資料中心中的AI轉型：
利用AMD解決方案釋放計算潛力

AMD 台灣區商用業務處
技術顧問
陳信宏Jeremy Chen

October 2025

AI Innovation is Accelerating

Training is Evolving

Inference Scaling Accelerates

Explosion of Models

Reasoning & Agents Surge

AMD

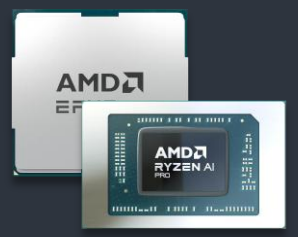
AI Strategy

Leadership Compute Engines

Open Ecosystem

Full Stack Solutions

End-to-end portfolio of AI offerings



CPUs

AMD EPYC™
AMD Ryzen™ AI PRO

GPUs

AMD Instinct™

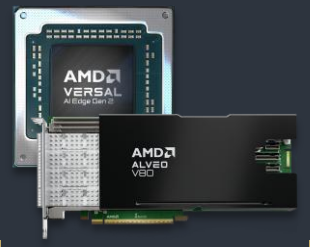
AMD
AI Software

AMD ROCm
AMD RYZEN AI PRO
AMD ZenDNN
AMD Vitis
AMD Vivado

AMD PENSANDO
Ultra Ethernet Consortium
UA Link | Ultra Accelerator Link

Networking

DPUs, UALink and
Ultra Ethernet Networking



Adaptive Computing

AMD Versal™
AMD Alveo™

zt Systems

Cluster Level
Systems Design

ISV, AI infrastructure partnerships: applications, models, frameworks, architecture, optimization

OEM, CSP, GSI Partnerships

Solutions

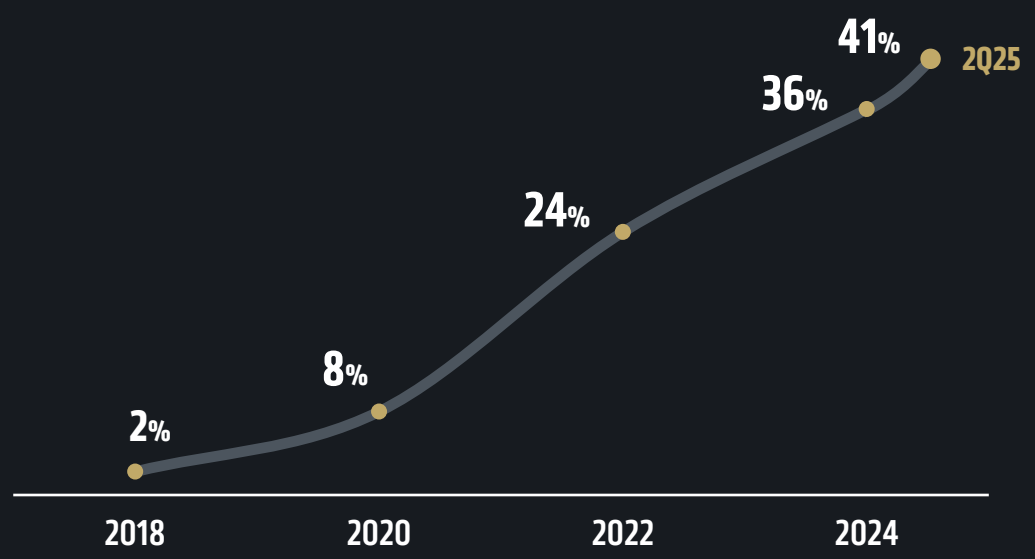
Workload validations, reference architectures, last mile support with AMD Silo AI™

Leadership Engines for Enterprise AI Workloads

From analytics to generative AI to agentic AI

EPYC Momentum Accelerates...

>18x Server CPU Market Share Growth



Industry Leaders Run on EPYC™

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

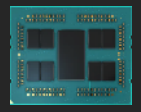
AMD EPYC™ Processors

Five generations of on time technology innovation

1st Gen AMD EPYC™

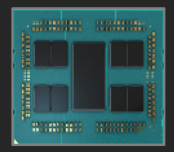
7001
SP3

2nd Gen AMD EPYC™



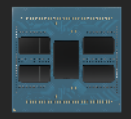
7002
SP3

3rd Gen AMD EPYC™

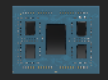


7003
SP3

4th Gen AMD EPYC™



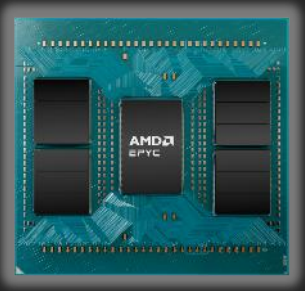
9004
SP5



8004
SP6

4004
AM5

5th Gen AMD EPYC™

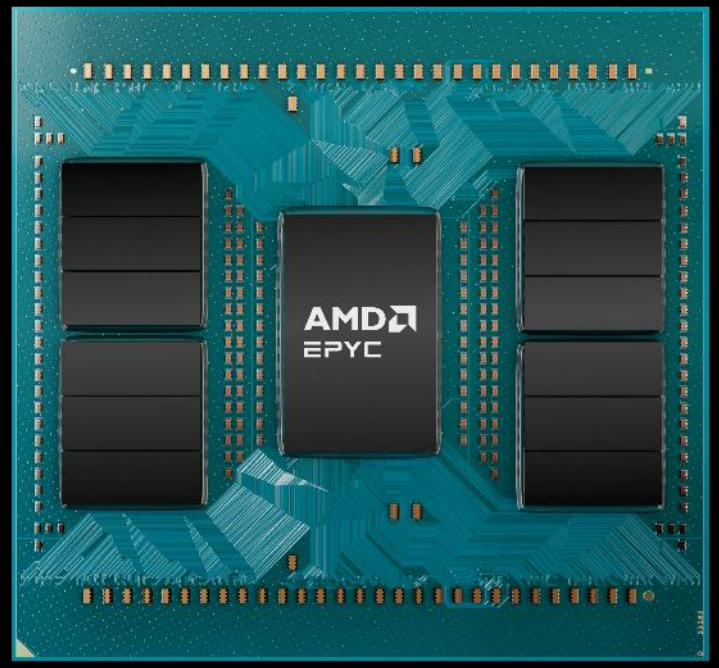


9005
SP5

All roadmaps are subject to change.

5th Gen AMD EPYC™ Processors

Formerly codenamed “Turin”



World’s best CPU for cloud, enterprise & AI

TSMC 3/4nm

Up to **192 cores**
Up to **384 threads**

Up to **5GHz**

AVX512
full 512b data path

17%
Enterprise IPC Uplift

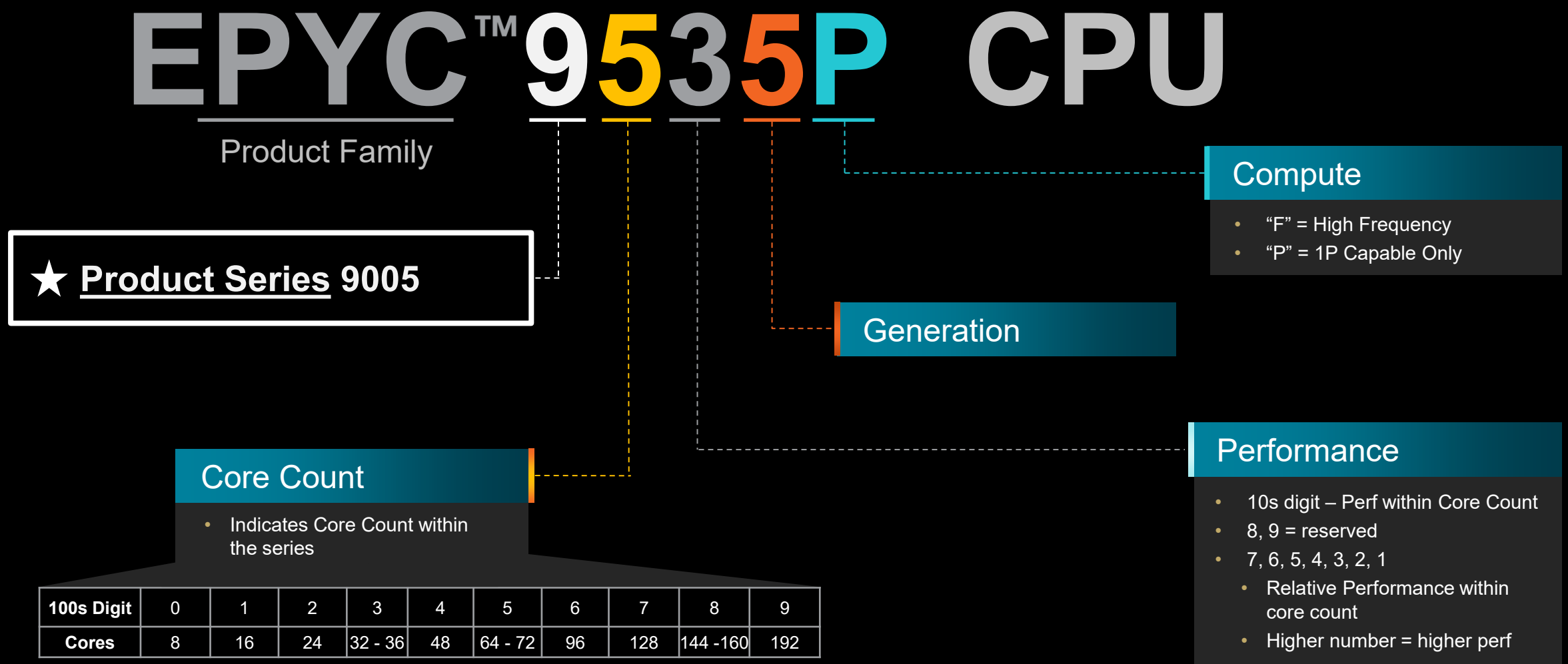
37%
HPC/AI IPC Uplift

SP5 Platform
Compatible with “Genoa”

See endnote: 9xx5-001, 048

AMD EPYC™ 9005 Series

Processor Naming Convention



100s Digit	0	1	2	3	4	5	6	7	8	9
Cores	8	16	24	32 - 36	48	64 - 72	96	128	144 -160	192

AMD EPYC™ 9005 Series Processors

- Increased core density
- Energy efficient
- Broad OPN stack

Cores	AMD EPYC	CCD (Zen5/Zen5c)	Base/Boost* (up to GHz)	Default TDP (W)	L3 Cache (MB)
192 cores	9965	“Zen5c”	2.25 / 3.7	500W	384
160 cores	9845	“Zen5c”	2.1 / 3.7	390W	320
144 cores	9825	“Zen5c”	2.2 / 3.7	390W	384
128 cores	9755 9745	“Zen5” “Zen5c”	2.7 / 4.1 2.4 / 3.7	500W 400W	512 256
96 cores	9655 9655P 9645	“Zen5” “Zen5” “Zen5c”	2.6 / 4.5 2.6 / 4.5 2.3 / 3.7	400W 400W 320W	384 384 256
72 cores	9565	“Zen5”	3.15 / 4.3	400W	384
64 cores	9575F 9555 9555P 9535	“Zen5” “Zen5” “Zen5” “Zen5”	3.3 / 5.0 3.2 / 4.4 3.2 / 4.4 2.4 / 4.3	400W 360W 360W 300W	256 256 256 256
48 cores	9475F 9455 9455P	“Zen5” “Zen5” “Zen5”	3.65 / 4.8 3.15 / 4.4 3.15 / 4.4	400W 300W 300W	256 256 256
36 cores	9365	“Zen5”	3.4 / 4.3	300W	192
32 cores	9375F 9355 9355P 9335	“Zen5” “Zen5” “Zen5” “Zen5”	3.8 / 4.8 3.55 / 4.4 3.55 / 4.4 3.0 / 4.4	320W 280W 280W 210W	256 256 256 128
24 cores	9275F 9255	“Zen5” “Zen5”	4.1 / 4.8 3.25 / 4.3	320W 200W	256 128
16 cores	9175F 9135 9115	“Zen5” “Zen5” “Zen5”	4.2 / 5.0 3.65 / 4.3 2.6 / 4.1	320W 200W 125W	512 64 64
8 cores	9015	“Zen5”	3.6 / 4.1	125W	64

Easily Upgrade to 5th Gen AMD EPYC™ CPUs

Modernize your data center – Add more capacity for your compute needs

1000 Old Servers
2P Intel® Xeon® Platinum 8280 servers

2P Intel® Platinum 8280 CPU
(28 core)

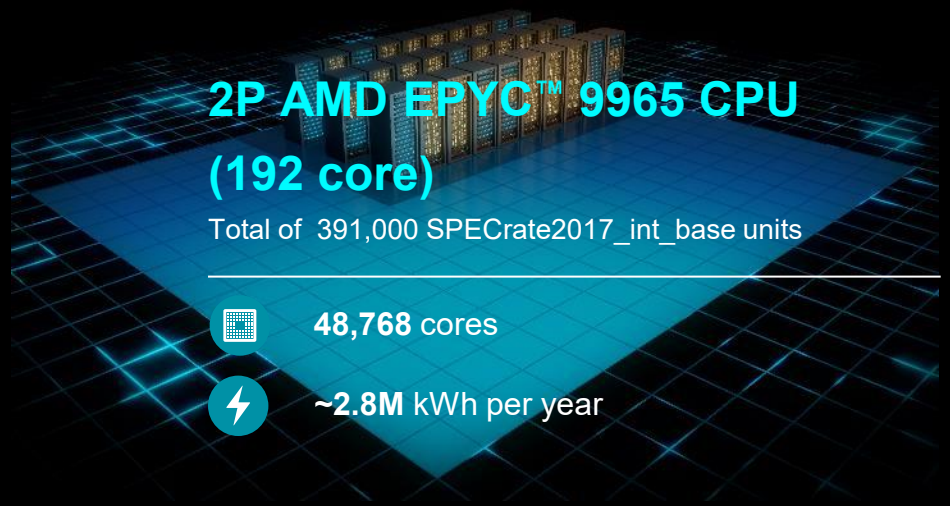
Total of 391,000 SPECrate2017_int_base units

- **56,000** cores
- **~9.0M** kWh per year

8 to 1

Easy to migrate to AMD
- X86 architecture
- Mature ecosystem
- Robust tools

127 Modern Servers
2P AMD EPYC™ 9965 servers



2P AMD EPYC™ 9965 CPU
(192 core)

Total of 391,000 SPECrate2017_int_base units

- **48,768** cores
- **~2.8M** kWh per year

Up to **69%**
Less power

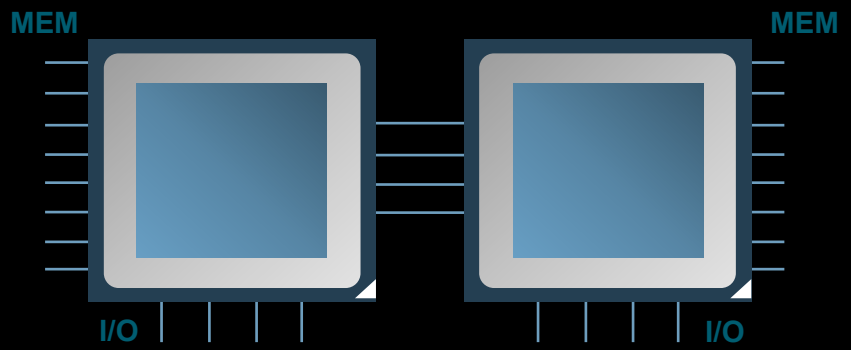
Up to **87%**
Fewer Servers

Up to **79%**
Lower 5-yr TCO

Servers required to achieve a **total of 39,100 SPECrate®2017_ int_base** performance score.
See endnotes 9xx5TCO-005

換一個思考方向

達成效率不需要妥協

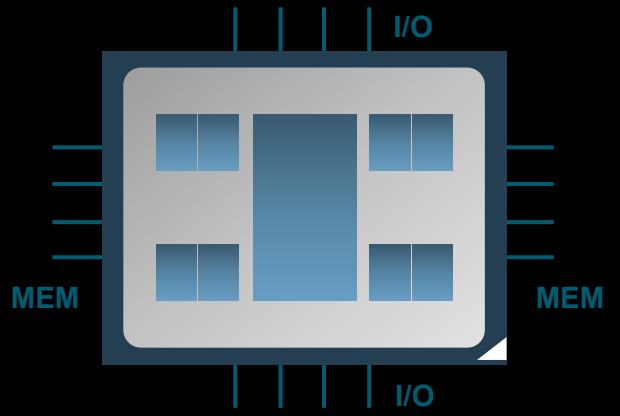


~~為何要買雙路伺服器?~~

- ~~運算效力需求~~
- ~~I/O或記憶體需求~~
- ~~一直以來都這樣買，為何要改變?~~
- ~~雙處理器可以提供提高可靠性~~

AMD EPYC™ 處理器
改變對單路伺服器的觀念:

- 一顆處理器就能提供雙處理器的效能和功能*
- 運算效率提高，減少跨CPU時產生的記憶體延遲
- 能源效益，且減少總持有成本(硬體費用和電力使用)
- 成本效益，軟體授權成本可能更低, 節省支出



AMD

AMD EPYC™ “Venice”

Highest Performance Server CPU

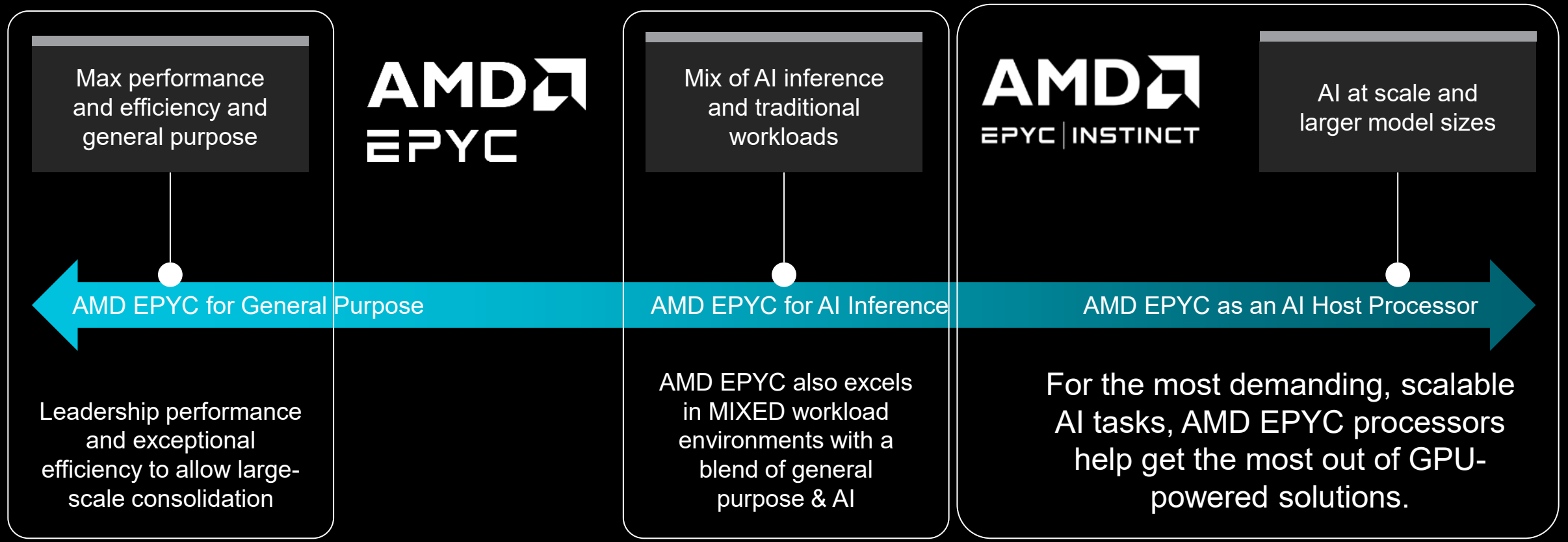
Up to **256 cores**	**2.0x**	**1.7x**	**1.6** TB/s
2nm • Zen 6	CPU to GPU Bandwidth	Gen vs. Gen Performance	Memory Bandwidth

Coming in 2026

See endnote: VEN-003

AMD EPYC™ CPUs Enable Customer AI Initiatives

Spanning traditional compute, mixed AI & AI at Scale with optimized CPU + GPU solutions



AMD

5 AI WORKLOADS
THAT CAN RUN ON A CPU

1 **CLASSIC MACHINE LEARNING**
Traditional machine learning algorithms don't benefit from parallel computing GPUs

2 **COMPUTER VISION**
Pattern recognition and deep learning vision models perform well on CPUs

3 **MEMORY-INTENSIVE GRAPH ANALYSIS**
For graph analysis on large datasets, CPUs often outperform GPUs

4 **SMALL TO MID-SIZED RECOMMENDATION SYSTEMS**
CPUs are a good fit for real-time recommendation engines

5 **EXPERTLY TUNED, INTERACTIVE AI AGENTS**
Tuning models for specific tasks can significantly reduce their footprint

Up to 192 cores, 5 GHz max frequency
Processing power for classic machine learning, recommendation systems, and AI inference

12 channels of DDR5 memory running at up to 6400 Mbps
Hold large databases, AI models, and training data in memory

Up to 160 PCIe® Gen5 lanes (2P)
Move large datasets faster for more responsive AI

Origin: 5 AI Workloads That Can Run on a CPU.PDF

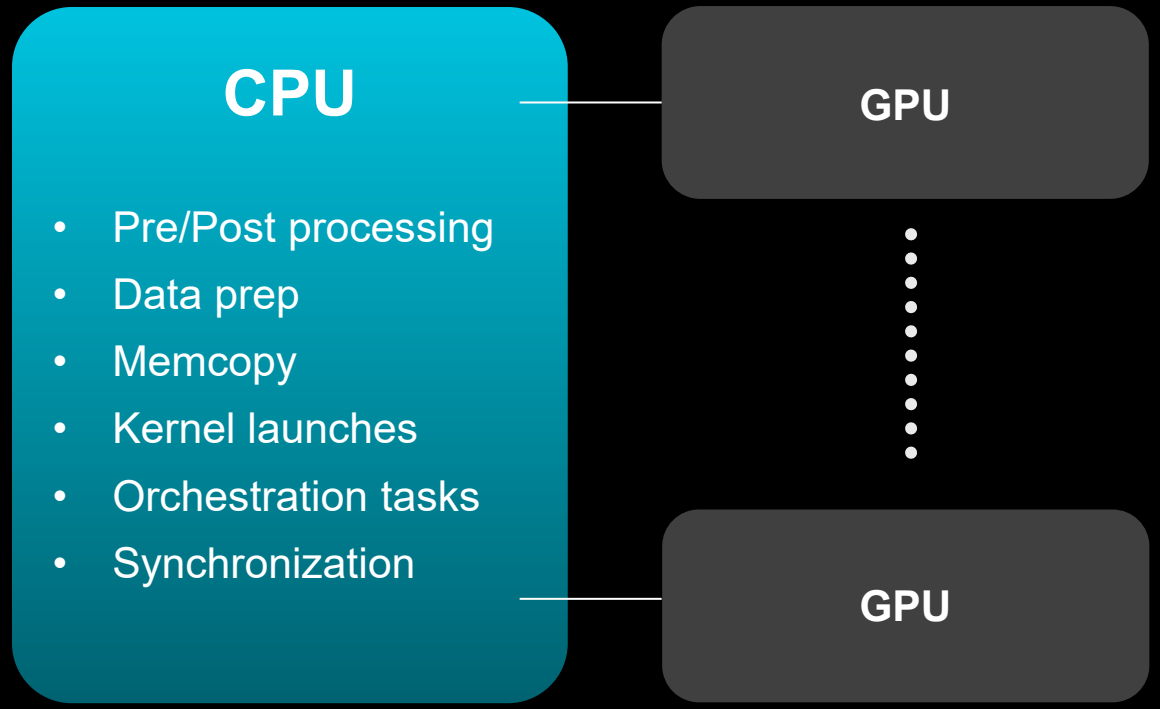
5th Gen AMD EPYC™ 9575F – 5.0 GHz High Frequency 64 Core SKU

Designed for GPU Accelerated AI Inference & Training

Faster Processing for GPU orchestration tasks

AMD EPYC™ 9575F (64C), 5.0 GHz max frequency vs.
Intel® Xeon® 8592+ (64C), 3.9 GHz max frequency

See endnotes GD-150



5th Gen AMD EPYC™ 9575F

Enabling Maximum GPU System Performance as an AI Host Processor

+8%
GPU System Performance Inference

Llama3.1-70B Inference Benchmark (8xGPU)

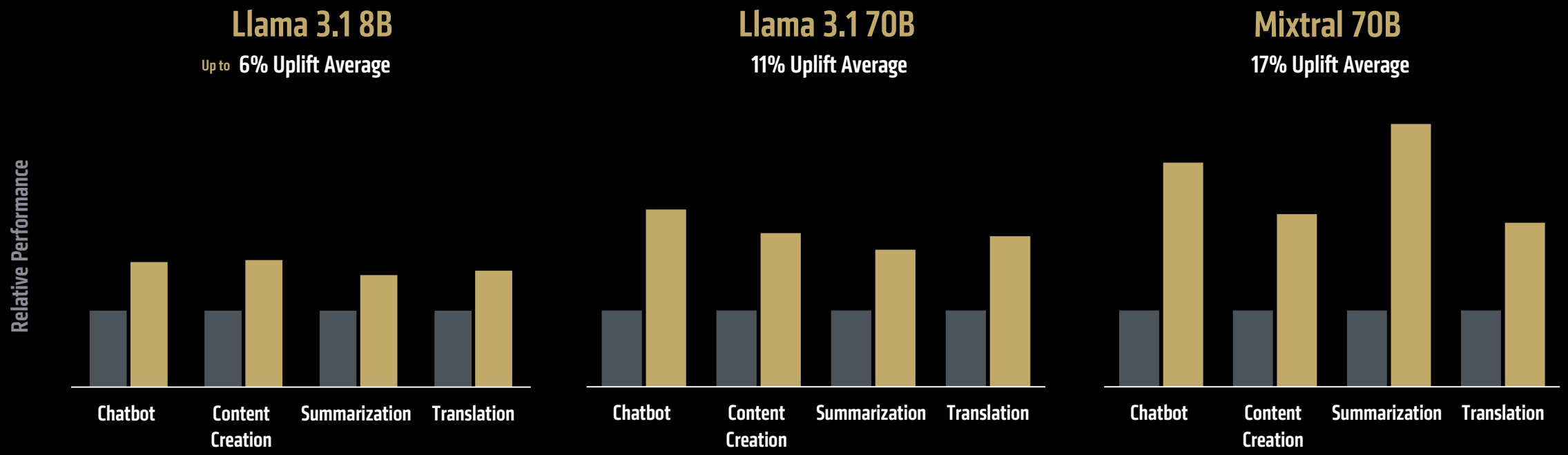
+20%
GPU System Performance Training

Stable Diffusion XL v2 Training Benchmark (8xGPU)

700K More Tokens/Second From 1K Node AI Cluster for Inference

See endnotes 9xx5-056A, 059, 087

AMD EPYC™ Driving End-to-End System Performance



https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/maximize-ai-gpu-efficiency-with-amd-epyc-processors.pdf

5th Gen AMD EPYC™ 9575F

Enabling Maximum GPU System Performance as an AI Host Processor

+20%
GPU System Performance Inference

Llama3.1-70B Inference Benchmark (8xGPU)

+15%
GPU System Performance Training

Llama3.1-8B Training Benchmark (8xGPU)

Up to 20% more requests and 15% better time to train with AMD EPYC™ 9575F

See endnotes 9xx5-014, 015

AMD EPYC™ CPUs and AMD Instinct™ GPUs

Address the full spectrum of AI workloads

Instinct™ Accelerators

- AI Training
- Dedicated AI deployments
- Medium to large Gen AI models
- Large-scale real-time inference

EPYC™ CPUs

- Mixed workload inference deployments
- Classical machine learning
- Small to medium models
- Batch, offline & small-scale real-time inference

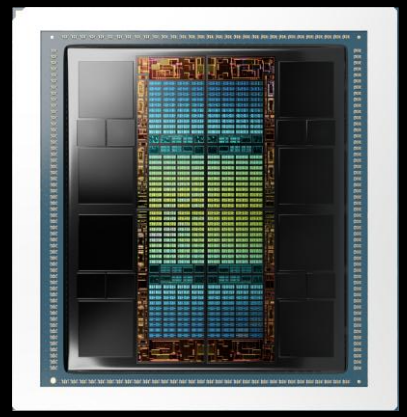
Increasing AI:
- **Performance**
- **Cost**
- **Energy Consumption**

Cost and infrastructure requirements make choosing the right solution for each workload critical

AMD INSTINCT

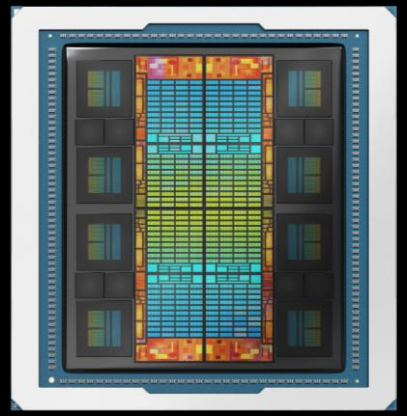
Delivering on Annual Roadmap Commitment

AMD Instinct™
MI300A/X



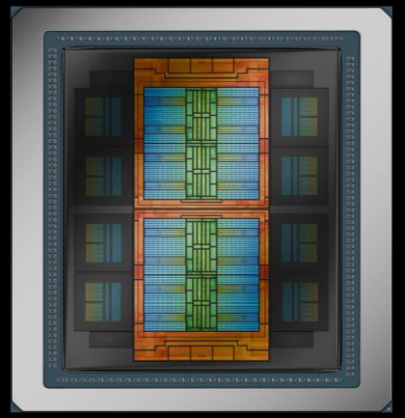
2023

AMD Instinct™
MI325X



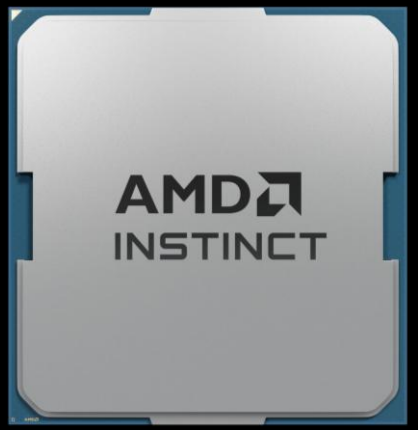
2024

AMD Instinct™
MI350 SERIES



2025

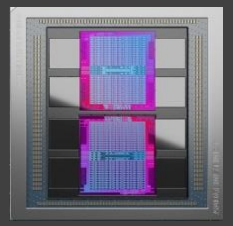
AMD Instinct™
MI400 SERIES



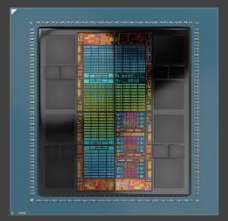
2026

Roadmap subject to change

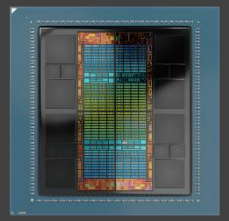
Delivering On Leadership GPU Commitment



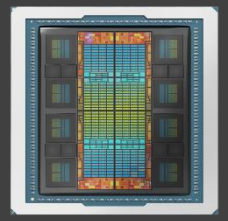
AMD Instinct™
MI250X



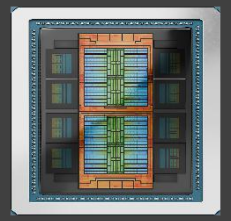
AMD Instinct™
MI300A



AMD Instinct™
MI300X



AMD Instinct™
MI325X



AMD Instinct™
MI350X

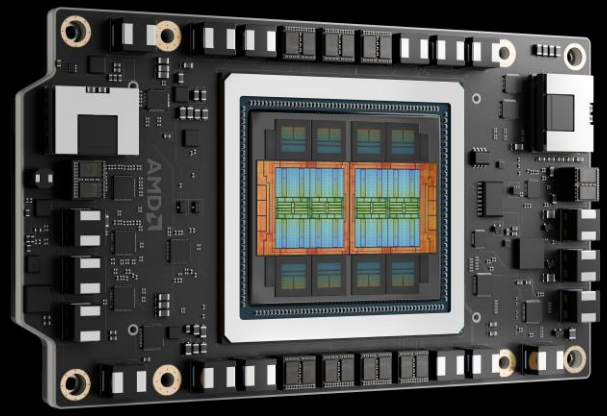
2021 — 2025

7 of 10 Largest AI Companies Use AMD Instinct

Microsoft · Meta · OpenAI · TESLA · xAI · Lawrence Livermore National Laboratory · QST · CSCS · SDSC San Diego Supercomputer Center · ORACLE · HUMAIN · LinkedIn · cohere · G42 · DELL Technologies · SUPERMICRO · Hewlett Packard Enterprise · Lenovo · IBM · SAMSUNG · CISCO · Luma AI · essential AI · HLRS High-Performance Computing Center Stuttgart · LUMI · VULTR · ASUS · GIGABYTE · COMPAL

AMD Instinct™ MI350 Series GPUs

AMD Instinct MI350X GPU



AMD Instinct MI355X GPU

20 PF | 10 PF Flops
FP4 | FP8

288 GB
HBM3E Capacity

8 TB /s
Memory Bandwidth

UBB8 Design
in Air Cooled or Liquid Cooled

Leadership Performance | Cost Efficient | Fully Open-Source

AMD
together we advance_

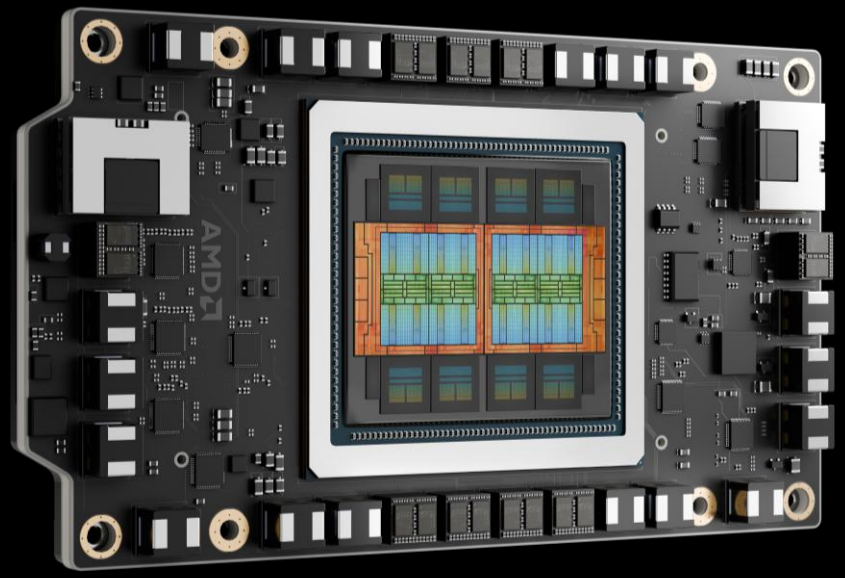
MI350 Series Accelerates Your Gen AI Outcomes

Faster AI Inference & Training	Larger AI Model Support	Rapid AI Infrastructure Deployment
20PF	**288GB**	**UBB8**
FP4 & FP6	HBM3E	Industry Standard GPU Node
4x Gen-on-Gen AI Compute Increase	Supports up to 520B Parameter AI Model	Available in Air Cooled & Direct Liquid Cooled

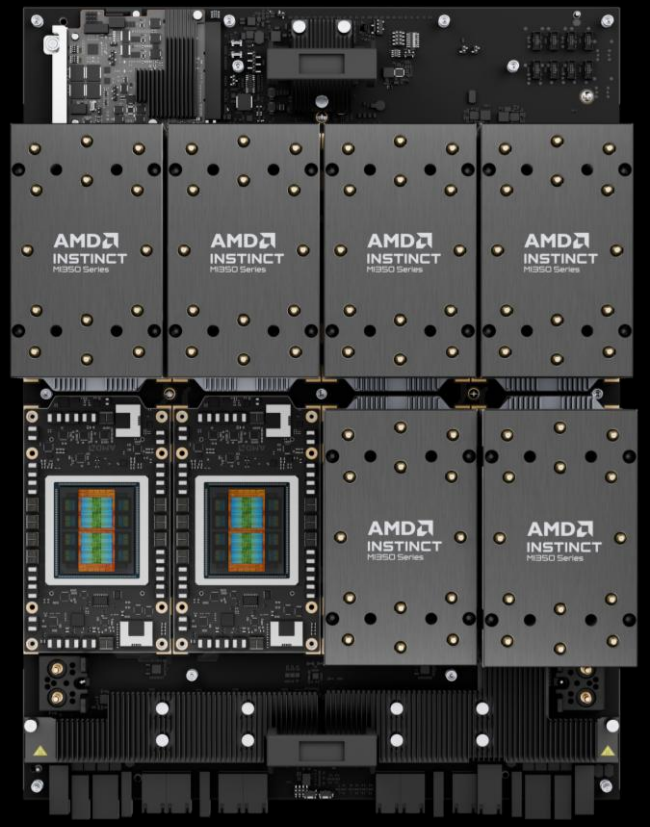
See endnote: MI350-004

AMD Instinct™

MI350 Series

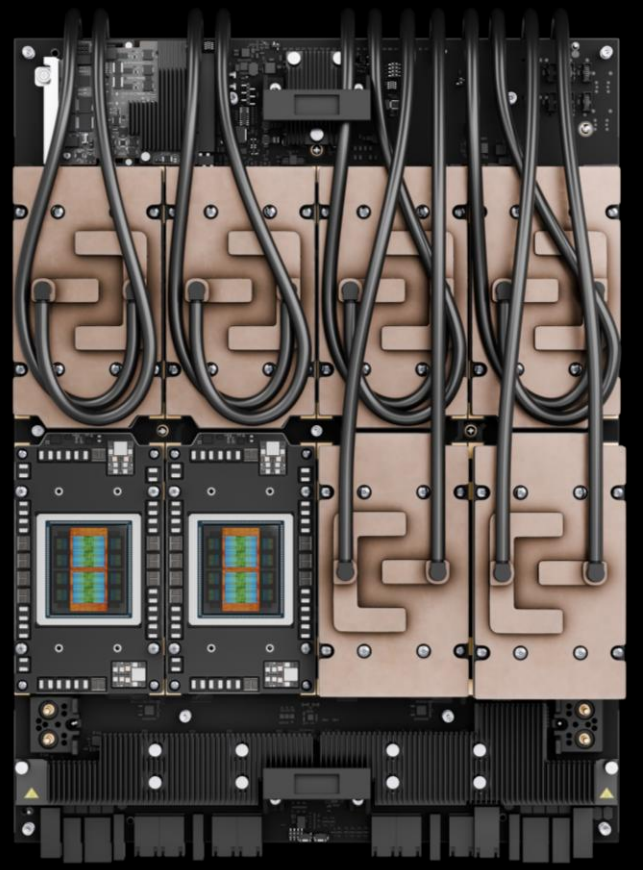


	Instinct™ MI355X
MEMORY	288 GB HBM3E
MEMORY BANDWIDTH	8 TB/s
FP64	79 TF
FP16	5 PF
FP8	10 PF
FP6	20 PF
FP4	20 PF
TBP	1400W



Air Cooled

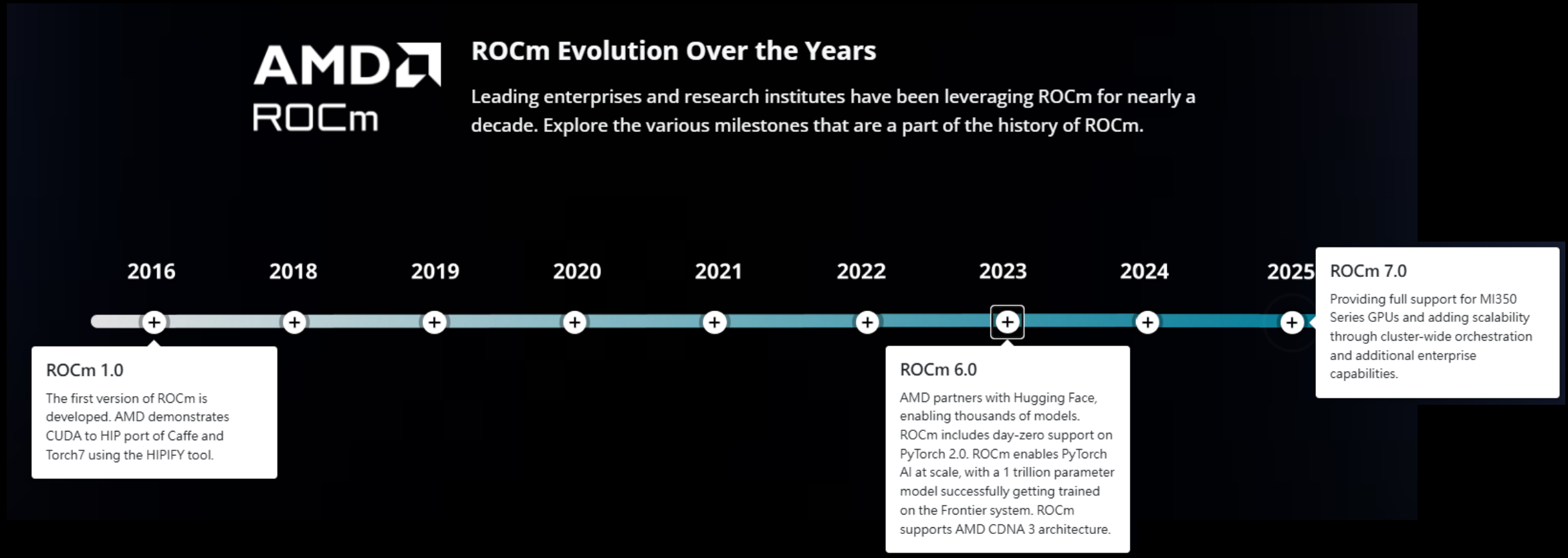
AMD Instinct™ **MI350 Series**



Liquid Cooled

AMD Instinct™ **MI350 Series**

ROCm Evolution Over the Years



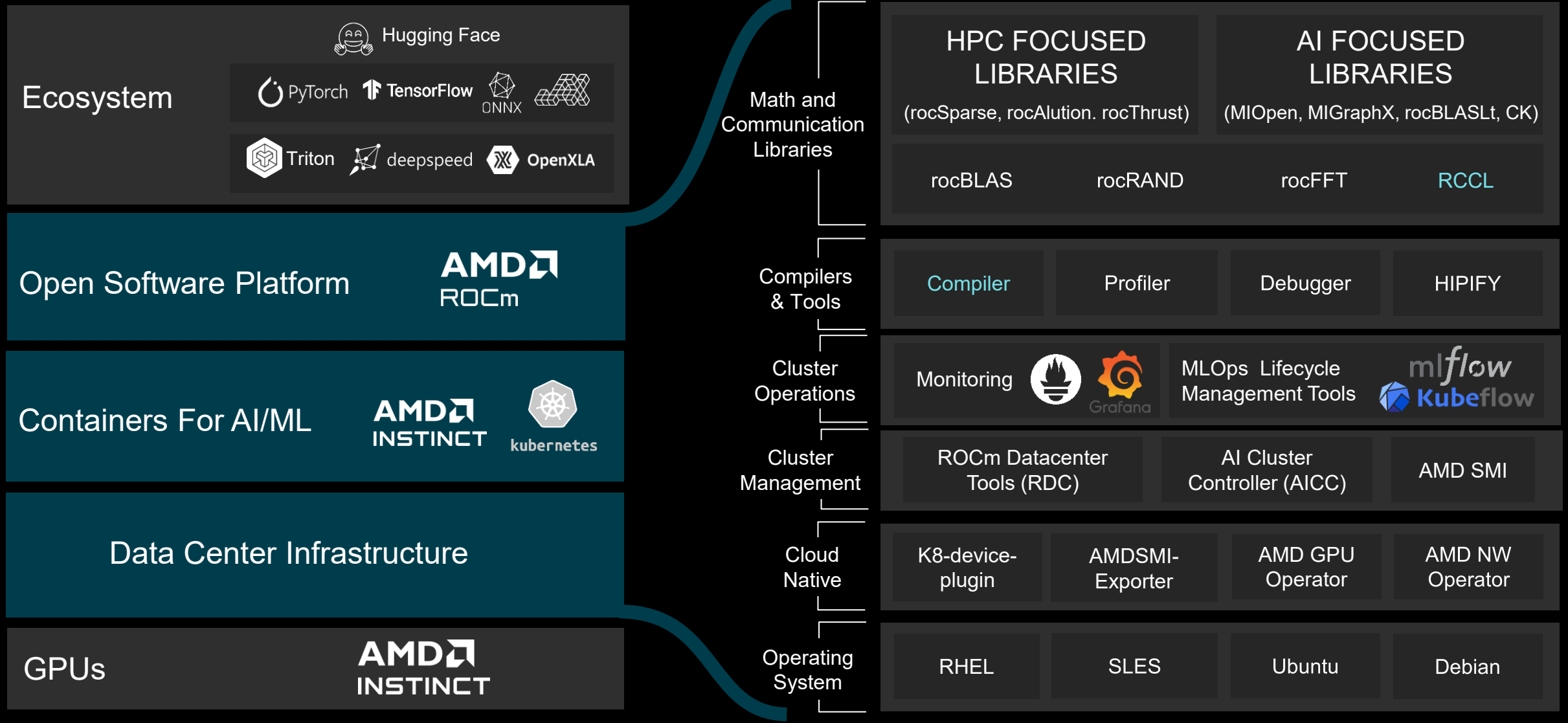
https://www.amd.com/en/products/software/rocm.html

Introducing AMD ROCm™ 7

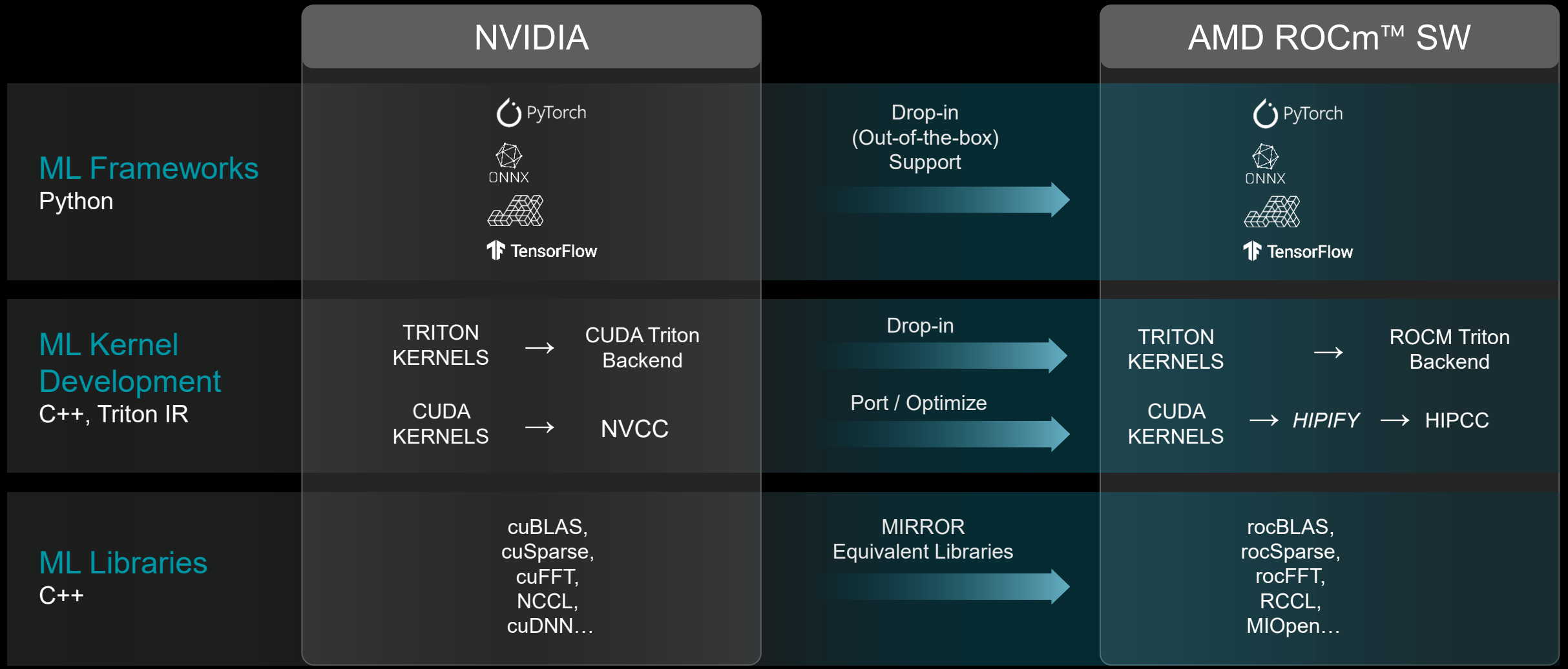
Accelerating AI Innovation & Developer Productivity

Latest Algorithms & Models	Advanced Features for Scaling AI	MI350 Series Support	Cluster Management	Enterprise Capabilities

AMD SOFTWARE OFFERING



Transitioning AI Workloads to AMD GPUs



ROCm™ Software: Can You Spot a Difference?

NVIDIA CUDA

```
import torch
import torch.nn as nn

# Get cpu or gpu device for training.
device = "cuda:0" if torch.cuda.is_available() else "cpu"
print(f"Using {device} device")

# Define model
class Network(nn.Module):
    def __init__(self):
        super().__init__()
        self.flatten = nn.Flatten()
        self.linear_relu_stack = nn.Sequential(
                nn.Linear(28 * 28, 512),
                nn.ReLU(),
                nn.Linear(512, 512),
                nn.ReLU(),
                nn.Linear(512, 10)
        )

    def forward(self, x):
        x = self.flatten(x)
        logits = self.linear_relu_stack(x)
        return logits

model = Network().to(device)
print(model)
```

AMD ROCm™ Software

```
import torch
import torch.nn as nn

# Get cpu or gpu device for training.
device = "cuda:0" if torch.cuda.is_available() else "cpu"
print(f"Using {device} device")

# Define model
class Network(nn.Module):
    def __init__(self):
        super().__init__()
        self.flatten = nn.Flatten()
        self.linear_relu_stack = nn.Sequential(
                nn.Linear(28 * 28, 512),
                nn.ReLU(),
                nn.Linear(512, 512),
                nn.ReLU(),
                nn.Linear(512, 10)
        )

    def forward(self, x):
        x = self.flatten(x)
        logits = self.linear_relu_stack(x)
        return logits

model = Network().to(device)
print(model)
```

Distributed Inference at Scale with Open Ecosystem

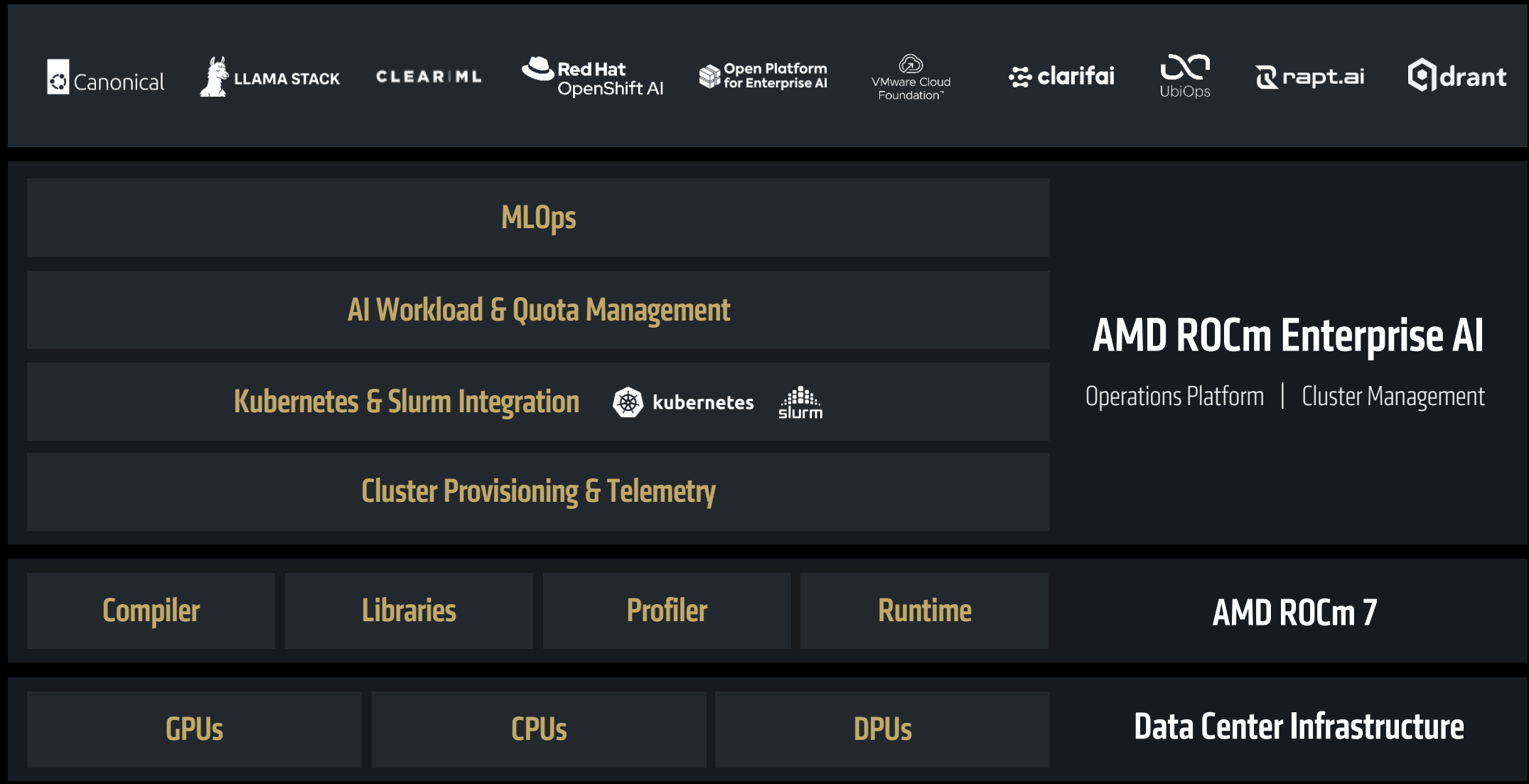
AI Serving Throughput, Multiplied

Orchestration Framework	vLLM · llm-d · SGL
Key Functions	PD KVcache Transfer | Cross-node Communication | Cross-PD Group Schedule
Key Technologies	Mooncake · GPU Direct Access · DeepEP · Distributed Triton · SHMEM

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


Canonical
LLAMA STACK
CLEARML
Red Hat OpenShift AI
Open Platform for Enterprise AI
VMware Cloud Foundation
clarifai
UbiOps
rapt.ai
drant
MLOps
AI Workload & Quota Management
Kubernetes & Slurm Integration
kubernetes
slurm
Cluster Provisioning & Telemetry
AMD ROCm Enterprise AI
Operations Platform | Cluster Management
Compiler
Libraries
Profiler
Runtime
AMD ROCm 7
GPUs
CPUs
DPUs
Data Center Infrastructure

HPE PORTFOLIO
WITH AMD EPYC FOURTH AND FIFTH GENERATION PROCESSORS

Compute

HPE ProLiant DL325 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

HPE ProLiant DL345 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

HPE ProLiant DL365 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

HPE ProLiant DL385 Gen11
4th Gen and 5th Gen AMD EPYC CPUs

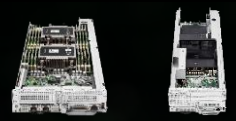
HPE ProLiant DL145 Gen11
4th Gen AMD EPYC CPUs

DX385 Gen11
4th Gen AMD EPYC CPUs

DX365 Gen11
4th Gen AMD EPYC CPUs

DX325 Gen10 Plus v2
4th Gen AMD EPYC CPUs

Supercomputing



HPE Cray XD2000
4th Gen AMD EPYC CPUs & MI210

HPE Cray Supercomputing EX4000 and EX2500
EX4252 4th Gen AMD EPYC CPUs
EX255a MI300a
EX235n 3rd Gen AMD EPYC CPUs
EX235a 3rd Gen AMD EPYC CPUs & MI250X

HPE Cray Supercomputing XD675
4th Gen AMD EPYC CPUs
AMD Instinct MI300X

HPE Cray XD665
4th Gen AMD EPYC CPUs

HPE Cray XD685
5th Gen AMD EPYC CPUs
8x AMD Instinct Accelerator

Storage

HPE Cray Supercomputing Storage Systems E2000 4th Gen AMD EPYC CPUs

SimpliVity 325 Gen11
4th Gen AMD EPYC CPUs

AMD PENSANDO | HPE aruba networking

HPE Aruba CX10000 powered by AMD Pensando

HPE GREENLAKE

AMD

HPE ProLiant Compute DL325 Gen12

QuickSpecs | QuickSpecs Changes

Overview

Shape the Future of QuickSpecs – Your Input Matters

HPE ProLiant Compute DL325 Gen12

Do you need a 1U single socket rackmount server, that maximizes rack utilization while mitigating virtualization risks in power-constrained environments?
The HPE ProLiant DL325 Gen12 maximizes your rack utilization while mitigating virtualization risks in power-constrained environments. Power your workloads with a server providing greater memory capacity compared to previous generations. With 5th Gen AMD EPYC processors up to 192 cores, increased memory capability (up to 6 TB), and new iLO7 help provide a high-performance solution with better datacenter efficiency.

HPE ProLiant Compute DL345 Gen12

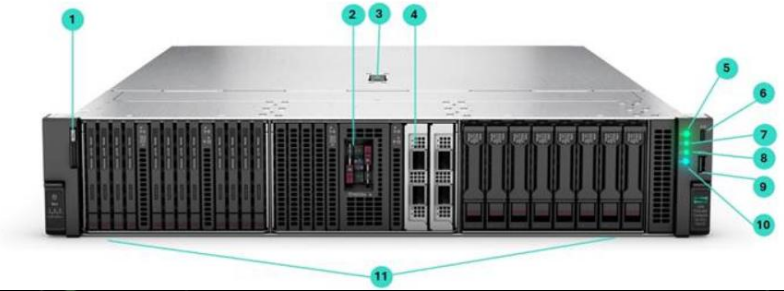
QuickSpecs | QuickSpecs Changes

Overview

Shape the Future of QuickSpecs – Your Input Matters

HPE ProLiant Compute DL345 Gen12

Are you looking for a single-socket scalable server solution to power your virtualized data-intensive, large-capacity memory workloads?
The HPE ProLiant DL345 Gen12 server is a scalable 2U 1P solution that delivers exceptional compute performance and large capacity storage options at 1P economics. This efficient and workload-optimized solution is ideal for Virtualization, SDS, and Data management.
Powered by 5th Generation AMD EPYC™ Processors with up to 192 cores, increased memory bandwidth (up to 6 TB), high-speed PCIe Gen5 I/O and EDSFF storage, up to 12LFF/ 24 SFF/ 36 EDSFF, and up to 4 double-width GPUs at the front, this server is a superb single-socket 2U solution for your data-intensive workloads



Thank You

Speaker contact information