SUPERMICRO® | AMD

Exploring Supermicro's H14 Solutions Powered by New AMD Instinct™ MI350 Series GPUs

Speakers

- Cyril Barranta, Digital Marketing, Supermicro (Host)
- Ted Marena, Business Development Director, AMD
- Tharun Kuppireddy, Sr. Solution Manager, Supermicro

Housekeeping

- Attachments
- Q & A
- On-demand

- Introduction and Housekeeping
- AMD Instinct™ MI350 Series Overview
- Supermicro H14 GPU Servers
- Q & A

AMD | Advancing the AI datacenter with Supermicro

AMD

Advancing the AI Data Center with Supermicro

AMD ROCm

Software Solution

Open Software Stack
Upstreamed Pytorch, Tensorflow, JAX, ONYX

+

CPUs

World's Best Server CPU for Cloud, Enterprise, AI

+

GPUs

Annual Roadmap, Leadership Memory and Performance

+

AMD PENSANDO

Ultra Ethernet Consortium

UA Link | Ultra Accelerator Link

Networking

DPUs, UALink + Ultra Ethernet

+

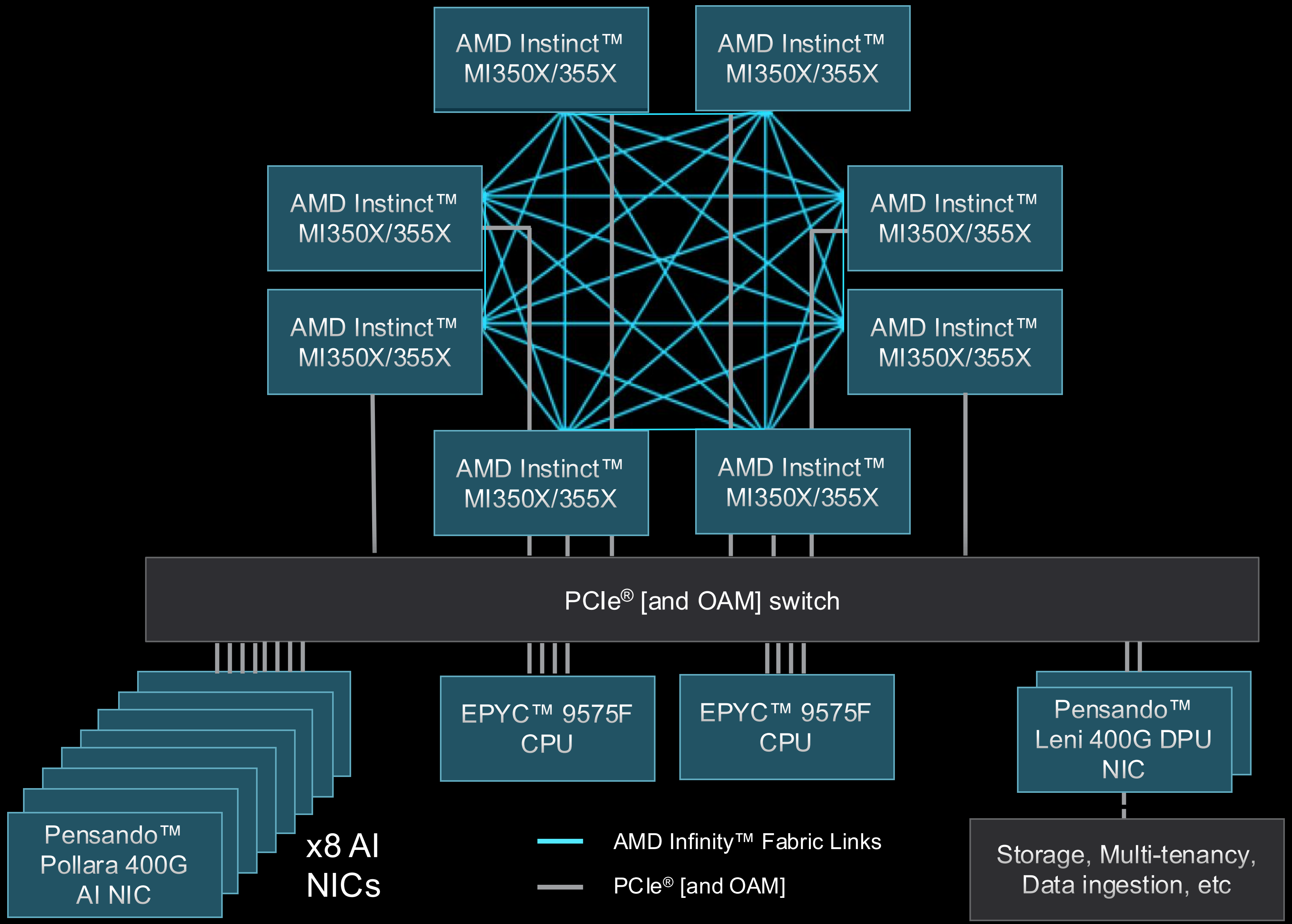
SUPERMICRO

Systems Design

Rack Level products
DLC & Air-cooled Servers

See endnote EPYC-029D

Supermicro H14 Block Diagram


AMD Instinct™ MI350X/355X
AMD Instinct™ MI350X/355X
AMD Instinct™ MI350X/355X
AMD Instinct™ MI350X/355X
AMD Instinct™ MI350X/355X
AMD Instinct™ MI350X/355X
AMD Instinct™ MI350X/355X
AMD Instinct™ MI350X/355X
PCIe® [and OAM] switch
Pensando™ Pollara 400G AI NIC
x8 AI NICs
EPYC™ 9575F CPU
EPYC™ 9575F CPU
Pensando™ Leni 400G DPU NIC
Storage, Multi-tenancy, Data ingestion, etc
AMD Infinity™ Fabric Links
PCIe® [and OAM]

AMD INSTINCT

Leadership roadmap commitment continues

2023

Gen AI Leadership

2024

HBM3E Memory Increased Compute

2025

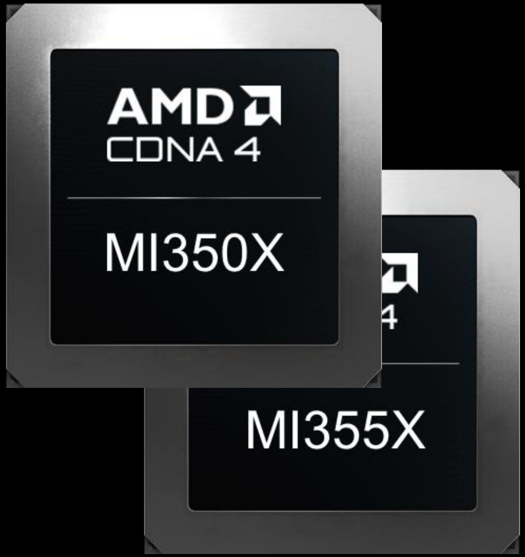
Compute and Memory Leadership

2026

Next-Gen Architecture

Roadmap subject to change

MI350X / MI355X Product Overview



- CDNA 4 architecture using N3P process technology
  - Support for new 4- and 6-bit data formats
- Improved HBM subsystem
  - Up to 288GB of HBM3E, 8.0 TB/s HBM BW
- Upgraded performance for inference and fine-tuning
- Orderable now from Supermicro!
  - DLC MI355X: AS-4126GS-NMR-LCC
  - Air cooled MI350X: AS-8126GS-TNMR

Product specification data is projected based on best available data, and are subject to change

AMD ROCm

Open, Modular Software Stack

AI Models and Algorithms

PyTorch ONNX JAX TensorFlow

Support all major frameworks and models

Libraries

Compilers and Tools

Runtime

Expanded Gen AI optimizations

New algorithms

New libraries

Expanding platform support

AMD GPUs

AMD INSTINCT

AMD RADEON

Extended developer support

TRANSITIONING WORKLOADS TO AMD INSTINCT^{TM} GPUS

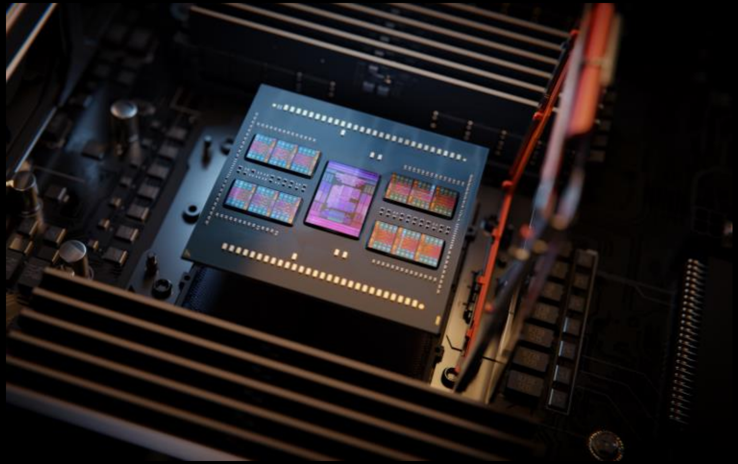
LOW FRICTION SOFTWARE PORTING USERS TO AMD

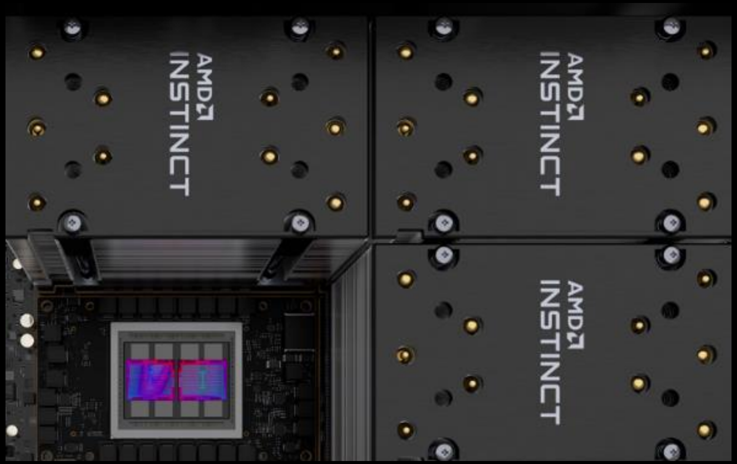
Use or mention of third-party marks, logos, products, services, or solutions herein is for informational purposes only and no endorsement by AMD is intended or implied. GD-83

The Most Demanding Data Center Workloads

Requires End-to-End Leadership in CPU, GPU, and Networking

AMD EPYC + AMD INSTINCT + AMD PENSANDO





SMC was the first OEM to qualify Instinct GPUs with EPYC 9575F and Pollara NICs

Summary Slide

- Supermicro is the lead time to market OEM for AMD Instinct™ GPUs

- Both air-cooled MI350X and DLC MI355X are orderable now

- Supermicro’s AMD Instinct based servers feature AMD EPYC CPUs and Pensando AI NICs

- Get started now! MI325X available today for remote POC evaluations
  - Access now at Vultr or via Supermicro Jumpstart

Supermicro H14 GPU Solutions

w/ AMD Instinct™ MI350 Series GPUs

A+ Accelerated MI350 Series Ready Systems

Air - Cooled

AS -8126GS-TNMR

w/ AMD Instinct™ MI350X

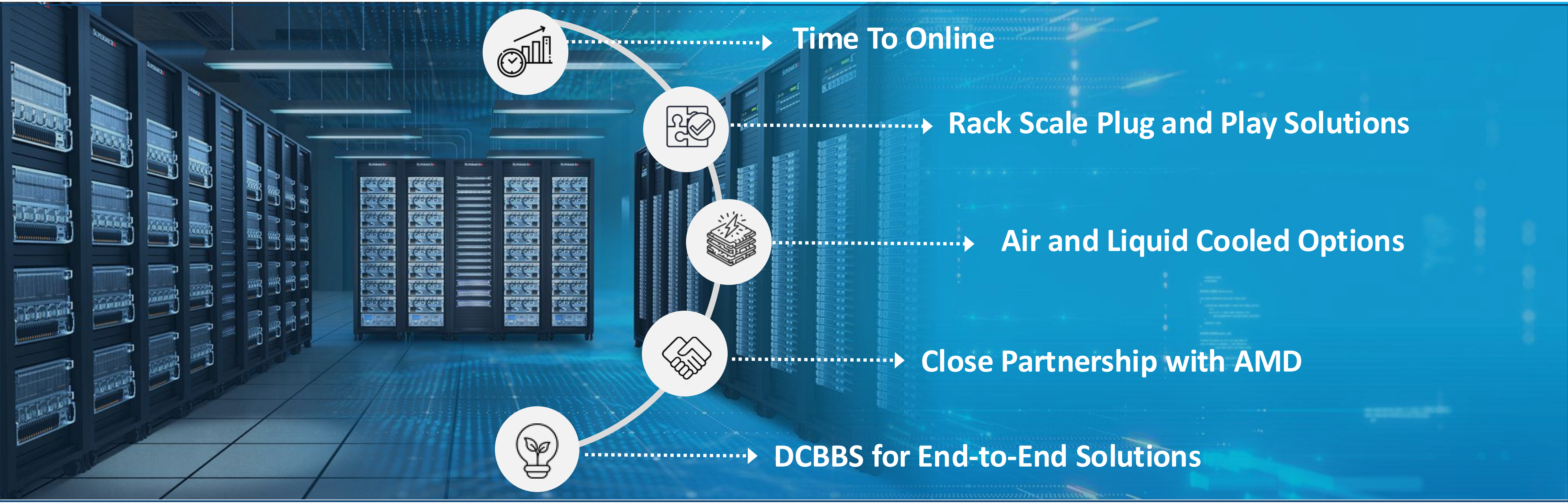
Liquid-Cooled

AS -4126GS-NMR-LCC

W/ AMD Instinct™ MI355X

AVAILABLE NOW!

Supermicro | AMD Data Center Solutions



- Time To Online
- Rack Scale Plug and Play Solutions
- Air and Liquid Cooled Options
- Close Partnership with AMD
- DCBBS for End-to-End Solutions

Datacenter Building Block Solutions

System Level

Rack Level

Rack-Scale PnP Level

Datacenter Infrastructure Level

H14 8U OAM GPU System AS -8126GS-TNMR

Integrated Performance, OAM MI350X 8-GPU AIR-COOLED

Key Features

- Dual AMD 5th Gen EPYC processors with up to 192 cores each and high-frequency SKUs up to 5GHz gives better performance when combined with MI350 series GPUs
- 8x PCIe Gen 5.0 X16 LP, and up to 4 PCIe Gen 5.0 X16 FHFL Slots
- 1-1 GPU to NIC direct-connect for low latency and fast IO
- Flexible networking options –ethernet or infiniband
- 2x NVMe M.2 with RAID support

AS -8126GS-TNMR

Key Applications:

- AI/Deep Learning Training
- High Performance Computing
- Industrial Automation, Retail
- Healthcare
- Conversational AI

- Business Intelligence & Analytics
- Drug Discovery
- Climate and Weather Modeling
- Finance & Economics simulations

H14 4U GPU System AS -4126GS-NMR-LCC

Performance Density, AMD MI355X 8-GPU LIQUID-COOLED

Key Features

- Dual AMD EPYC processors with up to 192 cores each and high-frequency SKUs up to 5GHz gives better performance when combined with MI350 series GPUs
- 8x PCIe Gen 5.0 X16 LP, and up to 4 PCIe Gen 5.0 X16 FHFL Slots
- 1-1 GPU to NIC direct-connect for low latency and fast IO
- Flexible networking options –ethernet or infiniband
- 2x NVMe M.2 with RAID support

Key Applications:

- AI/Deep Learning Training
- High Performance Computing
- Healthcare
- Conversational AI
- Business Intelligence & Analytics
- Drug Discovery
- Climate and Weather Modeling
- Finance & Economics

*Offered as Rack-level deployment

AS -4126GS-NMR-LCC

Broadest Portfolio of AMD Server Families

CloudDC

All-in-One Servers with Flexible I/O Options
for Cloud Scale Data Centers

Hyper

Industry Leading IOPS Server
with Energy Efficiency and Flexibility

Hyper-U

Enterprise-Focused Servers Delivering Memory Density, Flexibility, and Power Efficiency

GrandTwin™/FlexTwin™

Leading Multi-Node Architecture with High Power Efficiency

Petascale Storage

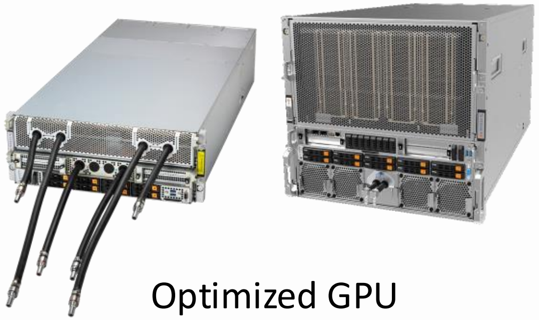
All-Flash EDSFF storage server for Software-Defined Data Center Workloads

MicroCloud

Max Density Multi-Node for Enterprise and dedicated Cloud Hosting

PCIe GPU

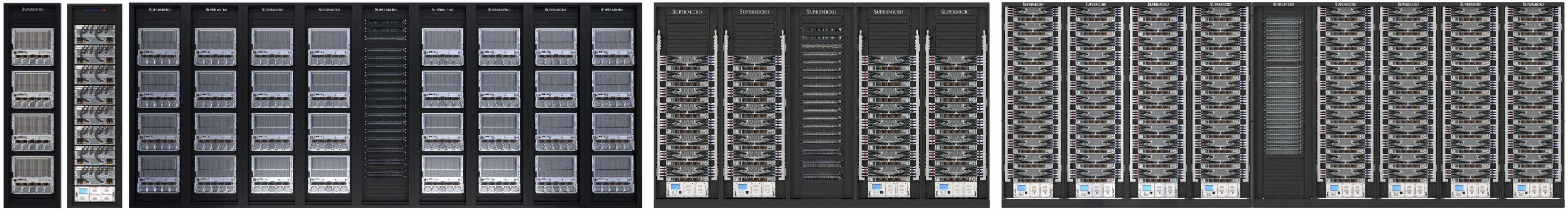
Maximum Acceleration
for AI/ Deep Learning and HPC



Optimized GPU

Maximum Acceleration
for AI/ Deep Learning and HPC

Supermicro + AMD Scalable AI Cluster



AI Workloads	Mid-Range Inference	Large Model Inference Mid-Range Training	Large Training Models
Platform Scale	Small GPU Pod	Medium GPU Pod	Large GPU Pod
Typical Pod Size	8 GPUs	32 - 96 GPUs	>256 GPUs
GPU-GPU Bandwidth	560GB/s	560GB/s	560GB/s
Scale Out Bandwidth	50GB/s	50GB/s	50GB/s
Rack	Air or Liquid Cooled	Air or Liquid Cooled	Air or Liquid Cooled
Rack Power	14kW	60-120kW	Up to 120kW / Rack

Reference GPU: AMD Instinct MI350 and MI355

Key Takeaways

Supermicro is first-to-market with AMD Instinct™ powered AI data-center solutions

Supermicro servers with MI350 series are available. Remote POC with H14 + MI350 series GPU systems coming soon!

Supermicro has the largest portfolio of AMD’s A+A+A systems across all regions

Q & A

DISCLAIMER

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

www.supermicro.com/aplus