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

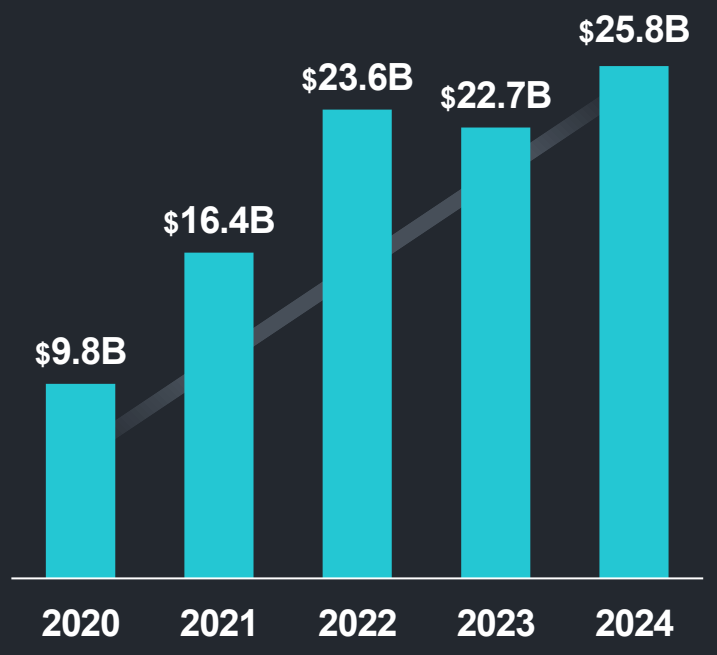
拠点数 100+

グローバルな事業展開

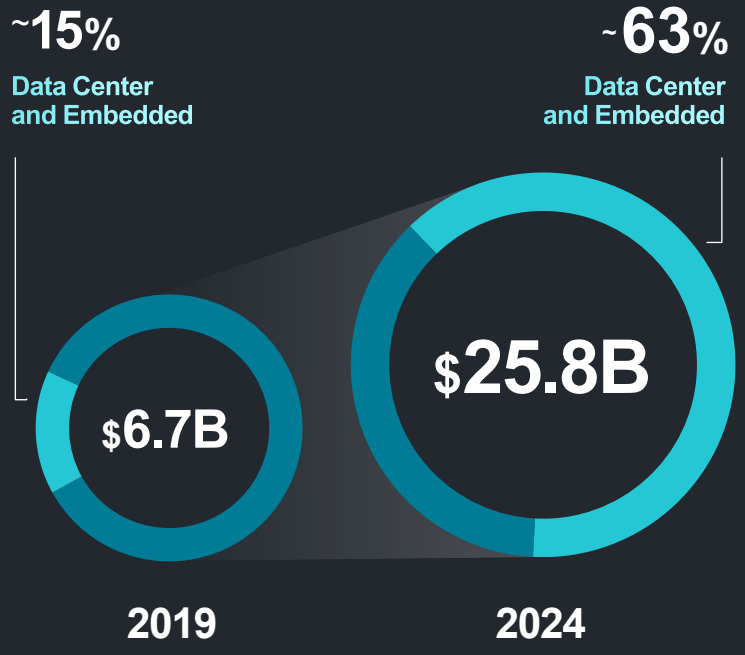
AMD
together we advance_

強力な財務基盤に支えられた成長

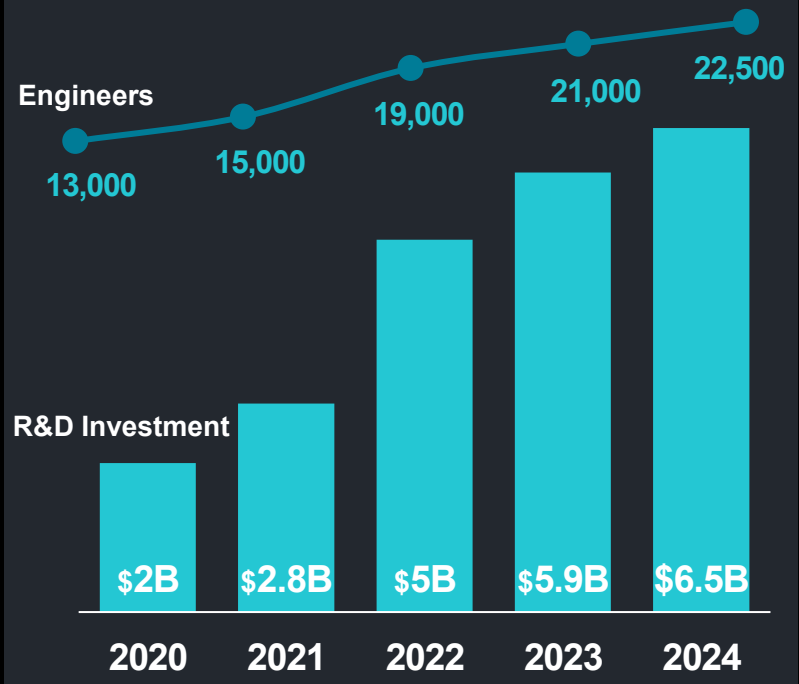
売上高成長率の加速



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

MI300
AI-Centric Design

Leadership 8-Stack HBM3 Capacity & Bandwidth

3.5D Chiplet Construction

Industry Standard Subsystem

MI325
Gen AI Performance

Increased HBM3e Capacity & Bandwidth

Higher Compute Throughput

Platform Compatibility

MI350 Series
Gen AI Inferencing & Training

Increased HBM3e Capacity & Bandwidth

Block Scaled AI Formats

Optimized Platform Solutions

MI400 Series
Rack-Scale Architecture

Increased HBM4 Capacity & Bandwidth

Expanded AI Formats with Higher Throughput

Standard-Based Rack-Scale Networking (UALoE, UAL, UEC)

MI500 Series

Next-Gen Compute, Memory & Interconnect

2023 — **2026**

*All roadmaps subject to change.

AMD Instinct™ MI450 Series

Most Advanced AI Accelerator

40 PF 20 PF	432 GB 19.6 TB/s	3.6 TB/s	300 GB/s
FP4 • FP8 Flops	HBM4 Memory	Scale Up Bandwidth	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

5nm • 4nm
Performance & Density

AVX-512 AI ISA Support
Dual Core Design Points
Larger L2 Cache

4nm • 3nm
New Foundation

Wider/Deeper Compute
Full 512 Bit AI Vectors
Re-optimized Cache Hierarchy

Industry-First 2nm
Extension of Leadership

New AI Data Type Support
More AI Pipelines

Future Node
Next-Generation

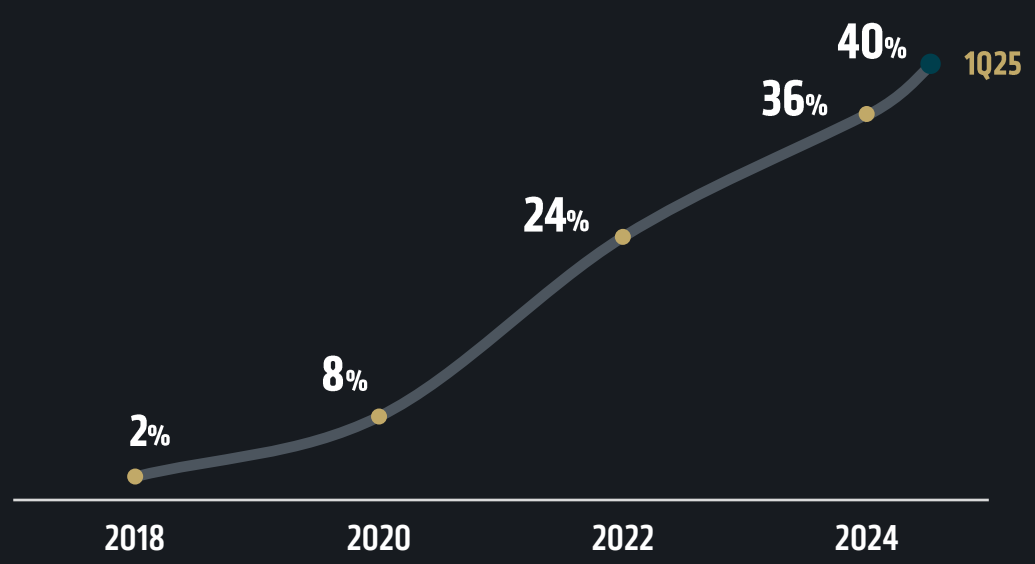
New Matrix Engine
AI Data Format Expansion

2022 **2026**

*All roadmaps subject to change.

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

Source: Mercury

AMD EPYC™ “Venice”

Highest Performance Server CPU

Up to **256 cores**	**2.0x**	**1.7x**	**1.6** TB/s
2nm • Zen 6	CPU to GPU Bandwidth	Gen vs. Gen Performance	Memory Bandwidth

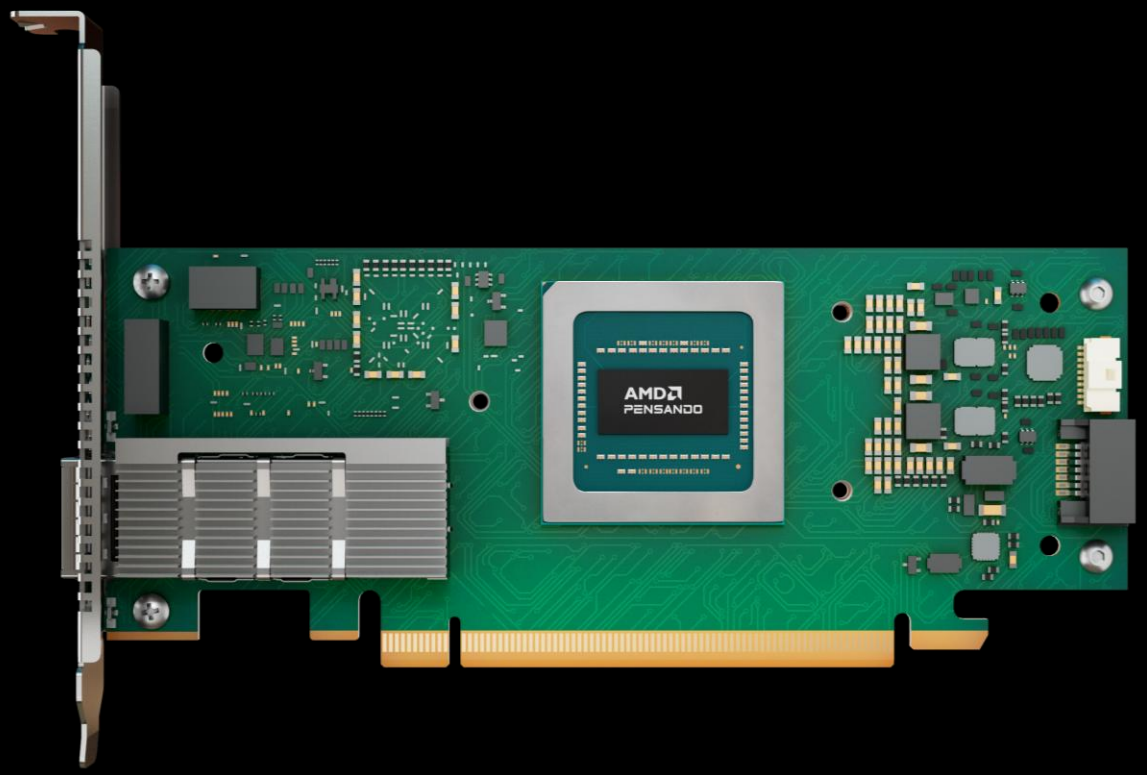
Coming in 2026

See endnote: VEN-003

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

* See Endnotes: PEN-016

AMD Pensando™ “Vulcano”

Next Gen NIC for AI Clusters

3nm	800G	Up to 8x	UAL | PCIe®
Process Node	Network Throughput	Scale Out Bandwidth per GPU	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**

OPEN Compute Project · ULTRA ACCELERATOR LINK™ · Ultra Ethernet Consortium · AMD ROCm · Hugging Face · PyTorch · Triton · vLLM · SGL

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

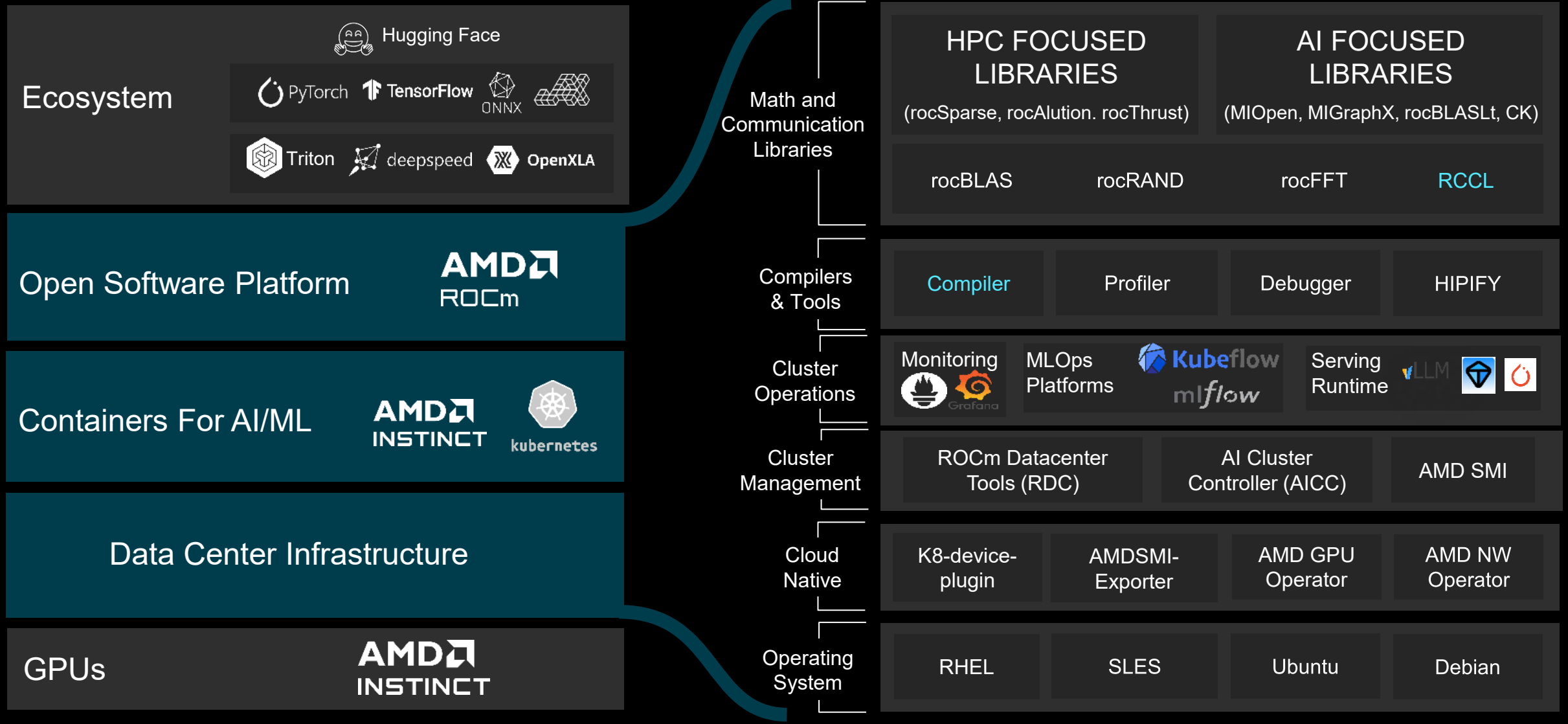
HPE Cray Supercomputing GX5000
AMD Instinct MI430 GPUs
Next-Generation AMD EPYC™ CPUs
Next-Generation HPE Slingshot

AMD ROCm

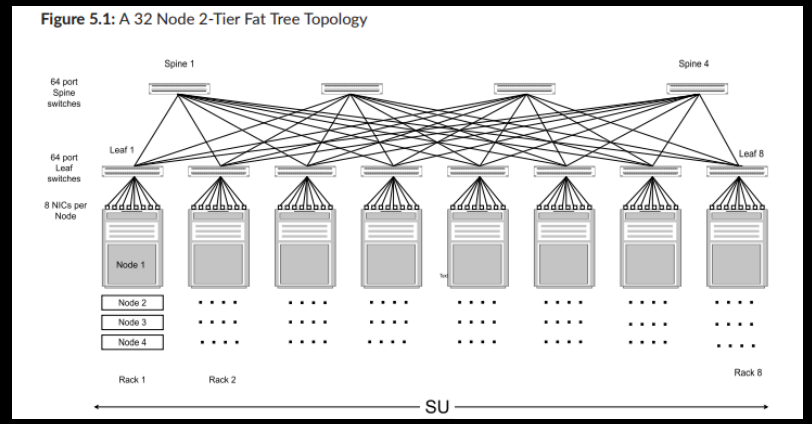
Deepening Ecosystem Collaboration

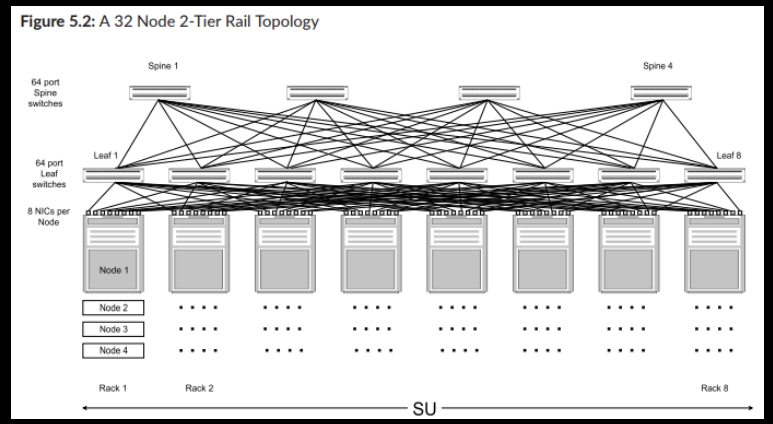
Pytorch	Triton v3.3	Hugging Face 1.8 million models	vLLM v1, SGL, llm-d	LLaMA 4, Gemma 3, deepseek, QwQ-32B, Command R+, Grok, MISTRAL AI_	ONNX, deepspeed, TensorFlow, OpenXLA, MLIR
Day 0 support daily performance CI	Performance focus	Nightly CI/CD, finetuning support	Serving leadership Distributed inference	Support for SOTA models	Expanding open-source footprint

AMD SOFTWARE OFFERING

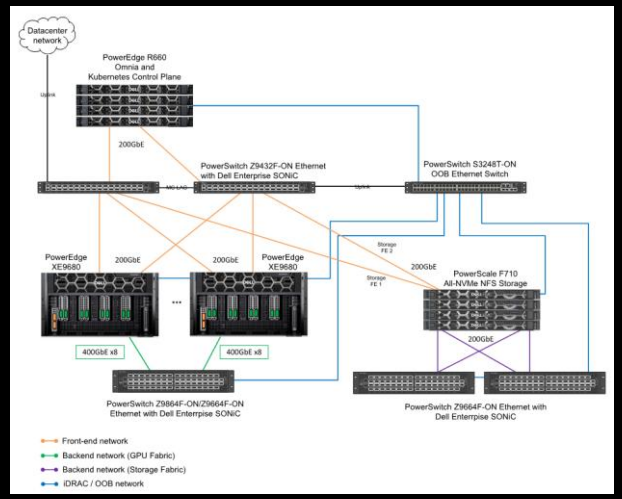


AMD Instinct™ MI300 Series Cluster Reference Architecture Guide

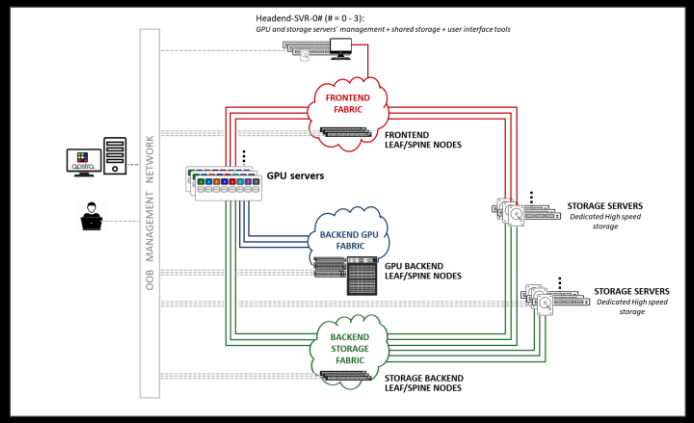




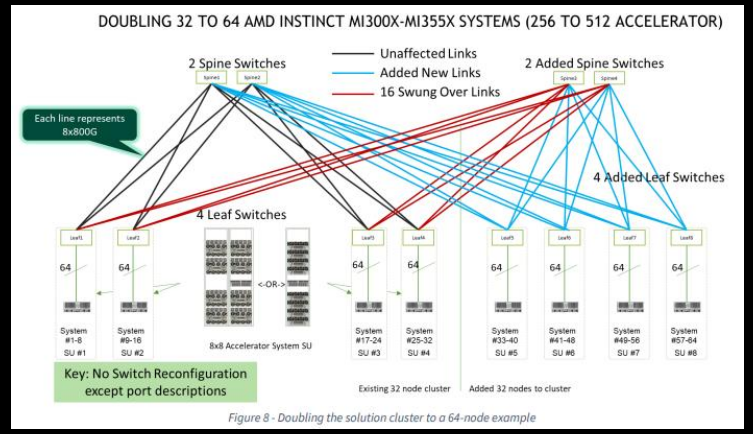
AMD Instinct™ MI300 Series Cluster Reference Architecture Guide



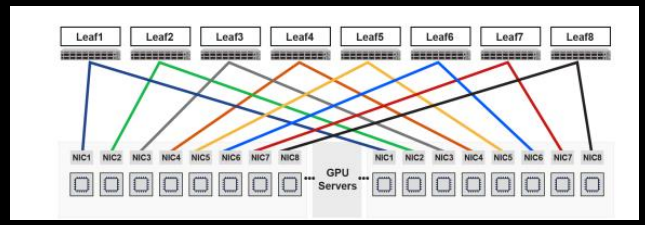
Networking design | Generative AI in the Enterprise with AMD Accelerators | Dell Technologies Info Hub



AI Use Case and Reference Design | Juniper Networks



Supermicro AI Cluster Architecture wth AMD CPU, GPUs, Networking



3HE21915AAAATQZZA_V1_Lenovo-Nokia AI-DC Validated Design.pdf

AMD INSTINCT

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

CPUs
AMD EPYC™

GPUs
AMD Instinct™

AMD ROCm

AMD ROCm™ Software

AMD PENSANDO

Ultra Ethernet Consortium

ULTRA ACCELERATOR LINK™

ESUN

Networking
DPUs, UALink™ & Ultra Ethernet

zt Systems

Cluster Level Systems Design

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

Roadmap subject to change

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


AMD
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.

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.