Press Release

Oracle and AMD Collaborate to Help Customers Deliver Breakthrough Performance for Large-Scale AI and Agentic Workloads

Oracle will be among the first hyperscalers to offer an AI supercomputer with AMD Instinct MI355X GPUs

OCI to deploy new zettascale AI cluster with up to 131,072 MI355X GPUs to enable customers to build, train, and inference AI at scale

Austin, Texas and Santa Clara, Calif.—Jun 12, 2025

Oracle and AMD today announced that AMD Instinct™ MI355X GPUs will be available on Oracle Cloud Infrastructure (OCI) to give customers more choice and more than 2X better price-performance for large-scale AI training and inference workloads compared to the previous generation. Oracle will offer zettascale AI clusters accelerated by the latest AMD Instinct processors with up to 131,072 MI355X GPUs to enable customers to build, train, and inference AI at scale.

“To support customers that are running the most demanding AI workloads in the cloud, we are dedicated to providing the broadest AI infrastructure offerings,” said Mahesh Thiagarajan, executive vice president, Oracle Cloud Infrastructure. “AMD Instinct GPUs, paired with OCI’s performance, advanced networking, flexibility, security, and scale, will help our customers meet their inference and training needs for AI workloads and new agentic applications.”

To support new AI applications that require larger and more complex datasets, customers need AI compute solutions that are specifically designed for large-scale AI training. The zettascale OCI Supercluster with AMD Instinct MI355X GPUs meets this need by providing a high-throughput, ultra-low latency RDMA cluster network architecture for up to 131,072 MI355X GPUs. AMD Instinct MI355X delivers nearly triple the compute power and a 50 percent increase in high-bandwidth memory than the previous generation.

“AMD and Oracle have a shared history of providing customers with open solutions to accommodate high performance, efficiency, and greater system design flexibility,” said Forrest Norrod, executive vice president and general manager, Data Center Solutions Business Group, AMD. “The latest generation of AMD Instinct GPUs and Pollara NICs on OCI will help support new use cases in inference, fine-tuning, and training, offering more choice to customers as AI adoption grows.”

AMD Instinct MI355X Coming to OCI

AMD Instinct MI355X-powered shapes are designed with superior value, cloud flexibility, and open-source compatibility—ideal for customers running today’s largest language models and AI

workloads. With AMD Instinct MI355X on OCI, customers will be able to benefit from:

- **Significant performance boost:** Helps customers increase performance for AI deployments with up to 2.8X higher throughput. To enable AI innovation at scale, customers can expect faster results, lower latency, and the ability to run larger AI workloads.
- **Larger, faster memory:** Allows customers to execute large models entirely in memory, enhancing inference and training speeds for models that require high memory bandwidth. The new shapes offer 288 gigabytes of high-bandwidth memory 3 (HBM3) and up to eight terabytes per second of memory bandwidth.
- **New FP4 support:** Allows customers to deploy modern large language and generative AI models cost-effectively with the support of the new 4-bit floating point compute (FP4) standard. This enables ultra-efficient and high-speed inference.
- **Dense, liquid-cooled design:** Enables customers to maximize performance density at 125 kilowatts per rack for demanding AI workloads. With 64 GPUs per rack at 1,400 watts each, customers can expect faster training times with higher throughput and lower latency.
- **Built for production-scale training and inference:** Supports customers deploying new agentic applications with a faster time-to-first token (TTFT) and high tokens-per-second throughput. Customers can expect improved price performance for both training and inference workloads.
- **Powerful head node:** Assists customers in optimizing their GPU performance by enabling efficient job orchestration and data processing with an AMD Turin high-frequency CPU with up to three terabytes of system memory.
- **Open-source stack:** Enables customers to leverage flexible architectures and easily migrate their existing code with no vendor lock-in through AMD ROCm. AMD ROCm is an open software stack that includes popular programming models, tools, compilers, libraries, and runtimes for AI and HPC solution development on AMD GPUs.
- **Network innovation with AMD Pollara™:** Provides customers with advanced RoCE functionality that enables innovative network fabric designs. Oracle will be the first to deploy AMD Pollara AI NICs on backend networks, providing advanced RoCE functions such as programmable congestion control and support for open industry standards from the Ultra Ethernet Consortium (UEC) for high-performance and low latency networking.

Additional Resources

- Learn more about OCI AI infrastructure
- Learn more about OCI Compute
- Learn more about AMD GPUs

Contact Info

Ben Wolfson

Oracle PR

ben.wolfson@oracle.com
+1 510.480.5230

About Oracle

Oracle offers integrated suites of applications plus secure, autonomous infrastructure in the Oracle Cloud. For more information about Oracle (NYSE: ORCL), please visit us at oracle.com.

Trademarks

Oracle, Java, MySQL, and NetSuite are registered trademarks of Oracle Corporation. NetSuite was the first cloud company—ushering in the new era of cloud computing.

AMD, the AMD Arrow Logo, AMD Instinct, Pollara, ROCm, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other names are for informational purposes only and may be trademarks of their respective owners.

© 2025 Oracle Privacy / Do Not Sell My Info Cookie Preferences Ad Choices Careers