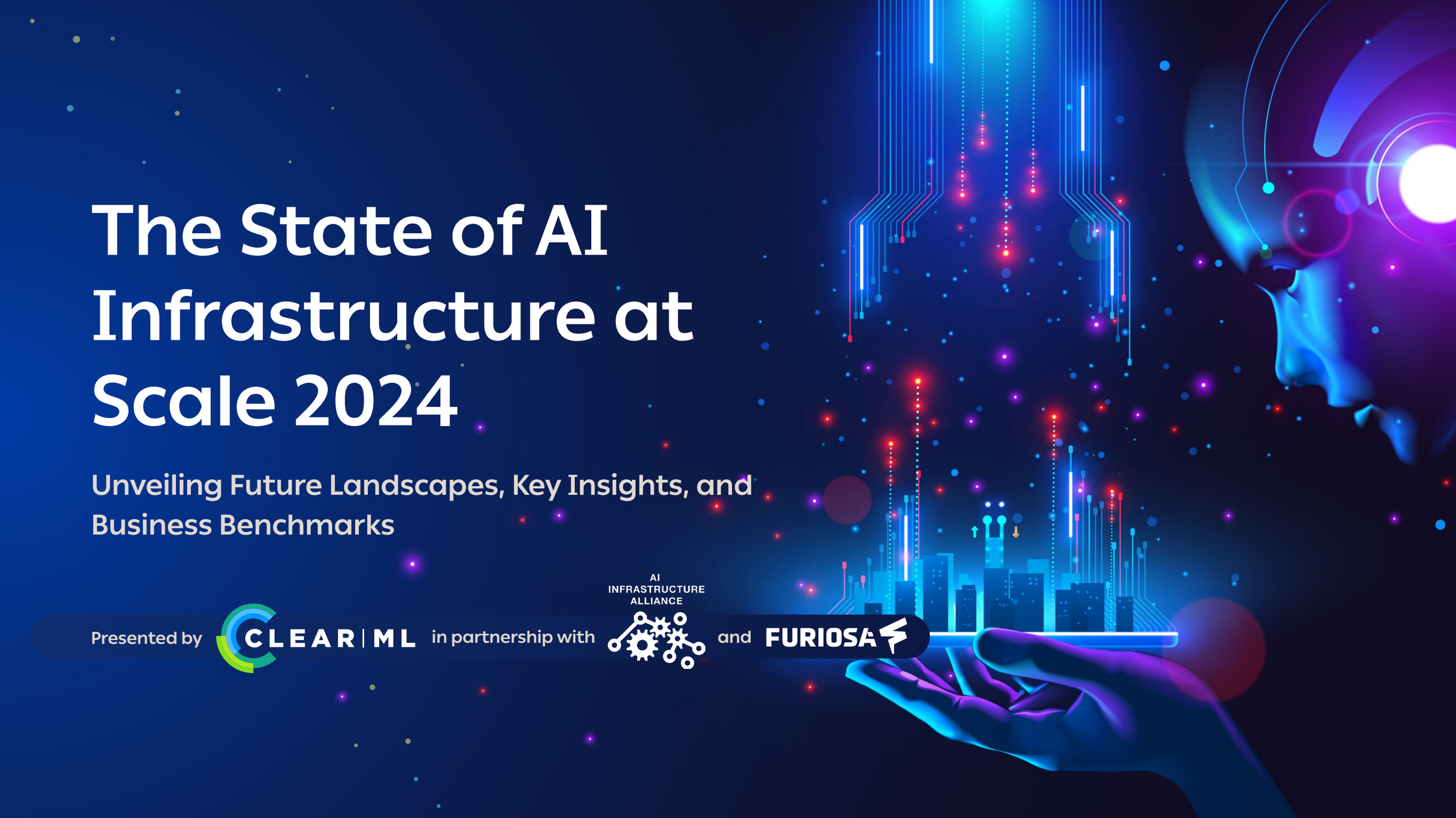
The State of AI Infrastructure at Scale 2024

Unveiling Future Landscapes, Key Insights, and Business Benchmarks

Presented by CLEAR | ML in partnership with AI INFRASTRUCTURE ALLIANCE and FURIOSA



The Artificial Intelligence (AI) sector has experienced an unprecedented surge in recent years. As the applications and use cases of Generative AI expand and more companies shift from research and evaluation to production, we expect that the growing need for robust infrastructure and computational capabilities will drive hyper market demand.

According to new research published by **Allied Market Research and reported by CIO News**[1], the trajectory of the Global AI Infrastructure Market reflects this burgeoning demand. In fact, the report notes that the AI infrastructure market was valued at $23.5 billion in 2021 and is estimated to soar to an astounding $309.4 billion by 2031, growing at a Compound Annual Growth Rate (CAGR) of 29.8% from 2022 to 2031.

One of the primary drivers of growth in the AI infrastructure market is the realization among enterprises of how AI can elevate their operational efficiency and enhance productivity, as well as expand revenue and reduce costs through the automation and orchestration of AI/ML workflows.

There is a wide array of new AI infrastructure tools, leaving prospective buyers grappling with the challenge of sorting through their critical AI infrastructure business needs for scaling AI into production. That may be why prior research conducted in 2023 indicated that only 5-10% of enterprises had started to move Gen AI into production. As organizations navigate the AI infrastructure market, they are actively seeking clarity on AI/ML platforms that can support scale while optimizing their compute utilization.

To deliver that clarity, ClearML, along with the AI Infrastructure Alliance and FuriosaAI, conducted a global AI Infrastructure research survey of AI/ML and technology leaders at 1,000 companies across multiple geographies (North America, Europe, and Asia Pacific) and various company sizes: we wanted to explore and map out new market trends and insights after the first year of preliminary Generative AI mainstream adoption.

One thing is clear: scalable AI infrastructure is crucial for global businesses commercializing AI, as it ensures that their AI systems can handle growing computational demands and AI workloads. That's why ClearML recently announced newly enhanced orchestration and scheduling capabilities along with GPU Partioning and MiG capabilities driving GPU maximal utilization and an enterprise-grade unified platform offering state-of-the-art LLMs called ClearGPT. Meanwhile, FuriosaAI offers its first-gen NPU WARBOY and next-gen LLM-optimized NPU with High Bandwidth Memory 3 (HBM3) for cutting-edge distributed inference.

1 - https://www.linkedin.com/pulse/ai-infrastructure-market-expected-achieve-usd-3094-billion-booming-2ri1f/

Both companies provide solutions that allow organizations to develop and host LLMs in a cost-efficient and performant way tailored to the organization's internal data and running securely on their network to power Enterprise AI transformation. As global startups that believe in providing companies with information that enables hardware awareness and clarity into the AI/ML space, we partnered with the AI Infrastructure Alliance (AIIA), a nonprofit whose mission is to help the AI/ML community make informed decisions about their AI infrastructure decisions.

Together, we looked at how AI and technology leaders are approaching the build of their AI infrastructure, the key challenges and considerations they face, and how they rank priorities when evaluating AI infrastructure solutions against their current needs and business use cases.

This survey is ClearML's third global AI research survey, following two previous surveys that covered generative AI adoption and hidden costs, challenges, and the TCO of Gen AI adoption in the enterprise. In the first survey, 1,000 executives and tech leaders (CTOs, VPs of AI, Chief Data and Analytics Officers, and others) from Fortune 1000 companies reported wasted opportunities and missed financial goals as a result of poor AI/ML operationalization or commercialization[1], some incurring losses of more than $200 million[2].

In our second survey, when asked about key Generative AI cost drivers, the top response was tools, systems, and infrastructure integration costs, followed by GPU and compute costs of model development and training[3]. Despite these challenges, 56.8% of companies surveyed expected double-digit increases to revenues from AI/ML investments and enterprise AI transformation in 2024. Based on these recent surveys, industry metrics, and data from the AI Infrastructure Alliance, we expect that compute infrastructure, especially AI chips, will continue to be in high demand as Generative AI and the number of Large Language Models (LLMs) increase in production and at scale. Optimizing a company's current tech stack to maximize existing compute resources is an efficient way to get more for less, but AI leaders and executives need to look ahead at future-proof technology that is flexible and scalable to support future AI/ML compute needs.

When thinking about AI/ML in production, model training, where models are trained on a test data set and learn how to make sense of data, is just one part of a more holistic workflow. Inference is a key part of moving AI into production. Inference involves taking a trained machine learning (ML) model and using it to make real-time predictions or to solve tasks. It powers use cases across a multitude of industries, including health care, automotive, and telecommunications. With the demand for Large Language Model (LLM)-powered products, inference can power real-time answers to enterprise end users.

In this report, we share our global AI Infrastructure survey results, including 1) respondents' compute infrastructure growth plans, 2) current scheduling and compute solutions experience, and 3) model and AI framework use and plans for 2024. Read on to dive into key findings!

1 - Page 10 and/or 14 of AIIA and ClearML, **Enterprise Generative AI Adoption: C-Level Key Considerations, Challenges, and Strategies for Unleashing AI at Scale** - https://go.clear.ml/new-research-report-on-enterprise-generative-ai-adoption
2 - Page 16 of AIIA and ClearML, IBID.
3 - Page 13 of AIIA and ClearML, **The Hidden Costs, Challenges, and Total Cost of Ownership of Generative AI Adoption in the Enterprise** - https://go.clear.ml/the_hidden_costs_challenges_and_tco_of_gen_ai_adoption

KEY FINDINGS

1 96% of companies plan to expand their AI compute capacity and investment with availability, cost, and infrastructure challenges weighing on their minds.

Nearly all respondents (96%) plan to expand their AI compute infrastructure, with 40% considering more on-premise and 60% considering more cloud, and they are looking for flexibility and speed. The top concern for cloud compute is wastage and idle costs.

When asked about challenges in scaling AI for 2024, compute limitations (availability and cost) topped the list, followed by infrastructure issues. Respondents felt they lacked automation or did not have the right systems in place.

The biggest concern for deploying generative AI was moving too fast and missing important considerations (e.g. prioritizing the wrong business use cases). The second-ranked concern was moving too slowly due to a lack of ability to execute.

2 A staggering 74% of companies are dissatisfied with their current job scheduling tools and face resource allocation constraints regularly, while limited on-demand and self-serve access to GPU compute inhibits productivity.

Job scheduling capabilities vary, and executives are generally not satisfied with their job scheduling tools, and report that productivity would dramatically increase if real-time compute was self-served by data science and machine learning (DSML) team members.

74% of respondents see value in having compute and scheduling functionality as part of a single, unified AI/ML platform (instead of cobbling together an AI infrastructure tech stack of stand-alone point solutions), but only 19% of respondents actually have a scheduling tool that supports the ability to view and manage jobs within queues and effectively optimize GPU utilization.

Respondents reported they have varying levels of scheduling functionality and features, leading with quota management (56%), and followed by Dynamic Multi-instance GPUs/GPU partioning (42%), and the creation of node pools (38%).

65% of companies surveyed use a vendor-specific solution or cloud service provider for managing and scheduling their AI/ML jobs. 25% of respondents use Slurm or another open source tool, and 9% use Kubernetes alone, which does not support scheduling capabilities. 74% of respondents report feeling dissatisfied or only somewhat satisfied with their current scheduling tool.

The ability for DSML practitioners to self-serve compute resources independently and manage job scheduling hovers between 22-27%. However, 93% of survey respondents believe that their AI team productivity would substantially increase if real-time compute resources could be self-served easily by anyone who needed it.

3 The key buying factor for inference solutions is cost.

To address GPU scarcity, approximately 52% of respondents reported actively looking for cost-effective alternatives to GPUs for inference in 2024 as compared to 27% for training, signaling a shift in AI hardware usage. Yet, one-fifth of respondents (20%) reported that they were interested in cost-effective alternatives to GPU but were not aware of existing alternatives.

This indicates that cost is a key buying factor for inference solutions, and we expect that as most companies have not reached Gen AI production at scale, the demand for cost-efficient inference compute will grow.

4 The biggest challenges for compute were latency, followed by access to compute and power consumption.

Latency, access to compute, and power consumption were consistently ranked as the top compute concerns across all company sizes and regions. More than half of respondents plan to use LLMs (LLama and LLama-like models) in 2024, followed by embedding models (BERT and family) (26%) in their commercial deployments in 2024. Mitigating compute challenges will be essential in realizing their aspirations.

5 Optimizing GPU utilization is a major concern for 2024-2025, with the majority of GPUs underutilized during peak times.

40% of respondents, regardless of company size, are planning to use orchestration and scheduling technology to maximize their existing compute infrastructure.

When asked about peak periods for GPU usage, 15% of respondents report that less than 50% of their available and purchased GPUs are in use. 53% believe 51-70% of GPU resources are utilized, and just 25% believe their GPU utilization reaches 85%. Only 7% of companies believe their GPU infrastructure achieves more than 85% utilization during peak periods.

When asked about current methods employed for managing GPU usage, respondents are employing queue management and job scheduling (67%), multi-instance GPUs (39%), and quotas (34%). Methods of optimizing GPU allocation between users include Open Source solutions (24%), HPC solutions (27%), and vendor-specific solutions (34%). Another 11% use Excel and 5% have a home-grown solution. Only 1% of respondents do not maximize or optimize their GPU utilization.

6 Open Source AI solutions and model customization are top priorities, with 96% of companies focused on customizing primarily Open Source models.

Almost all executives (95%) reported that having and using external Open Source technology solutions is important for their organization.

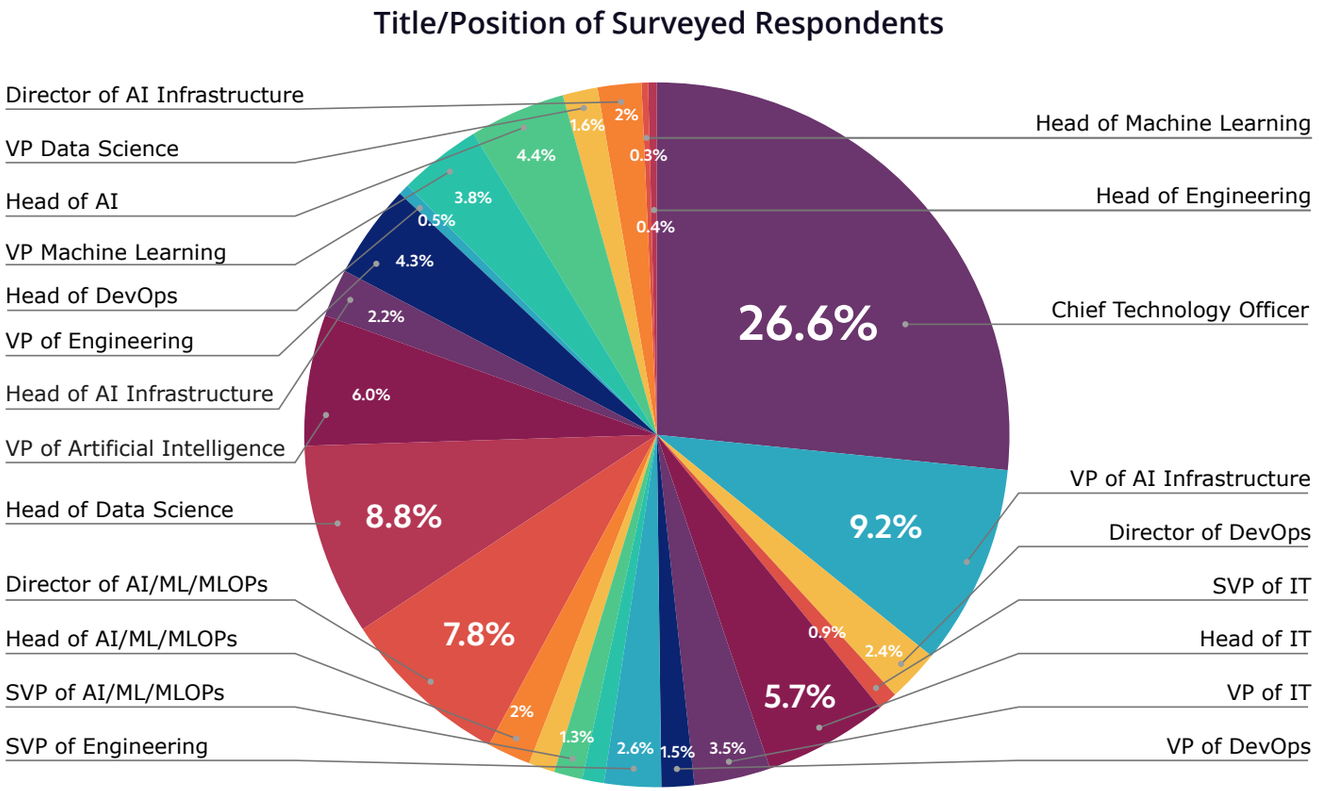
In addition, 96% of companies surveyed are currently or planning to customize Open Source models in 2024, with Open Source frameworks having the highest adoption globally. PyTorch was the leading framework for customizing Open Source models, with 61% of respondents using PyTorch, 43% using TensorFlow, and 16% using Jax. Approximately one-third of respondents currently use or plan to use CUDA for model customization.

DEMOGRAPHY

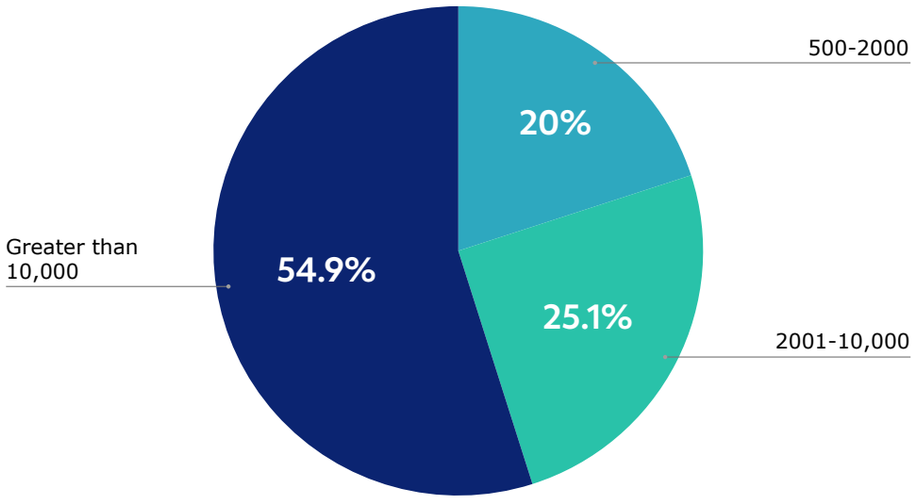
We surveyed a mix of company sizes, with 20% of respondents working in companies with 500-2000 employees, 25% with 2,000-10,000 employees, and a majority of 55% skewing to enterprises with more than 10,000 employees. We included larger companies due to our hypothesis that they would have higher AI infrastructure maturity and thus be best suited to share relevant experiences in the Generative AI space.

We talked primarily with AI/ML and technology leadership and team leads, with job titles such as CTO, Head of AI, VP of Data, or VP of Artificial Intelligence. We also included Directors of Engineering and Heads of Data Science. Therefore our results primarily represent the C-suite and more senior engineers with decision-making power.

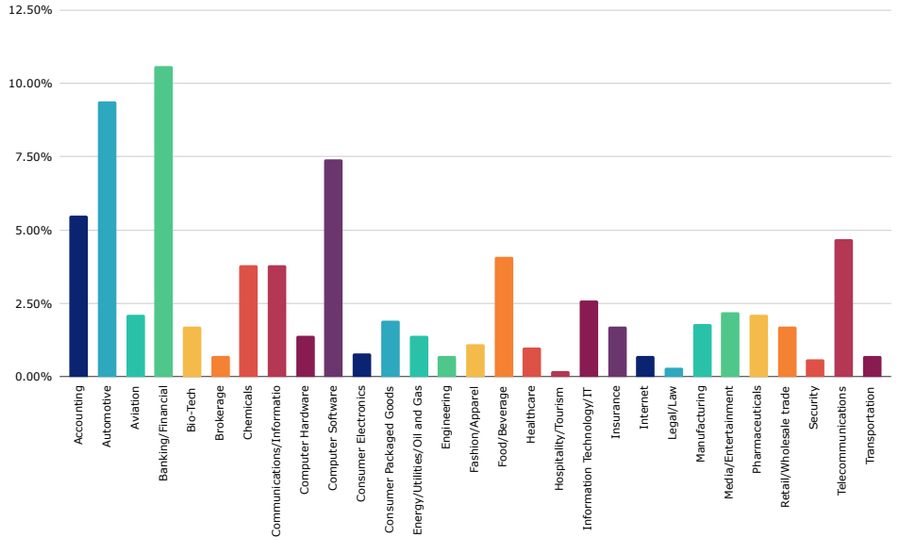
We targeted major economies in North America, Europe, and Asia-Pacific. We included a large range of verticals, including manufacturing, telecommunications, energy, food, healthcare, legal, and more. The largest representations came from Information Technology companies, but no vertical surveyed represented more than 8% of the total respondents.



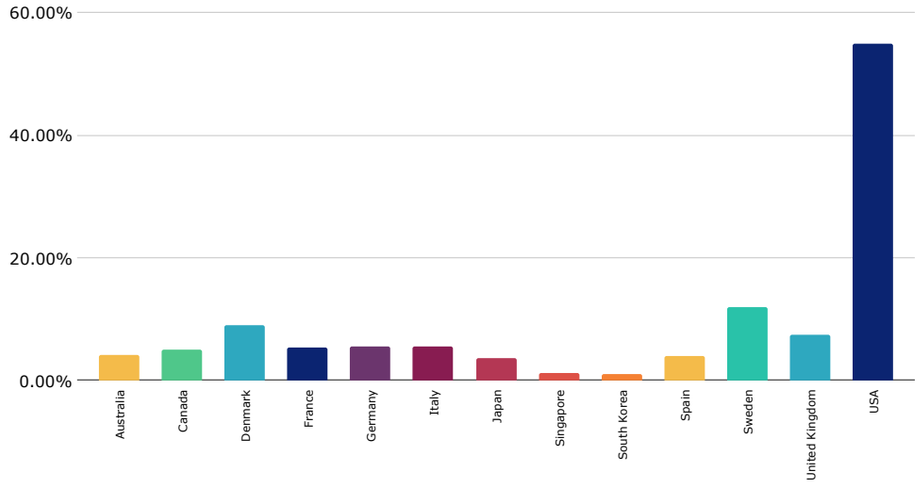
Company Size



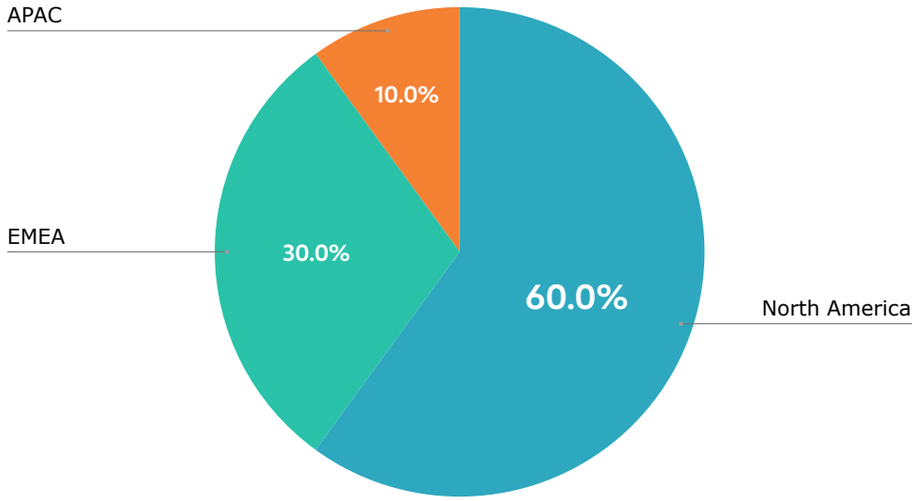
Industries Represented by Respondents



Headquarters of Responding Businesses



Geographic Regions of Headquarters



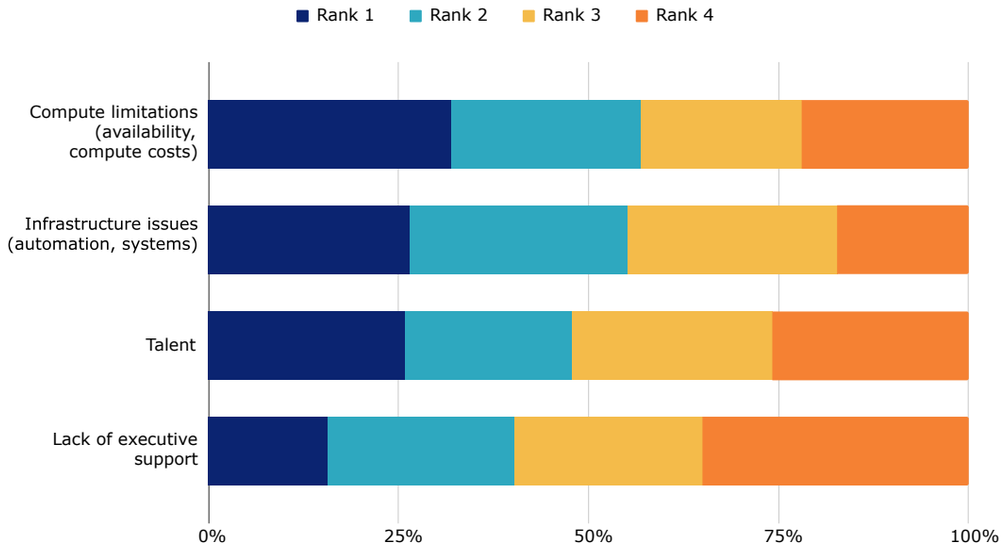
QUESTIONS AND INSIGHTS

SECTION I

Planning for 2024: Key Drivers of Expanding AI Infrastructure & Challenges in Scaling AI and Compute

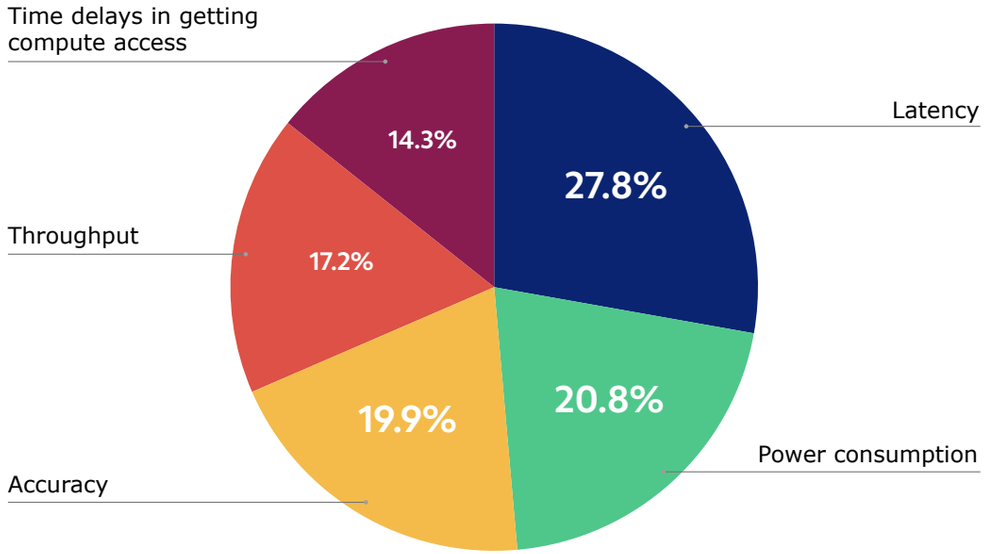
1) What are your biggest challenges in scaling AI at your organization?

Companies' biggest challenge in scaling AI for 2024 is compute limitations (availability and cost); it's the top-ranked issue for 32% of respondents. The next biggest challenge was infrastructure issues, which was the top challenge for 27% of respondents and second-ranked challenge for 29% of respondents. Respondents felt they lacked automation or did not have the right systems in place for scale.



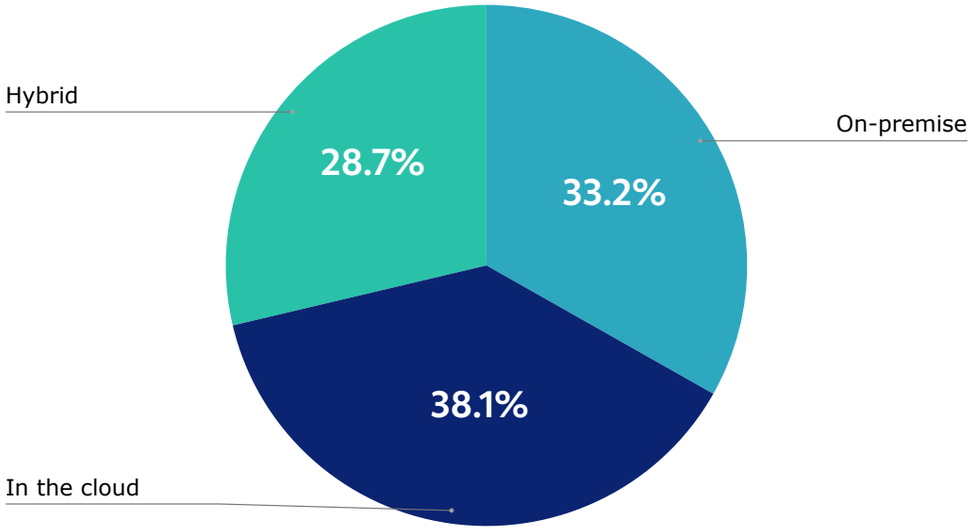
2) Rank your organization's compute concerns for 2024

When asked about their organization's compute concerns, latency was the top-ranked answer for 28% of respondents, followed by power consumption which was 21% of respondents' top-ranked issue. Time delays in getting access to compute is also weighing on respondents' minds; although it was top-ranked for only 14% of respondents, it received 30% of the votes as the second-ranked concern.



3) Cloud, On-premise, or Hybrid? What type of AI Infrastructure setup does your organization currently have for AI compute resources?

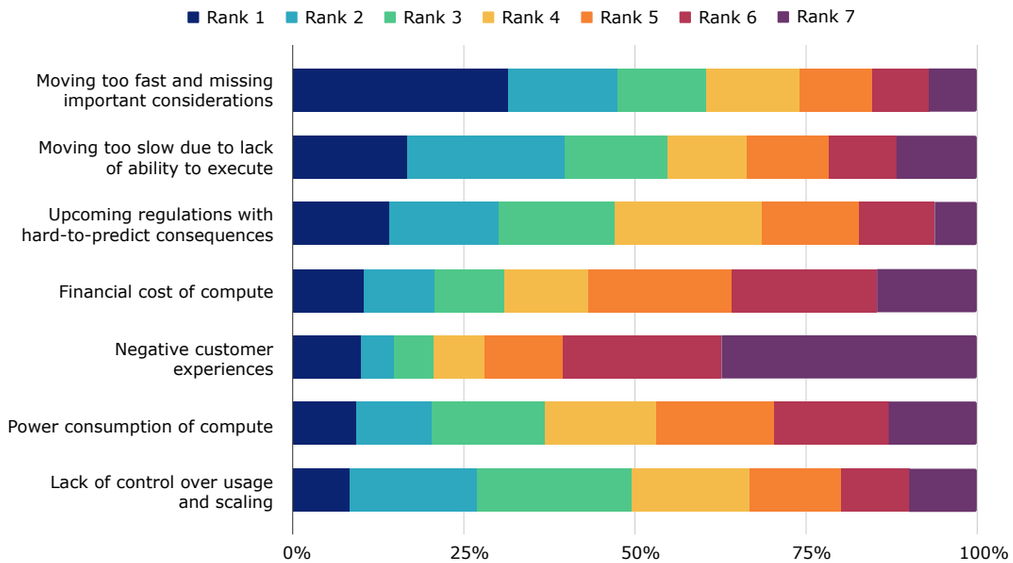
Respondents are fairly evenly divided between their current infrastructure setup. 33% have compute fully on-premise, 38% are fully cloud, and a little less than 29% have hybrid environments of both on-prem and cloud.



4) What is your organization's greatest concern about deploying Generative AI?

The biggest concern for deploying Generative AI was moving too fast and missing important considerations (e.g. prioritizing the wrong use cases), whereas the second most-important concern was moving too slow due to lack of ability to execute, exposing ambiguity amidst leadership. It appears that executives are caught between the desire to move quickly and the danger of costly mistakes.

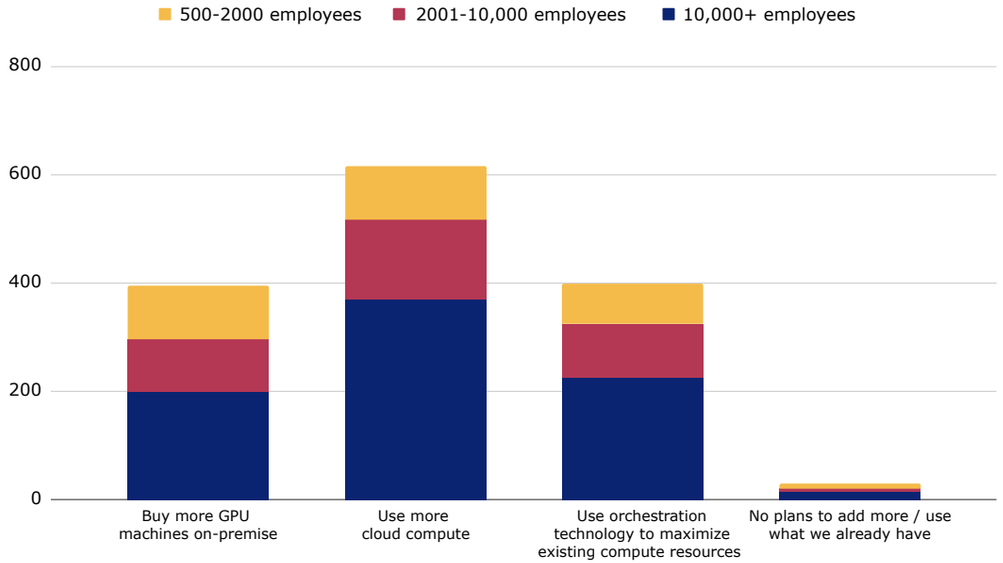
Governance also weighed in the back of respondents' minds, with upcoming regulations and lack of control over usage and scaling as the next two most-important concerns.



5) In 2024, what are your plans for expanding your AI compute infrastructure?

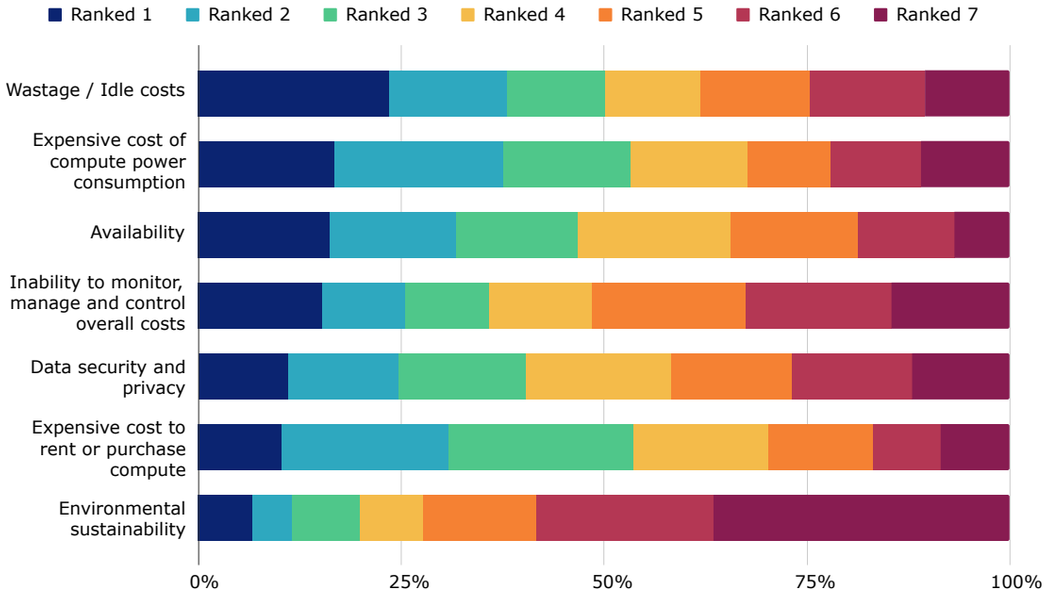
96% of companies surveyed plan to expand their AI Infrastructure in 2024. Overall, more than 60% plan to use more cloud compute, and 40% of respondents are planning to buy more GPU machines on-premise in 2024.

40% plan to use orchestration technology to maximize existing compute resources. Respondents from companies of all sizes are planning to use orchestration technology to get more from existing resources.



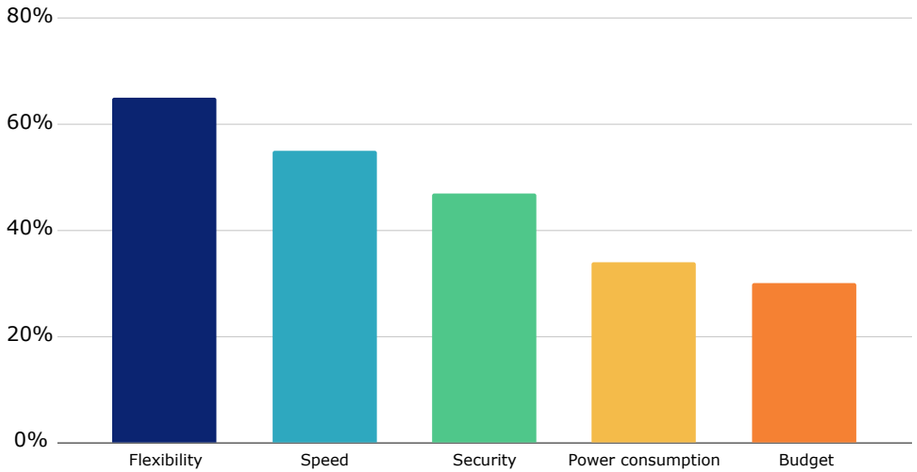
6) What are your organization's biggest concerns about cloud compute?

With 60% planning to expand AI infrastructure with more cloud compute, they will need to plan to face challenges regarding cost. Wastage / idle costs were executives' biggest concern with cloud compute, followed by the expensive cost of compute power consumption.



7) What are the key drivers and considerations in your 2024 plans for expanding your AI compute infrastructure?

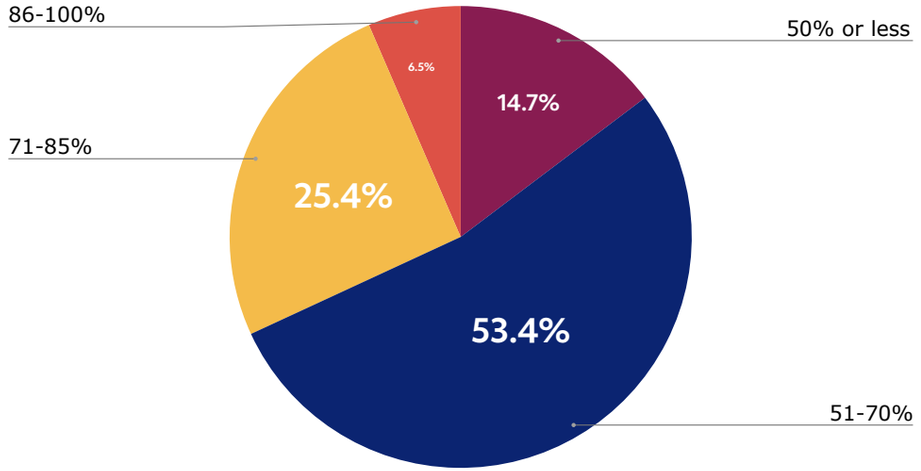
Respondents reported that flexibility and speed are the top drivers, with 65% and 55% citing these factors (respectively) when expanding AI infrastructure. This supersedes security and even budget, suggesting that respondents may be willing to pay a premium for infrastructure options that are more extensible and facilitate AI/ML output, even if it means overspending.



8) Estimate your current allocation of existing GPU resources (i.e. non-idle GPUs) during peak periods.

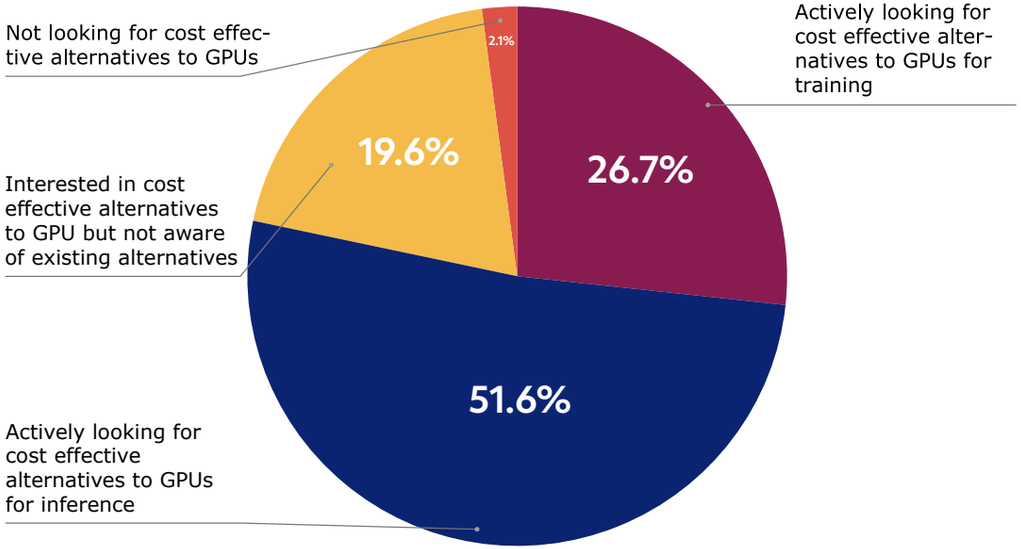
When asked about peak periods for GPU usage, 15% of respondents report that fewer than 50% of available GPUs are in use. 53% believe 51-70% of GPU resources are utilized, and 25% believe their GPU utilization reaches 85%. Only 7% of companies believe their GPU infrastructure achieves more than 85% utilization during peak periods.

Most respondents (78%) are using more than 50% of their total allocation of existing GPU resources during peak periods, indicating the need to better manage their existing compute and/or expand their compute with alternatives.



9) How is your company planning to address GPU scarcity in 2024?

To address GPU scarcity, approximately 52% of respondents reported actively looking for cost-effective alternatives to GPUs for inference as compared to 27% for training in 2024, signaling a significant shift in AI hardware usage. One-fifth of respondents (20%) reported that they were interested in cost-effective alternatives to GPU, but were not aware of existing alternatives. Cost appears to be a key driver of buying decisions for inference solutions.

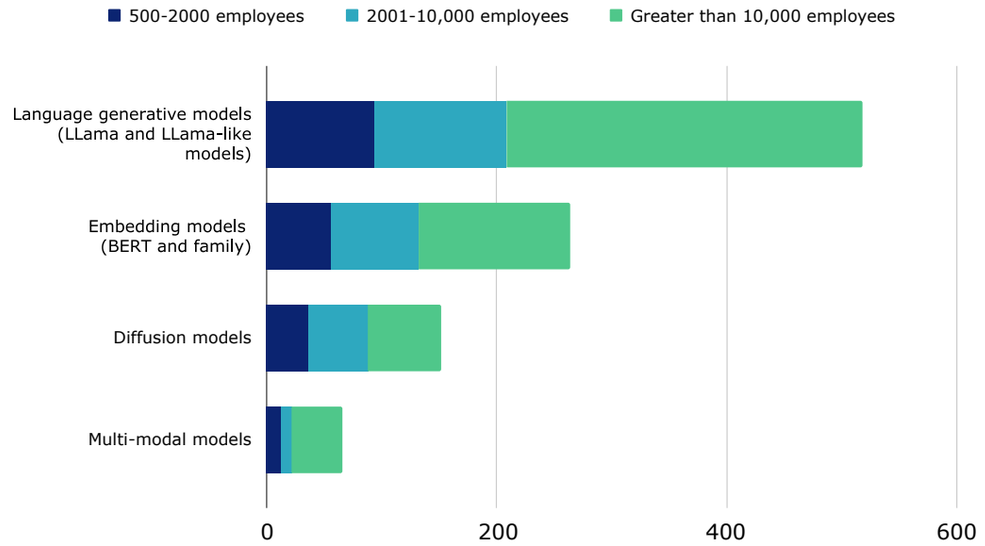


10) In 2024, what LLM models do you plan to use in your commercial deployments?

More than half of respondents plan to use LLMs (LLama and LLama-like models) in 2024, followed by embedding models (BERT and family) (26%) in their commercial deployments in 2024. This indicates that inference workloads are expected to grow substantially - more than half of respondents plan to use LLMs in production.

Notably, only 15% reported plans to use diffusion models and only 7% had plans to use multi-modal models. Why might this be? One potential explanation is that tech leaders are leveraging approaches such as model chaining to enable the use of multiple models in production. Another potential reason could be that when thinking about LLMs, tech leaders are thinking about multimodal.

In future research, we will delve into tech leaders' perceptions, plans, and use of ML models in production.

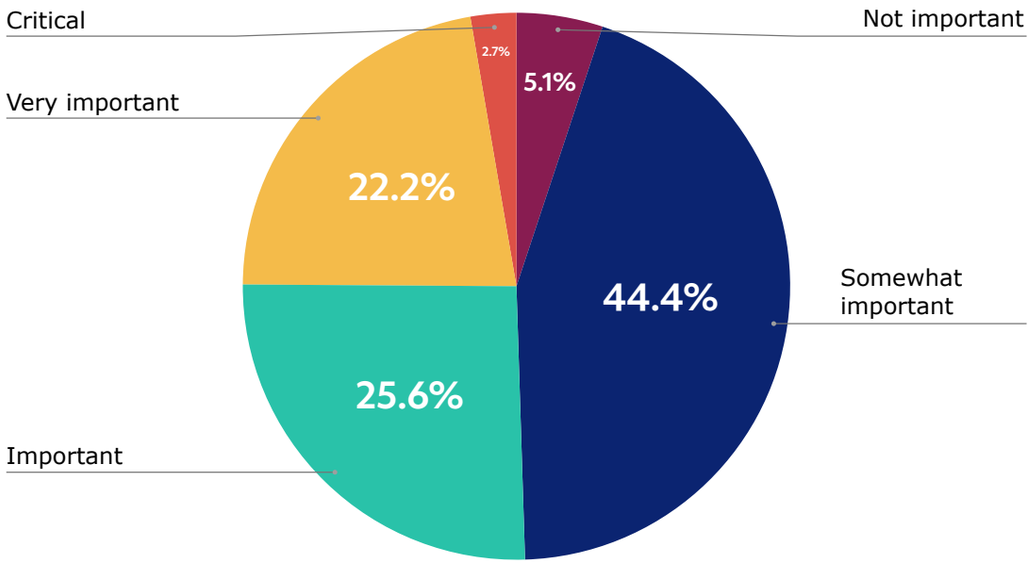


SECTION II

Open Source Demand & Model Customization Plans

1) How important is it for your organization's external technology solutions to be Open Source?

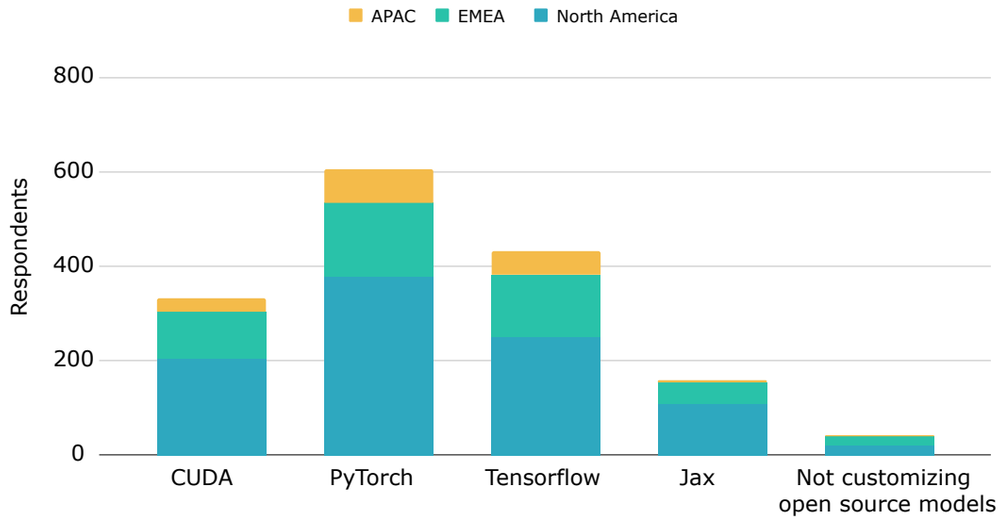
Almost all executives (95%) reported that having Open Source external technology solutions is at least somewhat important for their organization, with 26% deeming it very important or critical.



2) How is your company planning/what is your company currently using to customize your Open Source models in 2024?

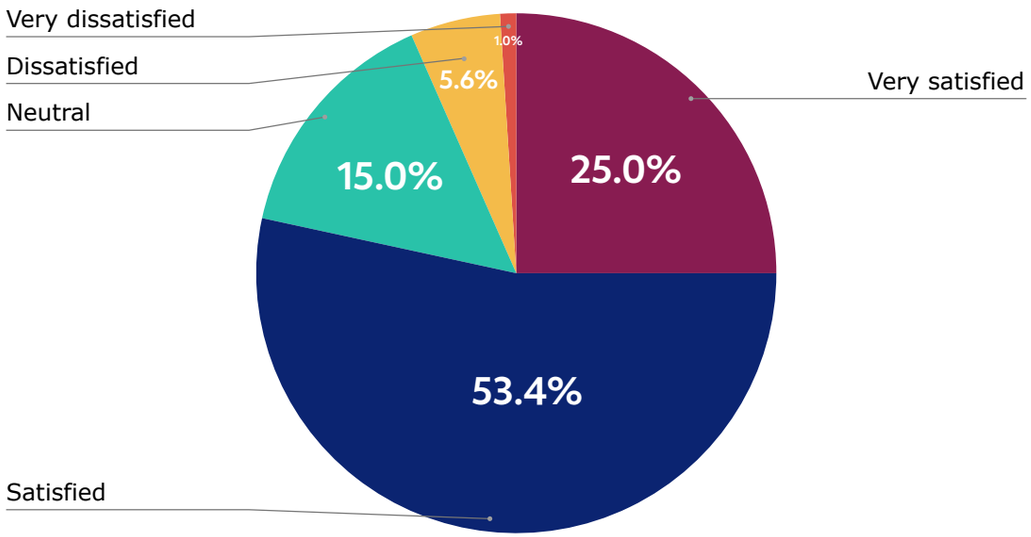
96% of companies surveyed are currently or planning to customize models in 2024, with Open Source frameworks having the highest adoption globally. Across survey responses, PyTorch is the leading framework for customizing Open Source models, with 61% of respondents using PyTorch, 43% using TensorFlow, and 16% using Jax. Approximately one-third of respondents currently use or plan to use CUDA for model customization. Only 4% do not currently or plan to customize models in 2024.

PyTorch and TensorFlow had significant market share in APAC, whereas CUDA adoption is generally equal across the regions, and Jax the most popular in North America.



3) How satisfied are you with your company's current solutions to customize Open Source models? (e.g. PyTorch, CUDA)

Most respondents use Open Source frameworks for model customization and were satisfied with their current solutions to customize Open Source models. More than 78% of respondents are satisfied or very satisfied with their current solution, indicating that Open Source frameworks are providing respondents with what they need.



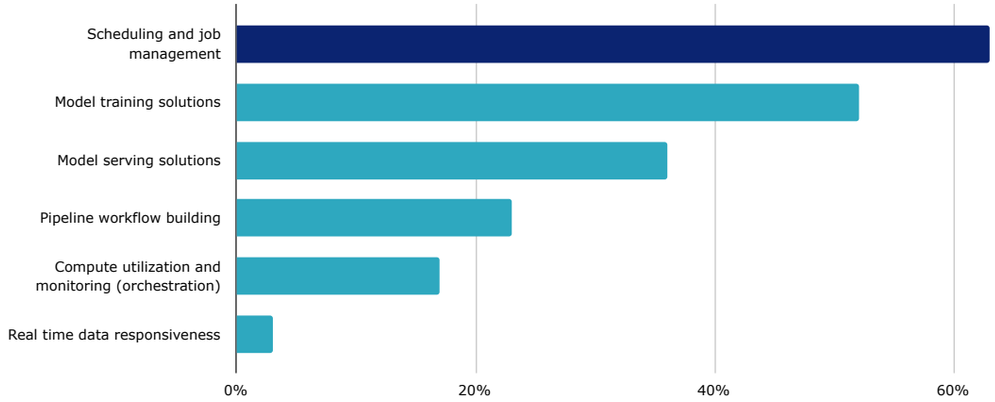
4) What types of solutions does your organization currently lack in your AI/ML tech stack?

Tech leaders and executives are dealing with issues with scheduling and job management (63%), model training solutions (52%), and model serving (36%). To effectively deal with these current issues while driving success in their plans for 2024, they will need to carefully manage their infrastructure expansion while planning for higher demand for compute – these are decisions that can not be easily changed and are costly if wrong.

With ambitious plans to use LLMs in production and to address GPU scarcity with alternatives for inference in the future, executives are making tough decisions that require balancing their current challenges with their readiness for future plans.

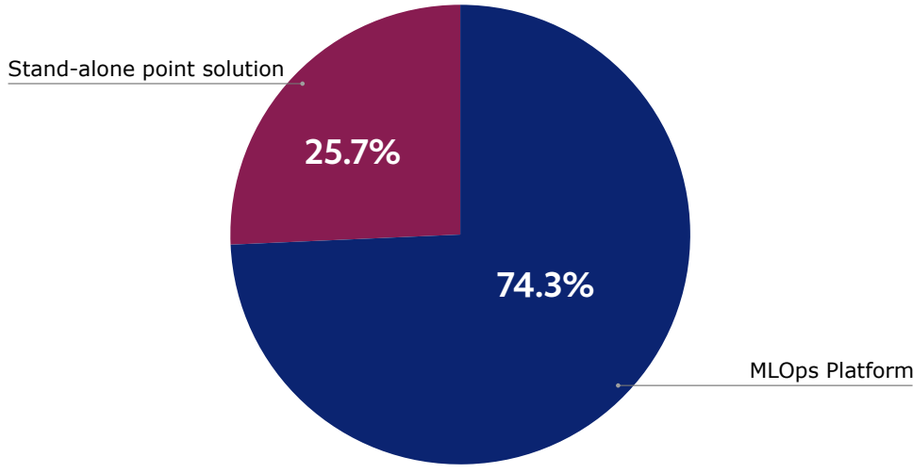
Scheduling remains a big pain point in AI stacks.

Model serving, which refers to hosting ML models and enabling access to model functionalities through APIs, is a core component of building applications that integrate AI (e.g. for AI-driven applications). Approximately one-third of respondents currently lack model serving solutions. Due to Generative AI models requiring highly performant inference workloads, we expect the need for model serving solutions to grow.



5) Do you see the value in having compute and scheduling functionality as part of an AI/ ML platform or as a stand-alone point solution?

Almost 75% of respondents see value in having compute and scheduling as part of an end-to-end platform. (74%). Compute is a fundamental part of such a platform, enabling fast and efficient model development and deployment. Coupling compute access with scheduling capabilities can be low-hanging fruit as a catalyst for AI/ML development.

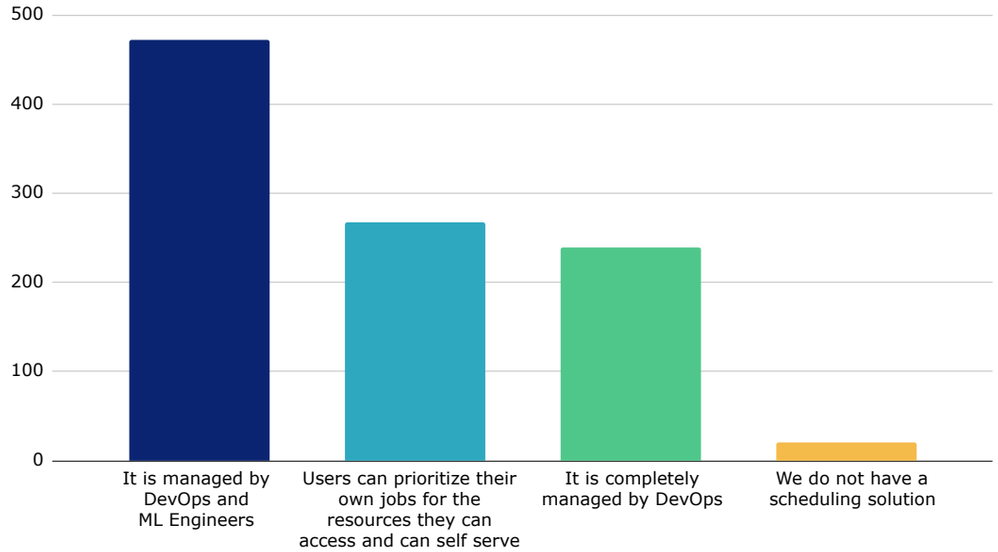


SECTION III

Job Scheduling

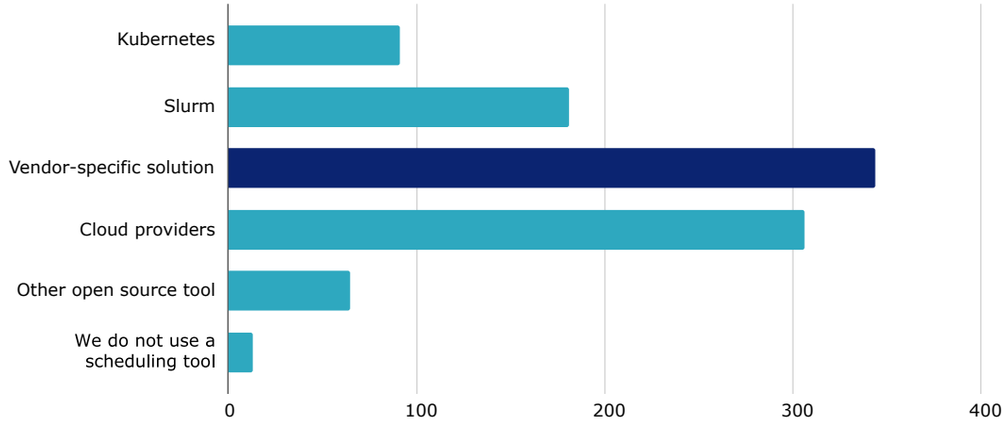
1) How is job scheduling managed at your organization (queues, job prioritization, etc.)?

Approximately half of respondents (47%) reported that job scheduling was managed by DevOps and ML Engineers, and only 27% of companies offer users the ability to self-serve, indicating a significant opportunity for companies to improve their AI/ML infrastructure to streamline development.



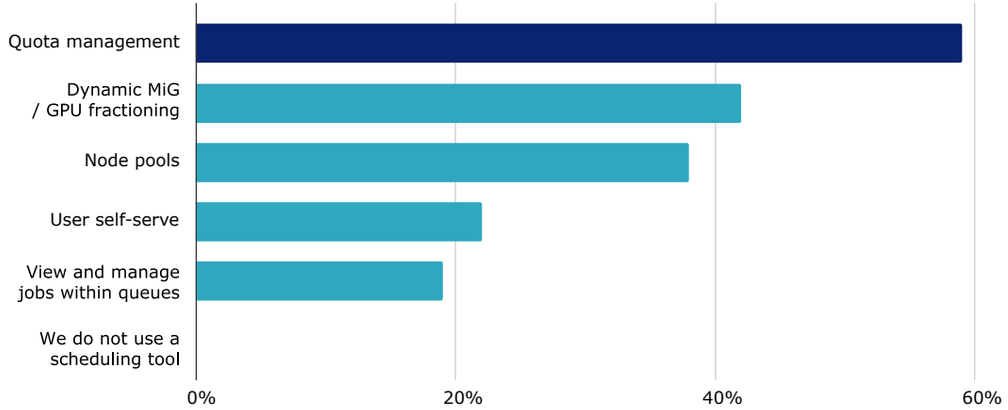
2) Which computing resource scheduling / job management tool do you currently use as part of your AI/ML tech stack?

65% of companies surveyed use vendor-specific solutions or cloud solution providers for managing and scheduling their AI/ML jobs. 25% use Slurm or another Open Source tool, and 9% use Kubernetes alone, which does not support scheduling.



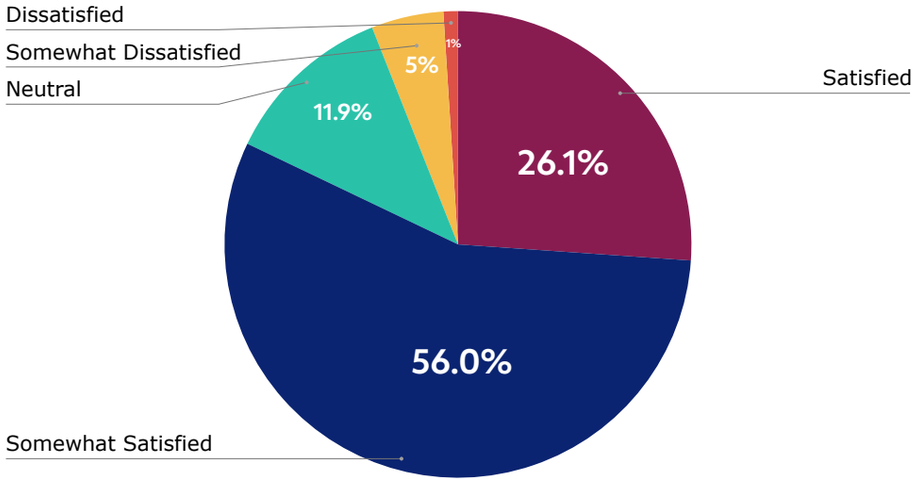
3) What are you able to do with the scheduling tool you currently have?

More than half of respondents (56%) can do quota management, followed by only 42% who have the capability to manage Dynamic MiG / GPU partioning (42%) capabilities to optimize GPU utilization. Only 19% of respondents are able to view and manage jobs within queues, potentially resulting in inefficiency.



4) Are you satisfied with the computing resource scheduling / job management tool you've chosen?

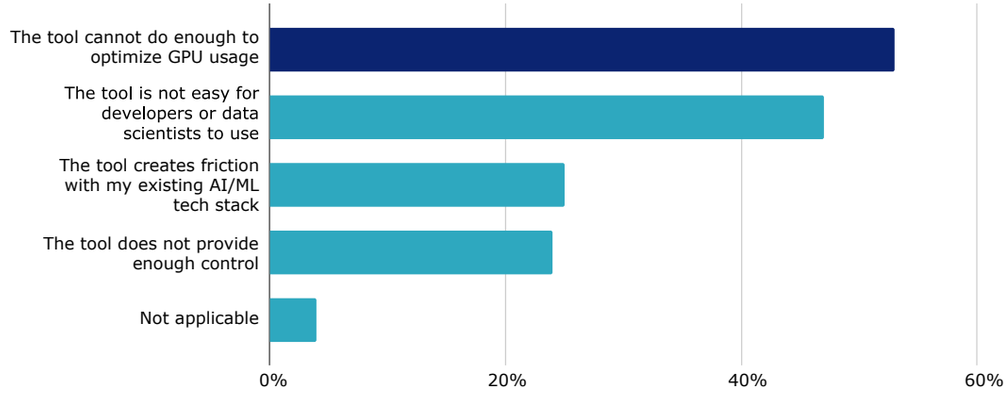
61% of respondents are either dissatisfied, somewhat dissatisfied, or partially satisfied with the scheduling tool they have chosen, with another 12% reporting they are neutral, indicating room for improvement.



5) If you chose Neutral, Somewhat Dissatisfied, or Dissatisfied in the previous question, what are the main reasons for your dissatisfaction?

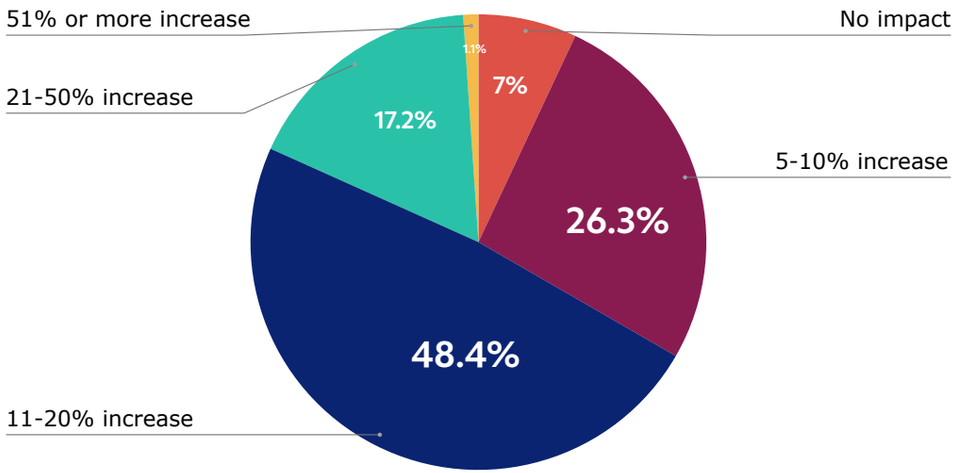
The main drivers of pain points were that the tool cannot do enough to optimize GPU usage (53%), followed by the tool is not easy for developers or data scientists to use (47%). It is also notable that approximately 25% reported lack of control and friction with existing AI/ML stacks as reasons for dissatisfaction.

Given these results, we recommend investment in a scheduling tool that offers optimization of GPU usage, is easy to use, and works well with other pieces of AI/ML infrastructure.



6) How would your organization's AI team productivity be impacted/ increased if real-time compute could be self-served and easily accessed by anyone who needed it in a seamless and cost-controlled way?

93% reported that their organizations' AI team productivity would increase if real-time compute could be self-served easily by anyone who needed it.

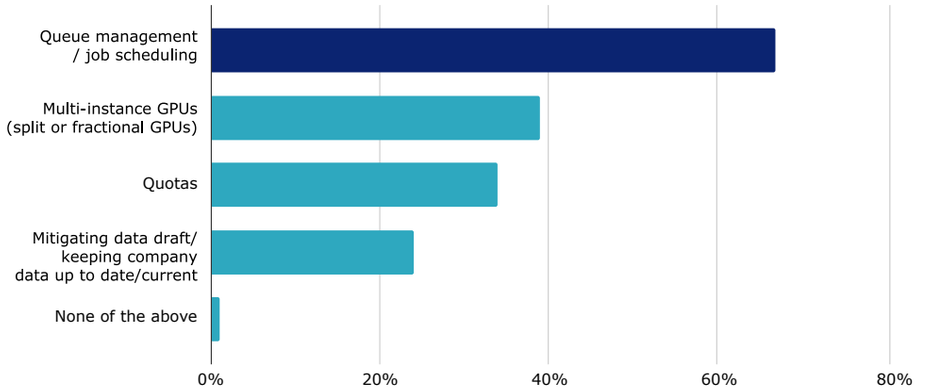


SECTION IV

Optimizing Compute Utilization

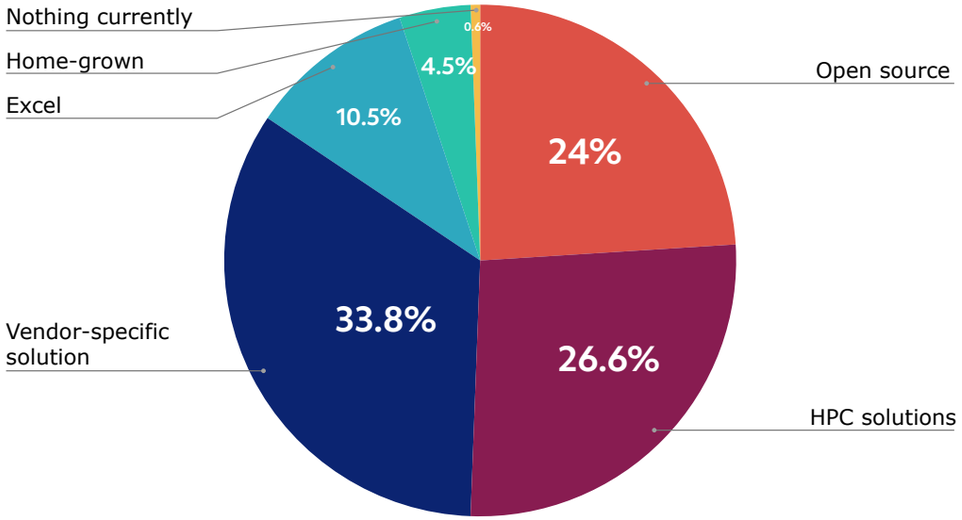
1) How does your organization currently maximize utilization of GPU usage?

We were impressed to see the sophistication of how companies are managing their compute infrastructure. For respondents, the top 3 methods being employed to maximize GPU utilization are queue management and job scheduling (67%), multi-instance GPUs (39%), and quotas (34%).



2) What tool do you use to optimize GPU allocation between users?

Methods of optimizing GPU allocation between users include Open Source solutions (24%), HPC solutions (27%), and vendor-specific solutions (34%). Another 11% use Excel and 5% have a home-grown solution. Only 1% of respondents are doing nothing to maximize or optimize their GPU utilization.

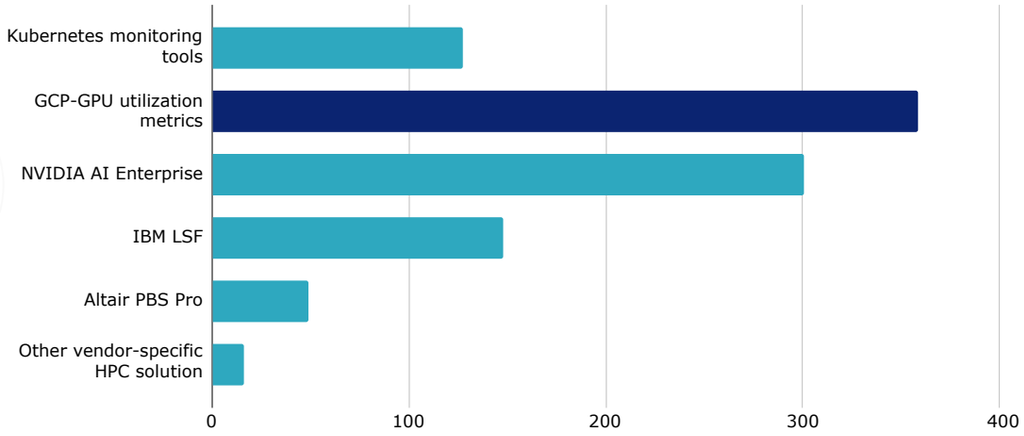


SECTION V

Monitoring Compute

1) What tool does your organization use to monitor GPU cluster utilization?

For monitoring GPU cluster utilization, using GCP-GPU utilization metrics 36% tops the list, followed by NVIDIA AI Enterprise (30%). IBM LSF and Kubernetes were selected by 15% and 13% of respondents, respectively.



CONCLUSION

As we've seen from these global survey results, while most organizations are planning to expand their AI infrastructure, executives and tech leaders face diverse challenges in their current workloads. Their ambitious plans for the future signal a **need for highly performant, cost-effective alternatives to GPUs and seamless, end-to-end AI/ML platforms**.

GPU utilization is a major concern for 2024, with the majority of respondents saying they are not maximizing their GPUs at peak times. Nearly all companies we surveyed reported that AI team productivity would increase if real-time compute could be self-served easily by anyone who needed it.

Tech leaders and executives have ambitious plans for LLMs -- and **mitigating compute challenges will be essential in realizing their aspirations**. More than half plan to use LLMs in commercial deployments and more than half are looking for cost-effective alternatives to GPUs for inference. We believe that highly performant inference workloads with low latency and efficient power consumption will be crucial to reducing the TCO of Generative AI deployments

Alongside this shift, executives and tech leaders value Open Source solutions. Nearly all reported that having Open Source solutions was at least somewhat important for their organization. Accordingly, we observed that Open Source AI frameworks are preferred for model customization – with PyTorch leading TensorFlow and Jax in production.

For successful deployment of AI at scale, taking a holistic view of AI workloads will be essential. We observed a lack of scheduling tools that support the ability to view and manage jobs within queues. While current gaps in AI tech stacks include training, model serving is top of mind. **Executives and tech leaders will need to balance solving their current pain points while executing on their future plans**.

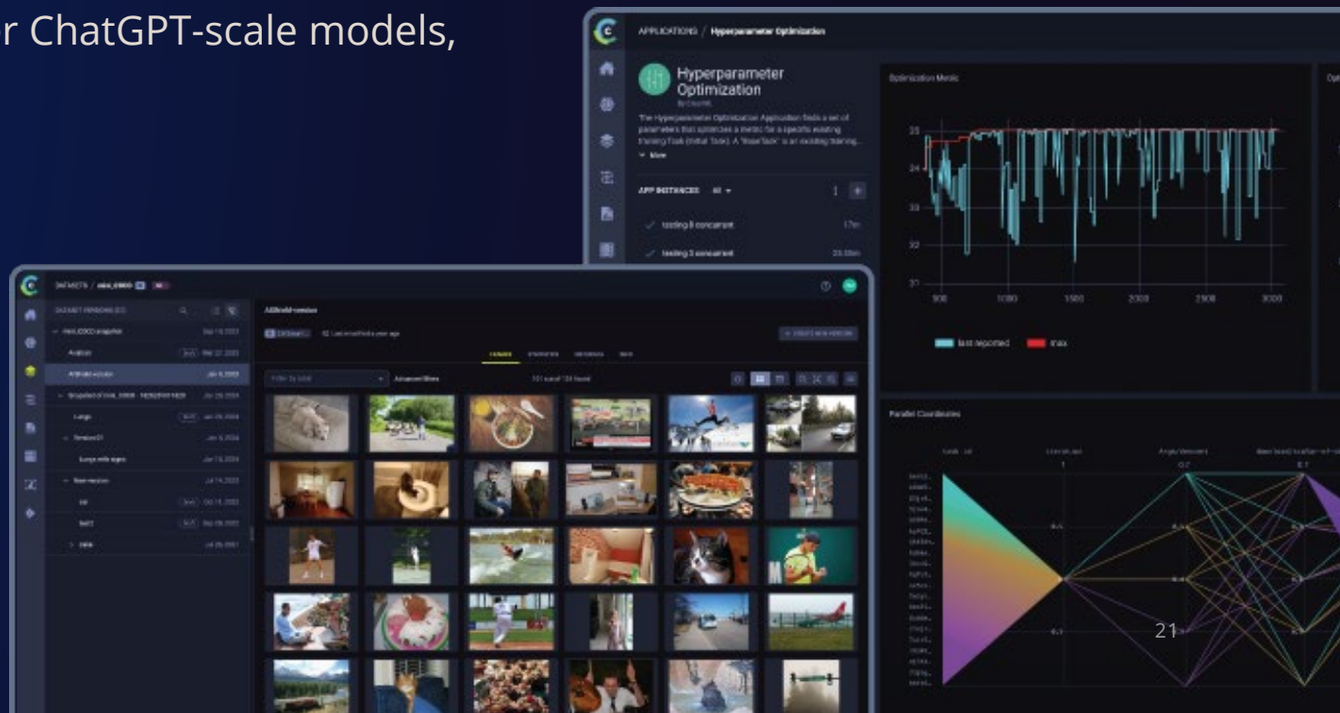
Time delays in getting compute access, and high latency can break product experiences. We recommend that **AI leaders consider the multiple factors that impact their TCO**, such as: compute, scheduling, and power-efficient inference with low latency when planning for Gen AI business adoption. Only then can they be confident in accurately predicting and forecasting the TCO for Gen AI in their organization.

We hope that the insights in this report have shed light into the experiences of leaders making decisions for today and in the future, and that these insights will empower you to find solutions that bring AI into production in a way that's aligned with your organization's vision.

NEXT STEPS

From AI/ML model development to training, and inference, ClearML's Dynamic Scheduling, Orchestration, and MiG helps you maintain optimal GPU/compute utilization for any cluster size and scale. To request a demo of ClearML, please visit https://clear.ml/demo.

FuriosaAI's highly efficient inference-focused products run cutting-edge distributed inference with low TCO. To request a demo of FuriosaAI's 1st-gen AI chip WARBOY for Computer Vision, please visit https://www.furiosa.ai/getstarted. For more information about FuriosaAI's 2nd-gen AI chip with High Bandwidth Memory 3 (HBM3), which provides H100-level performance to power ChatGPT-scale models, visit https://www.furiosa.ai/comingsoon.



About AIIA

The AI Infrastructure Alliance is dedicated to bringing together the essential building blocks for the Artificial Intelligence applications of today and tomorrow. The Alliance and its members bring striking clarity to this quickly developing field by highlighting the strongest platforms and showing how different components of a complete enterprise machine-learning stack can and should interoperate. They deliver essential reports and research, virtual events packed with fantastic speakers, and visual graphics that make sense of an ever-changing landscape. To learn more, visit https://ai-infrastructure.org/.

About FuriosaAI

FuriosaAI is a semiconductor company designing high-performance data center AI accelerators with vastly improved power efficiency. Through our innovative architecture and products, we strive to unlock the transformative potential of AI and make its benefits accessible to all. Our first generation chip WARBOY, which runs computer vision applications for data centers and enterprise customers, is available now and our second generation chip for LLM and multimodal deployment will launch later this year. To learn more, visit the company's website at: https://www.furiosa.ai/.

About ClearML

As the leading Open Source, end-to-end solution for unleashing AI in the enterprise, ClearML is used by more than 1,600 enterprise customers to develop a highly repeatable process for their end-to-end AI model lifecycle, from product feature exploration to model deployment and monitoring in production. Use all of our modules for a complete ecosystem or plug in and play with the tools you have. ClearML is an NVIDIA DGX-ready Software Partner and is trusted by more than 250,000 forward-thinking Data Scientists, Data Engineers, ML Engineers, DevOps, Product Managers and business unit decision makers at leading Fortune 500 companies, enterprises, academia, and innovative start-ups worldwide. To learn more, visit the company's website at https://clear.ml.