



# 從雲到端，AMD 全方位產品藍圖

張歐佑豪 SIMON  
AMD 台灣區資深技術顧問

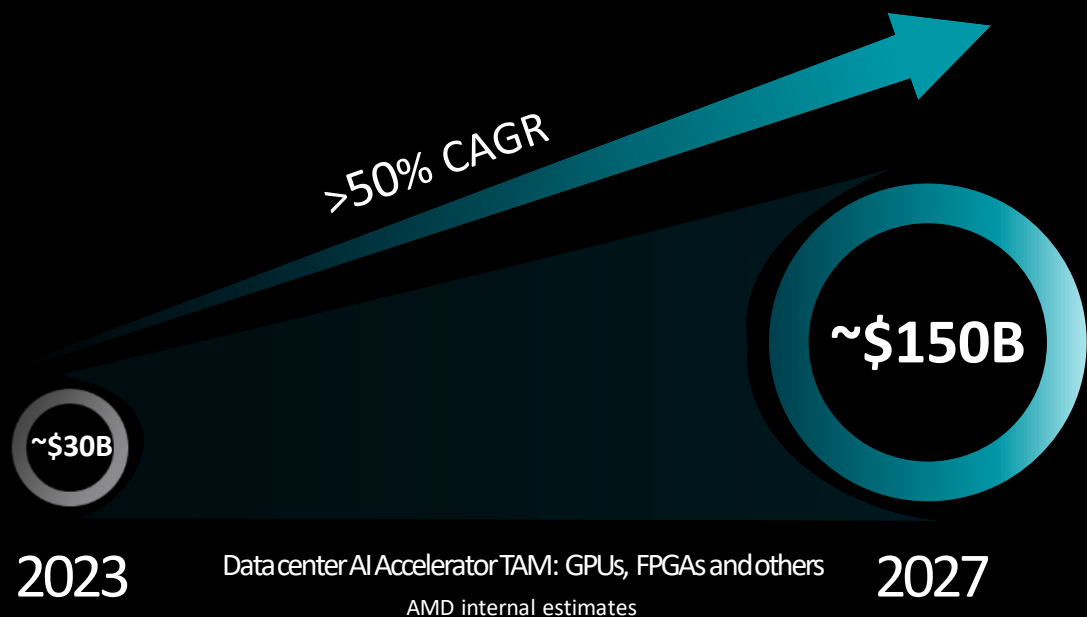
**AMD**  
together we advance\_









## CAUTIONARY STATEMENT

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products including the 4th Gen AMD EPYC™ processor family, AMD Instinct™ MI300 accelerator, AMD Instinct™ MI300X accelerator, AMD Instinct™ MI300A accelerator, and AMD Instinct™ Platform; AMD's AI strategy and AI Platforms; the Data center AI accelerator total addressable market; AMD CDNA 3 architecture; AMD "Zen 4c" architecture, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Material factors that could cause actual results to differ materially from current expectations include, without limitation, the following: Intel Corporation's dominance of the microprocessor market and its aggressive business practices; global economic uncertainty; cyclical nature of the semiconductor industry; market conditions of the industries in which AMD products are sold; loss of a significant customer; impact of the COVID-19 pandemic on AMD's business, financial condition and results of operations; competitive markets in which AMD's products are sold; quarterly and seasonal sales patterns; AMD's ability to adequately protect its technology or other intellectual property; unfavorable currency exchange rate fluctuations; ability of third party manufacturers to manufacture AMD's products on a timely basis in sufficient quantities and using competitive technologies; availability of essential equipment, materials, substrates or manufacturing processes; ability to achieve expected manufacturing yields for AMD's products; AMD's ability to introduce products on a timely basis with expected features and performance levels; AMD's ability to generate revenue from its semi-custom SoC products; potential security vulnerabilities; potential security incidents including IT outages, data loss, data breaches and cyber-attacks; potential difficulties in upgrading and operating AMD's new enterprise resource planning system; uncertainties involving the ordering and shipment of AMD's products; AMD's reliance on third-party intellectual property to design and introduce new products in a timely manner; AMD's reliance on third-party companies for design, manufacture and supply of motherboards, software and other computer platform components; AMD's reliance on Microsoft and other software vendors' support to design and develop software to run on AMD's products; AMD's reliance on third-party distributors and add-in-board partners; impact of modification or interruption of AMD's internal business processes and information systems; compatibility of AMD's products with some or all industry-standard software and hardware; costs related to defective products; efficiency of AMD's supply chain; AMD's ability to rely on third party supply-chain logistics functions; AMD's ability to effectively control sales of its products on the gray market; impact of government actions and regulations such as export administration regulations, tariffs and trade protection measures; AMD's ability to realize its deferred tax assets; potential tax liabilities; current and future claims and litigation; impact of environmental laws, conflict minerals-related provisions and other laws or regulations; impact of acquisitions, joint ventures and/or investments on AMD's business and AMD's ability to integrate acquired businesses; impact of any impairment of the combined company's assets on the combined company's financial position and results of operation; restrictions imposed by agreements governing AMD's notes, the guarantees of Xilinx's notes and the revolving credit facility; AMD's indebtedness; AMD's ability to generate sufficient cash to meet its working capital requirements or generate sufficient revenue and operating cash flow to make all of its planned R&D or strategic investments; political, legal, economic risks and natural disasters; future impairments of goodwill and technology license purchases; AMD's ability to attract and retain qualified personnel; AMD's stock price volatility; and worldwide political conditions. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q.

# AI is a leading technology megatrend.



Generative Models, Recommenders, Vision/Video	Gaming	Climate Change
		
Cloud	Healthcare	Scientific Research
		

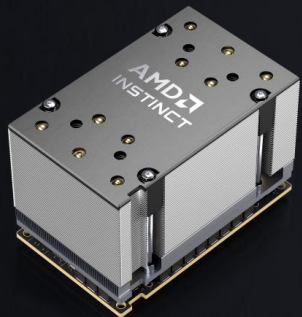
The explosive demand in compute that AI is driving requires not just one, **but multiple types of engines.**

# AMD

## AI Platforms

# Training and inference portfolio

Data center | Edge | End point



AMD Instinct™  
Accelerators

HPC and  
data center training  
and inference



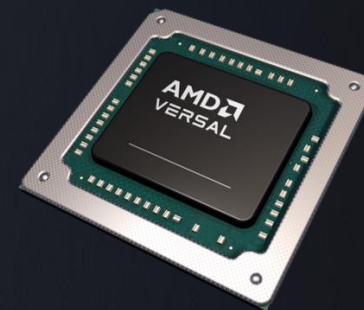
AMD Alveo™  
Accelerators

Data center and  
edge inference



4<sup>th</sup> Gen AMD EPYC™  
Processors

CPU AI leadership



AMD Embedded  
Versal™ AI Edge

AI + sensor embedded  
inference



AMD Ryzen™ 7040  
Mobile Processors

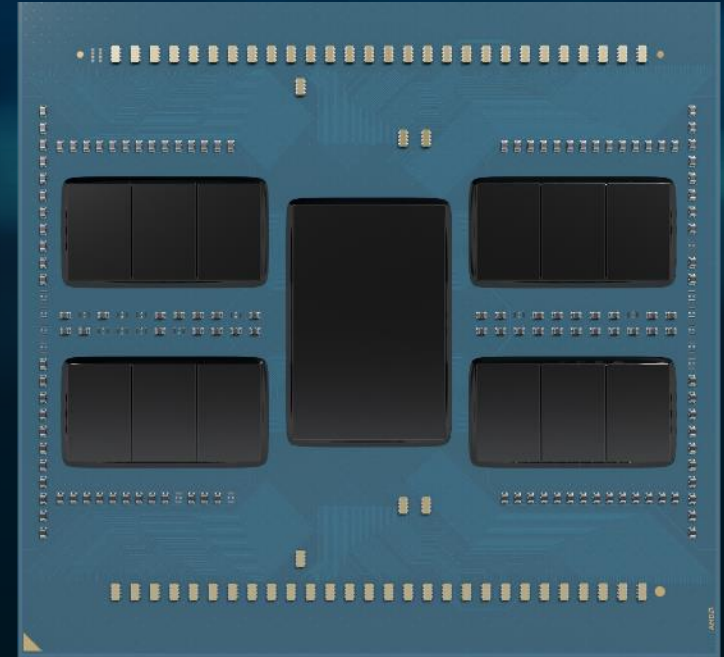
Ryzen™ AI inference  
for Windows PCs



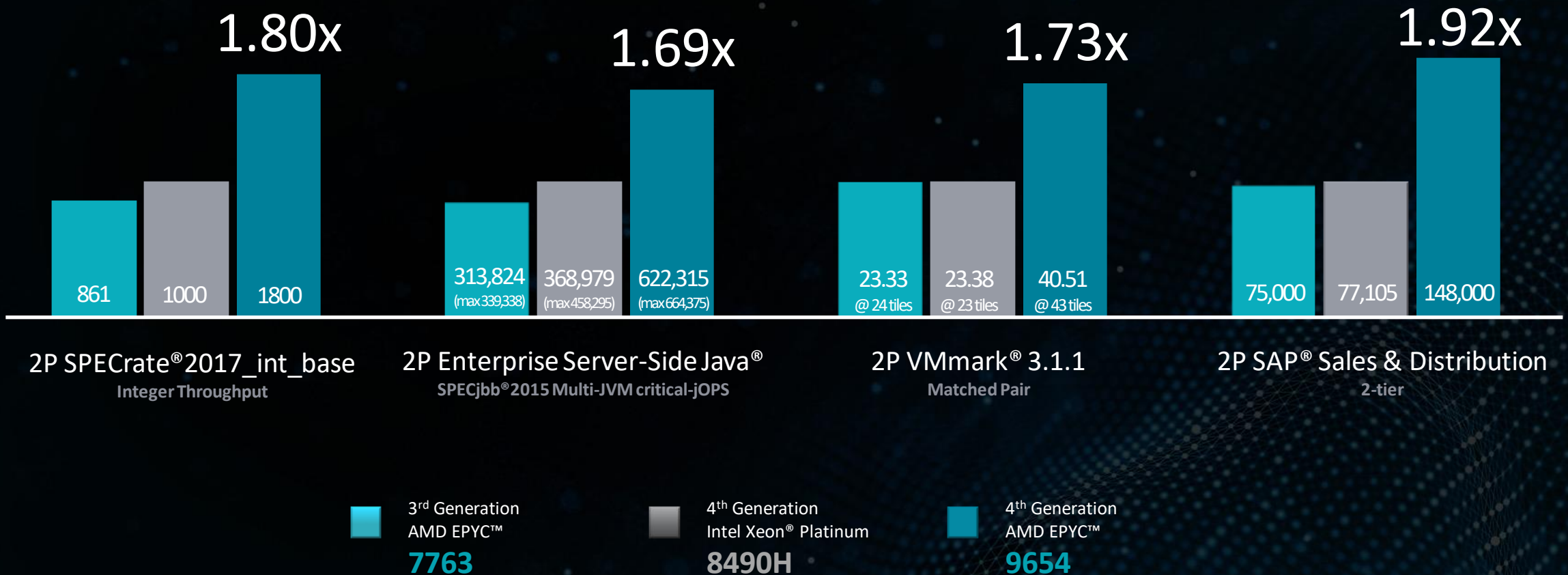
# 4th Gen AMD EPYC™ CPU

## Extending Compute Leadership

- Leadership Socket and Per-Core Performance  
Up to 96 “Zen 4” Cores in 5nm
- Leadership Memory Bandwidth and Capacity  
12 Channels DDR5
- Next Generation I/O  
Up to 160 Lanes of PCIe® Gen 5 (2P) | Memory Expansion with CXL™
- Advances in Confidential Computing  
~2X SEV-SNP Guests\* | Direct and CXL Attached Memory Encryption



# 第四代 AMD EPYC™ CPU 效能領導地位



# VMware 虛擬化環境：以單顆32核為一授權單位計價

## EPYC 有助節省

(Estimated)

60 cores per socket

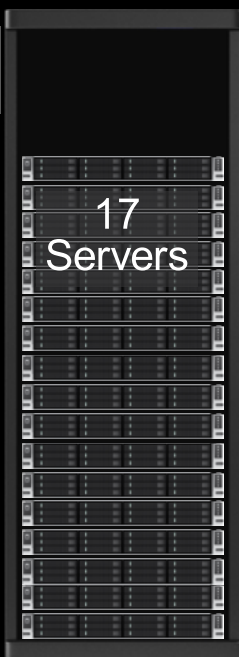
2P INTEL®  
Platinum 8490H

96 cores per socket

2P AMD  
EPYC™ 9654

2,000 VMs

Integer score  
1000  
per server

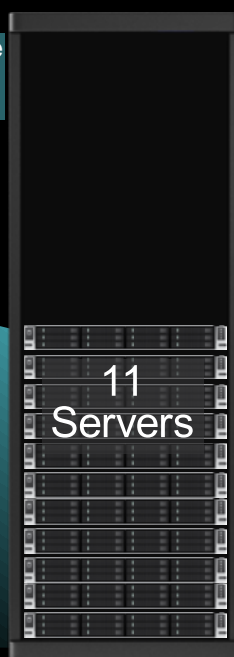


17  
Servers

2040 Cores

~27.3萬度電/一年

Integer score  
1800  
per server



11  
Servers

2112 Cores

~17.5萬度電/一年

35%<sup>up to</sup> 伺服器 and 處理器減少

36%<sup>up to</sup> 每年電力支出下降

## 擔心

- 硬體 & 軟體授權成本?
- 空間?
- 電力?

With: 6 台更少的伺服器<sup>1</sup>  
12 顆更少的處理器<sup>1</sup>  
2 個更少的軟體授權<sup>1</sup>

## EPYC 解決方案提供

(Estimated)

47%<sup>up to</sup> 硬體成本下降<sup>1</sup>

21%<sup>up to</sup> 降低每VM第一年成本<sup>1</sup>

Analysis based on the AMD EPYC™ Server Virtualization & Greenhouse Gas Emission TCO Estimation Tool - version 12.15 as of 05/19/2023. AMD processor pricing based on 1KU price as of Jan 2023. Intel® Xeon® Scalable CPU data and pricing from <https://ark.intel.com> as of Jan 2023. All pricing is in USD. Use of third-party marks /logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied. GD-83

Virtualization license cost are retail price for VMware® vSphere Enterprise Plus w/ Production support - 24x7 3yr support, calculated with one software license for every 32-core increment in a socket. VMware is a registered trademark of VMware in the US or other countries. <sup>1</sup> TCO time frame of 3-year and includes estimated costs for hardware, virtualization software, real estate, admin and power with power @ \$0.128/kWh with 8kW / rack and a PUE of 1.7. Networking and storage power external to the server are not included in this analysis. <sup>2</sup> Values are for USA.

See endnote SP5TCO-036A.

# AVX-512

## Extensions Supported

- AVX512F - Foundation
- AVX512DQ - Packed integer instructions
- AVX512\_IFMA - Integer fused mul-add
- AVX512CD - Conflict detection for vectorizing loops
- AVX512BW - Adds more packed integer instructions
- AVX512VL - Extends new instructions to 128 and 256 bits
- AVX512\_VBMI - vector byte permutation
- AVX512\_VBMI2 – more vector byte permutation
- GFNI - Galois Field New Instructions (SSE, AVX, and AVX512)
- AVX512\_VNNI - vector NN instructions
- AVX512\_BITALG
- AVX512\_VPOPCNTDQ
- AVX512\_BF16 – BFloat16 converts

## Neural Magic Sparse INT8 Inference

~4.2x

### NLP Throughput

2P AMD EPYC™ 9654 vs. 2P AMD EPYC™ 7763

---

~3x

### Image Classification Throughput

2P AMD EPYC™ 9654 vs. 2P AMD EPYC™ 7763

---

~3.5x

### Object Detection Throughput

2P AMD EPYC™ 9654 vs. 2P AMD EPYC™ 7763



# WORLD'S LARGEST HYPERSCALERS RUN ON AMD EPYC



Alibaba Cloud

Baidu

Meta

Google Cloud

IBM Cloud

ORACLE  
CLOUD

Microsoft Azure

Tencent Cloud

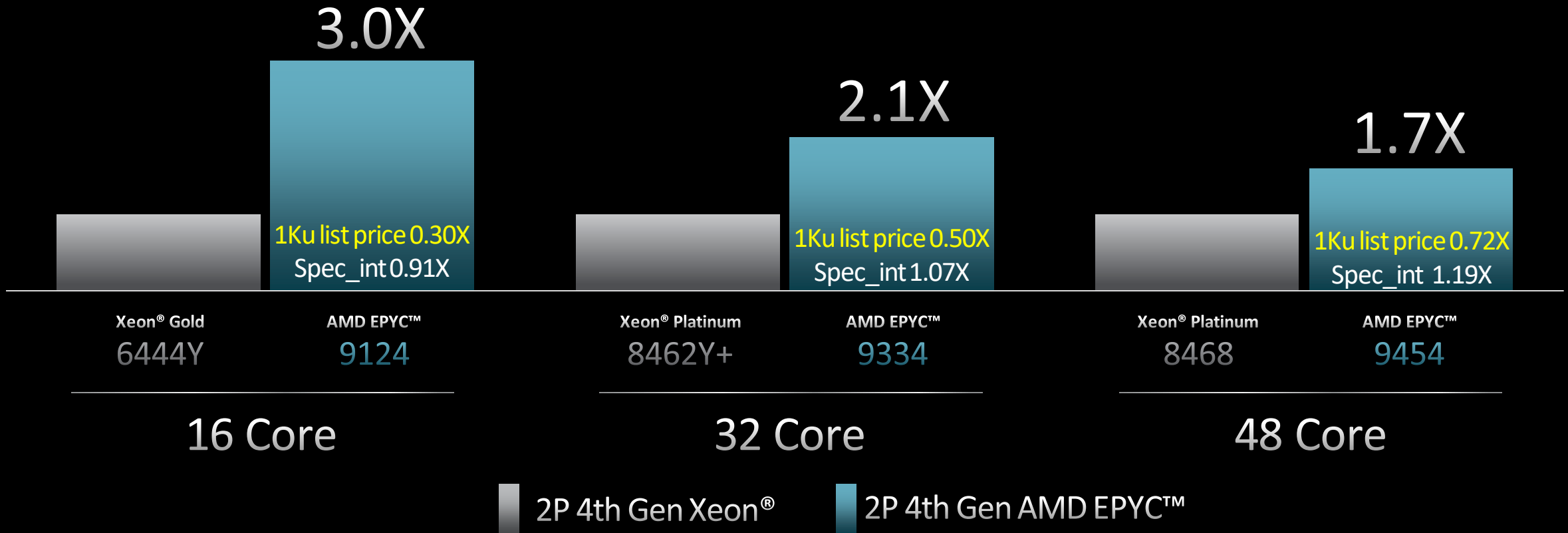


TODAY, EPYC™ PROCESSORS HAVE BEEN DESIGNED INTO DATA CENTERS BY TEN OF  
**THE WORLD'S LARGEST HYPERSCALE COMPANIES**

# 每核心整數運算效能 – 更優性價比

## 4th Gen AMD EPYC™ CPUs Offers The Best Performance/\$

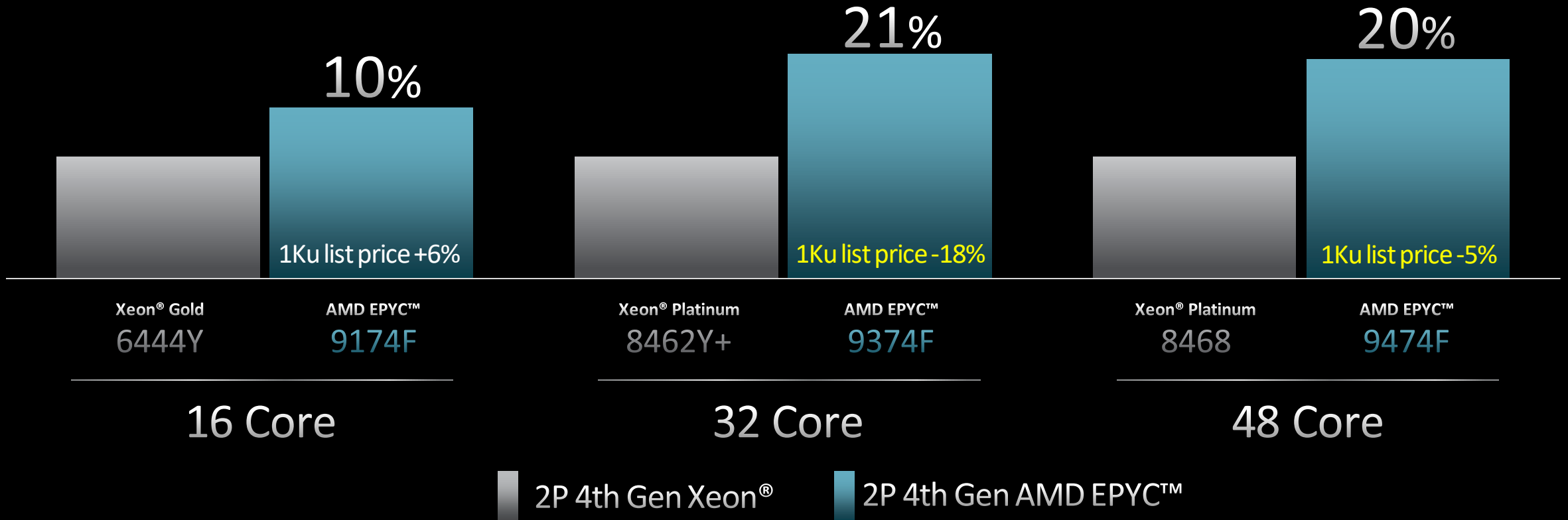
Estimated SPECrate®2017 Integer-Point Throughput Performance/\$ Comparison



# 每核心浮點運算效能領導地位

## 4th Gen AMD EPYC™ CPUs Outperform 4th Gen Xeon® Scalable Options

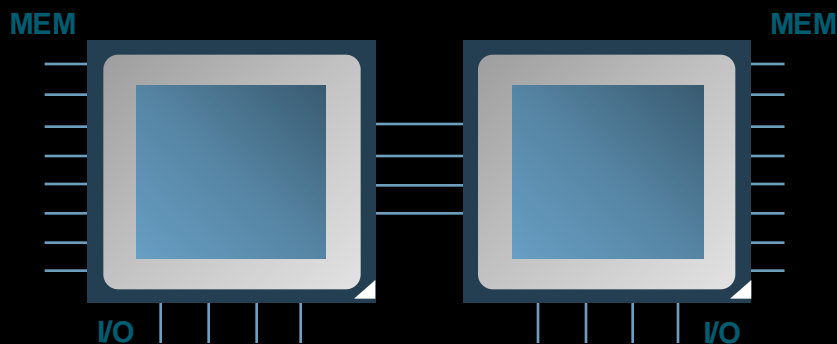
Estimated SPECrate®2017 Floating-Point Throughput Performance Uplift





# 換一個思考方向的時候到了

## 達成效率而且不需要妥協



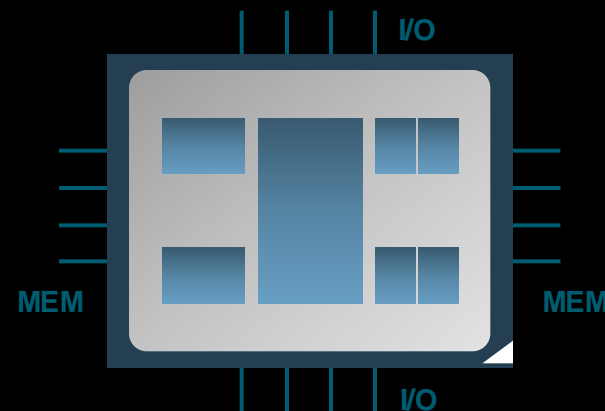
## 為何人們要買雙路伺服器?

- 運算效力需求
- IO或記憶體需求
- 一直以來都這樣買，為何需要改變?
- 雙路伺服器提供冗餘的錯誤認知

## AMD EPYC™ 處理器

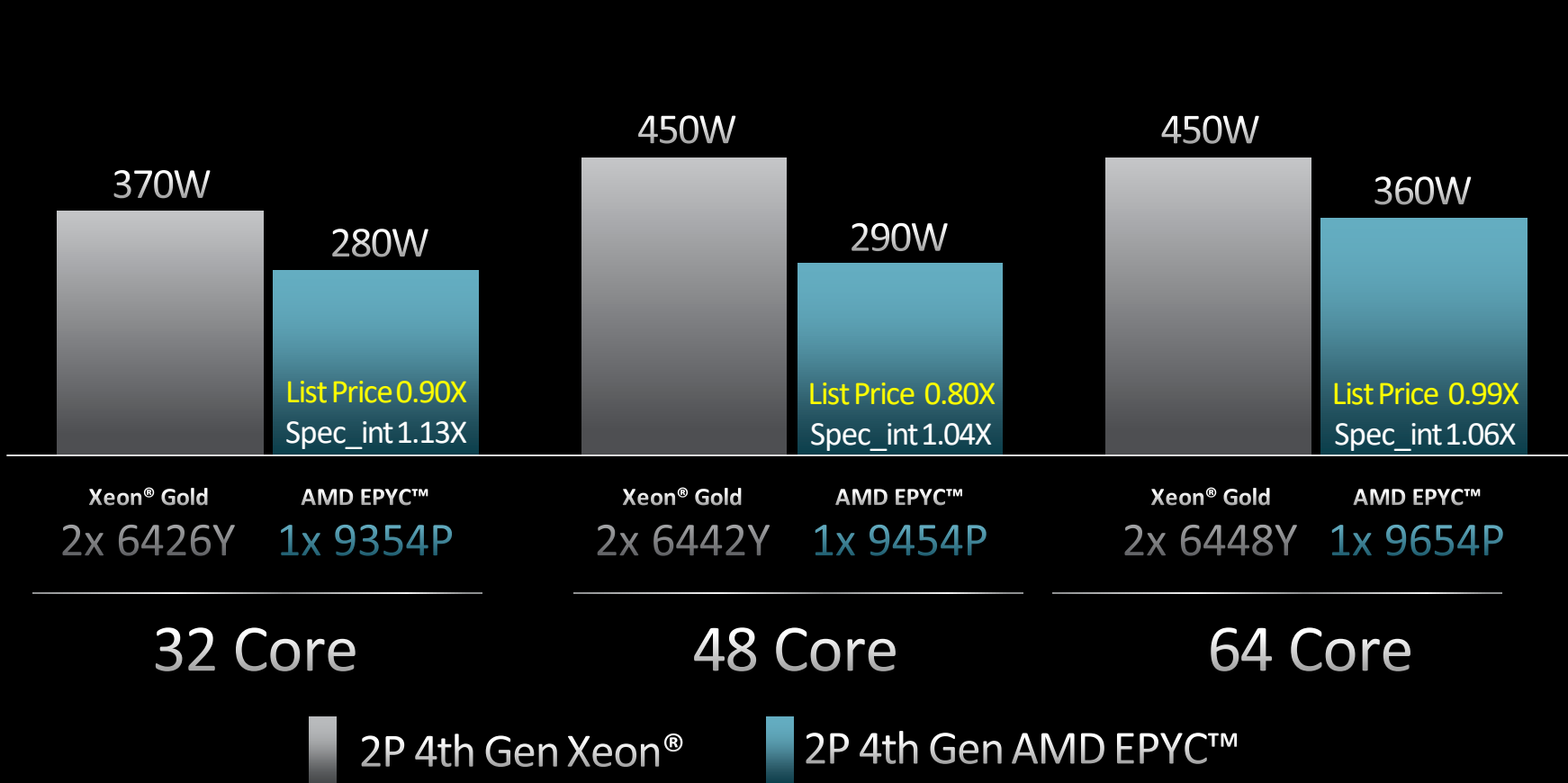
### 改變你對單路伺服器的觀念:

- 一顆CPU就能提供雙路的效能和功能\*
- 能源效益
- 減少跨CPU時產生的記憶體延遲
- 結構成本效益
- 運算效率



# 沒有任何妥協, 單路解決方案領導地位

## 1P 4<sup>th</sup> Gen AMD EPYC™ vs. 2P 4<sup>th</sup> Gen Xeon® Scalable Options



4th Gen AMD EPYC™ CPU

更多效能

更少功耗

更低CPU成本

vs. Competition

1P EPYC™ vs. 2P Xeon® CPUs

Integer Throughput

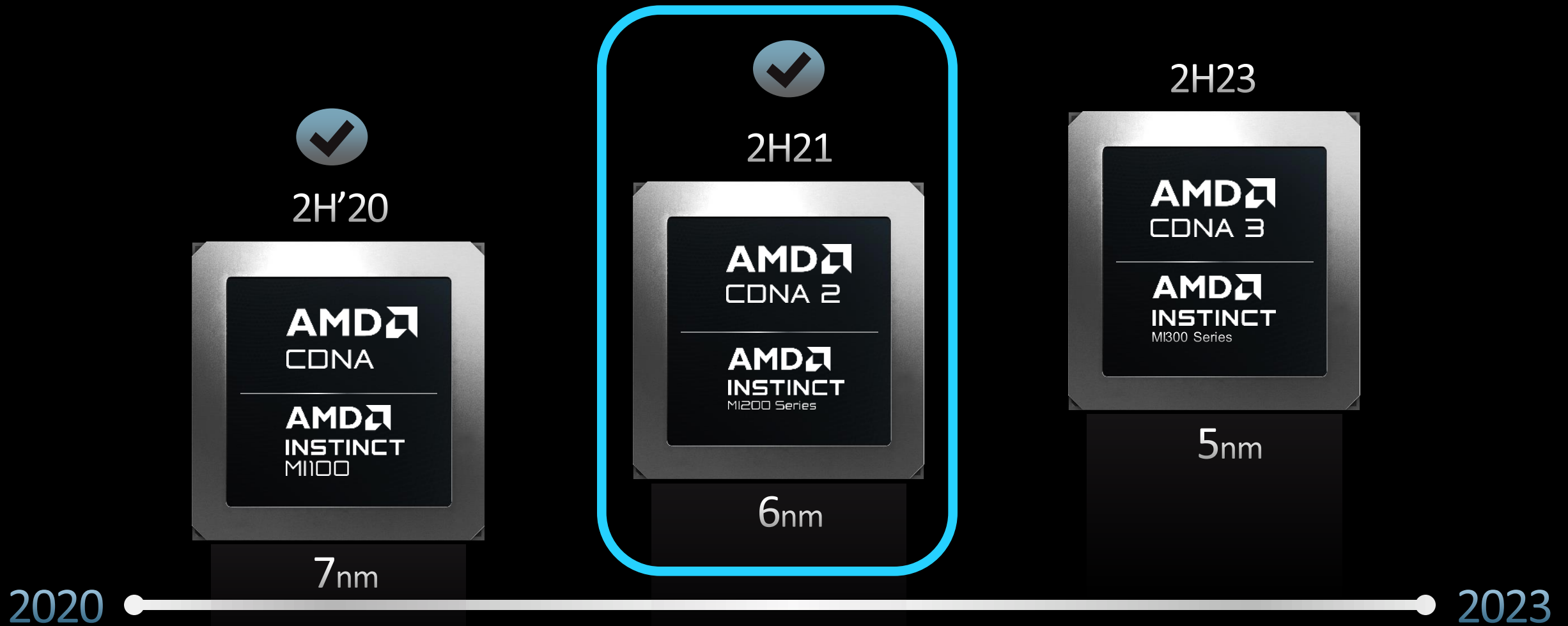
SPECrate®2017\_int\_base

# AMD 資料中心解決方案

解決方案	價值主張	應用場景
高核心數	高每機運算力，伺服器集縮比，硬體持有成本	虛擬化，公有雲，開源軟體，HPC/AI
通用型	高性價比，入門款價格，旗艦款性能	虛擬化，超融合，影音串流
AM EPYC™ 處理器	高基頻 / 高快取	資料庫，EDA，CAE/FEA，HPC
單路解決方案	軟體授權節省 (每CPU)，能源效益	虛擬化，VDI，CRM，容器化
第三代解決方案	最高性價比 (DDR4)	預算考量



# COMPUTE GPU ARCHITECTURE ROADMAP



# AMD INSTINCT™ MI200 SERIES

TACKLING YOUR MOST COMPLEX AND COMPUTE-INTENSIVE PROBLEMS



AMD INSTINCT™  
MI200 OAM SERIES

MI250, MI250X

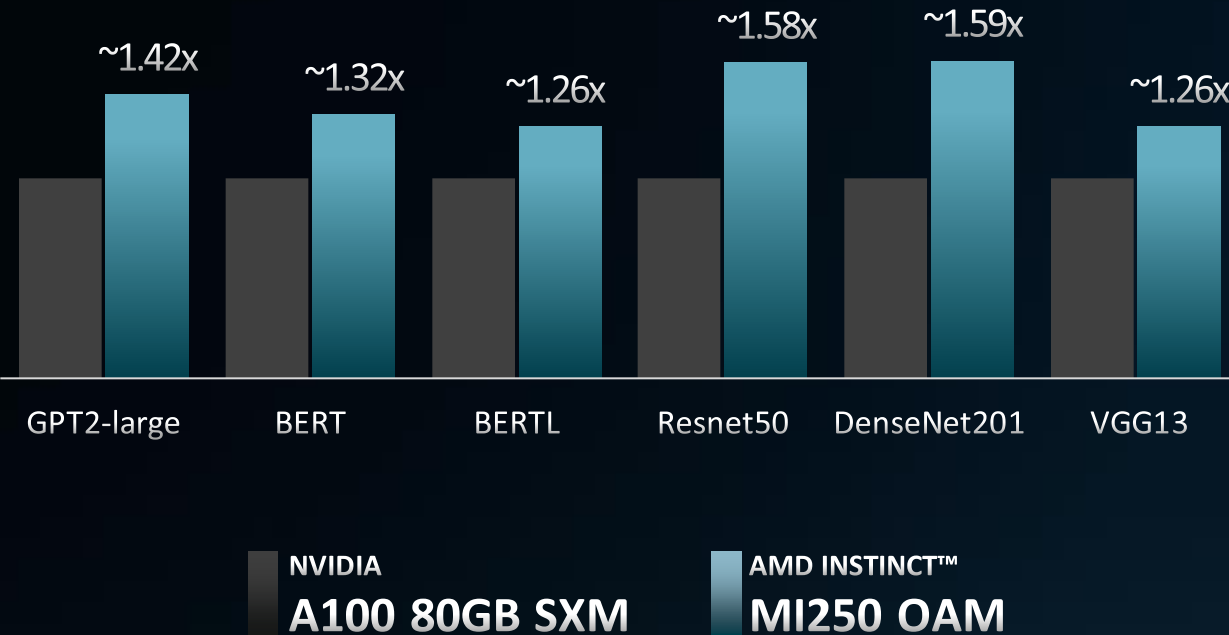


AMD INSTINCT™  
MI200 PCIe® SERIES

MI210

# ML TRAINING – PARTNERING WITH INDUSTRY LEADERS

## MICROSOFT SUPERBENCH



MICROSOFT SUPERBENCH SINGLE GPU MODEL TRAINING

## HUGGING FACE



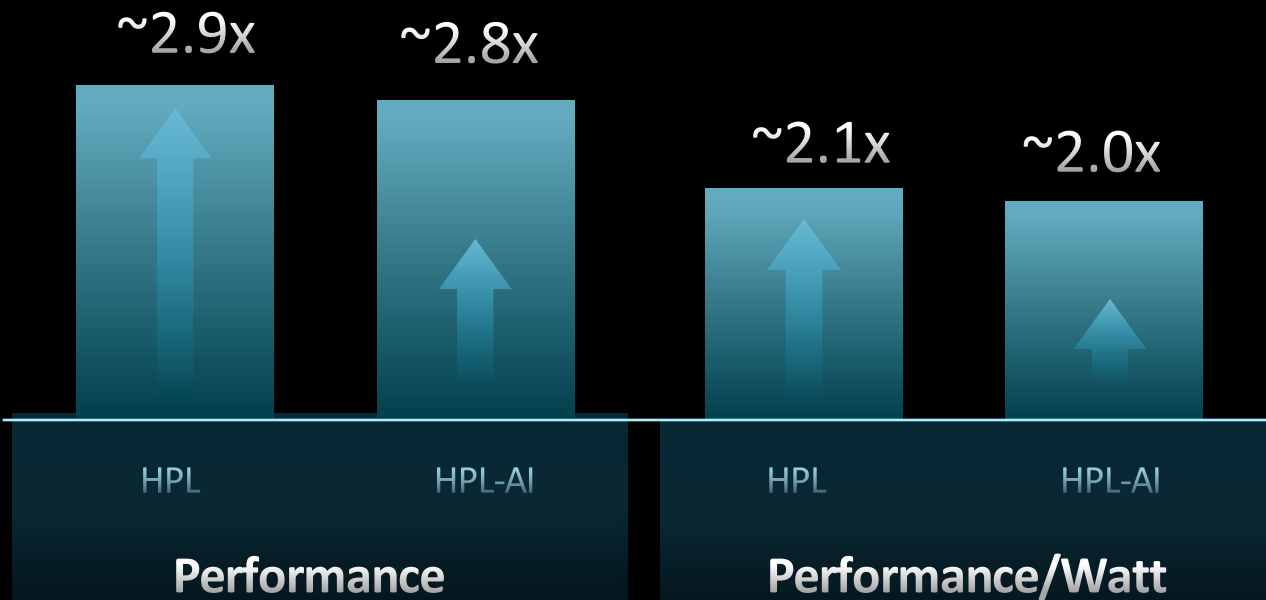
MULTI-GPU (4 GPU) MODEL TRAINING





# LEADERSHIP PERFORMANCE AND EFFICIENCY

## MI250X vs. A100 (4 GPU)

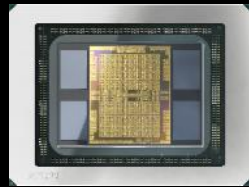


## AMD DATACENTER ENERGY EFFICIENCY

2020-2025: AMD has a goal to deliver 30X increase in energy efficiency for AMD processors and accelerators in HPC and AI training.

AMD Instinct accelerators power the #2 through #7 Green 500 systems.\*

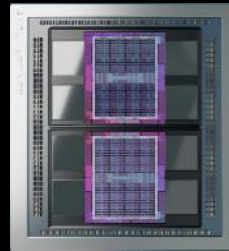
# OUR JOURNEY IN GPU ACCELERATION



**AMD Instinct™ MI100**  
AMD CDNA™

Ecosystem Growth

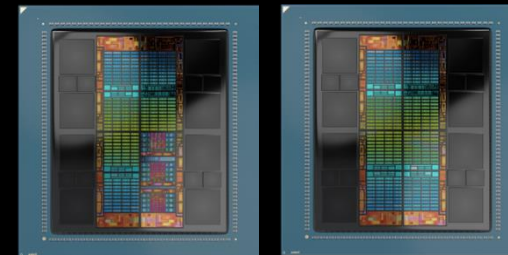
First purpose-built GPU architecture for the data center



**AMD Instinct™ MI200**  
AMD CDNA™ 2

Driving HPC and AI to a New Frontier

First purpose-built GPU powering discovery at Exascale



**AMD Instinct™ MI300**  
AMD CDNA™ 3

Data Center APU & Discrete GPU

Breakthrough architecture designed for leadership efficiency and performance for AI and HPC

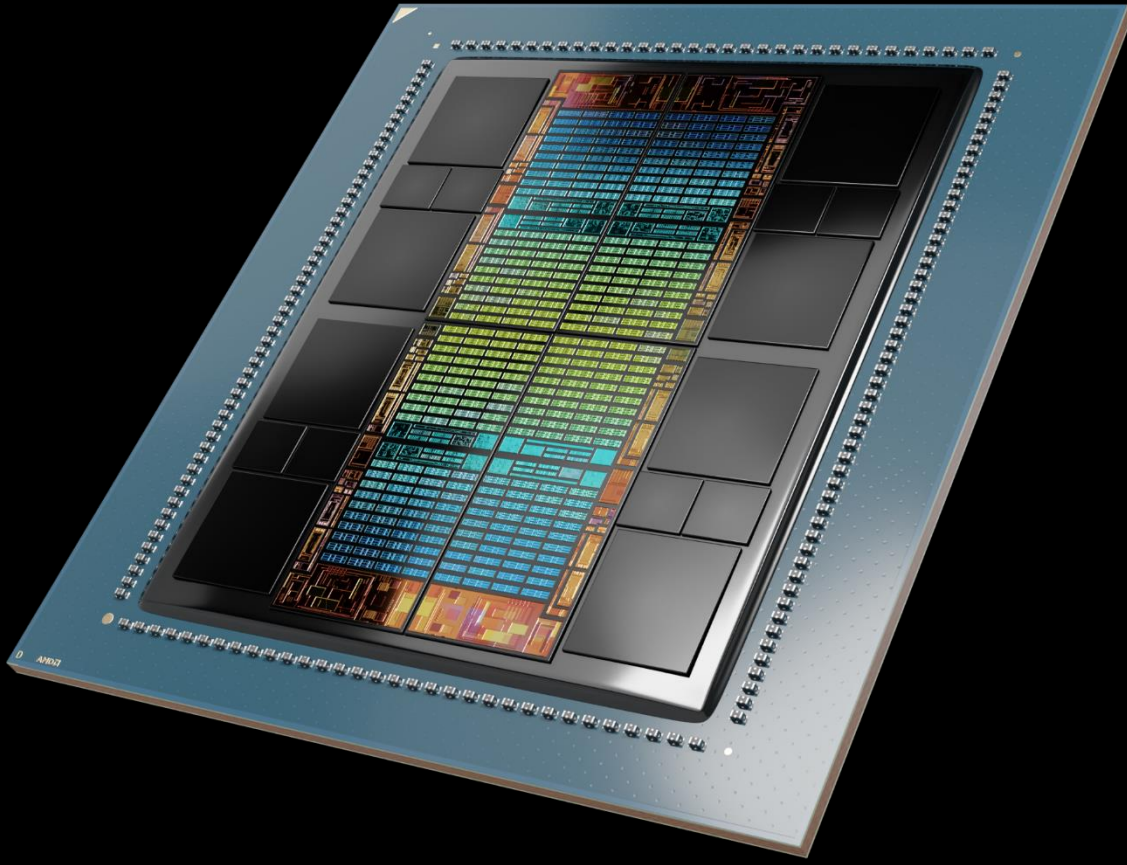
2020

2024

Sampling

Roadmaps Subject to Change

**AMD**  
together we advance\_



Introducing today

# AMD Instinct™ MI300X

Leadership generative AI accelerator



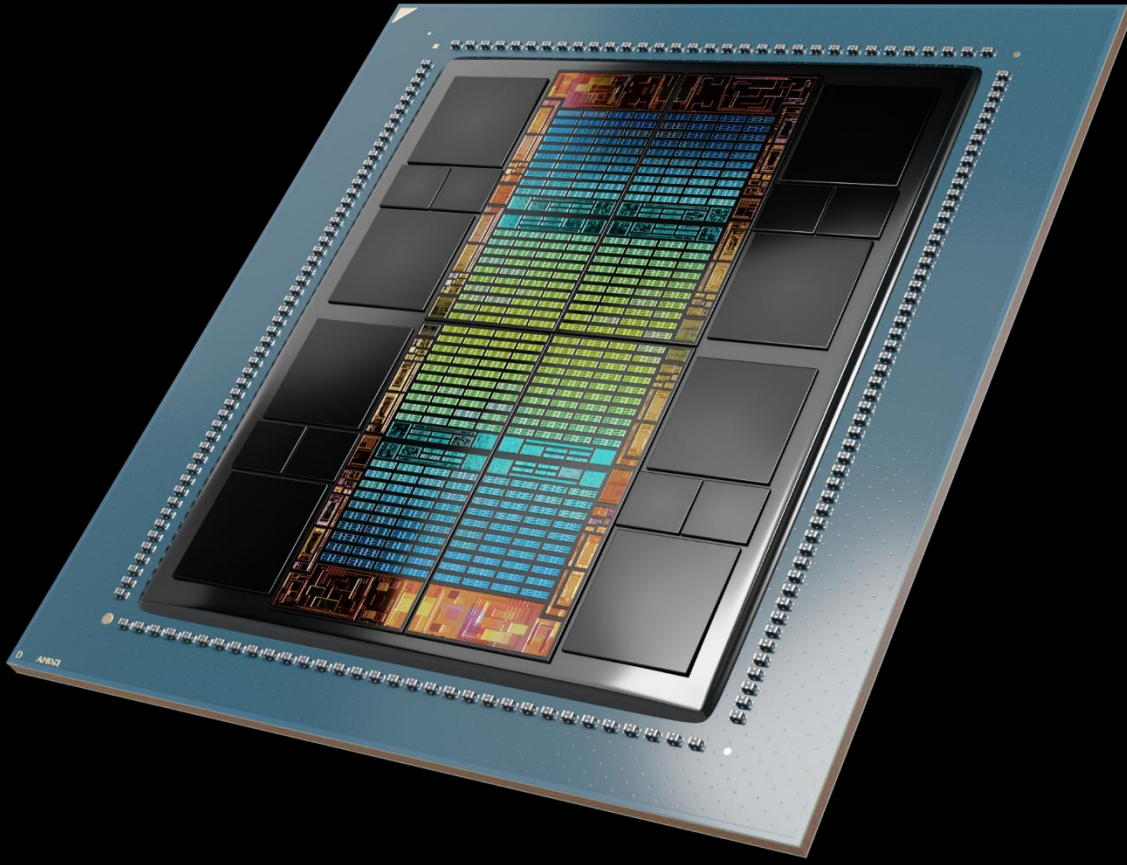
**AMD**  
CDNA 3

**192 GB**  
HBM3

**5.2 TB/s**  
Memory Bandwidth

**896 GB/s**  
Infinity Fabric™ Bandwidth

**153 B**  
Transistors



Introducing today

# AMD Instinct™ MI300X

Leadership generative AI accelerator



Up to

# 2.4x

HBM density  
compared to Nvidia H100

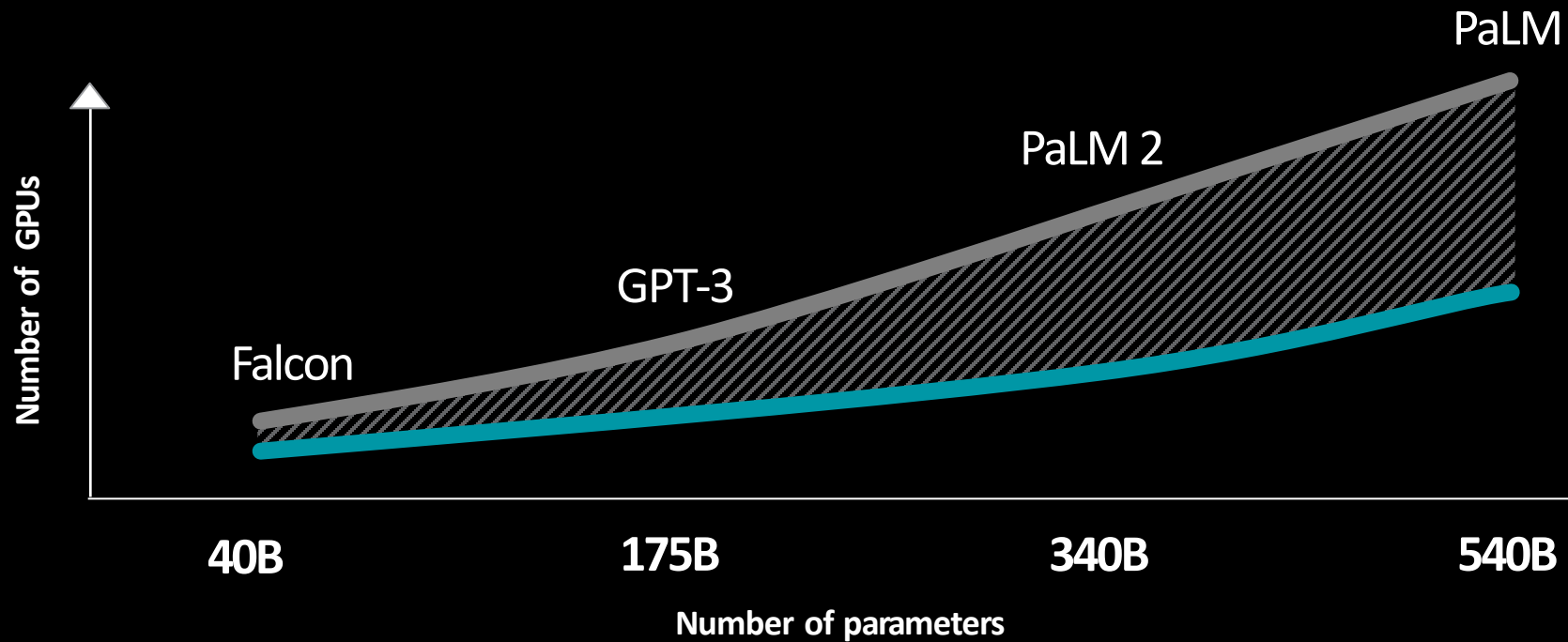
Up to

# 1.6x

HBM bandwidth  
compared to Nvidia H100

AMD Instinct™ MI300X

# Inference advantage



Competition 80GB | AMD Instinct™ MI300X 192GB



# CODE CONVERSION TOOLS

EXTEND YOUR APPLICATION PLATFORM SUPPORT BY CONVERTING CUDA<sup>®</sup> CODE

SINGLE SOURCE

MAINTAIN PORTABILITY

MAINTAIN PERFORMANCE

## Hipify-perl

- ▲ Easy to use; point at a directory and it will hipify CUDA code
- ▲ Very simple string replacement technique; may require manual post-processing
- ▲ Recommended for quick scans of projects

## Hipify-clang

- ▲ More robust translation of the code
- ▲ Generates warnings and assistance for additional analysis
- ▲ High quality translation, particularly for cases where the user is familiar with the make system

# TRANSITIONING CUDA® WORKLOADS

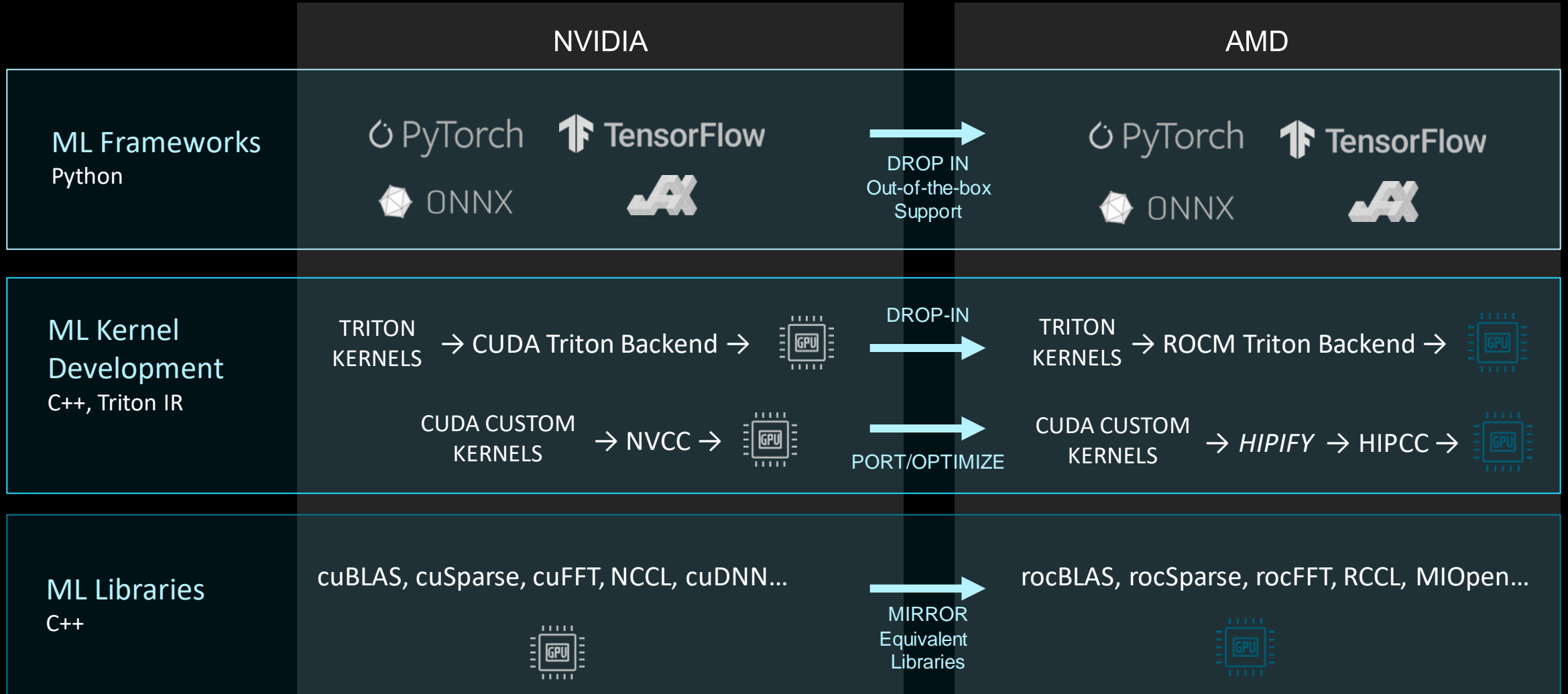
## FEATURE PARITY WITH COMMONLY USED MATH AND COMMUNICATION LIBRARIES

AMD DEVELOPS EQUIVALENT OPTIMIZED LIBRARIES  
TO MAXIMIZE PERFORMANCE OF COMMONLY USED HPC AND AI FUNCTIONS

	NVIDIA	AMD
Math Libraries	cuBLAS	rocBLAS
	cuBLASLt	hipBLASLt
	cuFFT	rocFFT
	cuSOLVER	rocSOLVER
	cuSPARSE	rocSPARSE
	cuRAND	rocRAND
Communication Library	NCCL	RCCL
C++ Core library	Thrust	rocThrust
	CUB	hipCUB
Wave Matrix Multiply Accumulate Library	WMMA	rocWMMA
Deep Learning / Machine Learning primitives	cuDNN	MIOpen
C++ templates abstraction for GEMMs	CUTLASS	Composable Kernel

# TRANSITIONING WORKLOADS TO INSTINCT GPUS

## LOW FRICTION SOFTWARE PORTING FOR EXISTING NVIDIA USERS TO AMD





# Powering datacenter AI at scale



#1 Frontier

National Cancer Institute  
and DOE accelerating  
cancer research  
and treatment



#3 LUMI

Largest Finnish language  
model (TurkuNLP-13B)

**A12 OLMo**

Allen Institute scientific LLM



#11 Explorer

WUS3 running  
AI and HPC workloads



1st Korean LLM

T5 NLP with  
11B parameters

# RESOURCES TO HELP IN THE JOURNEY

## Collateral



---

Product Brochures  
Whitepapers  
Qualified Server  
Catalog  
Case Studies  
Product Videos

[AMD.com/Instinct](https://www.amd.com/Instinct)

## AMD ROCm™ Info Portal



---

AMD ROCm Releases  
Tech Docs & FAQs  
Compilers, Libraries &  
Tools  
AMD ROCm Learning Ctr  
AMD Meet The Expert  
(MTE)

[DOCS.AMD.com](https://docs.amd.com)

## Software Containers & Install Guides



---

AMD ROCm  
AMD Infinity Hub  
Application Catalogue  
ML Frameworks

[Instinct Benchmarks](https://www.amd.com/Instinct/Benchmarks)  
[AMD.com/ROCm](https://www.amd.com/ROCm)  
[AMD.com/InfinityHub](https://www.amd.com/InfinityHub)

## Community & News



---

AMD ROCm  
Community  
AMD Instinct™ Blogs  
AMD ROCm Blogs  
AMD Instinct GPU  
Newsletter

[Community.AMD.com](https://community.amd.com)  
[AMD Lab Notes](https://www.amd.com/LabNotes)

## Funds & Initiatives



---

AMD AIER 2.0  
Initiative  
AMD HPC Fund

[AMD.com/AIER](https://www.amd.com/AIER)  
[AMD.com/HPC-FUND](https://www.amd.com/HPC-FUND)



# Unlocking the Power of AI

Dedicated AI Hardware on Endpoint Systems Enable A New Era For PCs



## Explosion of Generative AI

Bing > 100M users since AI launch

Microsoft use cases expanding rapidly

*This is just the  
Beginning !!*



## From Cloud to Hybrid

Cloud inferencing is costly for Gen-AI

ISVs need Client AI for Gen-AI features

*Client AI makes Gen-AI features  
affordable*



## Enhanced Experiences

Client = Personalized, Fast

Reliable and Protected

*“Generate insights from PPT that  
John forwarded me last week”*

# THE FUTURE STARTS NOW

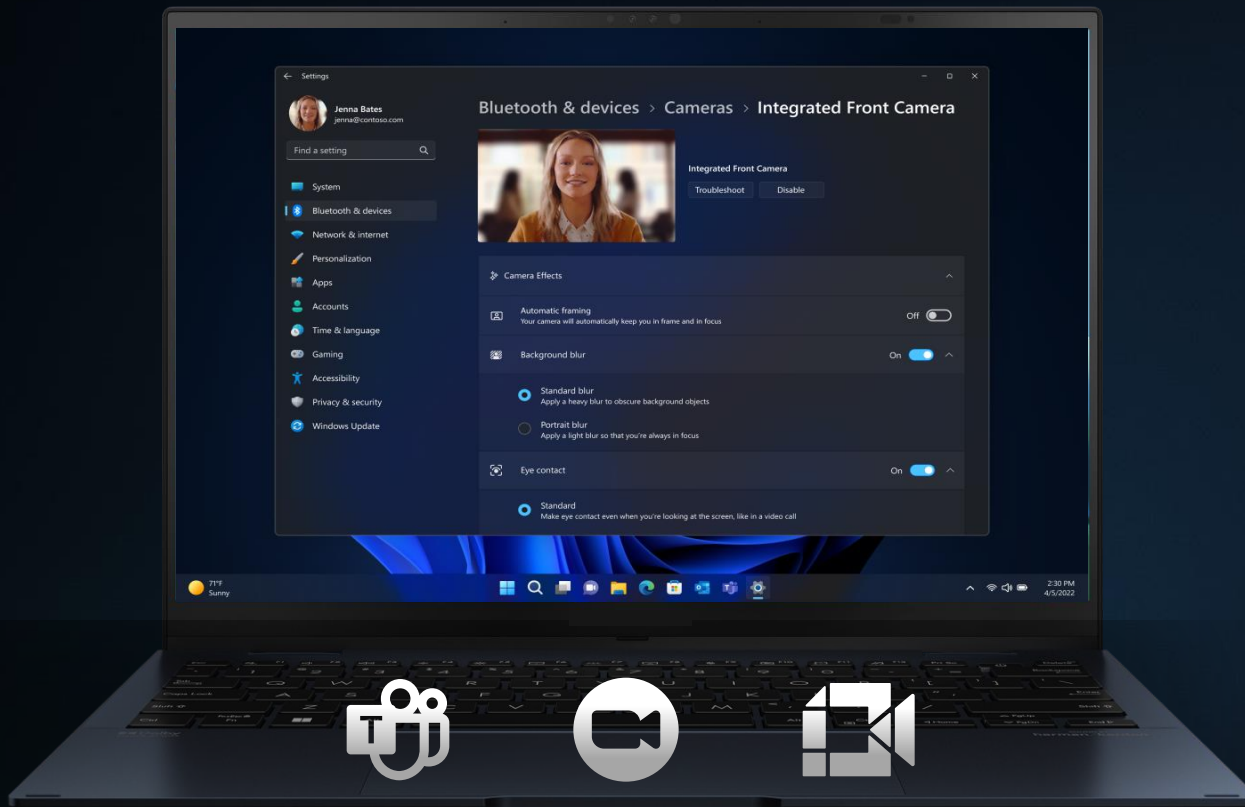
FOR WINDOWS LAPTOPS  
WITH AI TECHNOLOGY BUILT IN

Only with **AMD**  
**RYZEN AI**



# THE JOURNEY STARTS WITH ADVANCED VIDEO COLLABORATION

Windows Studio Effects uses AMD Ryzen™ AI using the integrated camera



**ENHANCED BACKGROUND BLURS**  
to help limit distractions while  
you are on a video call

**AUTO FRAMING**  
so the camera follows you while  
you are multitasking on a video call

**EYE GAZE CORRECTION**  
so that your audience knows  
you are focused on them



# GET READY TO EXPLORE A NEW WORLD OF POSSIBILITIES WITH FUTURE WINDOWS APPLICATIONS



## ACCELERATE DATA ANALYTICS

Work efficiently on data analysis, regression, and predictive modeling with your local data



## HAVE A PERSONAL AI ASSISTANT

Get help building a presentation, writing email responses, managing your budget, and more



## EXCEL IN COMPUTER VISION

Accelerate computation for object detection, image classification, and facial recognition



## CREATE LIKE NEVER BEFORE

Deliver a responsive experience to enable creators to dream bigger

**AMD** 



## Endnotes

- SP5-011E: SPECpower\_ssj®2008 comparison based on published 2P server results as of 6/13/2023. Configurations: 2P AMD EPYC 9654 (30,602 overall ssj\_ops/W, 2U, [https://spec.org/power\\_ssj2008/results/res2022q4/power\\_ssj2008-20221204-01204.html](https://spec.org/power_ssj2008/results/res2022q4/power_ssj2008-20221204-01204.html)) is 1.81x the performance of best published 2P Intel Xeon Platinum 8490H (16,902 overall ssj\_ops/W, 2U, [https://spec.org/power\\_ssj2008/results/res2023q2/power\\_ssj2008-20230507-01251.html](https://spec.org/power_ssj2008/results/res2023q2/power_ssj2008-20230507-01251.html)). SPEC® and SPECpower\_ssj® are registered trademarks of the Standard Performance Evaluation Corporation. See [www.spec.org](http://www.spec.org) for more information.
- SP5-013D: SPECrate®2017\_int\_base comparison based on published scores from [www.spec.org](http://www.spec.org) as of 06/2/2023. Comparison of published 2P AMD EPYC 9654 (1800 SPECrate®2017\_int\_base, 720 Total TDP W, \$23,610 total 1Ku, 192 Total Cores, 2,500 Perf/W, 0.076 Perf/CPU\$, <http://spec.org/cpu2017/results/res2023q2/cpu2017-20230424-36017.html>) is 1.80x the performance of published 2P Intel Xeon Platinum 8490H (1000 SPECrate®2017\_int\_base, 700 Total TDP W, \$34,000 total 1Ku, 120 Total Cores, 1,429 Perf/W, 0.029 Perf/CPU\$, <http://spec.org/cpu2017/results/res2023q1/cpu2017-20230310-34562.html>) [at 1.75x the performance/W] [at 2.59x the performance/CPU\$]. Published 2P AMD EPYC 7763 (861 SPECrate®2017\_int\_base, 560 Total TDP W, \$15,780 total 1Ku, 128 Total Cores, 1,538 Perf/W, 0.055 Perf/CPU\$, <http://spec.org/cpu2017/results/res2021q4/cpu2017-20211121-30148.html>) is shown for reference at 0.86x the performance [at 1.08x the performance/W] [at 1.86x the performance/CPU\$]. AMD 1Ku pricing and Intel [ark.intel.com](http://ark.intel.com) specifications and pricing as of 6/1/23. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See [www.spec.org](http://www.spec.org) for more information.
- SP5-049C: VMmark® 3.1.1 matched pair comparison based on published results as of 6/13/2023. Configurations: 2-node, 2P 96-core EPYC 9654 powered server running VMware ESXi 8.0b (40.66 @ 42 tiles/798 VMs, <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vmmark/2023-06-13-Lenovo-ThinkSystem-SR665V3.pdf>) versus 2-node, 2P 60-core Xeon Platinum 8490H running VMware ESXi 8.0 GA (23.38 @ 23 tiles/437 VMs, <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vmmark/2023-03-21-Fujitsu-PRIMERGY-RX2540M7.pdf>) for 1.74x the score and 1.75x the tile (VM) capacity. 2-node, 2P EPYC 7763-powered server (23.33 @ 24 tiles/456 VMs, <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/vmmark/2022-02-08-Fujitsu-RX2450M1.pdf>) shown at 0.98x performance for reference. VMmark is a registered trademark of VMware in the US or other countries.
- SP5-051: TPCx-AI SF3 derivative workload comparison based on AMD internal testing running multiple VM instances as of 6/13/2023. The aggregate end-to-end AI throughput test is derived from the TPCx-AI benchmark and as such is not comparable to published TPCx-AI results, as the end-to-end AI throughput test results do not comply with the TPCx-AI Specification. Configurations: 2 x AMD EPYC 9754 on Titanite (BIOS and Settings: AMI Core Ver. 5.25, Project Ver. RTI1000F and Default BIOS settings (SMT=on, Determinism=Auto, NPS=1)), 1.5TB (24) Dual-Rank DDR5-4800 64GB DIMMs, 1DPC, SK Hynix SHGP31-500GM 500GB NVMe, Ubuntu® 22.04 LTS (8-instances, 30 vCPUs/instance, 1841 AI test cases/min); 2 x AMD EPYC 9654 on Titanite (BIOS and Settings: AMI Core Ver. 5.25, Project Ver. RTI1000F and Default BIOS settings (SMT=on, Determinism=Auto, NPS=1)), 1.5TB (24) Dual-Rank DDR5-4800 64GB DIMMs, 1DPC, Samsung SSD 983 DCT 960GB, Ubuntu 22.04.1 LTS (6-instance, 28 vCPUs/instance, 1554 AI test cases/min); 2 x Intel(R) Xeon(R) Platinum 8490H on Dell PowerEdge R760 (BIOS and Settings: ESE110Q-1.10 and Package C1E, Default BIOS settings (C State=Disabled, Hyper-Threading, Turbo boost= enabled (ALL)=Enabled, SNC (Sub NUMA)=Disabled)), 2TB (32) Dual-Rank DDR5-4800 64GB DIMMs, 1DPC, Dell 1.7TB NVMe, Ubuntu 22.04.2 LTS (4-instance, 30 vCPUs/instance, 831 AI test cases/min). Results may vary due to factors including system configurations, software versions and BIOS settings. TPC Benchmark is a trademark of the TPC.
- SP5-056B: SAP® SD 2-tier comparison based on published results as of 6/13/2023. Configurations: 2P 96-core EPYC 9654 powered server (148,000 benchmark users, <https://www.sap.com/dmc/benchmark/2022/Cert22029.pdf>) versus 2P 60-core Xeon Platinum 8490H (77,105 benchmark users, <https://www.sap.com/dmc/benchmark/2023/Cert23021.pdf>) for 1.92x the number of SAP SD benchmark users. 2P EPYC 7763 powered server (75,000 benchmark users, <https://www.sap.com/dmc/benchmark/2021/Cert21021.pdf>) shown at 0.98x the performance for reference. For more details see <http://www.sap.com/benchmark>. SAP and SAP logo are the trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and in several other countries.
- SP5-104A: SPECjbb® 2015-MultiJVM Critical based on published scores from [www.spec.org](http://www.spec.org) as of 3/31/2023. Configurations: 2P AMD EPYC 9654 (664,375 SPECjbb®2015 MultiJVM max-jOPS, 622,315 SPECjbb®2015 MultiJVM critical-jOPS, 192 Total Cores, <https://www.spec.org/jbb2015/results/res2022q4/jbb2015-20221019-00860.html>) is 1.69x the critical-jOPS performance of published 2P Intel Xeon Platinum 8490H (458,295 SPECjbb®2015 MultiJVM max-jOPS, 368,979 SPECjbb®2015 MultiJVM critical-jOPS, 120 Total Cores, <http://www.spec.org/jbb2015/results/res2023q1/jbb2015-20230119-01007.html>).
- SP5-149: Container density throughput based on sustaining ~25k e-commerce Java Ops/sec/container until exceeding SLA utilizing >90% of the total cores on composite server-side Java workload as measured by AMD as of 6/13/2023. Common container settings: allocated 40GB memory, similar disks & NICs. 2P server configurations: 2P EPYC 9754 128C/256T SMT ON, Memory: 1.5TB = 24 x 64 GB DDR5 4800, OS Ubuntu 22.04, NPS Setting: L3 as NUMA running 16 vCPUs vs. 2P Xeon Platinum 8490H 60C/120T HT ON, Memory: 2TB = 32 x 64 GB DDR5 4800, OS Ubuntu 22.04, NPS Setting: NPS 2 running 16 vCPUs vs. 2P Ampere Altra Max 128-30, Memory: 1TB = 16 x 64GB DDR3200, OS Ubuntu 22.04, NPS Setting: NPS 1 running 25C. Results may vary due to factors including system configurations, software versions and BIOS settings.
- MI300-005: Calculations conducted by AMD Performance Labs as of May 17, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.218 TFLOPS sustained peak memory bandwidth performance. MI300X memory bus interface is 8,192 and memory data rate is 5.6 Gbps for total sustained peak memory bandwidth of 5.218 TB/s (8,192 bits memory bus interface \* 5.6 Gbps memory data rate/8)\*0.91 delivered adjustment. The highest published results on the NVidia Hopper H100 (80GB) SXM GPU accelerator resulted in 80GB HBM3 memory capacity and 3.35 TB/s GPU memory bandwidth performance.

## Endnotes

- MI300-07K: Measurements by internal AMD Performance Labs as of June 2, 2023 on current specifications and/or internal engineering calculations. Large Language Model (LLM) run or calculated with FP16 precision to determine the minimum number of GPUs needed to run the Falcon (40B parameter) models. Tested result configurations: AMD Lab system consisting of 1x EPYC 9654 (96-core) CPU with 1x AMD Instinct™ MI300X (192GB HBM3, OAM Module) 750W accelerator tested at FP16 precision. Server manufacturers may vary configuration offerings yielding different results.

MI300-08K - Measurements by internal AMD Performance Labs as of June 2, 2023 on current specifications and/or internal engineering calculations. Large Language Model (LLM) run comparisons with FP16 precision to determine the minimum number of GPUs needed to run the Falcon (40B parameters); GPT-3 (175 Billion parameters), PaLM 2 (340 Billion parameters); PaLM (540 Billion parameters) models. Calculated estimates based on GPU-only memory size versus memory required by the model at defined parameters plus 10% overhead.

Calculations rely on published and sometimes preliminary model memory sizes. Tested result configurations: AMD Lab system consisting of 1x EPYC 9654 (96-core) CPU with 1x AMD Instinct™ MI300X (192GB HBM3, OAM Module) 750W accelerator Vs. Competitive testing done on Cirrascale Cloud Services comparable instance with permission.

Results (FP16 precision):

Model:	Parameters	Tot Mem. Reqd	MI300X Reqd	Competition Reqd
--------	------------	---------------	-------------	------------------

Falcon-40B	40 Billion	88 GB	1 Actual	2 Actual
GPT-3	175 Billion	385 GB	3 Calculated	5 Calculated
PaLM 2	340 Billion	748 GB	4 Calculated	10 Calculated
PaLM	540 Billion	1188 GB	7 Calculated	15 Calculated

Calculated estimates may vary based on final model size; actual and estimates may vary due to actual overhead required and using system memory beyond that of the GPU. Server manufacturers may vary configuration offerings yielding different results.