GIGAIO BUILDS 'IMPOSSIBLE SERVERS' WITH AMD

CASE STUDY

GigaIO delivers AI training, inference, and HPC workloads on SuperNODEs powered by the AMD Instinct™ MI300X Platform

AMD × GIGAIO

"We make 'impossible servers,'" says Alan Benjamin, CEO of GigaIO. "And AMD is producing fantastic GPUs." This combination enables GigaIO customers to run compute-intense AI and high performance computing (HPC) workloads faster and with less expense on a GigaIO SuperNODE™ equipped with AMD Instinct™ MI300X GPUs.

> **"We make 'impossible servers.' And AMD is producing fantastic GPUs."**
> Alan Benjamin, CEO at GigaIO

"The SuperNODE is the world's most powerful and energy-efficient scale-up AI computing platform," says Benjamin. "Our AI fabric interconnect capabilities set GigaIO apart, effectively exploding the server because you're no longer constrained by what you can fit into a single node. Our strategic multi-fabric integration combines the low latency of our AI fabric, the memory efficiency of CXL, and the scale-out capabilities of Ultra Ethernet to deliver the ideal technology for each communication path within AI workloads."

"This enables us to build our SuperNODEs, which can attach vastly more GPUs and other accelerators to a given server so that customers can run workloads much faster and less expensively." The high bandwidth memory (HBM) and excellent price-performance ratio of the AMD Instinct MI300X Platform makes it a natural complement to GigaIO's SuperNODEs.

> **"With up to 32 AMD Instinct MI300X GPUs attached to a single server, we truly create an 'impossible server' that can be nearly any dimension."**
> Alan Benjamin, CEO at GigaIO

TURNING THE SERVER MODEL INSIDE OUT

Composability and scalability are two hallmarks of GigaIO's SuperNODEs. The AMD Instinct MI300X Platform delivers 42 petaFLOPs of peak theoretical FP8 with sparcity precision performance for generative AI and ML training and 1.3 petaFLOPs peak theoretical FP32 precision for the most challenging HPC codes. This leadership performance and efficiency dovetail with the GigaIO SuperNODE approach.

INDUSTRY
Cloud Service Provider

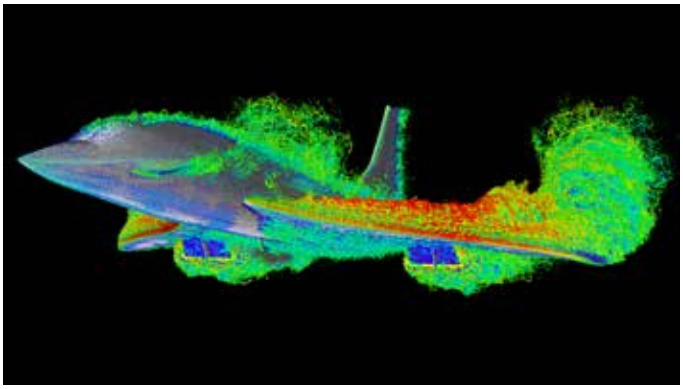
CHALLENGES
Constrained, rigid infrastructure makes it difficult to get the IT you need to scale and flex with workload demands

SOLUTION
Deploy up to 32 AMD Instinct™ MI300X GPUs per GigaIO SuperNODE™

RESULTS
SuperNODEs with AMD Instinct MI300X GPUs deliver processing power of more than 46,000 tokens per second for AI training and inference, providing excellent price-to-performance ratio

AMD TECHNOLOGY AT A GLANCE
AMD Instinct™ MI300X Platform

"At GigaIO, you're not constrained by the box – we can make a server with a single, unified memory space out of multiple boxes," says Benjamin. The company creates accelerator pooling appliances that can contain different types of accelerators including AMD Instinct GPUs. "We can attach up to 32 AMD Instinct GPUs to a single server, plus memory, plus petabytes of storage, to truly create an 'impossible server' that can be of nearly any dimension," says Benjamin.



GigaIO's SuperNODEs, powered by AMD Instinct™ MI300X GPUs, accelerate modeling and simulation

THE NEED FOR SPEED

GigaIO clients are turning to SuperNODEs equipped with AMD Instinct MI300X GPUs for speed, says Benjamin. "They want to train large language models (LLMs) faster, and they want to reach inference faster," he says.

Clients also run HPC workloads, such as computational fluid dynamics (CFD) on SuperNODEs, taking advantage of the same AMD Instinct MI300X GPU-powered performance. Benjamin shared, "Last year, a SuperNODE flawlessly executed one of the largest CFD simulations ever – the Concorde landing at 40 billion cells resolution – in a mere 33 hours."

> **"People want to run faster. If you give them the ability to access more GPUs, they can accomplish their task in half or one-third of the time."**
>
> Alan Benjamin, CEO at GigaIO

GigaIO recently conducted MLPerf testing with the Llama 2 70B model to simulate real-world AI inference workloads. In a SuperNODE setup featuring two AMD Instinct MI300X Platforms (totalling 16 GPUs), GigaIO demonstrated near-linear performance, scaling to 46,755.00 tokens per second – the highest number achieved for a single node in the MLPerf Inference: Datacenter benchmark database[1].

The results showed that for 16 GPUs, the FabreX-powered setup generates 12% more tokens per second than the next closest competitor using other interconnect technologies (MLPerf ID 4.1-0035)[1], further establishing GigaIO's solution as an exceptional choice for scale-up AI deployments.

"If you want your next-generation AI workloads to go fast, GigaIO and AMD can help you do that with simpler management and lower operational costs," says Benjamin.

> **"Between AMD and GigaIO, we're giving the users the ability to take existing containers that were written for an environment of four GPUs or eight GPUs, drop them into a SuperNODE with 32 AMD Instinct MI300X GPUs—and simply run four times faster without having to change a single line of code."**
>
> Alan Benjamin, CEO at GigaIO

GIGAIO AND AMD ADVANCE PERFORMANCE

Benjamin says cooperation with AMD engineering and software teams has enabled GigaIO to take the entire SuperNODE hardware and software stack completely up to the TensorFlow and PyTorch library levels. "Between AMD and GigaIO, we're giving users the ability to take existing containers that were written for an environment of four GPUs or eight GPUs, drop them into a SuperNODE with 32 AMD Instinct MI300X GPUs—and simply run four times faster without having to change a single line of code," he says.

"That gets a tremendous reaction," says Benjamin. "Most people don't believe it at first. Then we give them access to a SuperNODE equipped with 32 AMD Instinct MI300X GPUs and they run their job. Afterward, they tell us, 'You were right, we didn't have to change a line of code and it performed exactly as you said it would.'"

"That's been possible through cooperation with AMD. This is a case where we really have advanced together," says Benjamin. He's looking forward to still more leaps in the processing power of AMD Instinct GPUs and greater capabilities for AI and HPC applications.

"Don't let existing architecture constrain you," Benjamin urges. "If you are thinking, 'If only I could do this,' maybe we can do that for you today."

Accessible in the Cloud
AMD Instinct accelerators are available in the cloud to meet the scalability, flexibility, and performance demands of AI. For more information, **visit amd.com/instinct**

ABOUT GIGAIO

GigaIO builds composable, scalable servers called SuperNODEs that connect a variety of accelerators via the company's FabreX interconnect technology. Collaborating with AMD to deploy and finetune the AMD Instinct™ MI300X Platform performance on SuperNODEs enables GigaIO customers to complete HPC workloads and AI training and inference models faster, at lower costs, and without rewriting code. For more information visit gigaio.com.

ABOUT AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and X pages.

ENDNOTES

1. GigaIO press release, November 14, 2024: GigaIO Achieves Breakthrough MLPerf Inference Performance with SuperNODE.

DISCLAIMERS

All performance and cost savings claims are provided by GigaIO and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein are specific to GigaIO and may not be typical. GD-181

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes. GD-18.

COPYRIGHT NOTICE

© 2025 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, EPYC, Instinct, and combinations thereof are trademarks of Advanced Micro Devices, Inc.
Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.