AMD Data Center Portfolio

Michal Sztemon

together we advance_

AGENDA

- Moderní datacentra a přehled produktů AMD
- Procesory AMD EPYC™ 4. generace – to nejlepší pro general computing
  - AMD EPYC™ 97x4 – cloud
  - AMD EPYC™ s 3D V-Cache – nej pro technické výpočty
- AMD AI a grafické akcelerátory AMD Instinct™
- Dotazy, soutěž

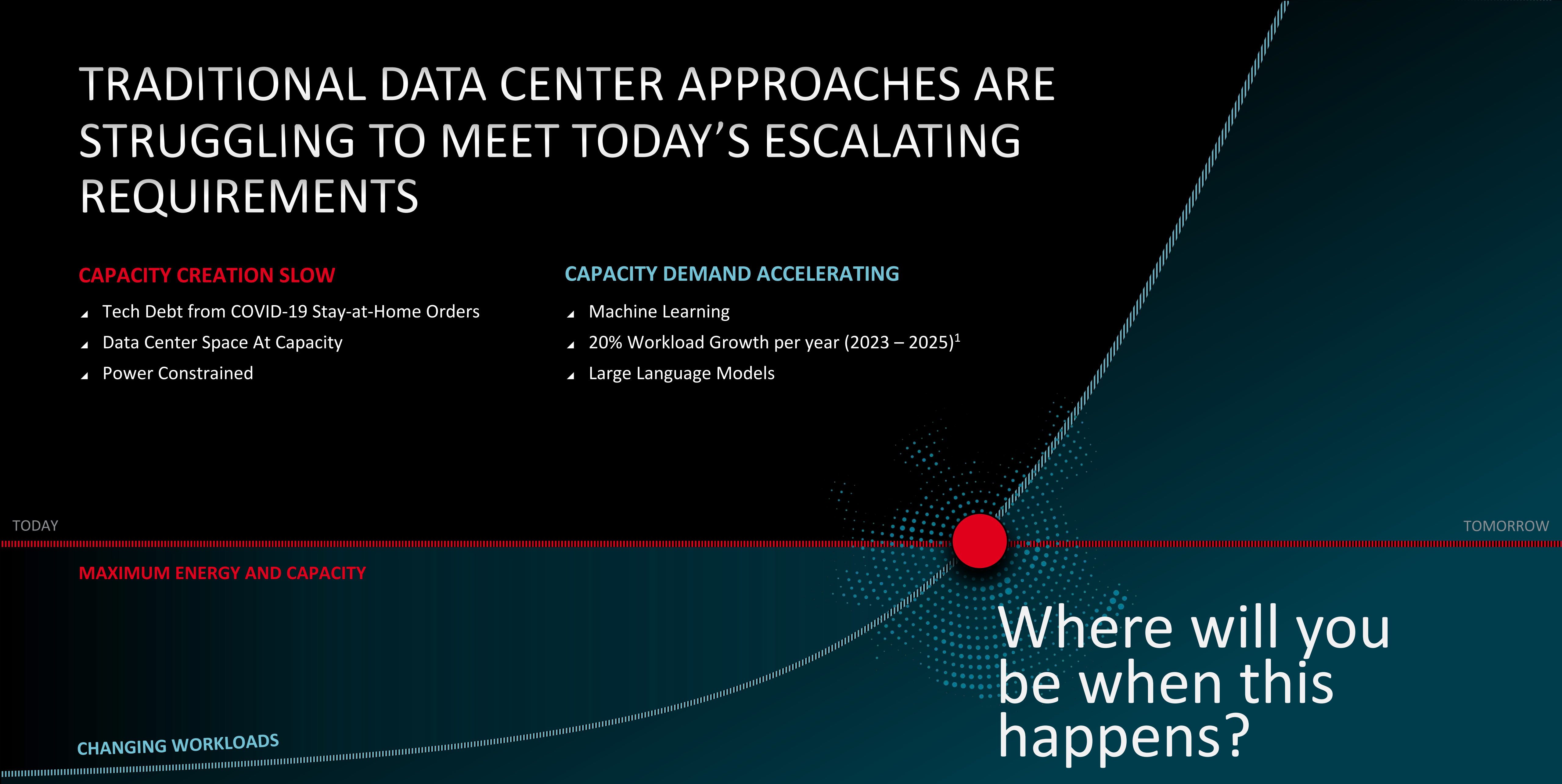
TRADITIONAL DATA CENTER APPROACHES ARE STRUGGLING TO MEET TODAY’S ESCALATING REQUIREMENTS

CAPACITY CREATION SLOW

- Tech Debt from COVID-19 Stay-at-Home Orders
- Data Center Space At Capacity
- Power Constrained

CAPACITY DEMAND ACCELERATING

- Machine Learning
- 20% Workload Growth per year (2023 – 2025)[1]
- Large Language Models



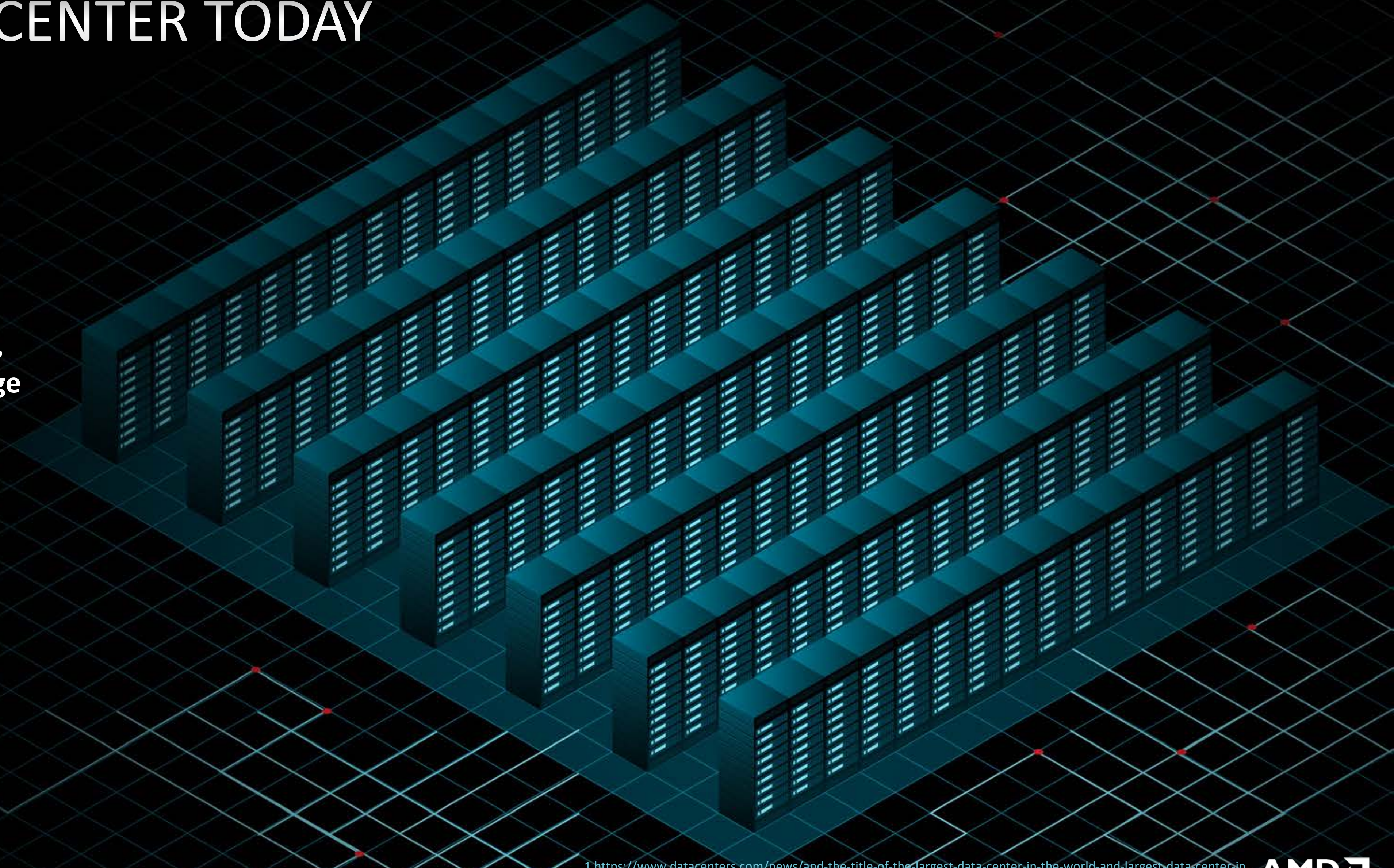
Where will you be when this happens?

[1] https://www.datacenterknowledge.com/industry-perspectives/how-ai-and-machine-learning-are-ready-change-game-data-center-operations

THE SPACE FOR YOUR IT INNOVATION IS IN YOUR DATA CENTER TODAY

The average data center size worldwide is 100,000 square feet.[1]

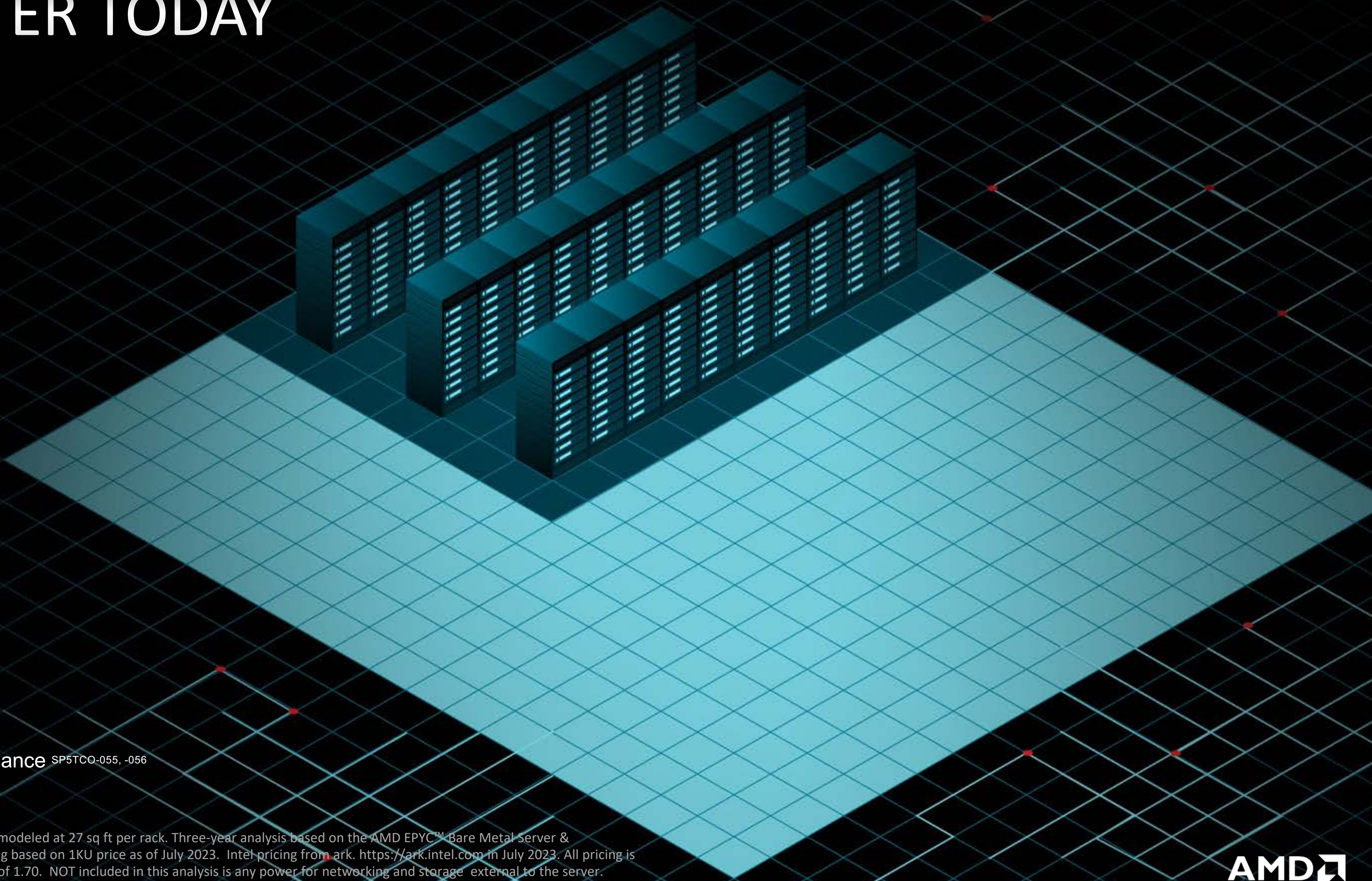
Much of it is dedicated to old, inefficient and hard-to-manage equipment[2]



[1] https://www.datacenters.com/news/and-the-title-of-the-largest-data-center-in-the-world-and-largest-data-center-in

[2] Analysis based on AMD internal data.

THE SPACE FOR YOUR IT INNOVATION IS IN YOUR DATA CENTER TODAY



INTEL® XEON® 6143
SKY LAKE CPU

-vs-

4th Gen AMD
EPYC™ 9334 CPU

73% Fewer Servers

70% Fewer Racks

65% Less Power

INTEL® XEON® 6242
CASCADE LAKE CPU

-vs-

4th Gen AMD
EPYC™ 9334 CPU

68% Fewer Servers

65% Fewer Racks

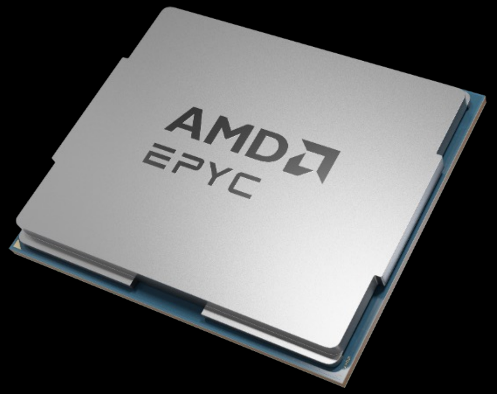
56% Less Power

Space & Power Comparisons Target: 80,000 Integer Performance SP5TCO-055, -056

Integer scores are the highest posted on SPEC.org for each server as of 06/01/2023 Rack space modeled at 27 sq ft per rack. Three-year analysis based on the AMD EPYC™ Bare Metal Server & Greenhouse Gas Emission TCO Estimation Tool - version 9.37 Pro Refresh. AMD processor pricing based on 1KU price as of July 2023. Intel pricing from ark. https://ark.intel.com in July 2023. All pricing is in USD. Servers / Rack limited by 42 RU and 10 kW. Cost per kW power $0.128/kWh; and PUE of 1.70. NOT included in this analysis is any power for networking and storage external to the server.
See Endnotes

[Public]

MODERN DATA CENTERS NEED
WORKLOAD-OPTIMIZED ENGINES



Server CPUs

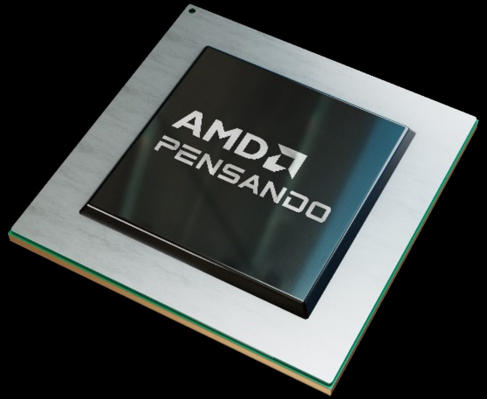
AMD EPYC

AI Accelerators

AMD INSTINCT AMD ALVEO

FPGAs and Adaptive SoCs

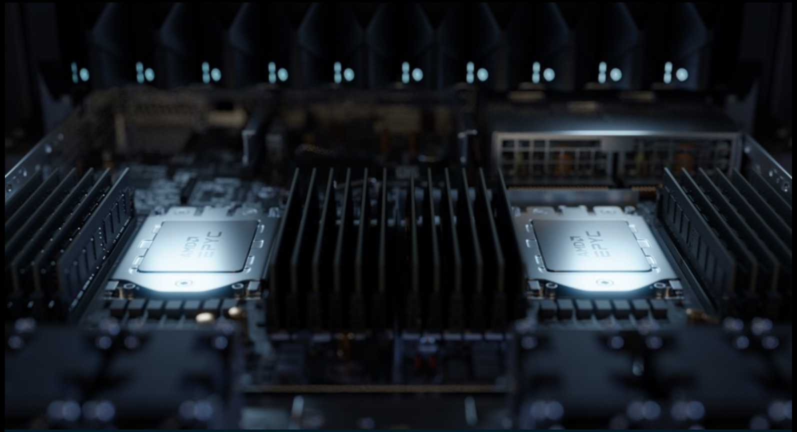
AMD ALVEO AMD VERSAL



SmartNICs and DPUs

AMD ALVEO AMD PENSANDO

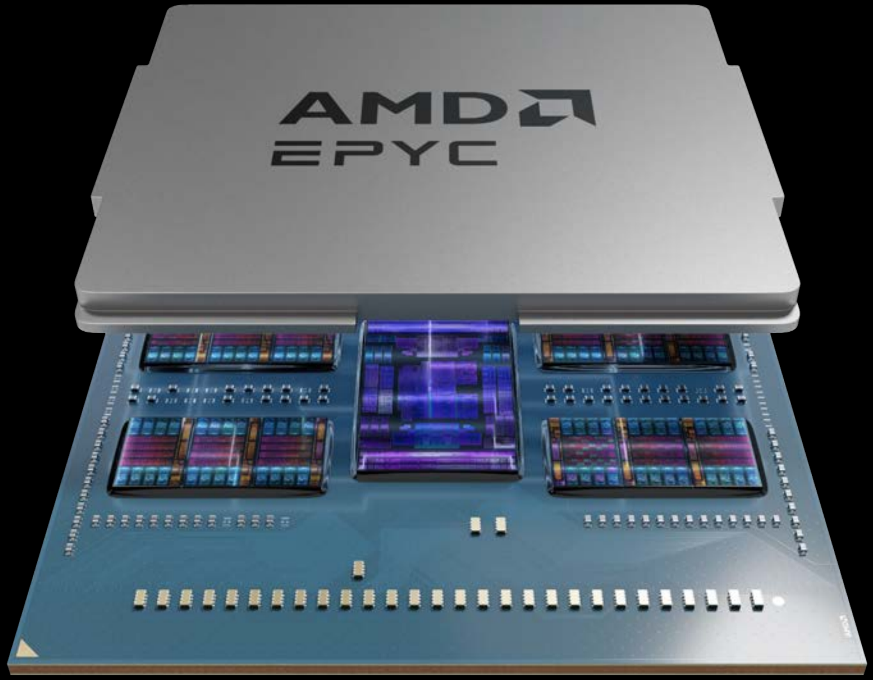
AMD SERVER STRATEGY



Highest performing general purpose data center CPU in the world

Optimized silicon for diverse workloads

Full stack solutions, ecosystem scale & partnerships to accelerate time-to-value



4TH GEN AMD EPYC™ CPU

The world’s best data center CPU

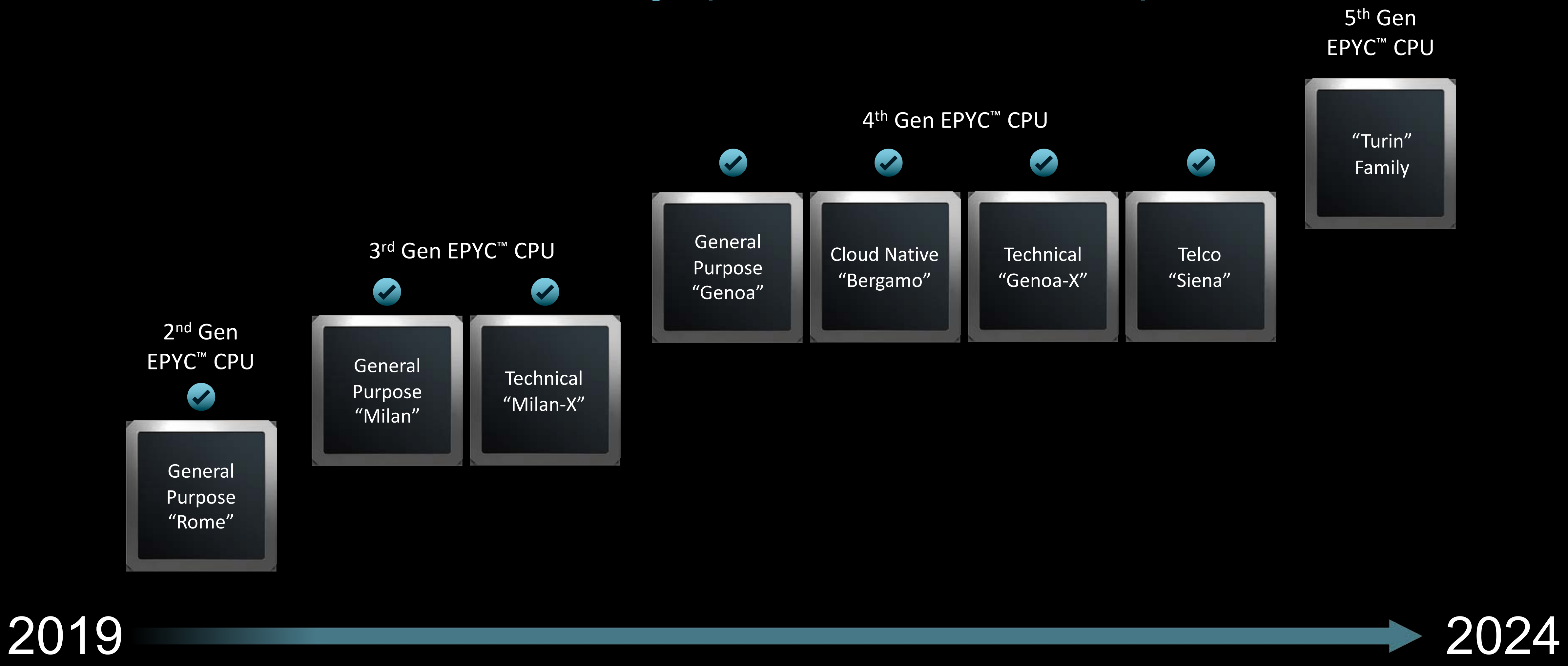
World’s Fastest Data Center Processor | Transformative Energy Efficiency | Leadership TCO Across Workloads and Industries | Robust Security Powering Confidential Computing | Rich Ecosystem of Solutions

See endnotes: SP5-143A; EPYC-028C

[Public]

AMD Data Center CPU Roadmap

Sustained High-performance Leadership



2nd Gen EPYC™ CPU

✓ General Purpose "Rome"

3rd Gen EPYC™ CPU

✓ General Purpose "Milan"

✓ Technical "Milan-X"

4th Gen EPYC™ CPU

✓ General Purpose "Genoa"

✓ Cloud Native "Bergamo"

✓ Technical "Genoa-X"

✓ Telco "Siena"

5th Gen EPYC™ CPU

"Turin" Family

2019 → 2024

Roadmaps subject to change.

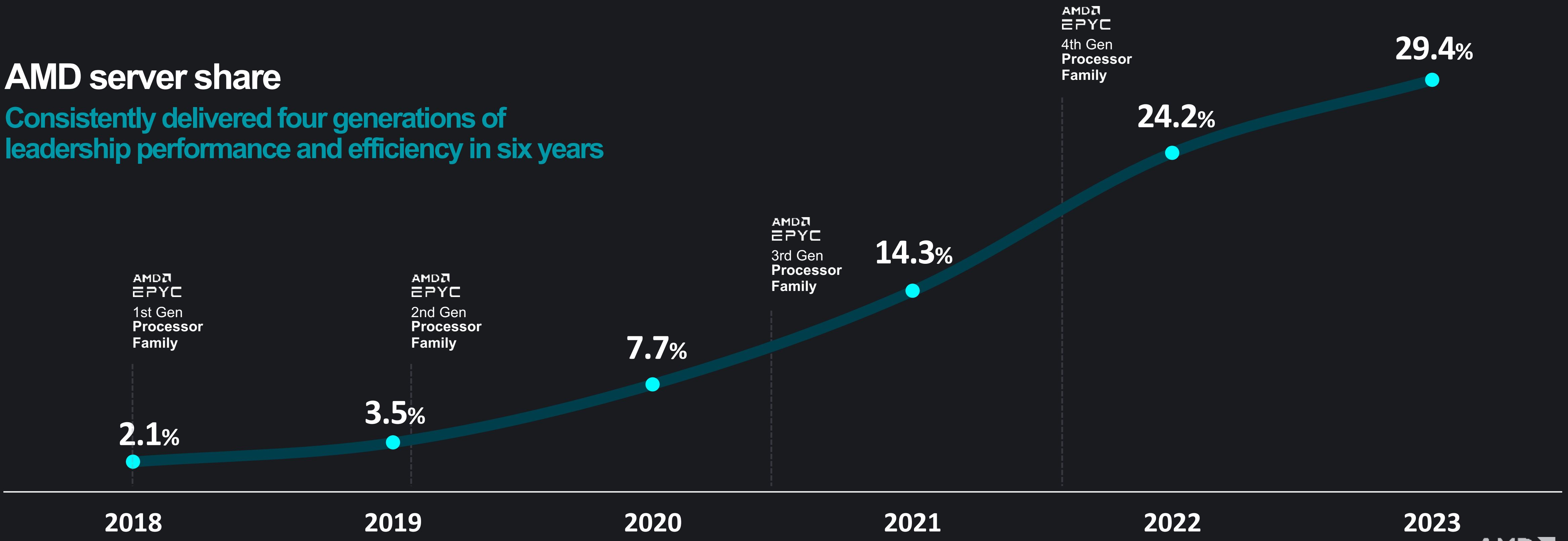
AMD EPYC™ Trusted to Power 30% of World's Servers

Broad range of platforms from all major OEMs

Hewlett Packard Enterprise | DELL | Lenovo | CISCO | SUPERMICRO

AMD server share

Consistently delivered four generations of leadership performance and efficiency in six years



Source: Mercury Research Sell-in Revenue Shipment Estimates, 2023 Q3

AMD together we advance_

4th Gen AMD EPYC™ CPU

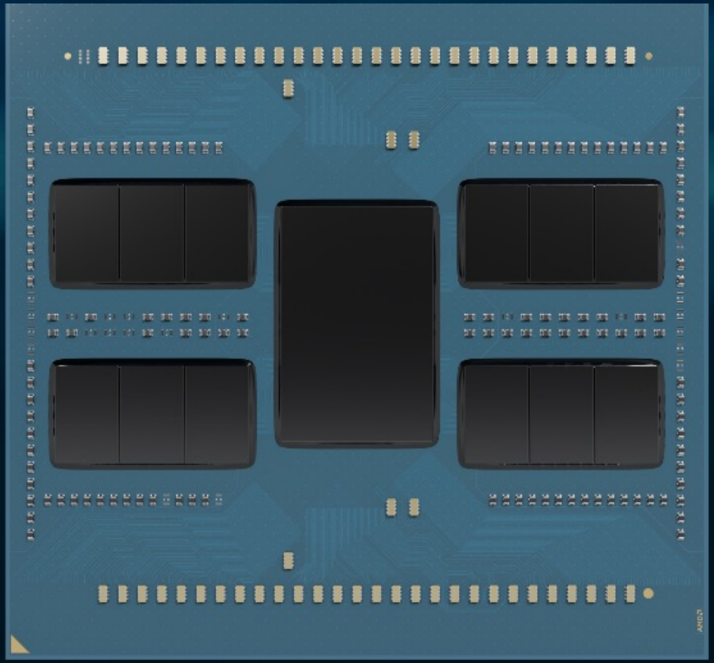
Extending Compute Leadership

- Leadership Socket and Per-Core Performance
  Up to 128 "Zen 4 & 4c" Cores in 5nm

- Leadership Memory Bandwidth and Capacity
  12 Channels DDR5

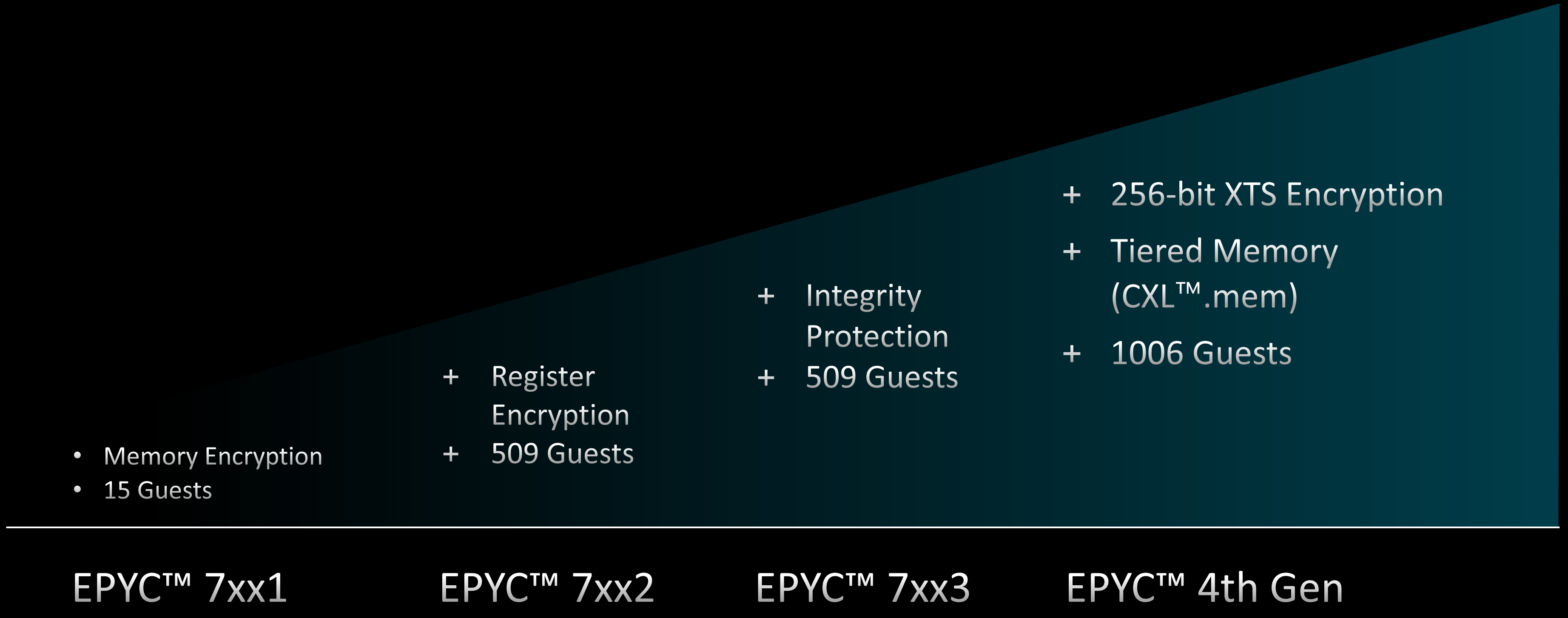
- Next Generation I/O
  Up to 160 Lanes of PCIe® Gen 5 (2P) | Memory Expansion with CXL™

- Advances in Confidential Computing
  ~2X SEV-SNP Guests* | Direct and CXL™ Attached Memory Encryption



See Endnotes EPYC-032A, EPYC-033, SP5-013B, SP5-014A.
* Based on generational increase in security keys.

Growing Ecosystem of Confidential Computing



EPYC™ 7xx1
- Memory Encryption
- 15 Guests

EPYC™ 7xx2
+ Register Encryption
+ 509 Guests

EPYC™ 7xx3
+ Integrity Protection
+ 509 Guests

EPYC™ 4th Gen
+ 256-bit XTS Encryption
+ Tiered Memory (CXL™.mem)
+ 1006 Guests

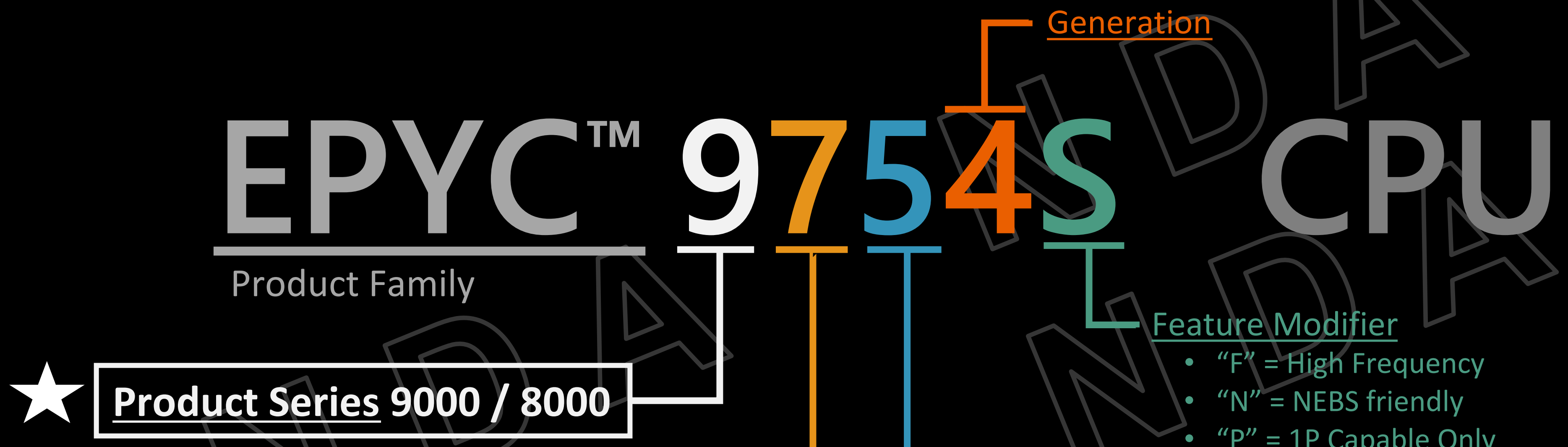
Data Encryption

- Data-In-Use
- Data-In-Flight
- Data-At-Rest

anjuna Azure Confidential Computing CANONICAL CONFIDENTIAL COMPUTING CONSORTIUM Confidential VMs IBM kata containers ORACLE Red Hat SUSE VMware Tanzu

GD-83: Use of third-party marks / logos/ products is for informational purposes only and no endorsement of or by AMD is intended or implied.

AMD EPYC™ 9004 / 8004 Series - Processor Naming Convention



Feature Modifier

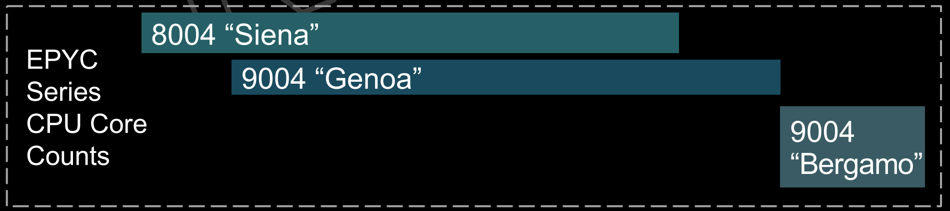
- “F” = High Frequency
- “N” = NEBS friendly
- “P” = 1P Capable Only
- “S” = Single Threaded
- “X” = 3D V-Cache™

Core Count

- Indicates Core Count within the Series

Heart of Market core numbering

100s Digit	0	1	2	3	4	5	6	7
Cores	8	16	24	32	48	64	84-96	112-128



Performance

10s digit – Perf w/in Core Count

- 9 = reserved
- 8, 7, 6, 5, 4, 3, 2, 1
  - Relative Performance w/in core count
  - Higher number = higher perf

AMD EPYC™ 9004 Series Processor

All-in Feature Set support

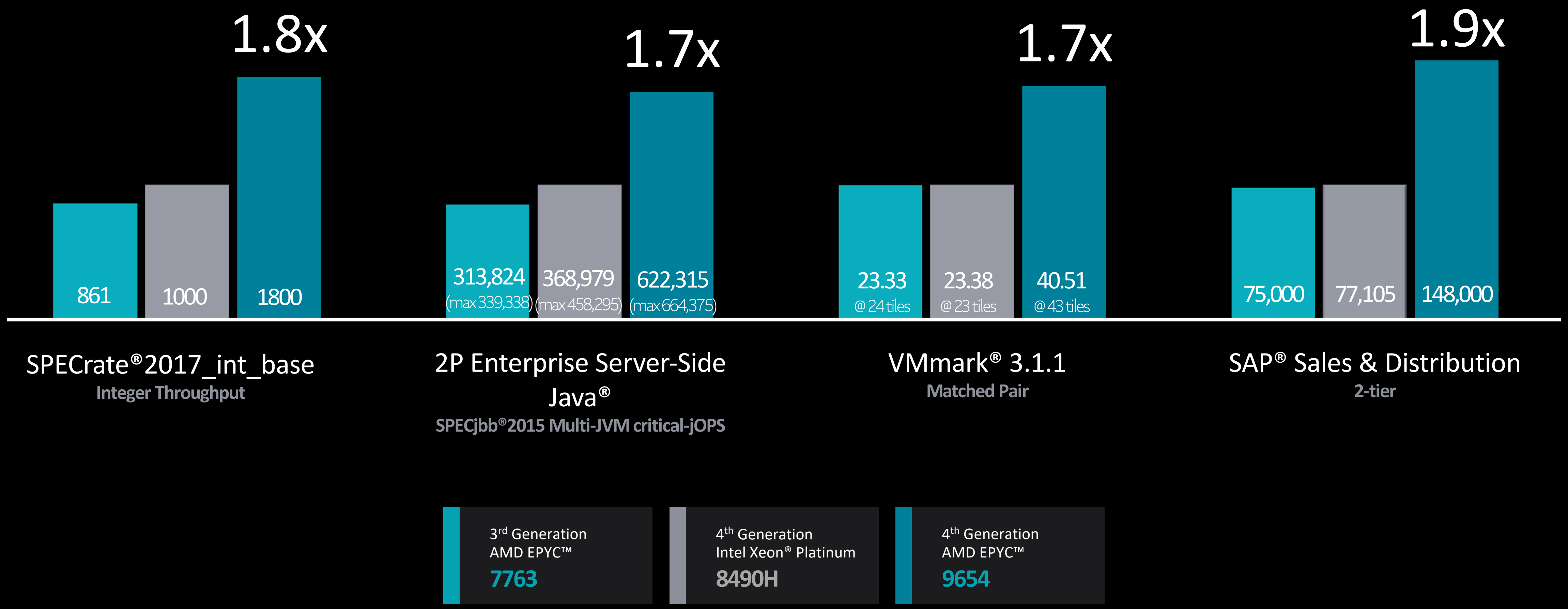
- 12 Channels of DDR5-4800
- Up to 6TB DDR5 memory capacity
- 128 lanes PCIe® 5
- 64 lanes CXL™ 1.1+
- AVX-512 ISA, SMT & core frequency boost
- AMD Infinity Fabric™
- AMD Infinity Guard[2]

Cores	AMD EPYC	Base/Boost[1] (up to GHz)	Default TDP (w)	cTDP (w)
96 cores	9654/P	2.40/3.70	360w	320-400w
84 cores	9634	2.25/3.70	290w	240-300w
64 cores	9554/P	3.10/3.75	360w	320-400w
64 cores	9534	2.45/3.70	280w	240-300w
48 cores	→ 9474F	3.60/4.10	360w	320-400w
	9454/P	2.75/3.80	290w	240-300w
32 cores	→ 9374F	3.85/4.30	320w	320-400w
32 cores	9354/P	3.25/3.80	280w	240-300w
32 cores	9334	2.70/3.90	210w	200-240w
24 cores	→ 9274F	4.05/4.30	320w	320-400w
	9254	2.90/4.15	200w	200-240w
	9224	2.50/3.70	200w	200-240w
16 cores	→ 9174F	4.10/4.40	320w	320-400w
	9124	3.00/3.70	200w	200-240w

[1] See Endnote EPYC-18. Max boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems. EPYC-018

[2] AMD Infinity Guard features vary by EPYC Processor generations. Infinity Guard security features on AMD EPYC processors must be enabled by server OEMs and/or cloud service providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard. GD-183.

4th Gen EPYC™ CPU - “Genoa”
PERFORMANCE LEADERSHIP



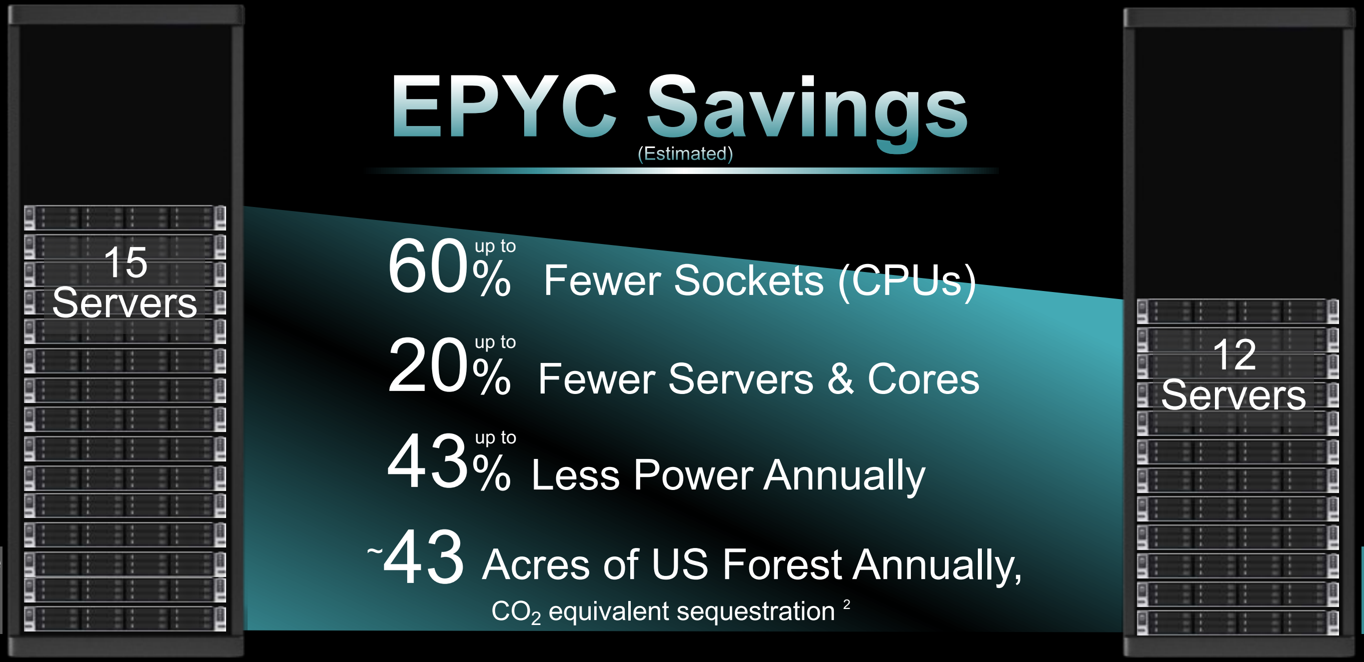
As of 6/13/2023, see SP5-104A, SP5-049C, SP5-056B

Fewer Servers, Less Power, Leading to Lower Emissions

7,500 SPECrate® 2017_int_base
64 Cores / Server – Head to Head Comparison

2P INTEL®
Platinum 8454H

1P AMD
EPYC™ 9554P



15 Servers

12 Servers

EPYC Savings
(Estimated)

60% up to Fewer Sockets (CPUs)

20% up to Fewer Servers & Cores

43% up to Less Power Annually

~43 Acres of US Forest Annually,
CO$_2$ equivalent sequestration [2]

Integer score
522
per server

960 Cores
~184k kWh per year

Integer score
631 est
per server

768 Cores
~105k kWh per year

PLUS AMD EPYC DELIVERS

up to 31% Lower Annual OPEX[1]

up to 36% Lower 3yr TCO[1]

SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

Analysis based on the AMD EPYC™ Bare Metal Server & Greenhouse Gas Emission TCO Estimation Tool - version 6.80.
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.
* Estimated AMD EPYC performance scores are based on AMD internal testing, Aug 2022 on AMD reference platforms.
[1] TCO time frame of 3-year and includes estimated costs for real estate, admin and power with power @ $0.16/kWh with 8kW / rack and a PUE of 1.7. Software cost as well as networking and storage power external to the server are not included in this analysis. [2] Values are for USA.

See endnote SP5TCO-029

[Public]

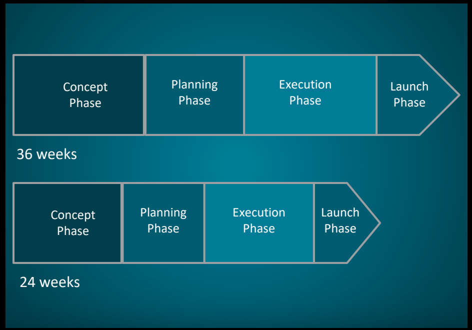
IMPROVE YOUR BUSINESS

with AMD EPYC™ CPU based Server solutions

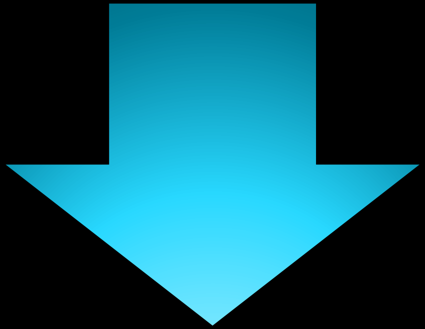
Increase Productivity

Reduce Power, Cooling, and Space

Free-up Resources



Shorten Development Time



Lower Costs

See endnote SP5TCO-045

Data Center Growth

Outstanding Momentum with AMD EPYC™ Processors

AWE CERN 中国电信 CHINA TELECOM FRONTIER
RESEARCH JÜLICH FORSCHUNGSZENTRUM Lawrence Livermore National Laboratory NERSC
Nikhef NSCC OAK RIDGE National Laboratory PAWSEY supercomputing centre
鹏城实验室通讯 Peng Cheng Laboratory Communications PURDUE UNIVERSITY SDSC SAN DIEGO SUPERCOMPUTER CENTER
UnitedHealthcare 之江实验室 ZHEJIANG LAB

HPC

Leading the Exascale Era
Consistently Winning Top
Deployments

Alibaba Cloud aws Microsoft Azure
BAIDU AI CLOUD EQUINIX METAL Google Cloud
IBM Cloud LinkedIn 美团 Meituan Meta
ORACLE CLOUD NETFLIX salesforce Tencent Cloud
UCLOUD优刻得 火山引擎

Cloud

Deployments
with Leading Providers

autox BASF DBS Dropbox
f5 Ford GE Goldman Sachs
INDUSTRIAL LIGHT & MAGIC MathWorks mastercard Micron
NIO PIXAR QUALCOMM salesforce
SAMSUNG WELLS FARGO weta DIGITAL

Enterprise

Large-scale Enterprise
Deployments

Use of third-party logos is for informational purposes only and no endorsement of or by AMD is intended or implied. GD-83

CUSTOMER SUCCESS STORIES

www.amd.com/stories

AMG PETRONAS FORMULA ONE TEAM

salesforce

browserless

casa systems

AIKON

CGG

UZ Universitair Ziekenhuis Brussel

i3D.NET PERFORMANCE HOSTING

universität uulm

tsmc

NOVOSERVE

CINESITE

logdna

PURDUE UNIVERSITY

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Blackboard

Durham University Institute for Computational Cosmology

ampleorganics

السوق المفتوح opensooq.com

market

HIVELOCITY DEDICATED SERVERS. IaaS. CLOUD.

wego

Vestas

GMO INTERNET GROUP

lytics

JELLYFISH PICTURES

keliweb [HOSTING DEDICATED TO YOU]

RIMAC

linode

DBS

BLUR

Bardel ENTERTAINMENT INC

TURTLE ROCK STUDIOS

Oregon State University

mml molecular modelling laboratory

Let's Encrypt

SYMMETRIC COMPUTING

OVHcloud

Cloudify.Asia

Nikhef

Scaleway

Washington University in St.Louis

SPINVFX

promarin

MEDICO POWERING & EMPOWERING

Dropbox

mml molecular modelling laboratory

COMPUTATION Lawrence Livermore National Laboratory

AXIS STUDIOS

VOGO

myLoc managed IT

JOST

UNIVERSITY OF NOTRE DAME

HUF HAUS Das Original · Seit 1912

HETZNER

THE MEDIA TEAM

UF UNIVERSITY of FLORIDA

fullstory

packet

NORTHERN DATA

ATEME Captivate your audience

Zdorovie

Richardson Electronics

AirShaper

THE UNIVERSITY OF AUCKLAND Te Whare Wānanga o Tāmaki Makaurau NEW ZEALAND

QTS

COMARCH

Supership

Qubit.

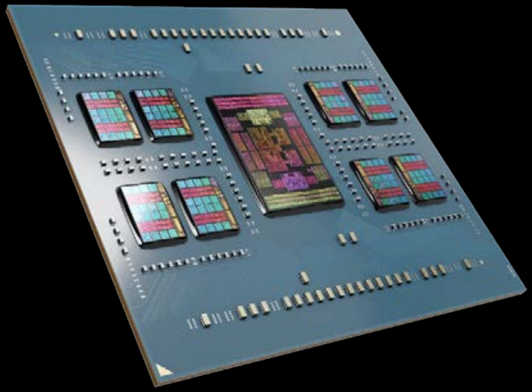
PIER Group Partners In Education & Research

CERN

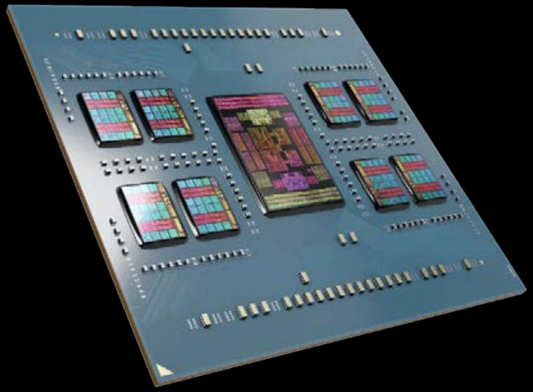
Fox VFX Lab

Use of third-party logos is for informational purposes only and no endorsement of or by AMD is intended or implied. GD-83

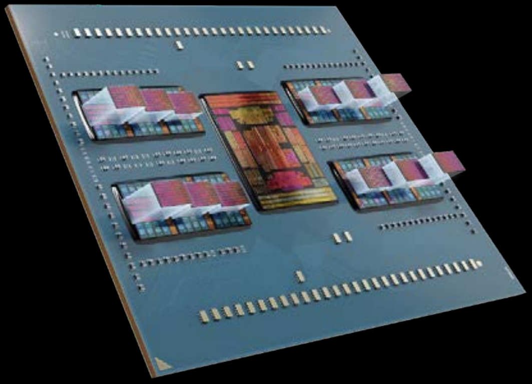
COMPUTING INFRASTRUCTURE
OPTIMIZED FOR DATA CENTER WORKLOADS



General Purpose Computing
AMD EPYC 9xx4



Cloud Native Computing
AMD EPYC 97x4



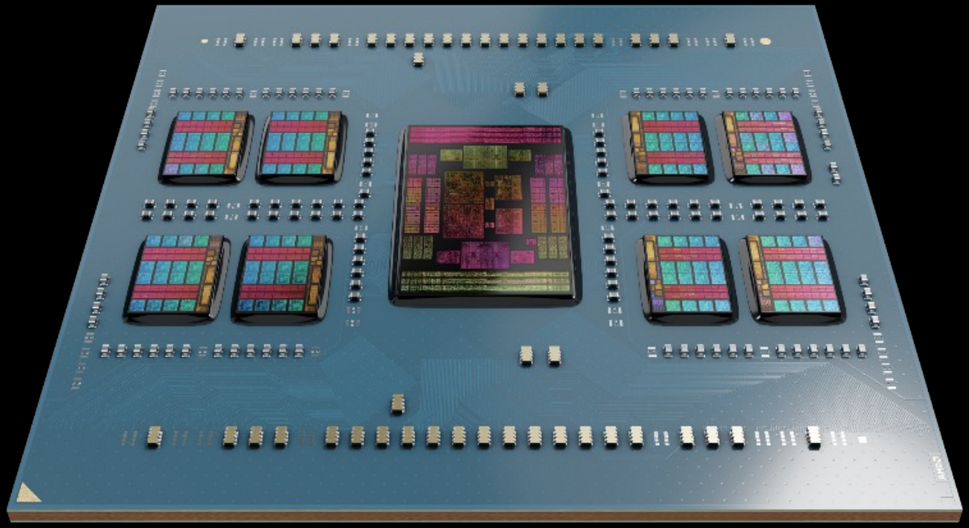
Technical Computing
AMD EPYC 9x84X

AMD EPYC

[Public]

4TH Gen AMD EPYC™ 97X4 CPU

Optimized for Cloud Native Workloads



Greatest vCPU Density

Leadership Cloud Performance

Best Energy Efficiency

Consistent x86 ISA

Up to 128 "Zen 4c" Cores

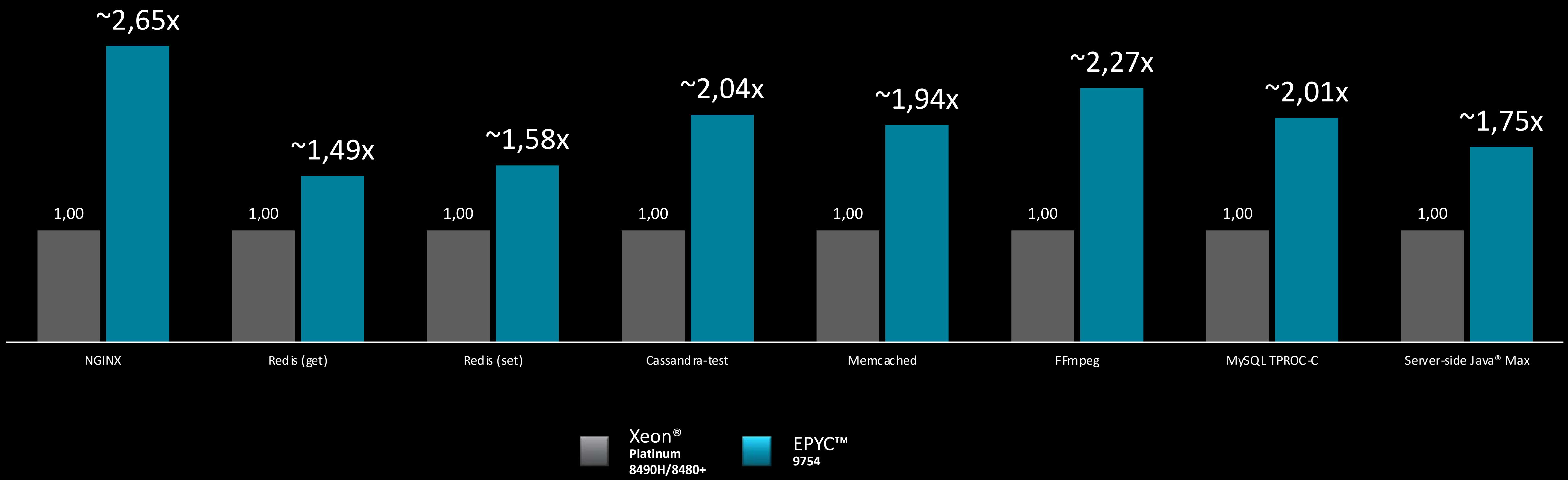
See Endnotes – EPYC-049 and https://www.amd.com/system/files/documents/amd-epyc-9004-pb-spec-power.pdf for Energy Efficiency

OPTIMIZED CLOUD NATIVE PERFORMANCE

Up to

2.6x throughput performance for a wide variety of cloud native workloads

vs Intel Xeon



2P servers: 128C AMD EPYC™ 9754 vs. 56C/60C Intel Xeon Platinum 8480+/8490H

Results may vary due to factors including system configurations, software versions and BIOS settings. As of 6/13/2023, see Cloud Native Workloads https://www.amd.com/system/files/documents/amd-epyc-9004-pb-cloud-native-workloads.pdf.

[Public]

MAXIMUM COMPUTE DENSITY

Reduce Power, Cooling, Space, Cost

2P INTEL®
PLATINUM 8490H, 60C

2P AMD
EPYC™ 9754, 128c

38
Servers
15
Servers

POWER[1]
594.0 Wh / Server
11.880kWh / Rack

POWER[1]
644.00 Wh / Server
11.592 kWh / Rack

AMD EPYC Delivers:

~23 FEWER SERVERS

~57% LESS POWER ANNUALLY[2]

~67% LOWER TCO

~79 US TONS LESS CO_2e ANNUALLY[2]

NGINX TARGET: Infrastructure delivering 375M Requests/Sec

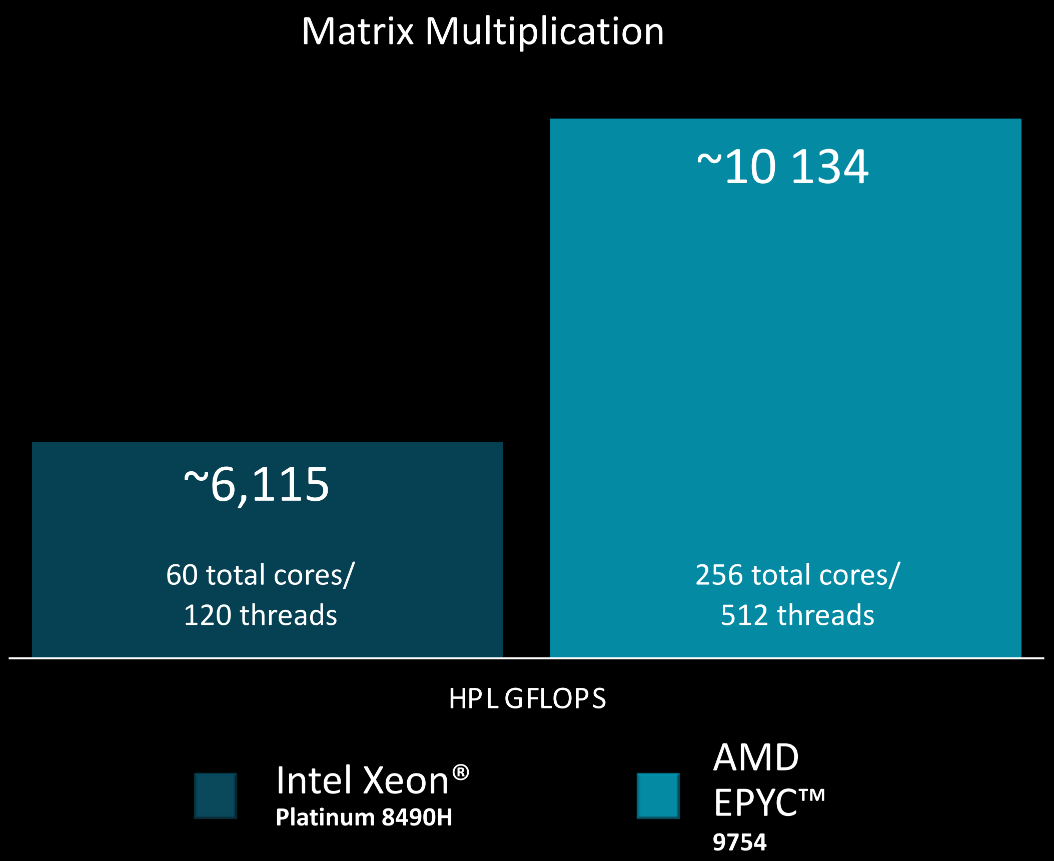
The power and Greenhouse Gas numbers above reflect a PUE of 1.70. Limit: 42RU rack & 12kW / Rack

All performance scores are estimates based on AMD internal testing in May & June 2023. AMD perf is on an AMD reference platform with a score of 26.248M requests / sec. Intel perf done on a Lenovo server with a score of 9,908,966. Ampere perf done on an Ampere Mt. Collins server with a score of 8.843M requests / sec. Analysis based on the AMD EPYC™ Bare Metal Server & Greenhouse Gas Emission TCO Estimation Tool - version 9.32 Pro. AMD processor pricing based on 1KU price as of April 2023. Intel pricing from ark. https://ark.intel.com in April 2023. Ampere C[1] Power and Server Cost only are included in this TCO. This is a power only OpEx and TCO analysis with a time frame of 3-year with power @ $0.128/kWh with 12kW / rack; and a PUE of 1.70. NOT included in this analysis are admin cost, real estate cost, software cost as well as power for any networking and storage external to the server. See endnote SP5TCO-050K, 051K PU data Phoronix.com May 2023. All pricing is in USD.

[Public]

SOLVING THE BIGGEST HPC PROBLEMS

DEMAND THE BEST COMPUTE PLATFORM TO SOLVE THE MOST CHALLENGING HPC PROBLEMS



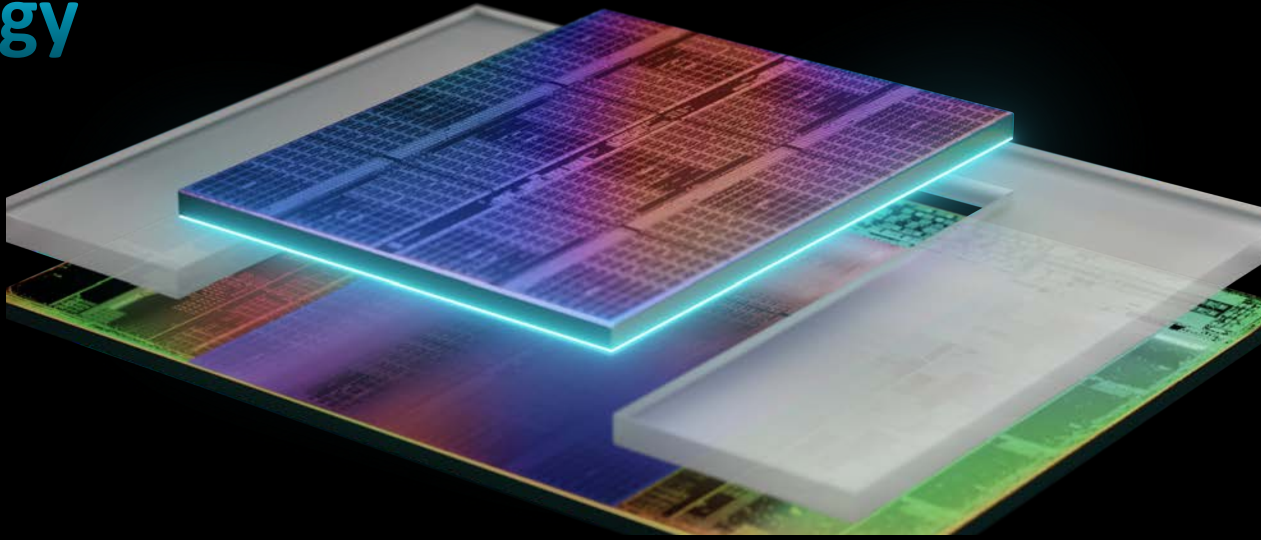
~**1.7x** more GFLOPS

vs. 60C Intel Xeon Platinum 8490H running the High Performance Linux (HPL) Benchmark

Results may vary due to factors including system configurations, software versions and BIOS settings. As of 6/13/2023, see SP5-154.

4TH GEN AMD EPYC

With AMD 3D V-Cache® Technology



ZEN 4 High Performance cores

Leadership 5nm Process Node

Up to 1.1 GB of L3 Cache

AMD Infinity Guard

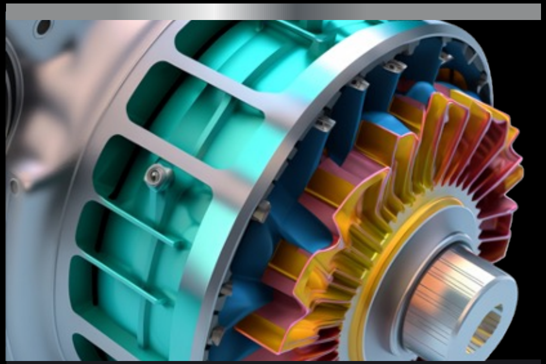
Rich Ecosystem of Solutions

World's highest performance x86 server CPU for technical computing

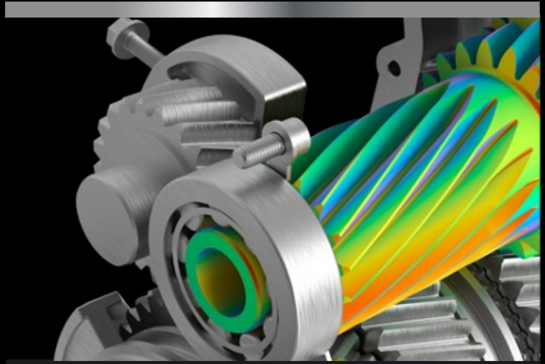
As of 6/13/2023, see SP5-165, GD-183, GD-204.

ENABLING BETTER PRODUCTS, FASTER

TECHNICAL COMPUTING

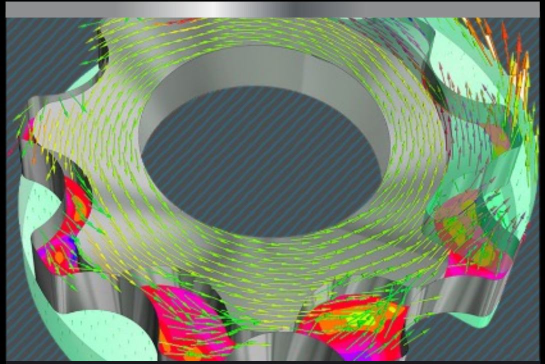


Finite Element Analysis



Structural Analysis

Electronic Design Automation



Computational Fluid Dynamics

ALTAIR Ansys cādence® DASSAULT SYSTEMES SIEMENS SYNOPSYS®

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

[Public]

LEADERSHIP EDA PERFORMANCE

~26.2

JOBS/HOUR

16-CORE 4th GEN AMD EPYC™
WITHOUT AMD 3D V-CACHE™

Up to 73%

FASTER RTL VERIFICATION

SYNOPSYS® VCS®

AMD graphics card

~45.4

JOBS/HOUR

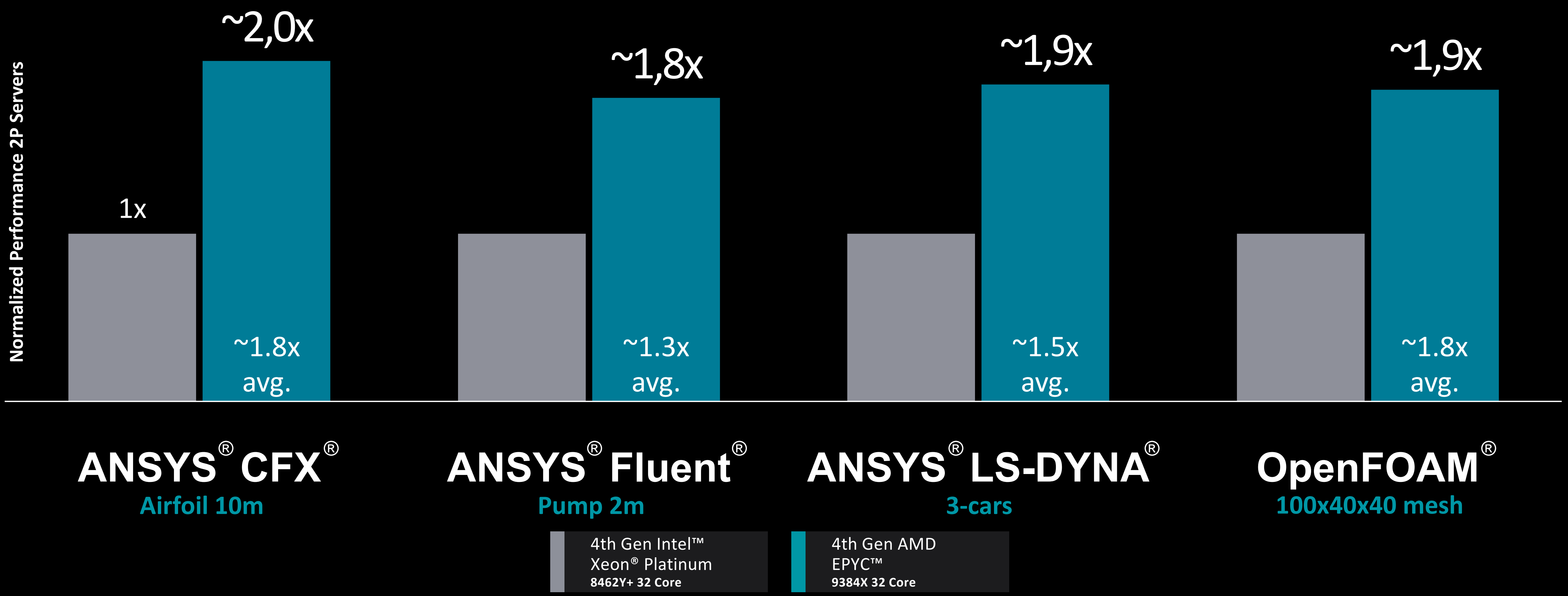
16-CORE 4th GEN AMD EPYC™
WITH AMD 3D V-CACHE

As of 4 May 2023. 1P servers: EPYC 9174F vs. EPYC 9184X. Results may vary due to factors including system configurations, software versions and BIOS settings. See SP5-050.

PERFORMANCE LEADERSHIP

TECHNICAL COMPUTING

CFD and FEA | 32-core Max/Avg. Result Comparison

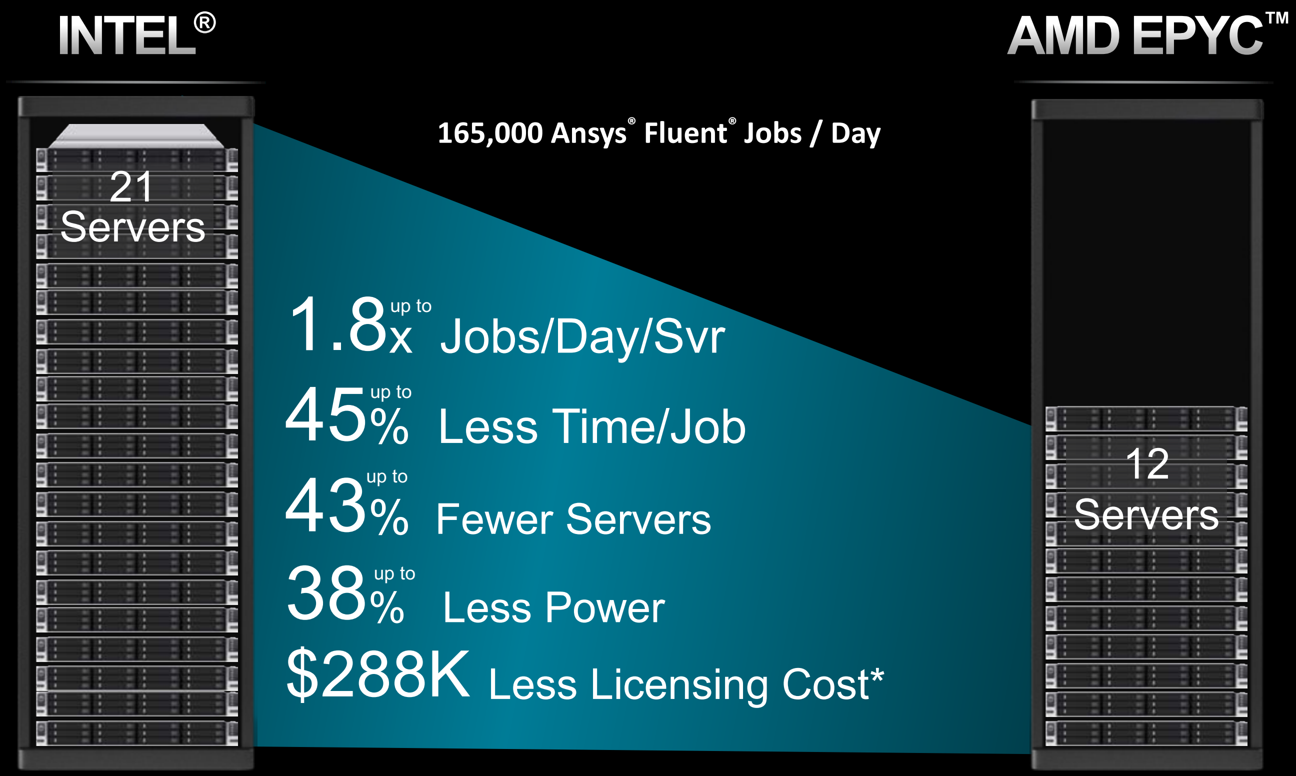


Results may vary due to factors including system configurations, software versions and BIOS settings.
As of 6/13/2023, see ANSYS CFX https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-cfx.pdf,
ANSYS LS-DYNA https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-ls-dyna.pdf, ANSYS Fluent https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-fluent.pdf and OpenFOAM https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-openfoam.pdf.

[Public]

ENABLING BETTER PRODUCTS, FASTER

Increase Productivity | Shorten Development Time | Lower Costs



INTEL®

AMD EPYC™

165,000 Ansys® Fluent® Jobs / Day

21 Servers

12 Servers

1.8x up to Jobs/Day/Svr

45% up to Less Time/Job

43% up to Fewer Servers

38% up to Less Power

$288K Less Licensing Cost*

Intel Platinum 32c 8462Y+
1,344 Cores
285k kWh per year
~8,006 jobs/day/server

AMD EPYC 32c 9384X
768 Cores
177k kWh per year
~14,482 jobs/day/server

SHORTEN DEVELOPMENT CYCLES

SPEED TIME TO REVENUE

REDUCE SPACE, POWER, AND COOLING

LOWER COSTS

All performance scores are estimates based on AMD internal testing in April and May 2023 and is shown in ‘jobs/day/server’. TCO Analysis based on the AMD EPYC™ Bare Metal Greenhouse Gas TCO Tool v9.32 Pro. AMD processor pricing based on 1KU price as of May 2023. Intel® Xeon® Scalable CPU pricing from https://ark.intel.com as of April 2023. All pricing is in USD. [1] **This is a power only** TCO with a time frame of 3-years. OpEx is power only, it does not include costs for real estate, admin and software. Power cost modeled @ $0.128/kWh with rack power of 10kW / rack having 10% reserved for non-server power use. The model uses a 1.7 PUE. [2] Values are for USA. See endnote SP5TCO-045

AMD AI

Broad portfolio of training and inference compute engines

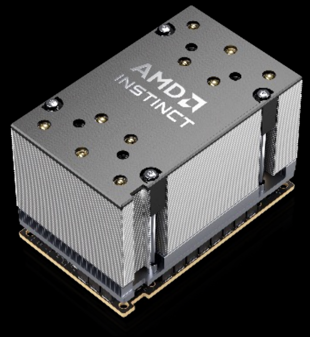
Open and proven software capabilities

Deep ecosystem of AI partners and co-innovation

AMD
AI Platforms

Training and inference portfolio

Data center | Edge | End point

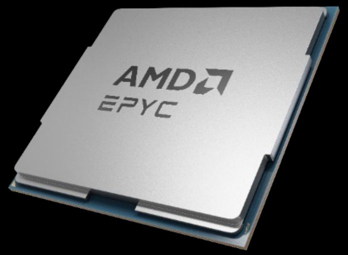


AMD Instinct™ Accelerators

HPC and data center training and inference

AMD Alveo™ Accelerators

Data center and edge inference



4th 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

AMD
EPYC | INSTINCT

Powering datacenter AI at scale

TOP 500 The List. | #1 Frontier

National Cancer Institute and DOE accelerating cancer research and treatment

TOP 500 The List. | #3 LUMI

Largest Finnish language model (TurkuNLP-13B)

AI2 OLMo

Allen Institute scientific LLM

TOP 500 The List. | #11 Explorer

WUS3 running AI and HPC workloads

1st Korean LLM

T5 NLP with 11B parameters

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

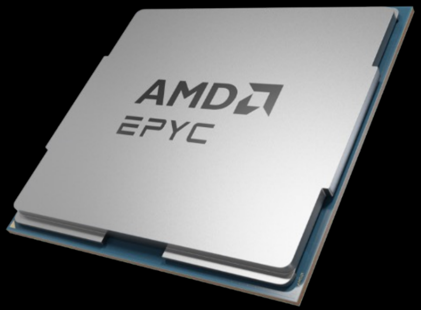
AMD AI Platforms

ROCm

Data center GPU

ZenDNN

Data center CPU



Vitis AI

Edge and end points

[Public]

AMD

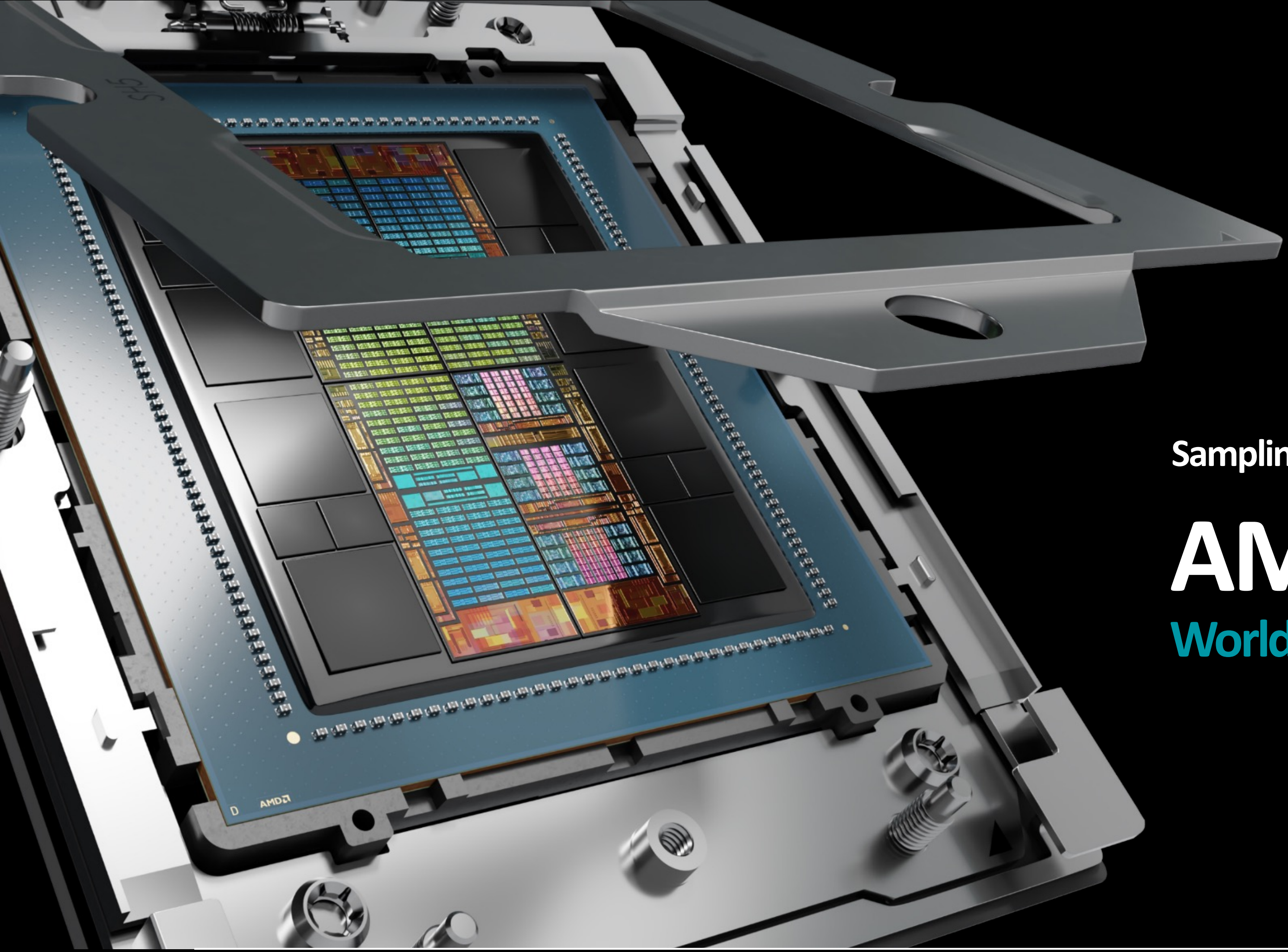
CDNA 3

Next-gen AI accelerator architecture

Dedicated accelerator engines for AI and HPC

3D packaging with 4th Gen AMD Infinity architecture

Optimized for performance and power efficiency



Sampling now

AMD Instinct™ MI300A

World's first APU accelerator for AI and HPC

AMD CDNA 3 Next-Gen Accelerator Architecture

ZEN 4 24 CPU Cores

128 GB HBM3

5nm and 6nm Process Technology

Shared Memory CPU + GPU

AMD Instinct™ MI300X

Leadership generative AI accelerator

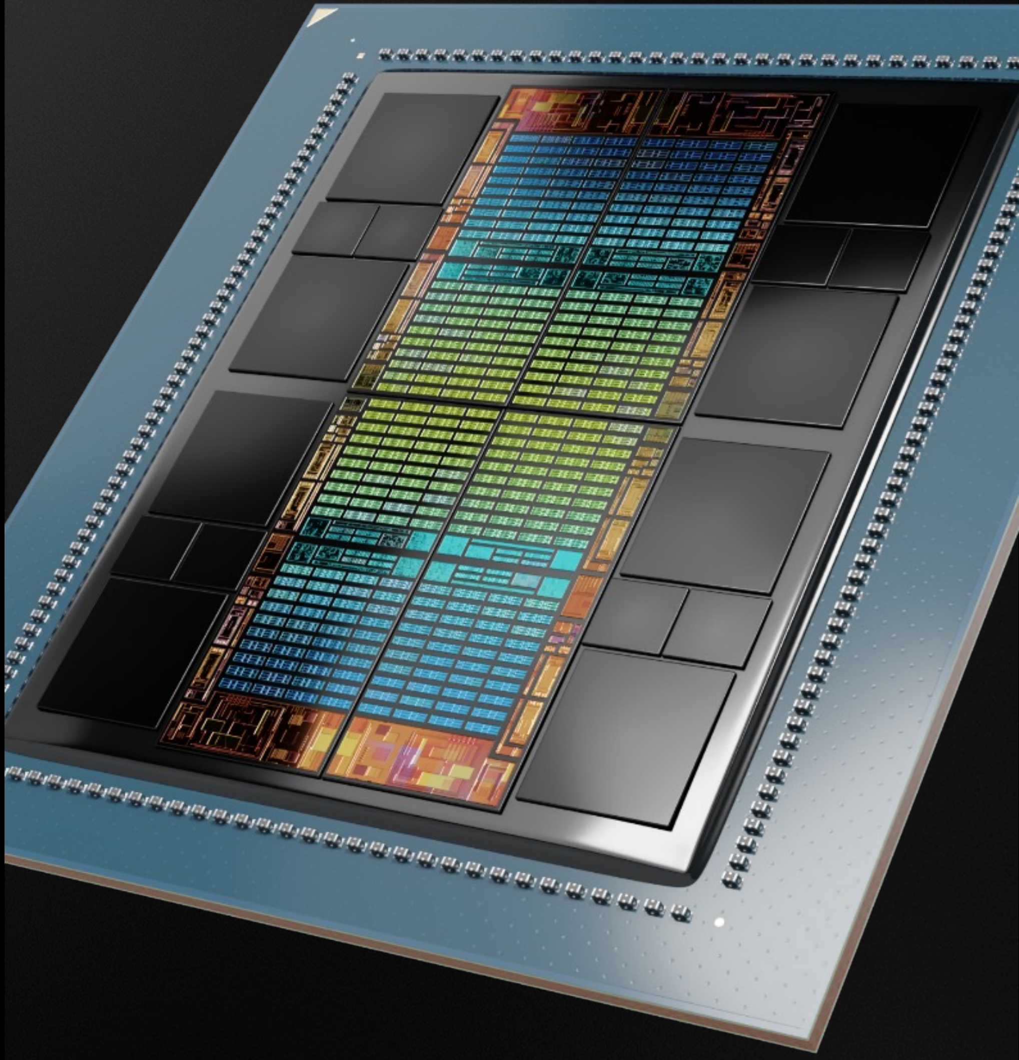
AMD CDNA 3

192 GB
HBM3

5.2 TB/s
Memory Bandwidth

896 GB/s
Infinity Fabric™ Bandwidth

3D Chiplet Architecture



AMD Instinct™ Platform

Available from leading OEMs & CSPs

8x MI300X	**21** PF BF16 | FP16	**1.5** TB/s HBM3

896 GB/s
Infinity Fabric™ Bandwidth

Industry-Standard
OCP Design

DELL Technologies

Hewlett Packard Enterprise

Lenovo

SUPERMICRO

Microsoft

ORACLE

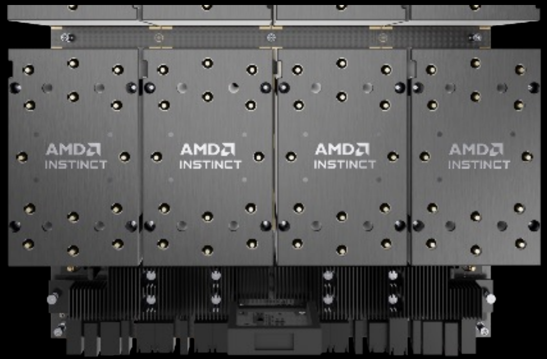
Cirrascale

AMD
together we advance_

AMD Instinct™ Platform
Performance and TCO Advantage

1 Nvidia H100 HGX

640 GB HBM3 | 26.4 TB/s



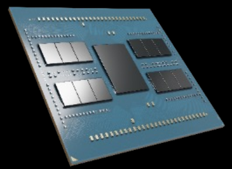
1 AMD Instinct™ MI300X Platform

1.5 TB HBM3 | 42.4 TB/s

Training & Inference		Training	Inference
1x	Performance per system	1x MPT-30B	1.6x Bloom 176B
1x	Models per system	2x ~30B	2x ~70B
1x	Max LLM model size per system	2x ~70B vs.~30B	2x ~680B vs 290B

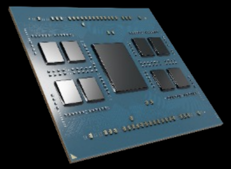
Results may vary. See endnotes:MI300-34, MI300-40, MI300-39, MI300-42

DELIVERING SOLUTIONS FOR THE MODERN DATA CENTER



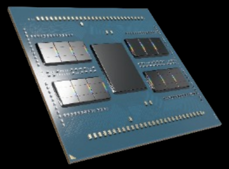
General Purpose Computing

4th Gen EPYC™ CPU
“Genoa”



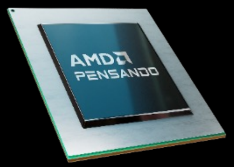
Cloud Native Computing

4th Gen EPYC™ CPU
“Bergamo”



Technical Computing

4th Gen EPYC™ CPU
“Genoa-X”



Networking Pensando P4 DPU

Cloud Efficiency for Enterprise

MI300A

MI300X

Open | Proven | Ready
AI Software

CPU AI leadership

100s of embedded
AI inference customers

Broadest
AI-powered
PC portfolio

[Public]

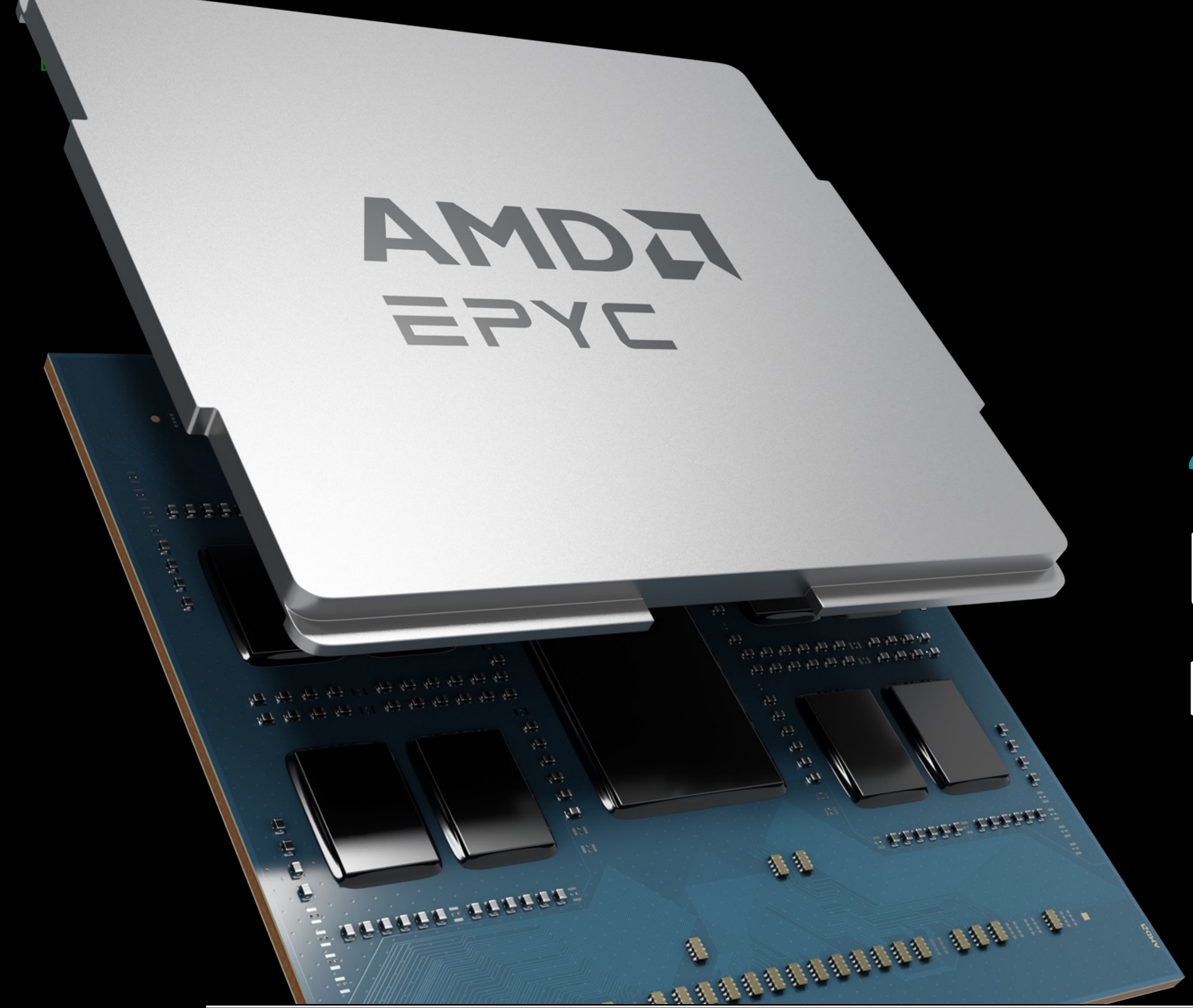
Back-up

AMD EPYC SKU Line-up

MODEL	# OF CPU CORES	# OF THREADS	MAX. BOOST CLOCK	ALL CORE BOOST SPEED	BASE CLOCK	L3 CACHE	DEFAULT TDP
AMD EPYC™ 9754	128	256	Up to 3.1GHz	3.1GHz	2.25GHz	256MB	360W
AMD EPYC™ 9754S	128	128	Up to 3.1GHz	3.1GHz	2.25GHz	256MB	360W
AMD EPYC™ 9734	112	224	Up to 3.0GHz	3.0GHz	2.2GHz	256MB	340W
AMD EPYC™ 9684X	96	192	Up to 3.7GHz	3.42GHz	2.55GHz	1152MB	400W
AMD EPYC™ 9384X	32	64	Up to 3.9GHz	3.5GHz	3.1GHz	768MB	320W
AMD EPYC™ 9184X	16	32	Up to 4.2GHz	3.85GHz	3.55GHz	768MB	320W
AMD EPYC™ 9654P	96	192	Up to 3.7GHz	3.55GHz	2.4GHz	384MB	360W
AMD EPYC™ 9654	96	192	Up to 3.7GHz	3.55GHz	2.4GHz	384MB	360W
AMD EPYC™ 9634	84	168	Up to 3.7GHz	3.1GHz	2.25GHz	384MB	290W
AMD EPYC™ 9554P	64	128	Up to 3.75GHz	3.75GHz	3.1GHz	256MB	360W
AMD EPYC™ 9554	64	128	Up to 3.75GHz	3.75GHz	3.1GHz	256MB	360W
AMD EPYC™ 9534	64	128	Up to 3.7GHz	3.55GHz	2.45GHz	256MB	280W
AMD EPYC™ 9474F	48	96	Up to 4.1GHz	3.95GHz	3.6GHz	256MB	360W
AMD EPYC™ 9454P	48	96	Up to 3.8GHz	3.65GHz	2.75GHz	256MB	290W
AMD EPYC™ 9454	48	96	Up to 3.8GHz	3.65GHz	2.75GHz	256MB	290W
AMD EPYC™ 9374F	32	64	Up to 4.3GHz	4.1GHz	3.85GHz	256MB	320W
AMD EPYC™ 9354P	32	64	Up to 3.8GHz	3.75GHz	3.25GHz	256MB	280W
AMD EPYC™ 9354	32	64	Up to 3.8GHz	3.75GHz	3.25GHz	256MB	280W
AMD EPYC™ 9334	32	64	Up to 3.9GHz	3.85GHz	2.7GHz	128MB	210W
AMD EPYC™ 9274F	24	48	Up to 4.3GHz	4.1GHz	4.05GHz	256MB	320W
AMD EPYC™ 9254	24	48	Up to 4.15GHz	3.9GHz	2.9GHz	128MB	200W
AMD EPYC™ 9224	24	48	Up to 3.7GHz	3.65GHz	2.5GHz	64MB	200W
AMD EPYC™ 9174F	16	32	Up to 4.4GHz	4.15GHz	4.1GHz	256MB	320W
AMD EPYC™ 9124	16	32	Up to 3.7GHz	3.6GHz	3.0GHz	64MB	200W

Designators

“F” = High Frequency
“P” = Single Socket
“S” = SMT Disabled
“X” = 3D V-Cache



4th Gen AMD EPYC™ CPU "Bergamo"

Leadership cloud native performance

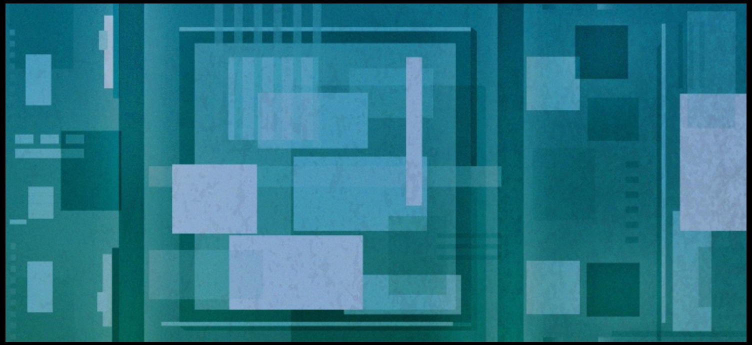
Up to 128 "Zen 4c" Cores	Consistent x86 ISA	82 B transistors	Greatest vCPU density	Best energy efficiency

Claims: EPYC-049, Source: https://www.amd.com/system/files/documents/amd-epyc-9004-pb-spec-power.pdf

"Zen 4c" cloud native core architecture

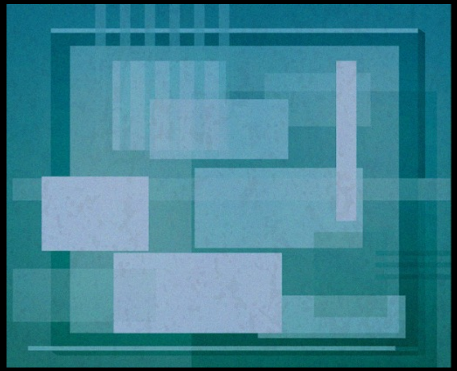
Designed for **density** and **power efficiency**

"Zen 4" core



Node	TSMC 5nm
Core + L2 Area	3.84 mm^2

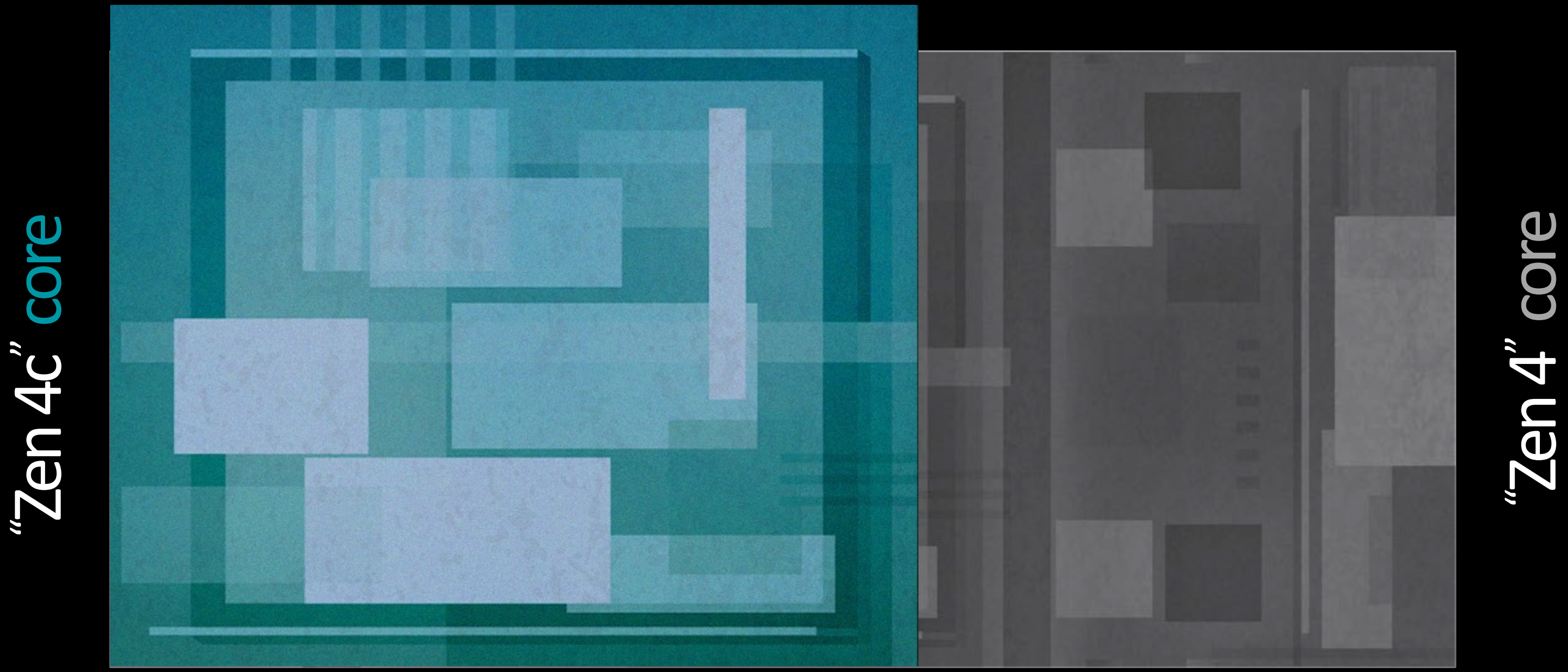
"Zen 4c" core



Node	TSMC 5nm
Core + L2 Area	2.48 mm^2

[Public]

~35% smaller core



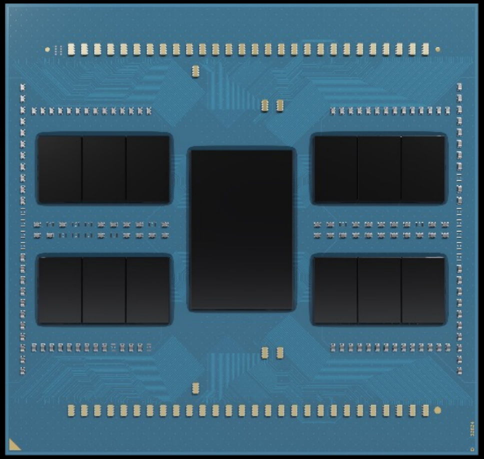
"Bergamo" with "Zen 4c"
8 CCDs, 16 cores per CCD

"Zen 4"

"Genoa" 4th Gen AMD EPYC™ CPU

Optimized for performance-per-core

12 x 8-core CCDs | Up to 96 cores

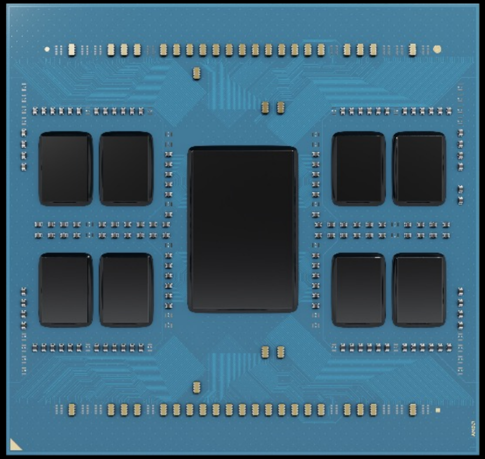


"Zen 4c"

"Bergamo" 4th Gen AMD EPYC™ CPU

Optimized for performance-per-watt

8 x 16-core CCDs | Up to 128 cores

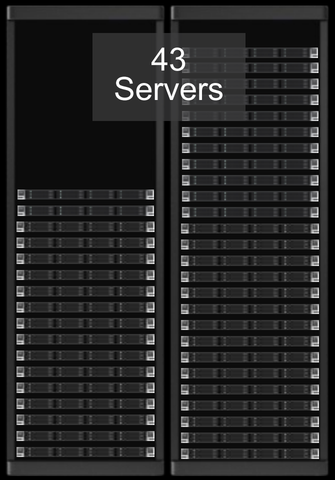


[Public]

AMD Cloud Native Advantage

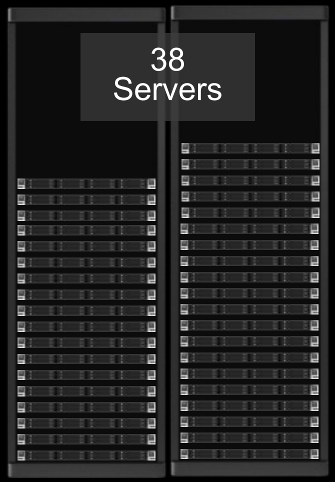
NGINX TARGET: Infrastructure delivering 375M Requests / Sec

2P Ampere[®] Altra Max M128-30, 128c



POWER[1]
444.50 Wh / Server
11.557 kWh / Rack

2P INTEL[®] PLATINUM 8490H, 60C



POWER[1]
594.0 Wh / Server
11.880kWh / Rack

2P AMD EPYC™ 9754, 128c

POWER[1]
644.00 Wh / Server
11.592 kWh / Rack

AMD EPYC 9754 Powered Servers Deliver

Vs. Ampere M128-30	Vs. Intel Platinum 8490H
~49% LOWER OPEX	~57% LOWER OPEX
~34% LOWER TCO	~67% LOWER TCO

2P EPYC 9754

Space & Power Savings

2P AMPERE[®] Altra Max M128-30	2P INTEL[®] Platinum 8490H
~28 FEWER SERVERS	~23 FEWER SERVERS
~49% LESS POWER ANNUALLY[2]	~57% LESS POWER ANNUALLY[2]
~57.9 US TONS LESS CO_2e ANNUALLY[2]	~79.09 US TONS LESS CO_2e ANNUALLY[2]

The power and Greenhouse Gas numbers above reflect a PUE of 1.70

All performance scores are estimates based on AMD internal testing in May & June 2023. AMD perf is on an AMD reference platform with a score of 26.248M requests / sec. Intel perf done on a Lenovo server with a score of 9,908,966. Ampere perf done on an Ampere Mt. Collins server with a score of 8.843M requests / sec. Analysis based on the AMD EPYC™ Bare Metal Server & Greenhouse Gas Emission TCO Estimation Tool - version 9.32 Pro. AMD processor pricing based on 1KU price as of April 2023. Intel pricing from ark. https://ark.intel.com in April 2023. Ampere C[1] **Power and Server Cost only are included in this TCO. This is a power only OpEx and** TCO analysis with a time frame of 3-year with power @ $0.128/kWh with 12kW / rack; and a PUE of 1.70. NOT included in this analysis are admin cost, real estate cost, software cost as well as power for any networking and storage external to the server. See endnote SP5TCO-050K, 051K
PU data Phoronix.com May 2023. All pricing is in USD.

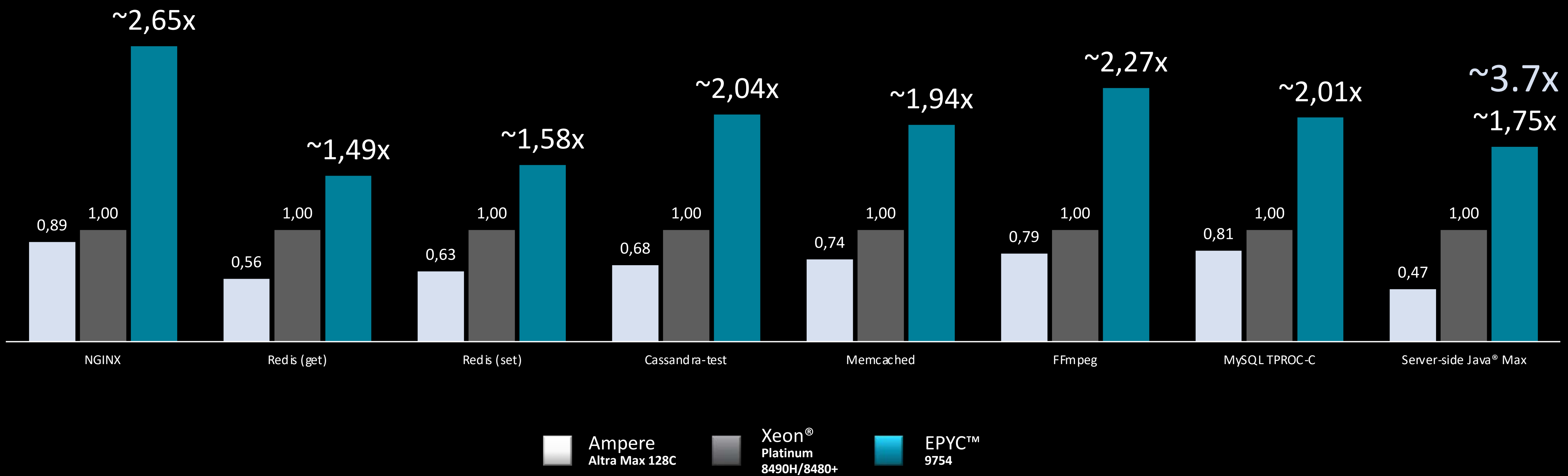
[Public]

Optimized Cloud Native Performance

Up to

3.7x throughput performance (~2.9x avg.)
for a wide variety of cloud native workloads

vs ampere



2P servers: 128C AMD EPYC™ 9754 vs. Ampere Altra® Max M128-30 and 56C/60C Intel Xeon Platinum 8480+/8490H

Results may vary due to factors including system configurations, software versions and BIOS settings. As of 6/13/2023, see Cloud Native Workloads https://www.amd.com/system/files/documents/amd-epyc-9004-pb-cloud-native-workloads.pdf.

Endnotes

- EPYC-018: Max boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems.

- EPYC-028: As of 2/2/22, of SPECpower_ssj® 2008 results published on SPEC's website, the 55 publications with the highest overall efficiency results were all powered by AMD EPYC processors. More information about SPEC® is available at http://www.spec.org. SPEC and SPECpower are registered trademarks of the Standard Performance Evaluation Corporation.

- EPYC-049: AMD EPYC 9754 is a 128 core dual threaded CPU and in a 2 socket server with 1 thread per vCPU delivers 512 vCPUs per EPYC powered server which is more than any Ampere or 4 socket Intel CPU based server as of 05/23/2023.

- SP5-013D: SPECrate®2017_int_base comparison based on published scores from www.spec.org as of 05/31/2023. 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 specifications and pricing as of 6/13/23. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See 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.

- SP5-050: EDA RTL Simulation comparison based on AMD internal testing completed on 4/13/2023 measuring the average time to complete a graphics card test case simulation. comparing: 1x 16C EPYC™ 9384X with AMD 3D V-Cache Technology versus 1x 16C AMD EPYC™ 9174F on the same AMD "Titanite" reference platform. Results may vary based on factors including silicon version, hardware and software configuration and driver versions.

- 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 8480+ (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 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). 2P AMD EPYC 7763 (339,338 SPECjbb®2015 MultiJVM max-jOPS, 313,824 SPECjbb®2015 MultiJVM critical-jOPS, 128 total cores, https://www.spec.org/jbb2015/results/res2021q3/jbb2015-20210701-00688.html) shown at 0.85x the performance and 2P Intel Xeon Platinum 8380 (269,094 SPECjbb®2015 MultiJVM max-jOPS, 213,195 SPECjbb®2015 MultiJVM critical-jOPS, 80 total cores, https://spec.org/jbb2015/results/res2021q3/jbb2015-20210810-00701.html) shown at 0.58x the performance for reference. SPEC® and SPECjbb® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

Endnotes

- SP5-143A: SPECrate®2017_int_base comparison based on performing system published scores from www.spec.org as of 6/13/2013. 2P AMD EPYC 9754 scores 1950 SPECrate®2017_int_base http://www.spec.org/cpu2017/results/res2023q2/cpu2017-20230522-36617.html is higher than all other 2P servers. 1P AMD EPYC 9754 scores 981 SPECrate®2017_int_base score (981.4 score/socket) http://www.spec.org/cpu2017/results/res2023q2/cpu2017-20230522-36613.html is higher per socket than all other servers. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

- SP5-145: SPECpower_ssj®2008 comparison based on published 2U, 1P results as of 5/13/2023. Configurations: 1P AMD EPYC 9754 (35,346 ssj_ops/W at 70% load , 29,124 overall ssj_ops/W, 2U, https://spec.org/power_ssj2008/results/res2023q2/power_ssj2008-20230521-01255.html) is 2.5x the performance/watt vs 1P Ampere Altra Max M128-30 (14,438 ssj_ops/W at 70% load, 11,497 overall ssj_ops/W, 2U, http://www.spec.org/power_ssj2008/results/res2023q2/power_ssj2008-20230522-01260.html). SPEC® and SPEcpower® are registered trademarks of Standard Performance Evaluation Corporation. Learn more at www.spec.org.

- SP5-149: 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.

- SP5-150: Memcached mem_tier 1:10 set/get ops/sec comparison based on median scores of AMD internal measurements as of 6/13/2023. See Memcached performance brief for more details https://www.amd.com/system/files/documents/amd-epyc-9004-pb-cloud-native-workloads.pdf. 2P EPYC 9754S added (configuration is same as 9754 in the paper) showing a throughput performance of 40,643,750 ops/sec at 256C/256T total (158,765/thread) is ~1.84x the ops/sec/thread compared to Altra Max M128-30 (22068452 ops/sec, 86205 ops/sec/thread). 2P 120C/240T Xeon 8490H (29893871 ops/sec, 124558 ops/sec/thread) and 2P 256C/512T EPYC 9754 (58129312 ops/sec, 113534 ops/sec/thread) shown for reference. Results may vary due to factors including system configurations, software versions and BIOS settings.

- SP5-154: HPL benchmark based on AMD internal testing as of 6/13/2023. 2P server configurations: 2P EPYC 9754, BIOS 1003F (Memory Target Speed = DDR4800, TSME = Disabled, IOMMU=Auto, TDP Control = Manual, TDP = 400, PPT Control=Manual, PPT=400, Determinism Control=Manual, Determinism Enable = Power, NUMA nodes per socket= NPS4, SMT Control=Disable), 768 GB (24x 32GB 2R DDR5-4800) scores an average 10,134 GFLOPS which is 1.66x the performance of AMD estimated   2P Xeon Platinum 8490H (6115 GFLOPS). 2P EPYC 9654, BIOS 1003F (Memory Target Speed = DDR4800, TSME = Disabled, IOMMU=Auto, TDP Control = Manual, TDP = 400, PPT Control=Manual, PPT=400, Determinism Control=Manual, Determinism Enable = Power, NUMA nodes per socket= NPS4, SMT Control=Disable), 768 GB (24x 32GB 2R DDR5-4800) scores 8856 GFLOPS for 45% better GFLOPS as reference. Results may vary due to factors including system configurations, software versions and BIOS settings.

Endnotes

- SP5-165: The EPYC 9684X CPU is the world's highest performance x86 server CPU for technical computing, comparison based on SPEC.org publications as of 6/13/2023 measuring the score, rating or jobs/day for each of SPECrate®2017_fp_base (SP5-009E), Altair AcuSolve (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-altair-acusolve.pdf), Ansys Fluent (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-ansys-fluent.pdf), OpenFOAM (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-openfoam.pdf), Ansys LS-Dyna (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-ansys-ls-dyna.pdf), and Altair Radioss (https://www.amd.com/en/processors/server-tech-docs/amd-epyc-9004x-pb-altair-radioss.pdf) application test case simulations average speedup on 2P servers running 96-core EPYC 9684X vs top 2P performance general-purpose 56-core Intel Xeon Platinum 8480+ or top-of-stack 60-core Xeon 8490H based server for technical computing performance leadership. "Technical Computing" or "Technical Computing Workloads" as defined by AMD can include: electronic design automation, computational fluid dynamics, finite element analysis, seismic tomography, weather forecasting, quantum mechanics, climate research, molecular modeling, or similar workloads. Results may vary based on factors including silicon version, hardware and software configuration and driver versions. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

- SP5TCO-034: This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The Bare Metal Server Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool - version 6.80, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 10,000 units of integer performance based on the published scores for these specific Intel Xeon and AMD EPYC CPU based servers as of January 10, 2023. This estimation reflects a 3-year time frame with a PUE of 1.7 and a power US power cost of $0.16 / kWh. This analysis compares a 2P AMD 64 core AMD EPYC_9554 powered server with a SPECrate2017_int_base score of ; to a 2P Intel Xeon 60 core Platinum_8490H based server with a SPECrate2017_int_base score of 991, https://spec.org/cpu2017/results/res2023q1/cpu2017-20221206-33039.pdf. Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. For additional details, see https://www.amd.com/en/claims/epyc4#SP5TCO-034

- SP5TCO-045: As of May 2023, based on AMD Internal analysis and using the AMD EPYC™ Bare Metal Server and Greenhouse Gas Emissions TCO Estimation Tool v9.33 PRO estimating the cost and quantity of 2P AMD 32 core EPYC™ 9384X powered server versus 2P Intel® Xeon® 32 core Platinum 8462Y+ based server solutions required to deliver 165,000 jobs / day with Ansys Fluent-pump2.Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'.This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing.

- SP5TCO-050K: As of June 2023, based on AMD Internal analysis and using the AMD EPYC™ Bare Metal Server and Greenhouse Gas Emissions TCO Estimation Tool v9.33 PRO estimating the cost and quantity of 2P AMD 128 core EPYC™ 9754 powered server versus 2P Ampere Max 128-30 based server solution required to deliver 325 million requests.Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'.This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. For more details see https://www.amd.com/en/claims/epyc4#SP5TCO-050K

Endnotes

- SP5TCO-051: As of June 2023, based on AMD Internal analysis and using the AMD EPYC™ Bare Metal Server and Greenhouse Gas Emissions TCO Estimation Tool v9.33 PRO estimating the cost and quantity of 2P AMD 128 core EPYC™ 9754 powered server versus 2P Intel Platinum 8490H based server solution required to deliver 325 million requests.Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'.This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. For more details see https://www.amd.com/en/claims/epyc4#SP5TCO-051K

- SP5TCO-052K: As of June 2023, based on AMD Internal analysis and using the AMD EPYC™ Bare Metal Server and Greenhouse Gas Emissions TCO Estimation Tool v9.33 PRO estimating the cost and quantity of 1P AMD 128 core EPYC™ 9754 powered server versus 1P Ampere Max 128-30 based server solution required to deliver 325 million requests.Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'.This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. For more details see https://www.amd.com/en/claims/epyc4#SP5TCO-052K

- SPTTCO-054K: As of June 2023, based on AMD Internal analysis and using the AMD EPYC™ Bare Metal Server and Greenhouse Gas Emissions TCO Estimation Tool v9.33 PRO estimating the cost and quantity of 2P AMD 128 core EPYC™ 9754S powered server versus 2P Ampere Max 128-30 based server solution required to deliver 375 million requests.Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'.This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. For details see https://www.amd.com/en/claims/epyc4#SP5TCO-054K

- GD-083: Use of third party marks / logos/ products is for informational purposes only and no endorsement of or by AMD is intended or implied.

- GD-183: AMD Infinity Guard features vary by EPYC™ Processor generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard.

- GD-204: “Technical Computing” or “Technical Computing Workloads” as defined by AMD can include: electronic design automation, computational fluid dynamics, finite element analysis, seismic tomography, weather forecasting, quantum mechanics, climate research, molecular modeling, or similar workloads. GD-204

Endnotes

- 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.

- 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. Competitve 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.

DISCLAIMER AND TRADEMARKS

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's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale.