Enterprise Strategy Group by TechTarget

WHITE PAPER

Driving Cost-effective GPU Performance With Hammerspace

Delivering Tier 0 Performance for AI At Scale, While Slashing Costs and Power Consumption, With Cutting-edge Software Capabilities From Data Platform Innovator Hammerspace

By Simon Robinson, Principal Analyst
Enterprise Strategy Group

December 2024

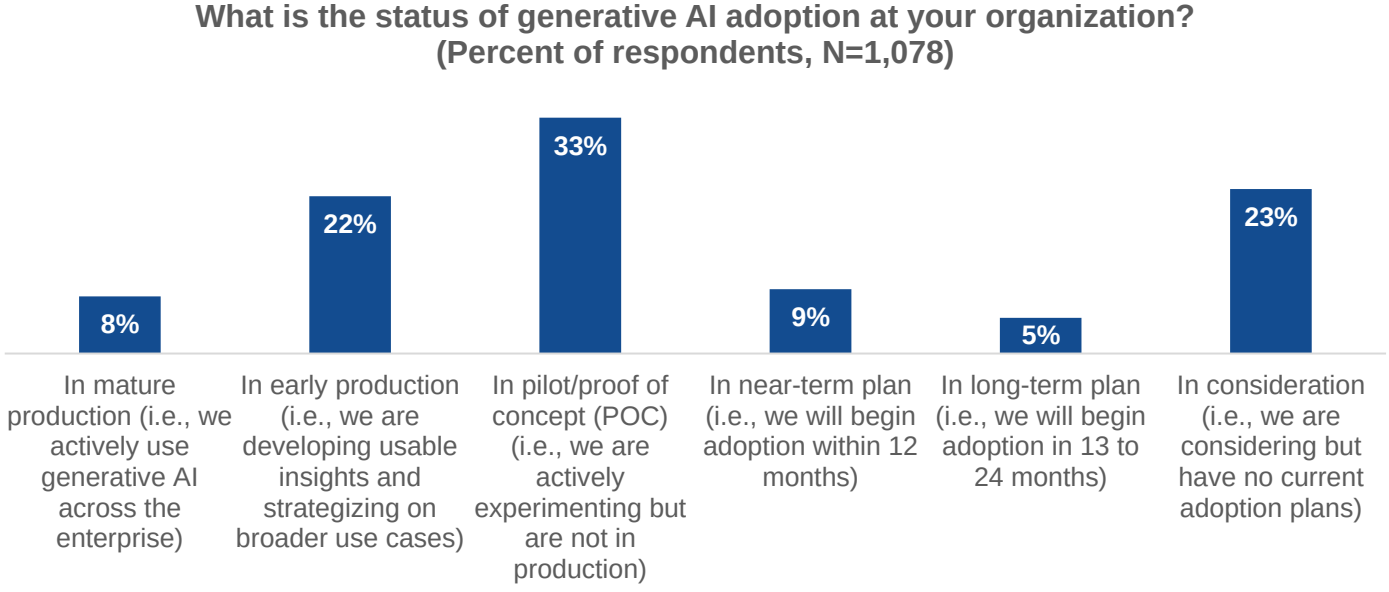
Contents

Infrastructure Challenges Abound in Realizing the AI Promise

The potential of emerging AI and generative AI (GenAI) workloads to create compelling new insights, automate complex processes, and transform practically every organization is beyond doubt. According to recent research from TechTarget's Enterprise Strategy Group, generative AI came third in a ranking of key IT and business initiatives, behind only digital transformation and cybersecurity initiatives[1]. For a technology that is barely two years old to occupy such a lofty position is unprecedented and speaks to the potential of the technology to be applied to almost every facet of an organization's operations.

Enterprise Strategy Group's research also suggested that this GenAI interest is already translating into real-world adoption. Though just 8% of organizations surveyed described their GenAI implementations as "mature"—hardly surprising given the nascent nature of the technology—a further 55% said they had GenAI in early production or proof of concept (see Figure 1).[2] A further 14% had GenAI in their near- or long-term plans, with just under a quarter (23%) considering their options.

Figure 1. GenAI Adoption Status



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Yet the challenges in turning this undoubted potential into transformational results are both multifaceted and profound. Unless organizations can overcome these obstacles, the promise of AI will remain tantalizingly out of reach, underwhelming in terms of results, or substantially over budget.

[1] Source: Enterprise Strategy Group Research Report, *The State of the Generative AI Market: Widespread Transformation Continues*, September 2024.
[2] Ibid.

65% of respondents believe they need to modernize supporting infrastructure before proceeding with GenAI initiatives.

These challenges are especially evident when it comes to building and deploying the infrastructure that will underpin an AI environment. According to Enterprise Strategy Group research, most organizations (65%) believe they will need to invest to modernize or change their supporting infrastructure before they can proceed with GenAI initiatives.[3] Most obviously, this includes the deployment of high-performance computing resources such as GPUs, especially at the initial training phase. However, this is just the starting point, and considerable attention should also be paid to exactly how this GPU infrastructure is deployed as part of the broader environment. Multiple factors driving this are discussed below.

An Idle GPU Is a Sin

The performance impacts of AI deployment have multiple dimensions, but essentially they boil down to how organizations can architect an environment that can effectively feed data quickly enough into GPU-based models to ensure this high-cost resource is utilized as effectively and efficiently as possible. Every minute or even tens of seconds that a GPU is idle is a waste of a data center's most costly resources. With such a high-performing compute environment, the bottleneck quickly shifts to other parts of the infrastructure, with the storage environment becoming especially key.

Here, organizations are quickly forced down one of many paths, such as deploying scale-out file system storage or more complex parallel file system technologies. Though these might be able to deliver the high-performance data ingestion that GPUs demand, there are tradeoffs in terms of cost and complexity.

Even then, performance might still not be adequate, especially when supporting necessary and frequent operations such as checkpointing, which might essentially take the GPUs offline for several minutes each hour due to latencies in writing checkpoint data to storage. In large GPU clusters, this can add up to millions of dollars of "wasted" GPU opportunity cost on an annual basis.

A painful irony here for many is that many GPU servers already contain plentiful amounts of super-fast NVMe storage. Indeed, for organizations deploying tens, hundreds, or even thousands of GPU servers, this in aggregate amounts to huge volumes of fast but stranded storage capacity. A single GPU server today can contain 60 to almost 500 TB of storage, and with NVMe drive capacities continuing to grow, maximum capacity could exceed 2PB in 2025. Yet technology limitations to date mean that compute and GPU resources cannot access this capacity as a pooled aggregate resource. As a result, this ample storage capacity is effectively going unused.

Power Consumption and Data Center Space Are Major Constraints

One unavoidable aspect of deploying GPUs at scale is that they consume substantial amounts of energy. This affects all organizations deploying AI at scale financially, especially in regions where power can be more expensive or might fluctuate in price—as has happened in Europe in recent years. Indeed, the broader issue of power availability might turn out to be a major gating factor in the global rollout and pace of adoption of AI. Lack of power can ultimately be a competitive brake for some organizations. Additionally, the power usage of GPUs might also have material impacts on an organization's commitments to reducing their carbon footprints and, more broadly, their sustainability targets.

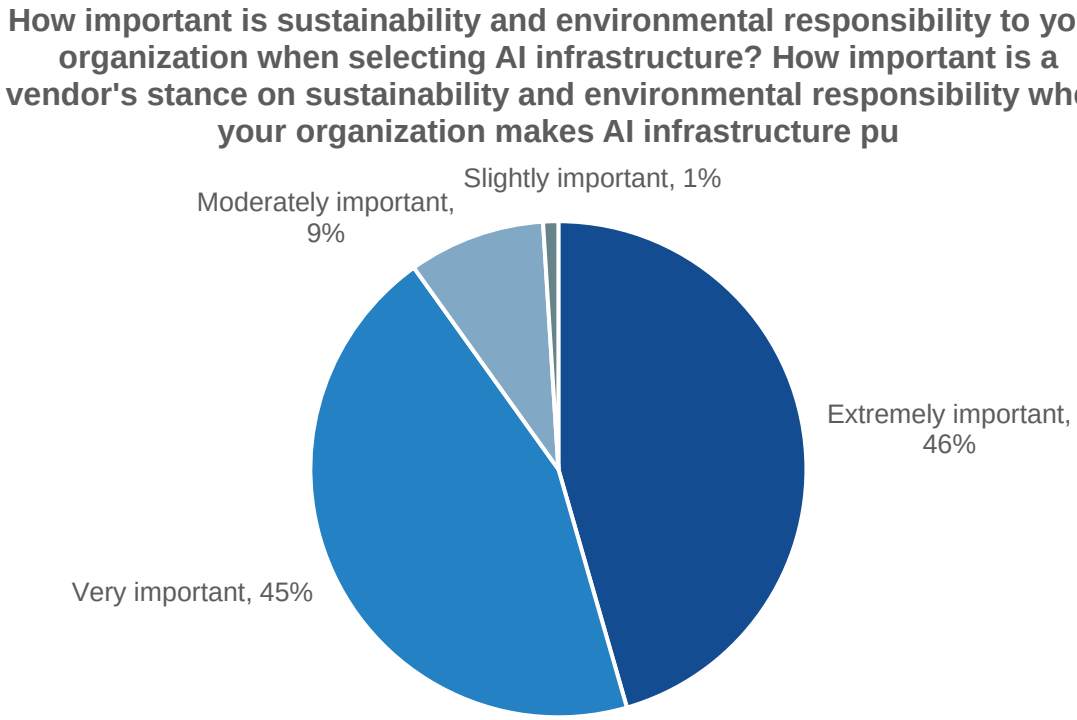
Enterprise Strategy Group research found this is a critical issue for many. As illustrated in Figure 2, 91% view sustainability as a very or extremely important factor when selecting AI infrastructure.[4] A similar proportion view a

[3] Ibid.
[4] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

vendor's stance on these issues as also very or extremely important.[5] Though GPU suppliers are investing to reduce the power consumption of their products, this is not likely to fall materially in the short term, so organizations should look elsewhere in their environment for savings.

Given the large data footprints in supporting AI models, the storage environment is one such candidate for savings, and once again, many organizations deploying GPU servers at scale are unable to tap into the internal storage resident within these servers. Instead, they must buy, deploy, and power additional external storage, which can be redundant to capacity they already have deployed. In addition to the cost of powering this external storage, which can be substantial if this storage is relying on more power-hungry, always-spinning hard disk drive (HDD) technology, organizations must find and fund the data center space for these systems to reside.

Figure 2. Importance of Sustainability to AI Initiatives



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Data Silos and Fragmentation – An Orchestration, Privacy, and Security Challenge

Training an AI model effectively requires providing access to a broad set of data types and resources, the logic being that the more data and greater variety of data a model can be trained on, and the more quickly it can be updated when the data changes, the better that model becomes.

[5] Ibid.

Even where organizations intend to utilize third-party models for AI, the overwhelming majority want to enrich these models with their own data. According to Enterprise Strategy Group research, 84% of organizations agreed that it is important to incorporate their own enterprise data to support generative AI.[6]

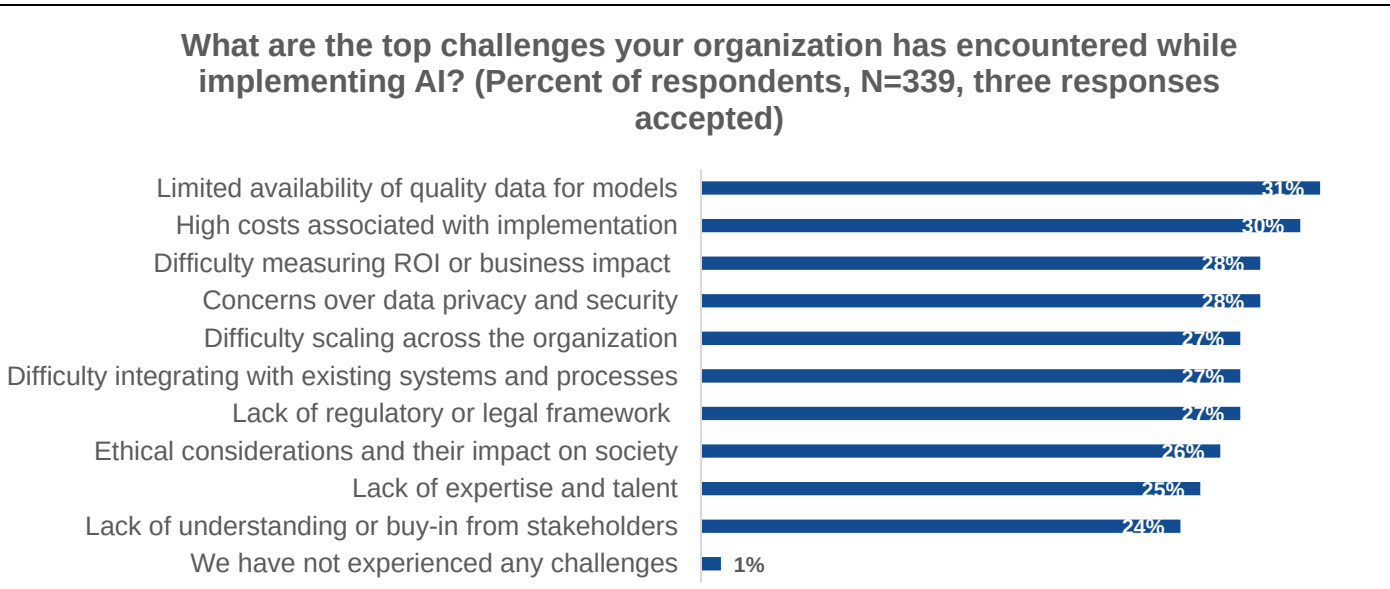
Unfortunately, for many enterprises, this is where the problems begin to mount. The reality is that organizational data continues to be fragmented across a range of silos and locations, including different islands of storage spread across a growing multitude of on-premises and off-premises locations.

Here, the task of building effective operational AI data pipelines can be substantial. In such a situation, the easiest way to get data into a model is to make a copy. But this approach further compounds an already large and growing problem of data growth and sprawl. Keeping track of the relationship between users and data is itself a growing challenge—and one that must increasingly be automated through software.

Additionally, the consequences of moving or copying the wrong data into an AI model can be severe. It can either lead to the model "hallucinating" and making incorrect inferences—which will erode or even destroy trust in it—or, worse, it could result in sensitive data such as personally identifiable information, intellectual property, and more leaking into a public model.

All of these challenges and risks perhaps explain why limited availability of quality data for models was highlighted as the No. 1 challenge that organizations have encountered when implementing AI, ahead of challenges such as high costs and difficulty measuring ROI (see Figure 3).[7] Organizations also highlighted the need for a robust and comprehensive approach to data orchestration that overcomes the data silos and can span multiple tiers of storage and data, from high-performance storage all the way out to the archiving tier. For many organizations, the challenge is not that there's not enough quality data; it's that it's too difficult to locate and direct into the model in a secure, efficient, and effective manner.

Figure 3. Top AI Implementation Challenges



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

[6] Source: Enterprise Strategy Group Research Report, *The State of the Generative AI Market: Widespread Transformation Continues*, September 2024.
[7] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

Taming Runaway AI Infrastructure Costs

As well as influencing the effectiveness of an AI initiative, the above factors might also have substantial cost implications for the organization, both directly and indirectly. As Figure 3 suggests, the high cost of implementation is already a top issue for organizations looking to implement AI at scale. The question for many organizations looking to scale their AI initiatives is how to do so cost-effectively. With GPU compute accounting for a large proportion of AI spend, many organizations are looking to optimize spending elsewhere in the infrastructure. Many will be looking at how to take advantage of sunk costs on resources they have purchased but are perhaps not using to full effect. Once again, attention will inevitably focus on better harnessing internal server storage contained within GPU servers. This can result in directly reducing storage costs and also might lead to savings in networking costs, since a smaller external network-switching infrastructure might be required across any shared environment.

Introducing Hammerspace Tier 0 – Deploy Ultra-fast Storage and Cut Costs

Thankfully, a resolution to all of these challenges is available in the form of Hammerspace's most recent update to its Global Data Platform software. The company's v5.1 release adds several highly innovative new capabilities that can help drive GPUs to peak performance, while at the same time reducing storage and power costs (see Figure 4).

Hammerspace is a private software company founded in 2018, and its software is used in very large GPU computing environments like at Meta, Jellyfish Pictures, and Los Alamos National Laboratory, where Hammerspace provides the data platform that feeds tens of thousands of GPUs in parallel.

The Hammerspace Global Data Platform combines a standards-based parallel file system architecture with data orchestration services that automate the protection and movement of data, both across tiers of storage and across sites and clouds as part of a global namespace.

The most notable innovation within the new software is a new Tier 0 capability that can transform GPU computing infrastructure into a new tier of ultra-fast shared storage. This enables organizations to turn the internal NVMe drives inside their GPU servers into a shared pool of Tier 0 storage, delivering microsecond latencies that can greatly exceed the performance of external NVMe shared storage systems—at a fraction of the cost. The result: multiple benefits for organizations deploying from just a few to thousands of GPU servers.

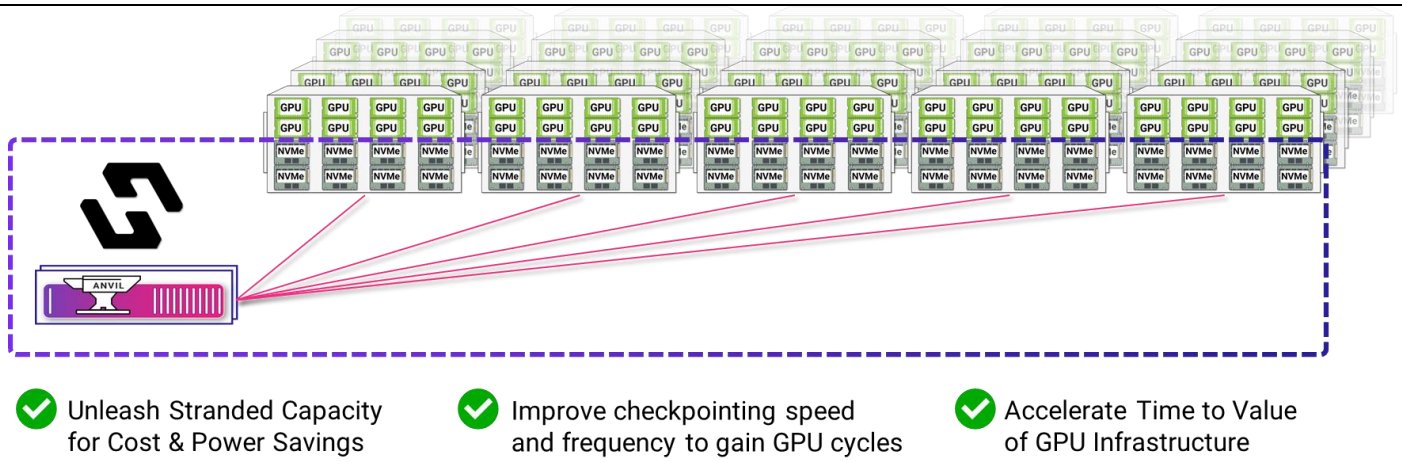
For example, Hammerspace notes that an organization deploying a large-scale AI training environment with 1,000 GPU servers and 100 PB of internal NVMe storage capacity, delivering 1 TB/sec throughput, could experience the following:

- **Significant cost savings.** In addition to avoiding spending as much as $40 million on additional storage systems, Hammerspace notes substantial cost savings associated with improved GPU utilization. For example, Hammerspace's software could help reduce checkpoint times from 600 seconds per checkpoint to 6 seconds. With GPUs idle during checkpointing using traditional approaches, this could add up to 929,000 GPU hours saved per year—the equivalent of 848 GPUs, which equates to $25 million to $30 million in gained GPU capacity (at $20,000 to $40,000 per GPU).
- **Enhanced performance.** Faster checkpointing also means faster time to value for the initiative overall.
- **Improved resource utilization.** Organizations that have already invested in GPU servers containing NVMe storage can leverage this resource without having to buy duplicate fast storage.
- **Lower operational costs.** A reduced hardware footprint minimizes both data center power requirements and real estate.

The benefits of deploying Hammerspace's Hyperscale NAS architecture with Tier 0 are not just reserved for those deploying AI and GPUs at massive scale; organizations running hundreds or even just tens of GPUs are able to benefit from its approach. The company notes that an organization deploying a much smaller 10 PB environment could see annual savings in the range of $3-4 million through avoidance of duplicate flash storage hardware costs as well as reduced energy consumption of around 3 million kWh over three years—equivalent to $500,000 in energy cost savings.

Importantly, Hammerspace's approach to building its Global Data Platform means it can manage this Tier 0 storage as part of a broader file and object environment, rather than as a standalone island. This enables users to manage their unstructured data across their entire environment, including Tier 0 ultrafast storage, Tier 1 NVMe storage, Tier 2 HDD-based storage, and archived data residing on cloud- and even tape-based solutions. This enables users to manage data for their AI environment right across the lifecycle, with Hammerspace providing the data orchestration with a policy-based approach that provides automated data movement and management across tiers. The software can also automate data protection, providing organizations with the data resilience and security they require to ensure that only the appropriate data is fed into an AI model. In addition, Hammerspace software is built on standard Linux and uses a NFS and parallel NFS standards-based approach, meaning no proprietary client software is required. Installation is also straightforward, with users able to start using Tier 0 capacity within 30 minutes or less.

Figure 4. Overview of Hammerspace Tier 0 Architecture



Source: Hammerspace

Though Enterprise Strategy Group has not independently validated these numbers, and we would always advise customers to verify savings based on their own requirements, we note these are potentially compelling savings that should appeal to a wide number of organizations looking to deploy—or, indeed, having already deployed—GPUs at scale.

Conclusion

Though the potential for AI to drive transformative benefits for an organization remains vast, the realities of building an AI infrastructure at operational scale can be formidable. Attention so far has disproportionately fallen on the GPU layer, but organizations ignore the broader infrastructure—in particular, the large and growing data and storage environment—at their peril. This could lead to numerous inefficiencies and cost effects that could damage the success of the overall AI initiative, as well as negatively affect environmental and sustainability goals.

Hammerspace's data platform architecture, together with its specific new capabilities around Tier 0 storage, offer organizations a compelling alternative to traditional storage approaches for large-scale AI deployments, avoiding costly additional and redundant storage purchases, boosting GPU utilization, and saving potentially meaningful amounts of time, effort, and energy. Though overall results will vary depending on the implementation, IT decision-makers driving AI infrastructure initiatives should take a closer look at Hammerspace before finalizing any storage investment.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com