



AMD Advancing AI&HPC

日本AMD株式会社
コマーシャル営業本部
皆川 直樹
2025/12/08





High Performance and AI Computing Leadership

Solving The World's Most Important Challenges





創業 55 年

設立: 1969 年 5 月 1 日
本社: カリフォルニア州サンタクララ

従業員数 28,000+

次世代コンピューティングを加速

2024 年の年間売上高 \$25.8B

25% 以上を研究開発に再投資

5 年間で 3 倍の時価総額

世界の時価総額ランキング Top 100 にランクイン

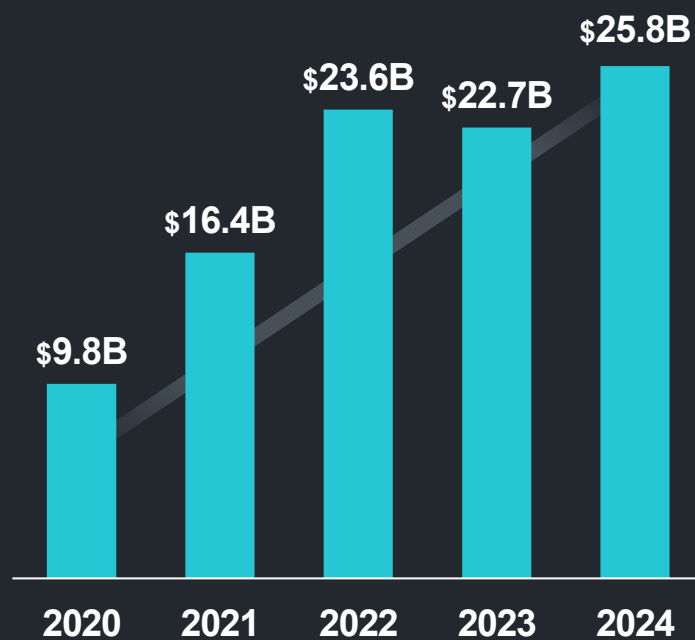
拠点数 100+

グローバルな事業展開

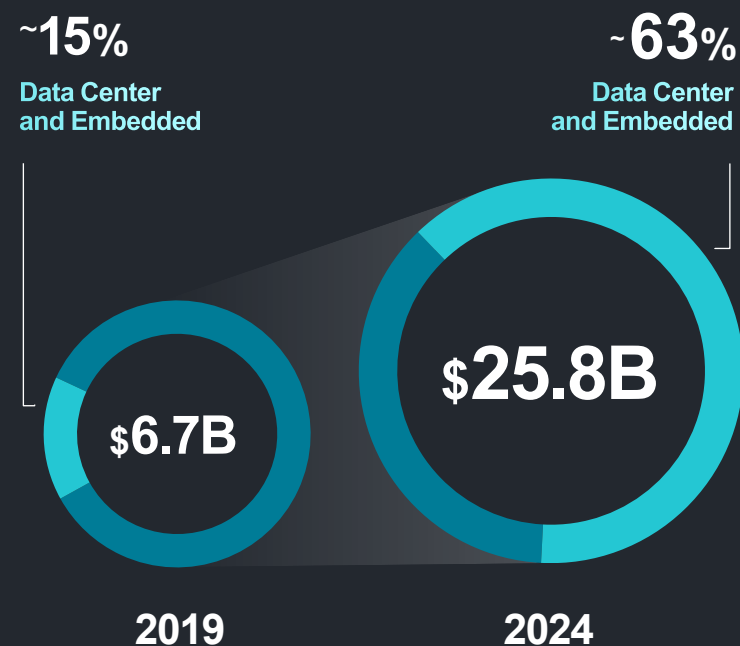
AMD
together we advance_

強力な財務基盤に支えられた成長

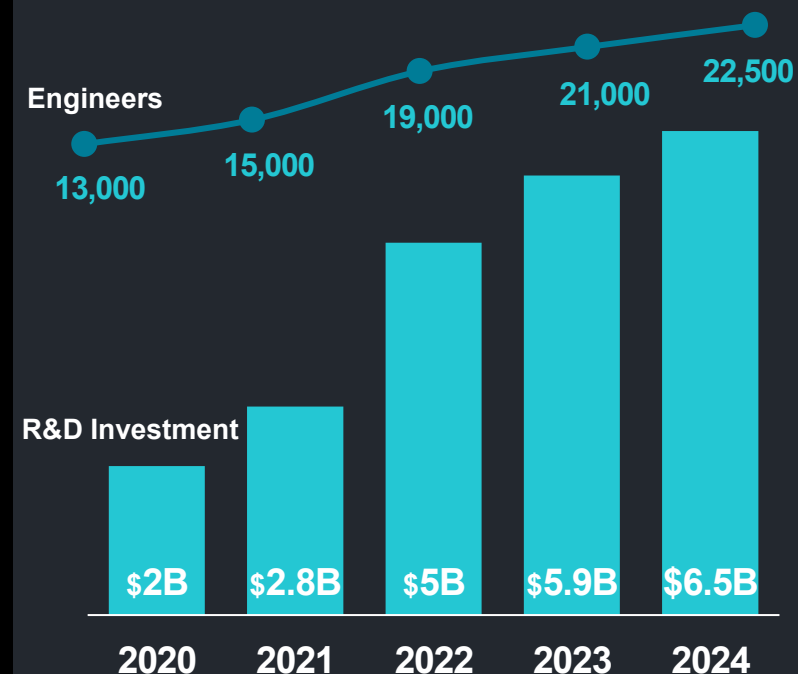
売上高成長率の加速



売上高構成比の変革



リーダーシップのための投資



AMD AI Strategy



**Leadership
Compute Engines**



**Open
Ecosystem**



**Full Stack
Solutions**

AMD Instinct™: Data Center GPU Architecture Roadmap



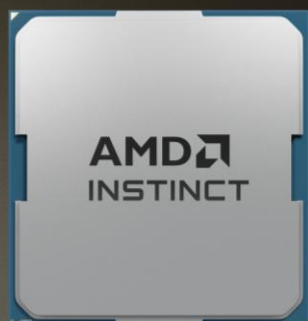
2023

2026

*All roadmaps subject to change.

AMD Instinct™ MI450 Series

Most Advanced AI Accelerator



40 PF
20 PF

FP4 • FP8 Flops

432 GB
19.6 TB/s

HBM4 Memory

3.6 TB/s

Scale Up Bandwidth

300 GB/s

Scale Out Bandwidth

Advanced Process
Technology

Leadership IP

Chiplet Architecture

3.5D Packaging

Co-Designed
Hardware & Software

Source: AMD specs based on engineering projections, results subject to change.

AMD Instinct™ MI400 Series Portfolio

Leadership Across AI & Scientific Computing

AMD Instinct™ MI455X

At Scale AI Training & Inference

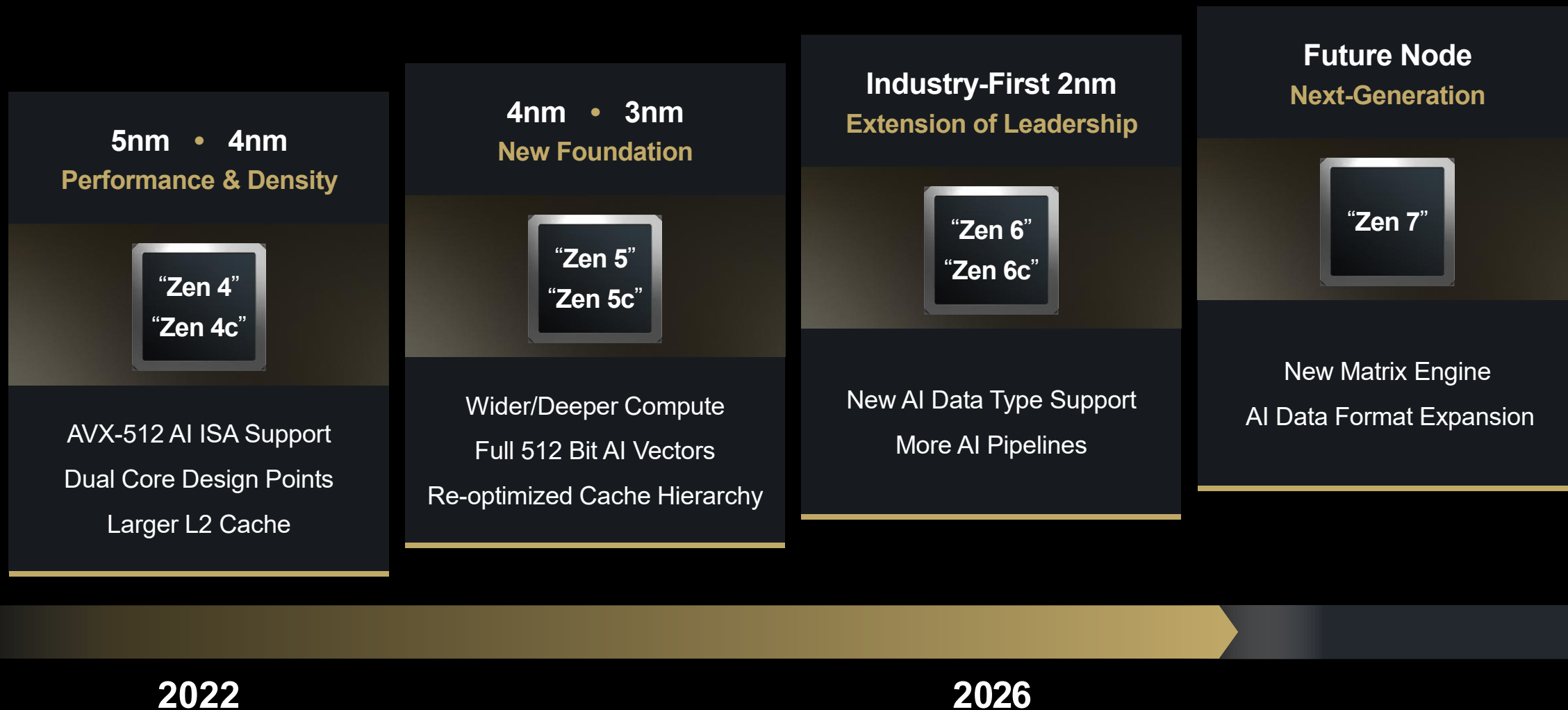
AI Compute | Scale Out Performance | HBM4 Memory

AMD Instinct™ MI430X

Sovereign AI & HPC

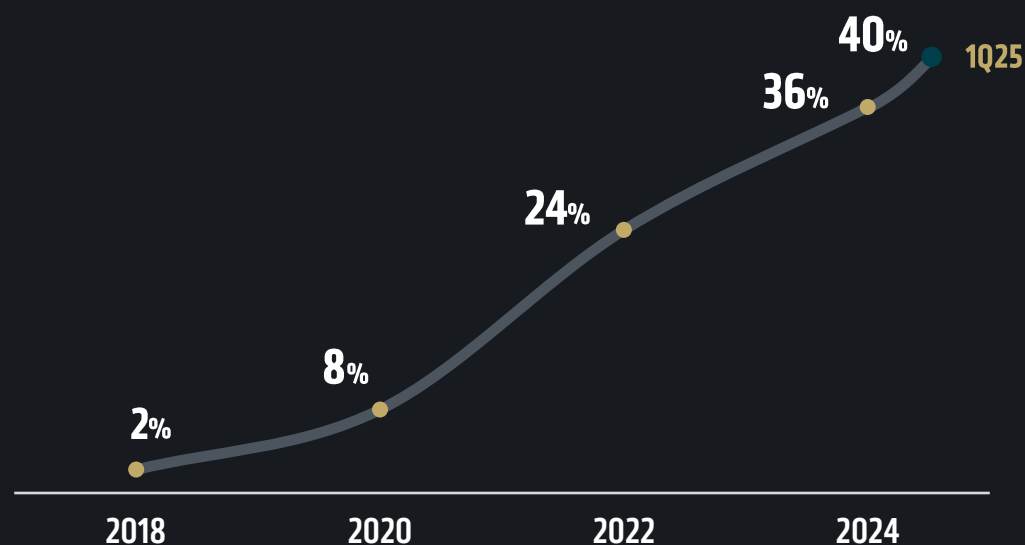
Hybrid Compute | Hardware-Based FP64 | HBM4 Memory

Leadership CPU Core Roadmap



EPYC 採用の流れが加速しています

>18倍 サーバーCPU市場シェア成長率



業界のリーダーはEPYC™を活用しています

Cloud

aws Microsoft Google ORACLE

Digital

NETFLIX Uber ∞ Meta zoom

Enterprise

BEST BUY IBM Emirates NBD NISSAN

OEM

DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO CISCO

AMD EPYC™ “Venice”

Highest Performance Server CPU

Up to **256 cores**

2nm • Zen 6

2.0x

CPU to GPU Bandwidth

1.7x

Gen vs. Gen Performance

1.6 TB/s

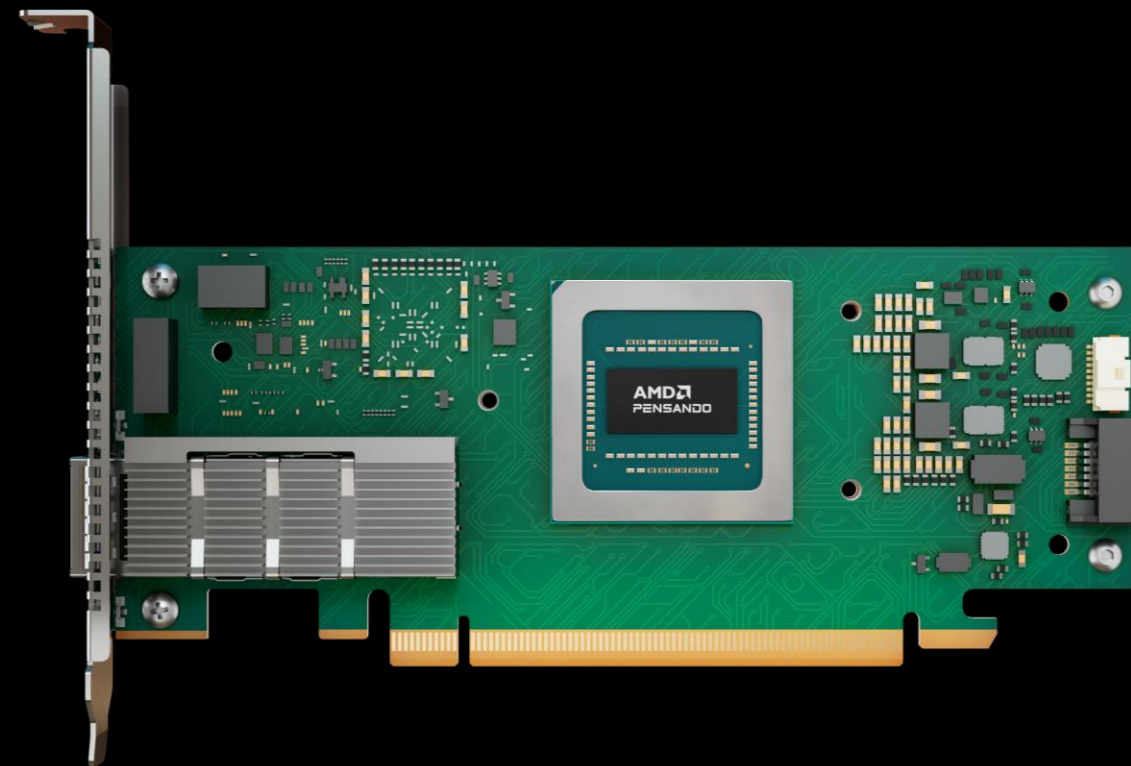
Memory Bandwidth

Coming in 2026

AMD Pensando™ Pollara 400 AI NIC

Industry's first Ultra Ethernet
Consortium ready AI NIC

- Programmable Hardware Pipeline
- Up to 1.25x Performance Boost*
- 400 Gbps
- Open Ecosystem
- UEC Ready RDMA
- Reduction in Job Completion Times
- High Availability



Ultra Ethernet
Consortium

AMD Pensando™ “Vulcano”

Next Gen NIC for AI Clusters

3nm

Process Node

800G

Network Throughput

Up to **8x**

Scale Out Bandwidth per GPU

UAL | PCIe®

Host Interface

Ultra Ethernet
Consortium

Coming in 2026

AMD AI Strategy



Leadership
Compute Engines



Open
Ecosystem



Full Stack
Solutions

Open Development Drives Value & Innovation

Open Hardware



Open Software



Open Ecosystem



Hugging Face



Triton



Choice

Flexibility

Rapid Co-Innovation

Portability

Proven

AMD Powers U.S. Sovereign AI Factory Supercomputers

Accelerating an Open American AI Stack



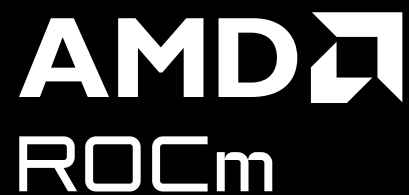
HPE to build two systems for Oak Ridge National Laboratory: Next-generation exascale supercomputer "Discovery" and AI cluster "Lux" | HPE

The Lux AI supercomputer

HPE ProLiant Compute XD685
AMD Instinct MI355X GPUs
AMD EPYC™ CPUs
AMD Pensando™ networking

The DISCOVERY AI supercomputer

HPE Cray Supercomputing GX5000
AMD Instinct MI430 GPUs
Next-Generation AMD EPYC™ CPUs
Next-Generation HPE Slingshot



Deepening Ecosystem Collaboration



Pytorch

Day 0 support
daily performance CI



Triton

v3.3

Performance focus



Hugging Face

1.8 million models

Nightly CI/CD,
finetuning support

vLLM_{v1}

SGL

llm-d

Serving leadership
Distributed
inference

LLaMA 4
by Meta

Gemma 3

deepseek

QwQ-32B

Command R⁺

Grok

MISTRAL
AI

Support for
SOTA models



ONNX

deepspeed

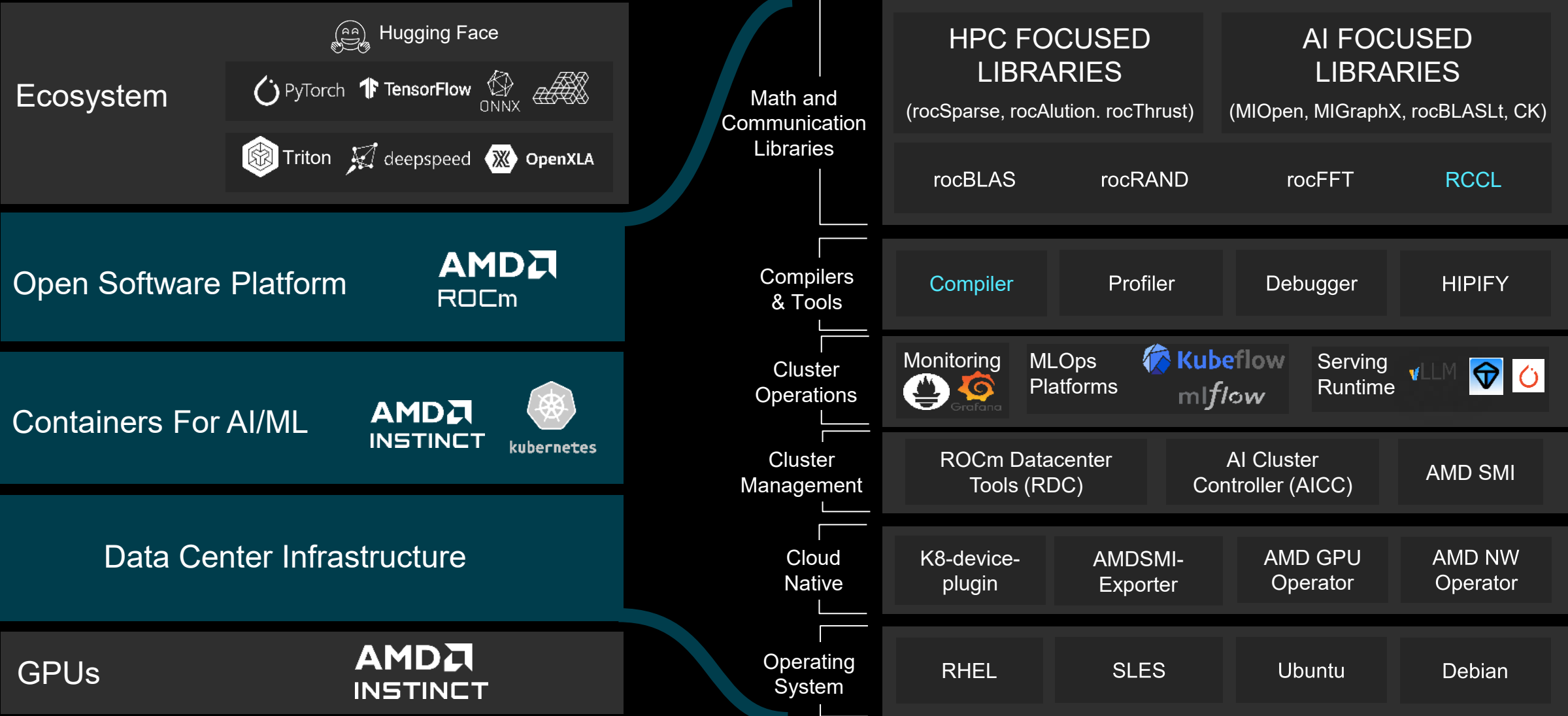
TensorFlow

OpenXLA

MLIR

Expanding open-
source footprint

AMD SOFTWARE OFFERING



AMD Instinct™ MI300 Series Cluster Reference Architecture Guide

Figure 5.1: A 32 Node 2-Tier Fat Tree Topology

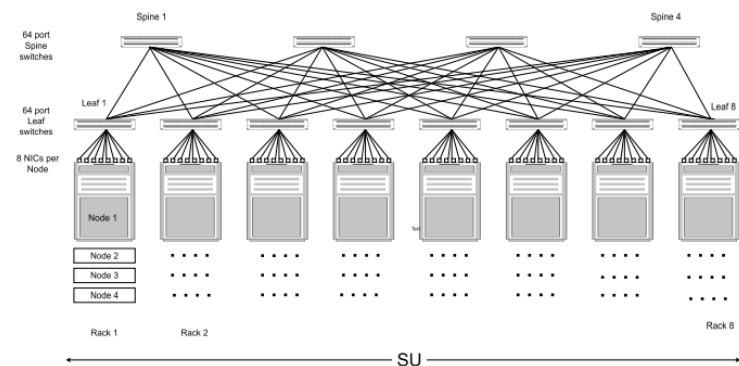
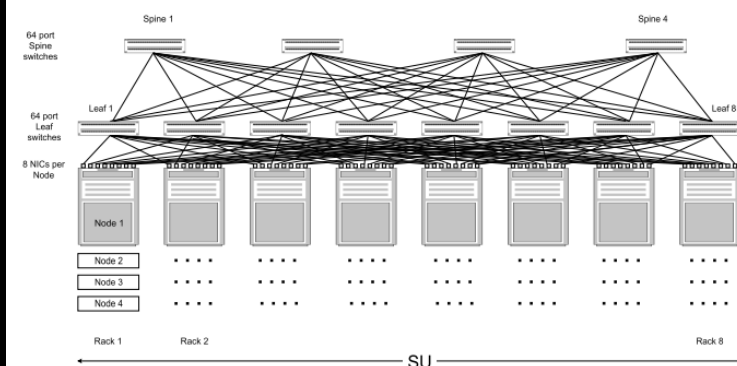
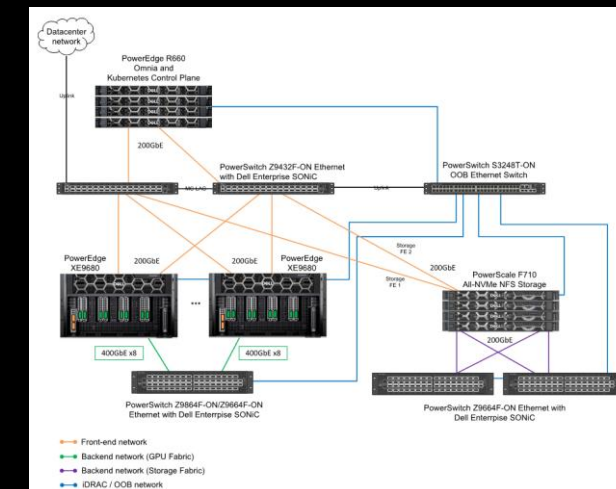


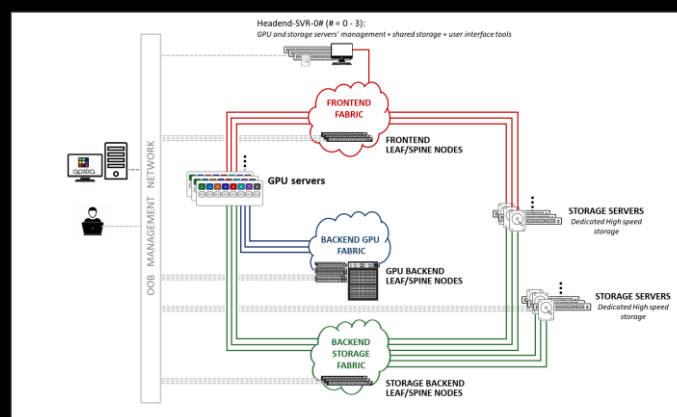
Figure 5.2: A 32 Node 2-Tier Rail Topology



AMD Instinct™ MI300 Series Cluster Reference Architecture Guide



Networking design | Generative AI in the Enterprise with AMD Accelerators | Dell Technologies Info Hub



AI Use Case and Reference Design | Juniper Networks

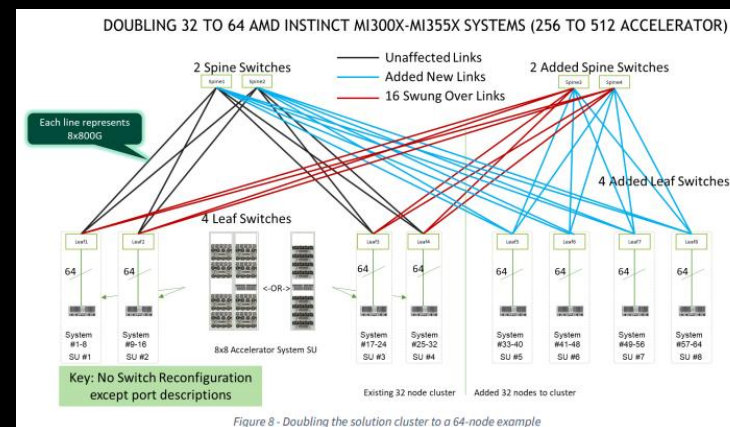
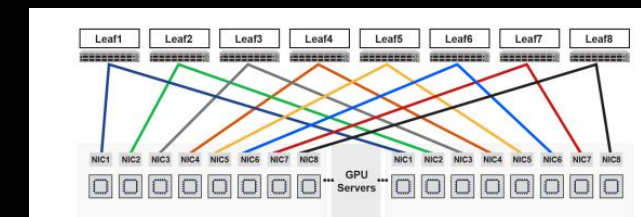


Figure 8 - Doubling the solution cluster to a 64-node example

Supermicro AI Cluster Architecture with AMD CPU, GPUs, Networking



3HE21915AAAATQZZA_V1_Lenovo-Nokia AI-DC Validated Design.pdf



Growing Industry Adoption

7 of 10 Largest AI Companies Use AMD Instinct



AMD AI Strategy



**Leadership
Compute Engines**

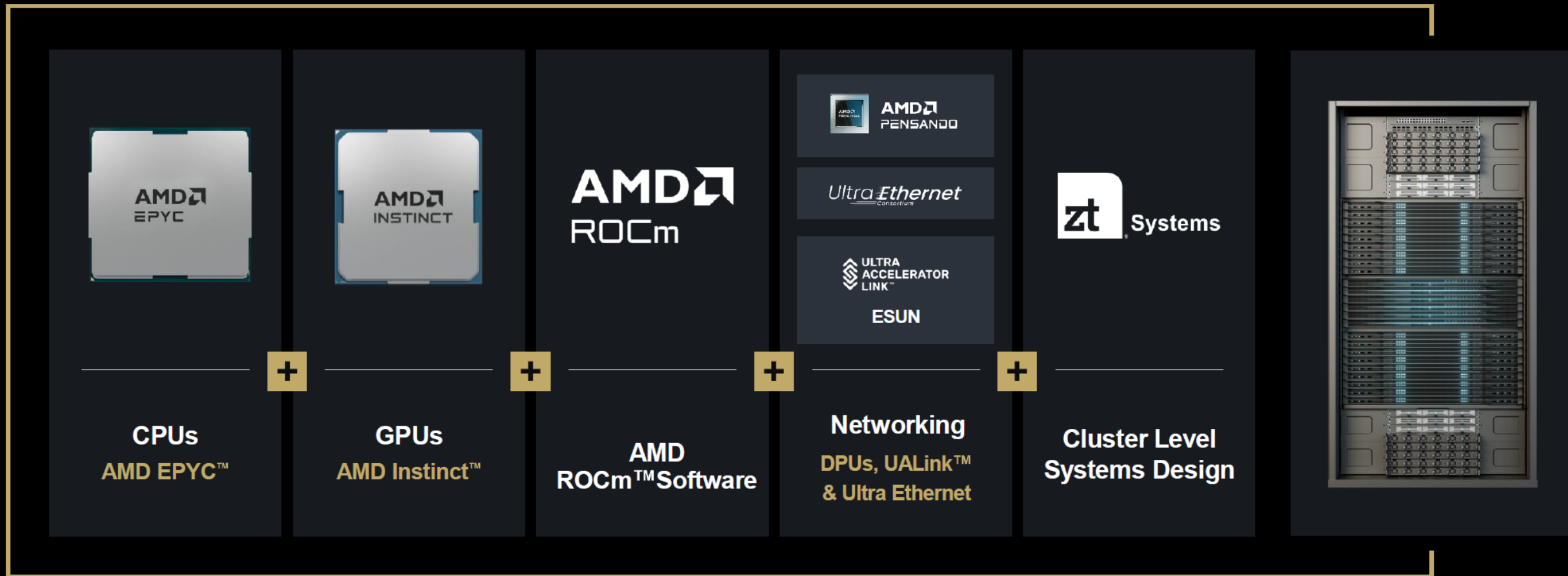


**Open
Ecosystem**



**Full Stack
Solutions**

AMD | Delivering AI Factories



AMD | Advancing AI Infrastructure on an Annual Cadence

2025

AMD EPYC
"TURIN"

AMD Instinct
MI350 SERIES

AMD Pensando
POLLARA 400



2026

AMD EPYC
"VENICE"

AMD Instinct
MI400 SERIES

AMD Pensando
"VULCANO"



"Helios"

2027

AMD EPYC
"VERANO"

AMD Instinct
MI500 SERIES

AMD Pensando
"VULCANO"



Next Gen AI Rack

次世代AIコンピューティングを 加速させる戦略パートナーシップ

AMD × OpenAI

AI開発のフロンティア

6GW

AMD Instinct™ アクセラレーター
2026年下半期より

AMD × ORACLE

ゼタスケール
コンピューティング

50,000

AMD Instinct™
MI450シリーズ GPUs
2026年下半期より

AMD × Meta

共同設計のオープンインフラ

Metaと共同で策定し、
OCPにプレビューされる
AMD “Helios” ラック

AMD × U.S. DEPARTMENT
of ENERGY

米国HPCリーダーシップの
拡張

Lux: 最初の” US AI Factory”
AMD Instinct MI355Xシリーズ

Discovery: AIスパコンの
フラッグシップ
AMD Instinct MI430Xシリーズ



together we advance_

Disclaimers and Attributions

The information contained herein is for informational purposes only, and is subject to change without notice. Timelines, roadmaps, and/or product release dates shown herein are plans only and subject to change. “Zen”, “Zen2”, “Zen3”, “Zen4”, “Zen5”, “Rome”, “Milan”, “Genoa”, “Bergamo”, “Siena”, “Turin”, “Raphael”, “Sorano” and “Venice” are codenames for AMD architectures, and are not product names. GD-122

DISCLAIMER: The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

© 2024 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, EPYC, 3D V-Cache, and combinations thereof are trademarks of Advanced Micro Devices, Inc. ANSYS, FLUENT and any and all ANSYS, Inc. brand, product, service and feature names, logos and slogans are registered trademarks or trademarks of ANSYS, Inc. or its subsidiaries in the United States or other countries. AWS is a trademark of Amazon.com, Inc. or its affiliates in the United States and/or other countries. Azure® is a registered trademark of Microsoft Corporation in the US and/or other countries. Canonical and the Canonical logo are registered trademarks of Canonical Ltd. Cisco is a registered trademark of Cisco Systems Inc. Citrix is a trademark of Citrix Systems, Inc. and/or one or more of its subsidiaries, and may be registered in the United States Patent and Trademark Office and in other countries. Dell is a trademark of Dell Inc. or its subsidiaries. GitLab is a registered trademark of GitLab, Inc. HPE is a registered trademark of Hewlett Packard Enterprise Company and/or its affiliates. IBM and IBM Cloud are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Google is a registered trademark of Google LLC. Intel is a trademark of Intel Corporation or its subsidiaries. Micron and the Micron orbit logo are trademarks of Micron Technology, Inc. Microsoft is a registered trademark of Microsoft Corporation in the US and/or other countries. Kubernetes is a registered trademark of The Linux Foundation. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. MongoDB is a registered trademark of MongoDB Inc. MySQL is a trademark of Oracle and/or its affiliates. NGINX is a trademark of Nginx Software Inc. Nutanix is a trademark of Nutanix Inc. Oracle is a registered mark of Oracle and/or its affiliates. PCIe is a registered trademark of PCI-SIG Corporation. Pixar is a registered trademarks of Pixar Animation Studios. Red Hat and the Shadowman logo are registered trademarks of Red Hat, Inc. www.redhat.com in the U.S. and other countries. Redis is a trademark of Redis Labs Ltd. Any rights therein are reserved to Redis Labs Ltd. Any use by AMD is for referential purposes only and does not indicate any sponsorship, endorsement or affiliation between Redis and AMD. Samsung is a trademark or registered trademark of Samsung Electronics Co., Ltd. SPEC® is a trademark or registered trademark of Standard Performance Evaluation Corporation (SPEC). NOTE: SPEC® and the benchmarks [list benchmarks with their appropriate trademark symbol found here: <https://www.spec.org/spec/trademarks.html>] are trademarks or registered trademarks of Standard Performance Evaluation Corporation (SPEC). Learn more at www.spec.org. Supermicro is a trademark or registered trademark of Super Micro Computer, Inc. or its subsidiaries in the United States and other countries. SUSE is a registered trademark of SUSE LLC in the United States and other countries. VMware is a registered trademark of VMware in the US or other countries. TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc. Ubuntu and the Ubuntu logo are registered trademarks of Canonical Ltd. VMmark is a registered trademark of VMware in the US or other countries. Xeon is a trademark of Intel Corporation or its subsidiaries. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners.

© 2025 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD Instinct, EPYC, Pensando, Radeon, ROCm, Ryzen, Versal, Xilinx, and combinations thereof are trademarks of Advanced Micro Devices, Inc. CXL is a registered trademark of Compute Express Link Consortium, Inc. OpenAI is a trademark of OpenAI, Inc. PCIe® is a registered trademark of PCI-SIG Corporation. UCIE is a trademark of Universal Chiplet Interconnect Express, Inc. Ultra Accelerator Link and UALink are trademarks of the UALink Consortium. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

Fortune content is From Fortune Magazine. © 2025 Fortune Media IP Limited. All rights reserved. Used under license. Fortune is a registered trademark of Fortune Media IP Limited and is used under license.

Endnotes

VEN-003: PCIe Gen comparison based on PCI-SIG published statements, <https://pcisig.com/pci-express-6.0-specification>. 2P 6th Gen EPYC CPU with 128 lanes of PCIe Gen 6 and 5th Gen EPYC with 128 lanes of PCIe Gen 5 as of 6/3/2025. PCIe is a registered trademark of PCI-SIG Corporation

PEN-016 - Testing conducted by AMD Performance Labs as of [28th April 2025] on the [AMD Pensando™ Pollara 400 AI NIC], on a production system comprising of: 2 Nodes of 8xMI300X AMD GPUs (16 GPUs): Broadcom Tomahawk-4 based leaf switch (64x400G) from MICAS network; CLOS Topology; AMD Pensando Pollara AI NIC – 16 NICs; CPU Model in each of the 2 nodes - Dual socket 5th gen Intel® Xeon® 8568 - 48 core CPU with PCIe® Gen-5 BIOS version 1.3.6 ; Mitigation - Off (default); System profile setting - Performance (default) SMT- enabled (default); Operating System Ubuntu 22.04.5 LTS, Kernel 5.15.0-139-generic. Following operation were measured: Allreduce; Average 25% for All-Reduce operations with 4QP and using UEC ready RDMA vs the RoCEv2 for multiple different message size samples (512MB, 1GB, 2GB, 4GB, 8GB, 16GB). The results are based on the average at least 8 test runs.