspace Tier 0 requires no custom protocols or proprietary client software that may introduce further complexity and delay.

Reduced Storage Costs

IT organizations deploying GPU data centers are additionally challenged with high costs, directly from both the GPU servers and the associated data storage. The storage requirements of GPU data centers running AI and HPC workloads are typically among the most demanding, requiring both high capacity and high performance. Storage for such environments is typically required to accommodate petabyte-scale capacities while additionally providing performance capable of fully utilizing high-end GPUs. Fulfilling these storage requirements often leads to significant costs for storage, rack space, networking, and most importantly, power, which may limit organizations from realizing their AI and HPC goals.

By leveraging Hammerspace Tier 0 storage, organizations can significantly lower their total storage costs by utilizing direct-attached storage capacity that has already been purchased. With Tier 0, this capacity, which often consists of petabytes of NVMe SSDs, becomes readily available to meet the storage needs of GPU-based applications. By unlocking this large storage capacity, Tier 0 can reduce the burden of purchasing additional storage and lower the overall TCO of the data center. Since each GPU server includes its own storage capacity, this solution scales linearly, allowing the storage savings to continue even as additional GPUs are added.

For environments in which the total capacity requirements exceed the total capacity accumulated across GPU servers, external storage can still be added to the Hammerspace environment. By utilizing Tier 0 storage, however, organizations may be able to lower the cost of any additional external storage required. The extreme performance of direct-attached storage, which significantly exceeds that of external flash arrays, enables Tier 0 to function as a high-performance storage tier, which can be backed by slower, more economic tiers as needed. As part of the Hammerspace Global Data Platform, data will be intelligently moved to the appropriate tier as needed. This enables organizations to minimize purchases of high-performance external storage solutions and instead leverage lower-cost tiers, such as object or cloud storage, to further optimize storage.

Greater GPU Utilization

Due to the high cost and power consumption of GPUs, IT organizations deploying AI and HPC applications must put an emphasis on maximizing GPU utilization. Hammerspace Tier 0 can help organizations reach this goal through its high performance, local storage.

For AI training workloads in particular, Tier 0 offers a compelling approach to model checkpointing, which can further aid in maximizing GPU utilization. Model checkpointing is a vital part of model training, which preserves model weights at various

stages of the process to ensure the availability of the model and enable access to previous versions as needed. Although checkpointing is crucial, it involves a tradeoff when considering GPU utilization, as GPUs are left idle while checkpoints are written to storage. Model checkpoints are often large, and writing them to external storage can be time-consuming, reducing GPU utilization and the overall efficiency of the training process.

By leveraging the performance of the local storage within GPU servers, Hammerspace Tier 0 offers an extreme-performance storage tier capable of accelerating checkpointing and reducing GPU idle time. With Tier 0, checkpoints can be quickly written to the tier of local NVMe storage, saving time and allowing GPUs to resume the training process as efficiently as possible. Checkpoints stored on Tier 0 storage can then be asynchronously migrated to lower storage tiers in the Hammerspace environment for long-term retention.

Hammerspace claims that Tier 0 storage can enable between 2.5x and 10x faster checkpointing compared to writing to external NVMe storage. This addresses a key challenge during model training and enables organizations to utilize their GPUs more efficiently. By reducing GPU idle time during checkpointing, it is estimated that GPU computation capacity can be increased by up to 15% annually. These savings allow organizations to do more with less GPUs and reduce the cost, time, and complexity associated with procuring additional GPUs.

While model checkpointing provides a clear use case where Tier 0 can increase GPU efficiency, it is only one such example. AI workloads are increasingly growing more complex and resource-demanding due to large models and long context lengths. For complex AI inferencing workloads where GPU memory is a constraint, storage can be leveraged to offload model weights and key-value caches. However, without local NVMe, relying on networked storage can introduce significant latency, reduce GPU utilization, and degrade inference performance. Hammerspace Tier 0 mitigates this bottleneck by enabling access to high-performance local NVMe storage, reducing network-induced latency, and improving GPU efficiency. This enables organizations to achieve faster, more responsive AI inferencing applications.

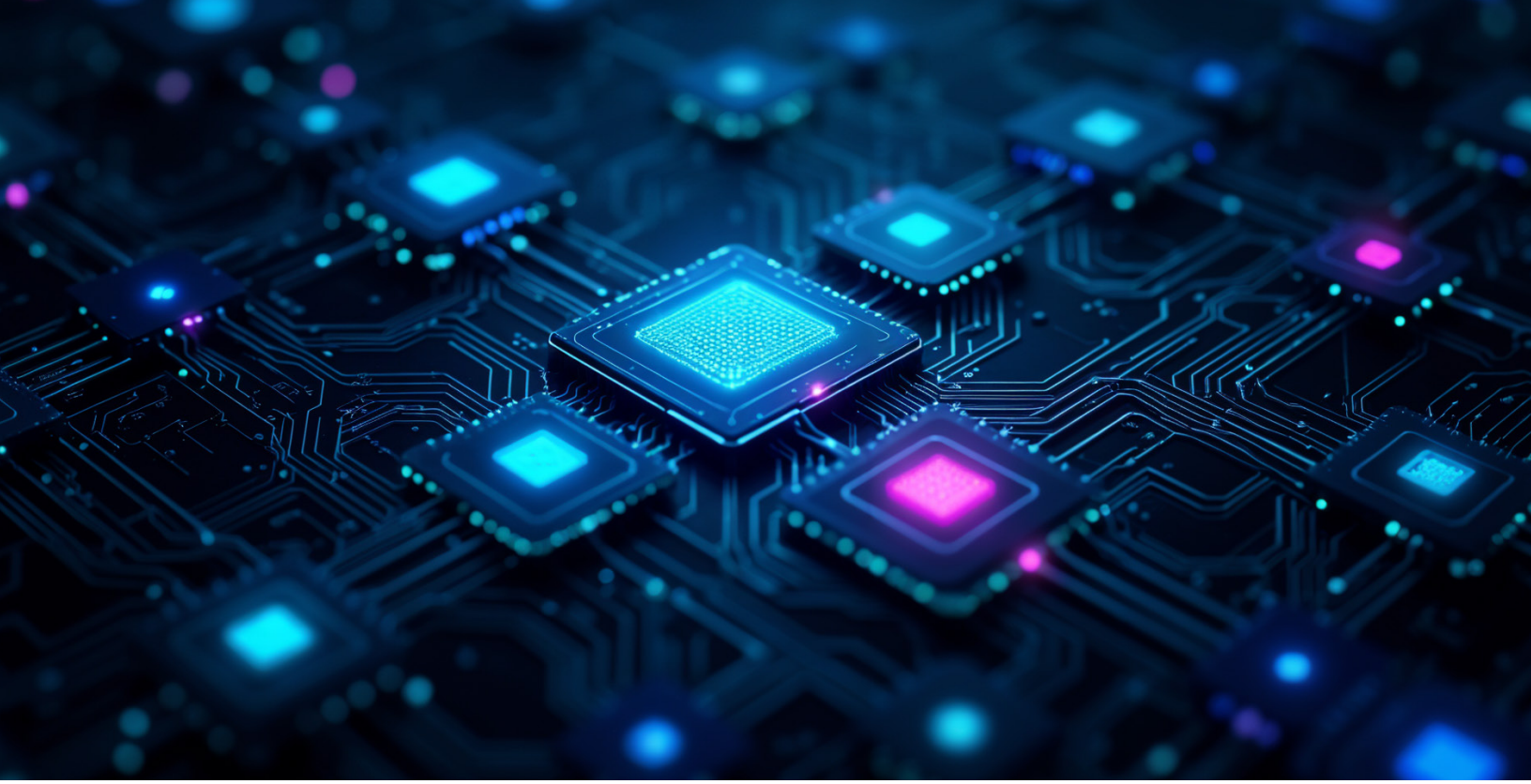
Greater Data Center Power Efficiency

In addition to standard data center criteria such as cost and performance, IT organizations deploying large GPU data centers are increasingly challenged by limitations of power and floorspace. The power efficiency of data centers has become a growing concern for many organizations, as it adds cost, introduces environmental concerns, and can become a limiting factor for the total infrastructure supported. Such considerations are of particular interest in AI and HPC environments, primarily due to the large power consumption of GPU-based computing.

While GPUs are typically the main focus of data center power consumption, large storage environments additionally play a key role. The large storage capacities required in AI and HPC environments require significant hardware, involving its own power consumption. Hammerspace Tier 0 minimizes the amount of external storage infrastructure required, reducing power consumption and freeing up additional floorspace. In addition to increasing the power efficiency of data storage, Tier 0 is capable of further optimizing the power consumption of the GPUs themselves by reducing GPU idle time. By improving power efficiency in large GPU data centers, Tier 0 can enable organizations to save costs, meet their environmental goals, and deploy more infrastructure as needed.

Tier 0 in CPU-based Computing

While this paper has primarily focused on the benefits of Hammerspace Tier 0 in GPU-driven computing environments, it should be noted that the technology can additionally be leveraged for CPU-based computing. For CPU-based applications, including CPU-based AI inferencing, Tier 0 storage provides many of the same benefits as previously discussed, including reducing external storage costs, accelerating performance, and reducing energy consumption.



Final Thoughts

Innovation in the fields of AI and HPC have led to increasingly demanding infrastructure requirements, including high-powered GPU compute, high-speed networking, and highly scalable, high-performance data storage. In addition, the possibilities of new AI workloads have raised the demand for GPU-based computing in traditional enterprise environments. The large infrastructure requirements of these environments present significant challenges to both HPC-focused and traditional enterprise organizations. IT organizations are challenged to get new infrastructure up and running as quickly as possible, while also meeting strict performance and efficiency requirements.

The increasing data capacities required in AI and HPC environments, combined with the high performance needed to support modern GPU computing, have become a significant barrier for many IT organizations. Procuring storage capable of meeting these demands adds significant cost and time, ultimately leading to greater inefficiency and delaying the utilization of valuable GPU resources. To meet the growing data challenges of GPU data centers, organizations must look to new strategies that meet their storage requirements while saving costs and enabling greater efficiency.

Hammerspace Tier 0 provides a practical approach to solving key data storage challenges found within GPU data centers by unlocking previously stranded storage capacity and leveraging it within a global data environment. This allows organizations to utilize petabytes of high-performance storage that they have already purchased, reducing the cost of additional network-attached storage and enabling quick deployment of new applications. In addition, unlocking this direct-attached storage with Tier 0 is uniquely suited for GPU-based workloads by optimizing both performance and cost savings, while also improving GPU utilization.

Hammerspace Tier 0 presents a compelling solution to utilize previously stranded capacity and address the demanding storage requirements of GPU computing. The Futurum Group believes that the capabilities of Tier 0 storage combined with the broader data orchestration capabilities of the Hammerspace Global Data Platform create a unique offering that is well suited to address the data challenges of modern AI and HPC workloads.

Important Information About This Report

AUTHORS

Mitch Lewis
Research Analyst | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

HAMMERSPACE

ABOUT HAMMERSPACE

Hammerspace stands apart as the first data platform built entirely around the performance needs of modern computing. From unmatched GPU throughput to instant access across file and object data, to deployment time measured in minutes—not weeks—every capability of the platform is designed to remove bottlenecks, maximize speed, and deliver scale. Hammerspace's award-winning software delivers a high-throughput, low-latency parallel global file system unifying data from edge to core to multi-cloud, accessible via pNFS, NFSv3, SMB and S3 standards. Instant, agentless deployment with native Linux support combined with migration-free data assimilation radically accelerates pipelines. Whether training thousands of GPUs on-premises or in the cloud, deploying large-scale inference or maximizing NVMe performance in local GPU servers, Hammerspace is purpose-built to unleash data performance at scale.

Futurum

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum

CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337