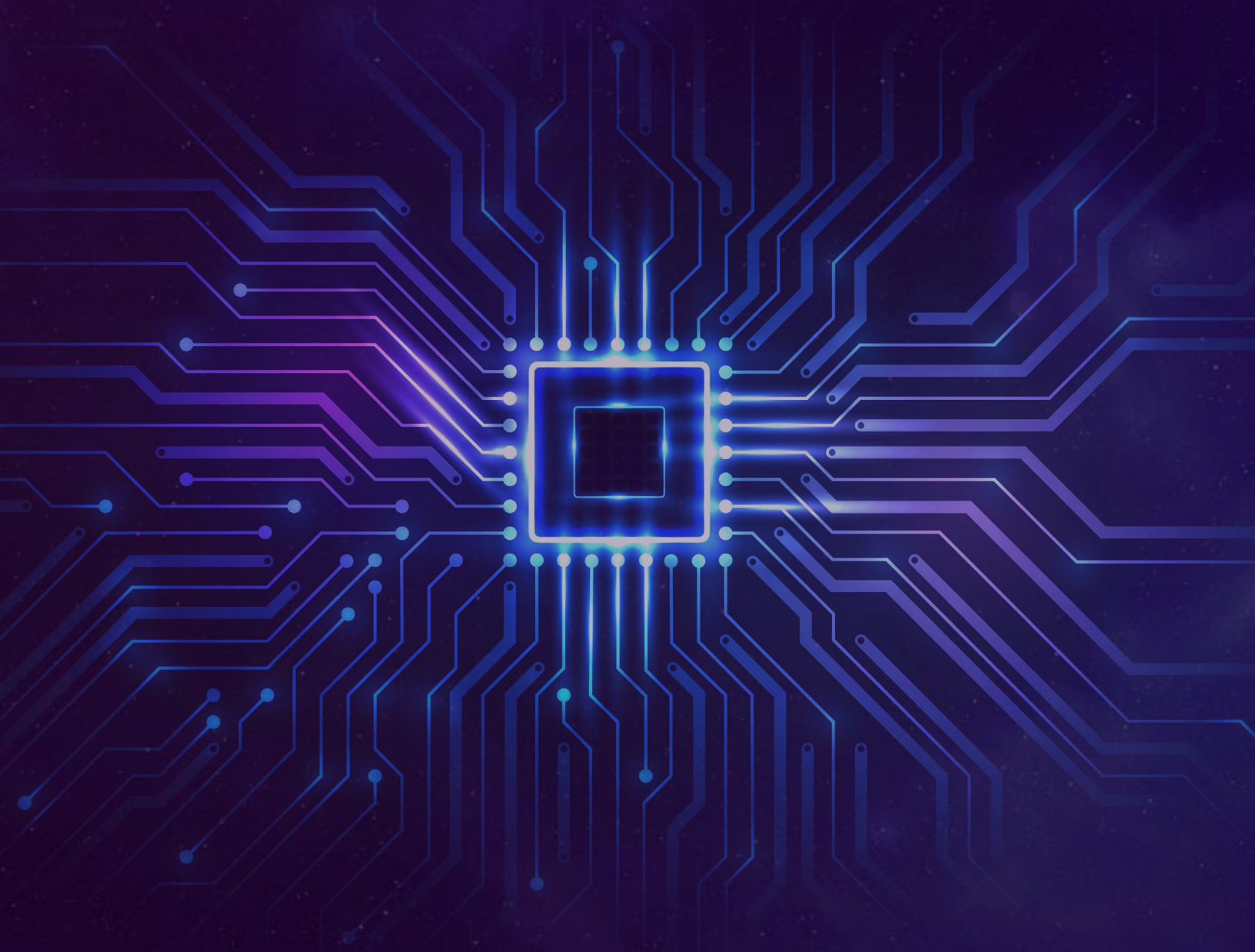
HAMMERSPACE

The State of the Next Data Cycle

How do You GPU?



Since the 2022 public debut of DALL-E and ChatGPT, artificial intelligence (AI) has remained at the center of technology discourse and has dominated the news day after day. One thing is clear: AI is the future, and everyone wants a piece of it. But beneath this enthusiasm lies a critical question: **How are enterprises deploying and scaling their AI solutions?**

It's broadly believed that AI has enormous potential. Analysts and business leaders are predicting a wave of fresh innovation that unlocks new levels of efficiency and productivity. Yet, there's a stark contrast between AI's future promise and its current progress. So far, AI is a concept with immense potential just beginning to generate real-world value. While executives speak often about AI's ability to improve their businesses, most use cases lie months or years in the future — or are entirely hypothetical. Real-world deployments have yet to match the ambitious predictions.

To better understand the current state of enterprise AI, Hammerspace commissioned an analysis of digital conversations among industry leaders and executives across several platforms, including LinkedIn, Twitter, Reddit, GitHub, and Discord. Based on nearly 17,000 conversations involving about 200 leaders representing enterprise AI customers, the data provides a unique insight into their opinions and experiences with AI.

The findings suggest that the market remains largely in an exploration phase. A close analysis of the conversations found that in the wake of big investments in new computing hardware including graphics processing units (GPUs) intended for AI, these companies have instead put these investments to unexpected and innovative uses in the near term as their AI plans firm up over the medium- to long-term. This is producing unanticipated ROI and uncovered another important stage of the AI journey. The potential is real, and possibilities are just beginning to unfold.

AI Reality Check:
Conversations vs. Implementation

Since 2022, public-facing **conversations in online communities about AI have increased 383%.** Among the 150 companies represented in the Hammerspace study, nearly 60% are actively and publicly discussing AI, with a primary focus on elevating company leadership as thought leaders in the space.

The majority of these conversations (56%) fall into five categories, highlighting where senior leaders believe AI's importance lies:

50% » **Executive Visibility and Leadership**
Projecting self, company, or colleagues as forward-minded and visionary within the tech industry

34% » **AI Transformation**
News or commentary on how AI is changing the world and society

33% » **Innovation advancements**
Announcements of new products, news-sharing about strengthened AI, and thoughts on how AI models can be improved

32% » **Ethics**
Ideas on how to ensure responsible, safe, and benign use of AI and technology

15% » **Education**
Circulating new understandings or trainings on AI and technology by thought leaders

Meta: Enhancing AI Training Efficiency with Hammerspace

Meta scaled its AI infrastructure to support Llama 2 and 3 model training by deploying over 24,000 NVIDIA H100 GPUs. These GPUs use high-performance networking like RoCE and InfiniBand to prevent bottlenecks during training. This infrastructure maintains performance even during outages, improving upon both efficiency and resilience.

Optimizing GPU performance was made possible by Meta's collaboration with Hammerspace. Using Hammerspace's standards-based parallel global file system coupled with automated data orchestration, engineers can then debug and update code, making changes accessible across the thousands of Meta's GPUs. This reduced the delays caused by data synchronization, helping large-scale training jobs run smoothly without stalling.

Meta integrated Hammerspace with its Tectonic-distributed storage, providing massive scalability, achieving over 90% bandwidth utilization after targeted optimizations. Together, Meta's GPU infrastructure and Hammerspace's data solutions are driving the development of next-generation models like Llama 3.

Online community-driven conversations about AI have Increased by

383%

Of discussions around implementing AI, a majority include accelerating, simplifying, or improving existing workflows. In fact, **59% of conversations covering innovation are largely focused on enhancing productivity**. However, even as AI transformation makes up 34% of the overall market conversation, only 18% of the community conversation on innovation and advancements is dedicated to achieving better AI outcomes.

Senior executives (C-level, EVPs, and SVPs) and VPs responsible for purchasing and implementing new technologies tend to dominate the conversation. Surprisingly, CTOs, traditionally responsible for digital transformation initiatives, were largely absent from the conversation, showing that while AI is still in the exploratory phase, it isn't reaching everywhere in the C-suite, perhaps even where it matters most. Despite tech companies holding tight to the AI narrative, **only 9% of the companies tracked have three or more executives actively engaged in the conversation.**

AI transformation makes up 34% of the overall market conversation, only 18% of the community conversation on innovation and advancements is dedicated to achieving better AI outcomes.

27%

of the conversation involves leaders at the VP Level

23%

of the conversations are engaged by CEOs

9%

of the conversations are engaged by CTOs

Some sectors, like healthcare and government, **account for 34% of discussions on data management**, revealing a need for improvements in AI data access. While the majority of AI conversations involve enterprise tech companies implementing and purchasing AI, **healthcare companies participated in a sizable portion (12%) of the total conversation**.

Ethics:

Innovating Within Guardrails

Ethics accounts for one-third of the AI conversation, hinting that amid the growth they also expect or desire the creation of ethical guardrails, either self-imposed or via regulation at the state or federal level. More than half, **51%, of the ethics conversation concerns policy and best practices**, revealing that many are pushing for guidelines to ensure responsible AI development — we see this in practice with OpenAI's ethics board. Several are concerned with data privacy issues, as AI's increasing reliance on personal data sparks discussions on how to manage and secure sensitive information. The data also reveals concerns on misuse and environmental costs.

19% is dominated by data privacy concerns

The remaining conversations centered on generative AI ethics (17%) and industry and environmental impact (13%).

Los Alamos National Laboratory: Optimizing Hybrid Compute with Hammerspace

Los Alamos National Laboratory (LANL) worked with Hammerspace to streamline its data architecture for traditional high-performance computing (HPC) and AI research. Its hybrid supercomputer, powered by NVIDIA Grace Hopper Superchips, integrates CPU and GPU processing to support national security projects, pandemic preparedness, climate change mitigation and defense research.

Hammerspace worked with LANL to unify siloed file systems for home directories, scratch, archive and AI training into a single platform. Using Hammerspace's parallel file system and Global Data Platform, this simplified workload management and improved efficiency.

With Hammerspace, LANL orchestrated data across CPU- and GPU-based workloads, maximizing resource use. This strengthens its collaboration with other labs, solidifying LANL's leadership in advanced computing and data architectures.

Successful Artificial Intelligence Requires Learning

Education, at only 10%, accounted for the smallest portion of the conversation. A significant portion of the education conversation focuses on efficiency. Specifically, **43% of discussions cover "thinking differently" about AI, primarily how to adapt and boost productivity. Only 24% of the education conversation centers on how AI works**.

This limited focus on education and training hints at how companies are still navigating their way through AI implementation. Though rather than a setback, this phase is an active exploration period. As businesses experiment with AI, they're finding new and creative ways to enhance existing processes and boost productivity.

As for the GPUs they bought initially for AI projects, they're proving valuable in advancing data science and analytics initiatives. Widespread AI may still be in its exploratory phase, but companies are uncovering unexpected returns on their investments, paving the way for the future.

GPUs: Enterprises' Best Kept Secret for Fast, Efficient AI… or Maybe Big Data?

The exploration stage for AI runs counter to the well-documented buying spree around graphic processing units, the chips used to drive AI computing. Despite significant investments in AI infrastructure, including enormous investments in powerful GPU chips, some companies have struggled to put them to use on AI workloads and used them on other better-understood use cases.

In 2024, the global GPU revenue rose to $65 billion, with forecasts suggesting it will grow to $274 billion by 2029. Growth aside, the expectation that a surge in AI projects following the increased sale of GPUs has not materialized. Analysts at Goldman Sachs predict that GPU supply, rather than demand, will dictate AI chip shipments into the first half of 2025, a bottleneck further complicating AI deployment.

But, for every innovative and powerful AI application, there is (usually) a GPU or two behind it. These are incredibly useful for many things other than AI: Initially designed to enhance PC games, GPUs also excel at big data and analytics applications. The challenge then is AI system development and the training of large language models: It can cost up to $200,000 for coding assistants, $1 million to embed generative AI in custom apps, $6.5 million to fine-tune gen AI models, and $20 million to build custom models from scratch, according to Gartner. AI is exceptionally expensive and demands tremendous amounts of data and a great deal of raw computing horsepower — exactly where GPU investments shine.

AI's exploration phase parallels prior ventures by large enterprises to unlock value from big data. Enterprises were once only activating about 10% of their data, primarily structured data living in drives. Now, they have the tools to access the remaining 90% of the unstructured data, innovating on this in new and impressive ways. This initiative has fueled demand for tools to make unstructured data readily accessible to improve both existing business applications and AI.

Enter GPUs.

Leading Streaming Platform: Optimizing Machine Learning for Media and Entertainment

A leading streaming platform has used machine learning (ML) to enhance industries such as healthcare, finance, and ecommerce. With the growing complexity of ML models, they focus on optimizing hardware architecture to balance CPU and GPU usage for efficient model training and deployment.

Memory optimization is essential for handling large models while conserving resources. The platform uses dynamic memory management techniques, such as model pruning, to reduce memory footprints by up to 90%, improving inference times and energy efficiency. CPUs are used for tasks requiring data locality, while GPUs and specialized accelerators handle large-scale training tasks.

By profiling memory usage and implementing hardware-aware strategies, the platform maximizes performance across diverse hardware environments. This approach reduces infrastructure costs and energy consumption, enabling efficient deployment of ML models.

GPUs in the Wild:
Innovating Within Guardrails

A large percentage of companies using GPUs do so not just for AI, but also for existing big data projects requiring substantial compute power to complete. The purchased GPUs initially intended for AI projects are instead being used for a wide variety of applications, including expediting existing big data and analytics projects.

GPUs are not operating alone, and in many use cases, businesses are combining them with powerful CPUs as well. A leading streaming provider uses this combined power to improve recommendation algorithms and optimize video streaming quality across millions of users. By tightly integrating CPUs and GPUs in its infrastructure, the provider has significantly improved the speed and accuracy of its content recommendation system, ensuring personalized, data-driven suggestions for each viewer at scale — impossible without a significant investment in GPUs.

CPUs manage the large-scale data processing, complex workloads, and orchestrate multiple tasks simultaneously, while the GPUs accelerate video encoding, enable faster processing and deliver higher-quality streams with reduced buffering. By optimizing both GPU and CPU usage, the company has continued to push the boundaries of innovation in entertainment, consistently setting new standards for content personalization and streaming quality.

Meta Platforms, Inc., more commonly known as Meta, and formerly known as Facebook, Inc. has also significantly scaled its infrastructure to support large-scale training for its Llama 2 and 3 models by deploying more than 24,000 GPUs. This volume of investment is extensive, but the payoff is clear: GPUs are enabling the company to achieve faster, more efficient training of AI models. To get more from its GPU infrastructure, Meta used a data platform built on a standards-based parallel global file system and coupled with automated data orchestration to allow engineers to quickly update, debug and code data as it becomes available. With seamless data access across thousands of GPUs, Meta can fully leverage its GPU investment, achieving bandwidth utilization above 90%.

A final example of successful GPU deployment comes from Los Alamos National Laboratory (LANL). The LANL team streamlined its data architecture, optimizing both traditional high-performance computing (HPC) and AI-driven research. LANL's hybrid supercomputer integrates CPU and GPU processing to address critical national security projects such as pandemic preparedness and climate change mitigation.

Using its GPU and CPU infrastructure, LANL has consolidated multiple file systems (previously used for home directories, scratch, archive and AI training) into a unified platform. This consolidation enhanced efficiency, allowing seamless data sharing across GPU workloads. LANL reduced the complexity of its systems and now is able to maximize resource utilization, reinforcing its lead in advanced computing and data architecture innovation.

[1] This statistic derived from a sampling of Hammerspace customers.

Exiting Exploration, Entering Innovation

Exploration is a pivotal part of developing technology — and AI's exploration phase has unlocked unexpected benefits from GPUs. The investment in this technology was initially aimed at driving AI, but it has since proven to benefit other applications like data management, analytics, and workflow optimization. The adaptability and power of GPUs allows enterprises to tackle their most pressing challenges while also building a foundation for future innovation.

HAMMERSPACE

www.HAMMMERSPACE.com