Enterprise Strategy Group by TechTarget

SHOWCASE

NOVEMBER 2024

Accelerate AI Time to Value With AMD Instinct™ MI300X Accelerators on Oracle Cloud Infrastructure

Scott Sinclair, Practice Director; and Monya Keane, Senior Research Analyst

Abstract: The rise in adoption of artificial intelligence (AI) initiatives has come with a corresponding increase in the size of AI models. As organizations seek better accuracy, they are placing an increased importance on leveraging the right infrastructure, accelerators, and cloud environment to ensure success.

Overview

The current state of AI initiatives is robust, and trends show that many AI models are increasing in size as organizations seek to improve both AI-related performance and the accuracy of the models' output. Many organizations now consider AI initiatives very helpful to achieving competitive success in today's environment.

TechTarget's Enterprise Strategy Group has conducted research into these trends, including surveying IT and business decision-makers who are familiar and involved with their organizations' 2024 IT budgeting and spending related to information management (i.e., data science, machine learning, and AI). According to the findings, 78% of those organizations planned to make a significant investment in data science this year.[1]

Multiple business drivers are propelling the rise of AI initiatives. In a survey of data and IT professionals tasked with strategizing, evaluating, purchasing, or managing infrastructure specifically to support their businesses' AI initiatives:

- 100% of them agreed that AI improves their business's ability to gain a competitive advantage in the marketplace.
- 100% agreed that AI improves processes and workflows.
- 99% agreed that AI improves employee productivity and job satisfaction.[2]

In an effort to achieve higher accuracy and improve the user experience, AI models are increasing in size. This growth places a heavier demand on the underlying cloud infrastructure. For example, generative AI workloads that require training and inferencing of large language models can strain the performance of a compute infrastructure and, thus, require clustering of GPUs to achieve the desired level of price-performance.

Any company facing these challenges should consider it good news that AMD and Oracle are partnering to offer the AMD Instinct MI300X Series accelerators on Oracle Cloud Infrastructure (OCI). These accelerators are equipped with high levels of memory capacity and bandwidth to support the growing demands of modern AI models fully.

[1] Source: Enterprise Strategy Group Complete Survey Results, *2024 Technology Spending Intentions Survey*, February 2024.
[2] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

Infrastructure's Role in ROI and Success With AI

Enterprise Strategy Group research highlights the importance of selecting the right infrastructure for AI and the fact that all cloud services are not created equally.

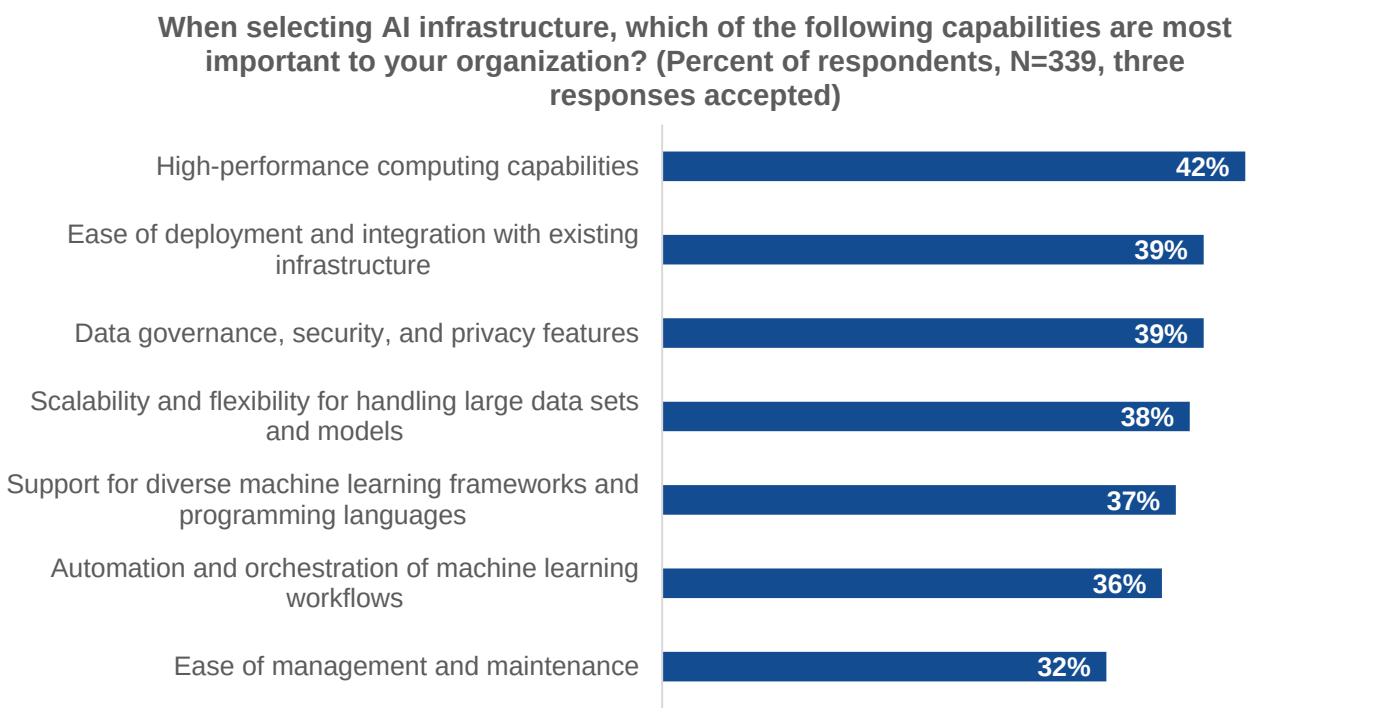
According to this research, 84% of organizations agree that the growth of AI, including generative AI, has them reevaluating their application deployment strategy. Enterprise Strategy Group also found that better support for AI projects and initiatives was the most commonly identified reason why organizations expect to have a different primary cloud provider in 24 months and/or expect to increase their budget for a secondary cloud provider (cited by 37% of respondents), beating out other decision criteria such as better support for application development, better alignment with existing skill sets, and security considerations.[3]

The takeaway is that infrastructure matters. It is fortunate that, prior to AI workload deployment, organizations are carefully evaluating not just capabilities, but also hardware requirements.

Figure 1 identifies the most important capabilities that organizations take into account when selecting infrastructure to support AI, with high-performance compute capabilities being cited most often.[4] But high performance requires more than accelerator technology—it also requires that the surrounding infrastructure (e.g., memory capacity and bandwidth) helps ensure full utilization of the environment. It is also important to be able to cluster GPUs to train large models (especially large language models). A cluster consists of multiple GPU-accelerated compute instances, connected by a dedicated ultra-high bandwidth network with access to high performance storage.

Rounding out the top considerations are ease of deployment and scalability for large data sets and models, highlighting that organizations desiring success in AI are seeking simplicity combined with performance and scalability.

Figure 1. Most Important Capabilities When Selecting Infrastructure for AI



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

[3] Source: Complete Survey Results, *Enterprise Strategy Group Research Report: Understanding Workload, App, and Data Deployment and Migration Decision-making*, July 2024.
[4] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

According to Enterprise Strategy Group research, 72% of respondents report that their organizations have seen value from their AI initiatives within the first three months.[5] The right infrastructure—such as the solution from AMD and OCI with its ability to leverage GPU clustering to train LLMs and other large models—could accelerate their success with AI even more.

The Value of AMD GPU Accelerators on OCI for AI

The OCI AI infrastructure developed through the AMD/Oracle alliance is unique. It offers:

- Bare-metal compute accelerated by AMD GPUs, which provides increased control and enables the organization to decide where and how it wants to run its AI workloads.
- AMD ROCm™—the open software stack that includes programming models, tools, compilers, libraries, and runtimes for AI solution development on AMD GPUs. It provides support for all major frameworks and models, including day zero support on PyTorch for the latest features and support for all models on Hugging Face.
- Powerful RDMA cluster networking, which enables microsecond latency and 3.2Tb/s bandwidth.
- High-performance storage with locally attached NVMe storage and clustered file systems.
- Distributed cloud solutions (such as dedicated regions and Alloy) from OCI, which enable AI infrastructure and AI workloads to be deployed anywhere.

OCI also offers a portfolio of 150+ public cloud services, including infrastructure services such as compute, storage, and networking, as well as application-layer services such as AI, containers, and database tools across all of its 50 public cloud regions and distributed cloud solutions (Dedicated Regions, Alloy, etc.).

To better enable AI workloads and support the scalability demands of growing models, Oracle and AMD engineers teamed up to integrate the AMD Instinct MI300X platform with OCI Compute. They designed the platform specifically to provide acceleration across multiple data types, giving it a large memory and ample I/O bandwidth to handle large data sets. As a result, this integrated platform offers numerous advantages to an organization, including:

- **Improved performance for generative AI.** According to Oracle and AMD, the MI300X accelerator delivers up to 1.3 times better AI performance compared with competitive accelerators, thereby providing a boost to AI inference and training workloads.
- **Increased memory density.** OCI integrates with the platform to provide 192 GB of HBM3 memory per GPU and a peak theoretical memory bandwidth of 5.3 TB/s per GPU. It does this by increasing both the size of the memory capacity and the bandwidth, meaning that the resulting solution can ensure efficient data access and reduced latency for AI processing. As a result, AI training initiatives are accelerated, risks are reduced, and time to value is improved. Additionally, the platform allows for increases in the size of the data and the model that can be used in AI initiatives.
- **Cloud delivery instead of DIY deployment.** By offering MI300X-based bare-metal instances in OCI, Oracle and AMD are able to simplify the adoption of the accelerator technology, reducing the burden on internal resources for sizing and planning, while simplifying the overall IT estate. The result further accelerates times to value for AI initiatives, particularly for large-scale clusters with hundreds or thousands of GPUs.
- **Access to AMD's ROCm library.** AMD ROCm is an open source software platform optimized to extract the best HPC and AI workload performance from AMD MI300X accelerators. The AMD ROCm Documentation website features the latest release notes, how-to guides, tutorials, examples, and other resources for developers. The ROCm Application Catalog contains an up-to-date list of applications supported by the ROCm platform, and the ROCm Developer Hub is home to developer resources including training webinars, videos, and blogs.

[5] Ibid.

Conclusion

The AI market is obviously increasing rapidly, as is the size of the AI models themselves. These trends are placing increasing demands on compute infrastructure. Simply put, AI workloads deserve the best hardware. The memory capacity and memory bandwidth advantages that Oracle and AMD have injected into Oracle's compute accelerated by AMD's MI300X series of GPUs translate into companies being able to power larger AI models using fewer server instances.

Oracle and AMD picked an excellent time to focus on innovating the MI300X series—with features such as high-bandwidth, high-density memory—to ensure those GPUs fit into the OCI and function as key elements not just in supporting AI initiatives, but in accelerating them.

Every business today is looking for any way possible to improve productivity and efficiency, to accelerate time to value, and to drive down energy costs. When it comes to AI initiatives, specifically, Oracle and AMD have given these businesses a powerful tool to meet those goals.

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