Transform GPU Computing Infrastructure by Unlocking a New Tier 0 of Ultra-Fast Shared Storage

Cut Storage Costs, Reduce Power Consumption, and Gain Valuable GPU Compute Cycles

EXECUTIVE BRIEF

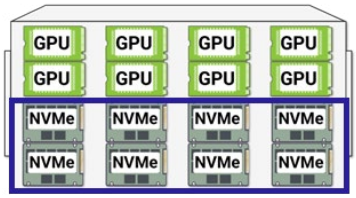
GPU Server Local Storage Has Become Stranded Capacity

Organizations are deploying large GPU computing environments for AI, enterprise HPC, and other forms of unstructured data processing. Typically, these same organizations also deploy external flash storage systems to provide the data to these GPUs, and these storage systems add cost, consume power, and take time to evaluate, purchase and deploy.

Meanwhile, the flash storage in GPU servers typically goes unused for a few reasons:

1. It is siloed - only available to the GPUs in the server and not other GPUs in the cluster
2. It is not protected - so there is risk of data loss if the GPU server goes down
3. It is difficult for data users to manually move data between local storage and shared storage

GPU Server Local Storage Typically Goes Unused

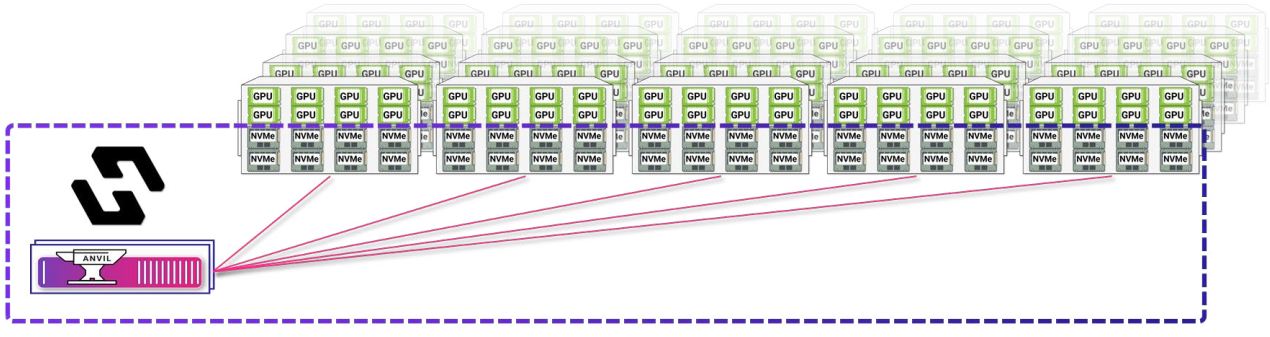


And the capacities are no longer small. A single GPU server today can have 100s of Terabytes of data, and by next year a single server could have 2 Petabytes of capacity.

So organizations today have Petabytes of available storage in their GPU cluster that ends up as 'stranded' capacity. And now, for the first time, Hammerspace allows you to use that ultra-high performance capacity in new ways by enabling the local storage in the GPU servers to become part of the Hammerspace Global Data Platform.

This creates a new tier of ultra-fast shared storage that we call "Tier 0."

Hammerspace Tier 0 is a New Tier of Ultra-Fast Shared Storage



- Unleash Stranded Capacity for Cost & Power Savings
- Improve checkpointing speed and frequency to gain GPU cycles
- Accelerate Time to Value of GPU Infrastructure

Tier 0 Transforms GPU Computing Infrastructure

Hammerspace Tier 0 transforms GPU computing infrastructure by turning the unused “stranded” capacity in your GPU servers into a new tier of ultra-fast shared storage.

Data placed in this local storage is no longer siloed within each GPU server, but instead is part of a unified l namespace that spans other on-premises and cloud storage types from any vendor.

Current Situation GPU server local capacity is stranded and unused	With Hammerspace Tier 0 Activate existing GPU server capacity as a new tier of ultra-fast shared storage
✕ **High Storage Costs:** External flash storage systems are expensive	✓ **Reduce Storage Costs:** Reduces external storage requirements to save millions
✕ **Power Constraints:** External flash servers and switches consume power and space	✓ **Reduce Power Consumption:** Save millions of kilowatt-hours per year
✕ **GPUs Sit Idle Too Often:** External storage systems are not fast enough	✓ **Increase GPU Utilization:** 20x faster checkpointing frees up GPU compute cycles
✕ **Months to Realize Value:** Evaluate, purchase, deploy external storage	✓ **Faster Time to Value:** Start using capacity in minutes; bring Petabytes online in days

Data is automatically protected, and data is seamlessly orchestrated between Tier 0 and external storage systems, including external NAS storage, cloud storage, and even storage systems in different locations.

By reducing the need for external storage, Tier 0 significantly reduces storage costs, power consumption, energy costs, and data center space.

And because Tier 0 is so much faster than external storage, critical AI and HPC operations like checkpointing occur 10x, 20x, even 100x faster. This means GPU servers spend less time checkpointing, and can spend more time doing other tasks. So not only are you making better use of your existing GPU storage resources, you end up getting more GPU computational power on an annual basis.

Lastly, Hammerspace Tier 0 capacity can be made available in minutes, and even Petabytes of Tier 0 capacity can be brought online in a matter of days, so you can accelerate time-to-value of your investment in GPU computing.

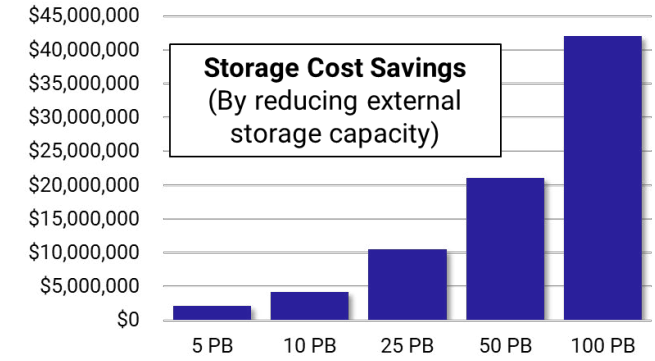
For GPU computing clusters that are hundreds or thousands of servers, such as NVIDIA SuperPOD architectures, organizations will realize:

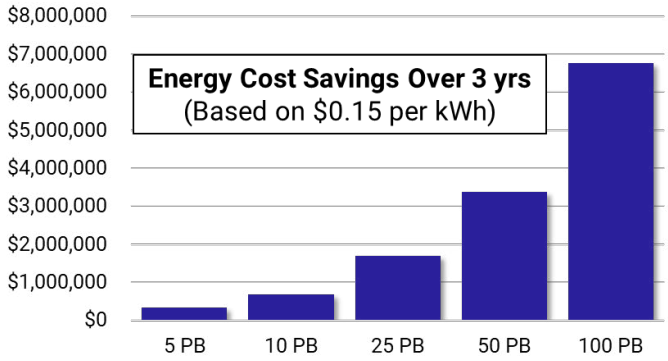
1. Millions of dollars of storage cost savings
2. Millions of kilowatt-hours of power savings
3. An increase of GPU computational capacity of 10-15% annually

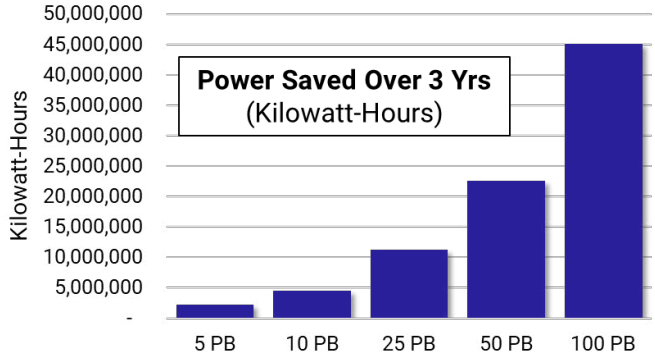
Tier 0 Delivers Significant Cost, Power and Rack Savings

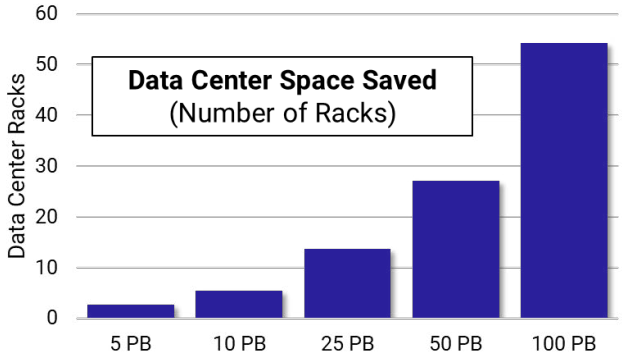
The charts below indicate the order of magnitude cost, power, and rack space savings that can be achieved by using this new tier of shared storage. These figures are approximate, and these results will vary based on geography as well as the costs and configurations of external storage systems.

For those interested in seeing what the savings would be for their particular environment, Hammerspace can adjust the model accordingly - Get Started here









Tier 0 Frees Up GPU Cycles to Unlock GPU Opportunity Cost

In addition to the savings above, which are driven by Tier 0 capacity reducing the need for external storage capacity, Tier 0 unlocks GPU opportunity cost and increases GPU computational capacity on an annual basis by speeding up checkpointing.

Checkpointing is used in AI training and HPC modeling, and is the process of periodically saving the state of an application to persistent storage. Because Tier 0 capacity is much faster than external storage, it speeds up checkpointing by 10x, 20x, even 100x in large scale environments. Checkpoints are completed in seconds, not minutes. This means GPU servers spend less time checkpointing, and on an annual basis this can add up to 100s of hours of extra time that each GPU server can be used for data processing.

This is equivalent to adding 10-15% GPU computing power to your existing cluster without adding a single GPU! And this in turn equates to millions of dollars - even tens of millions of dollars - in GPU opportunity cost that is unlocked by Tier 0.

Read this whitepaper for a detailed analysis of using Hammerspace Tier 0 for checkpointing.