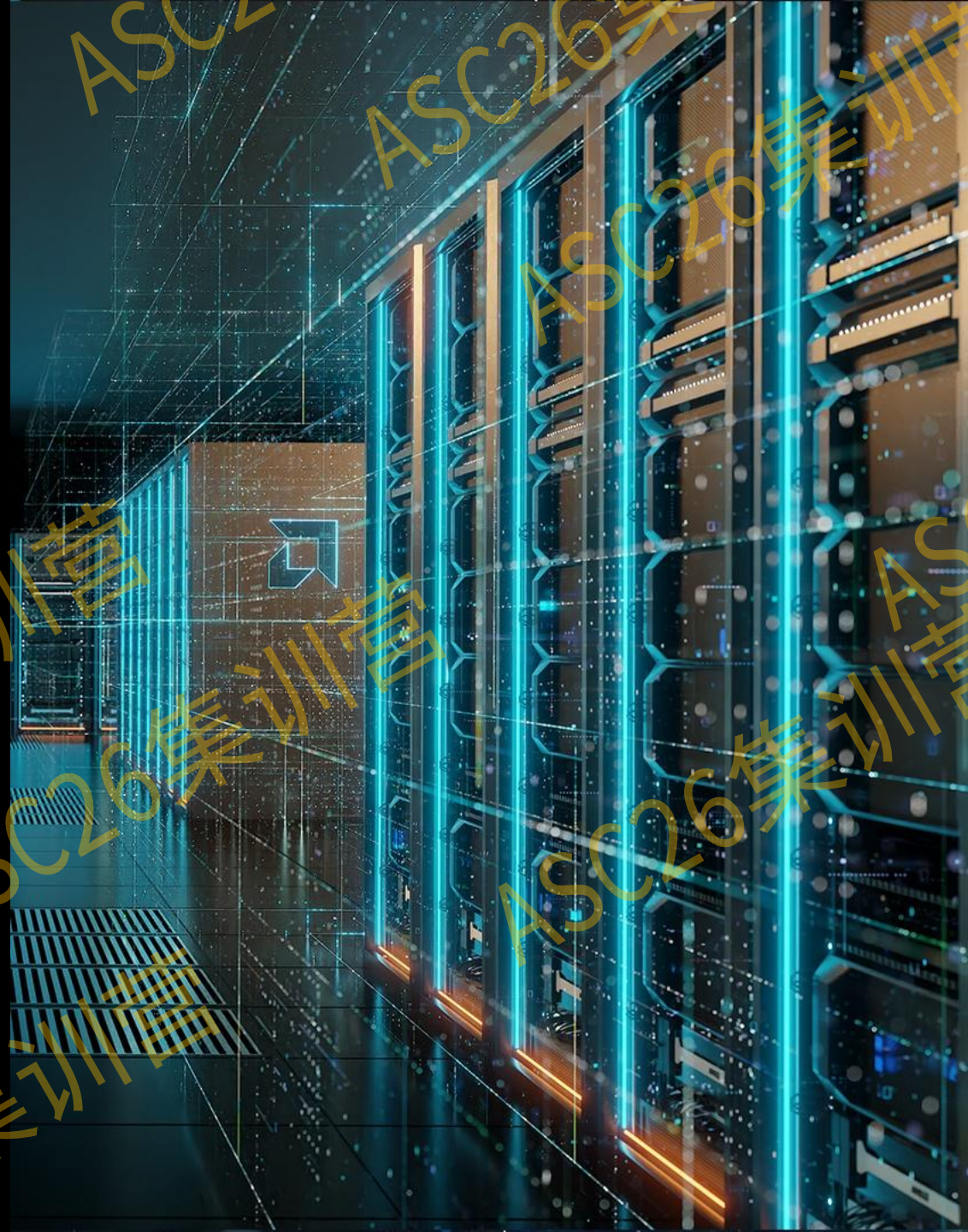
AMD

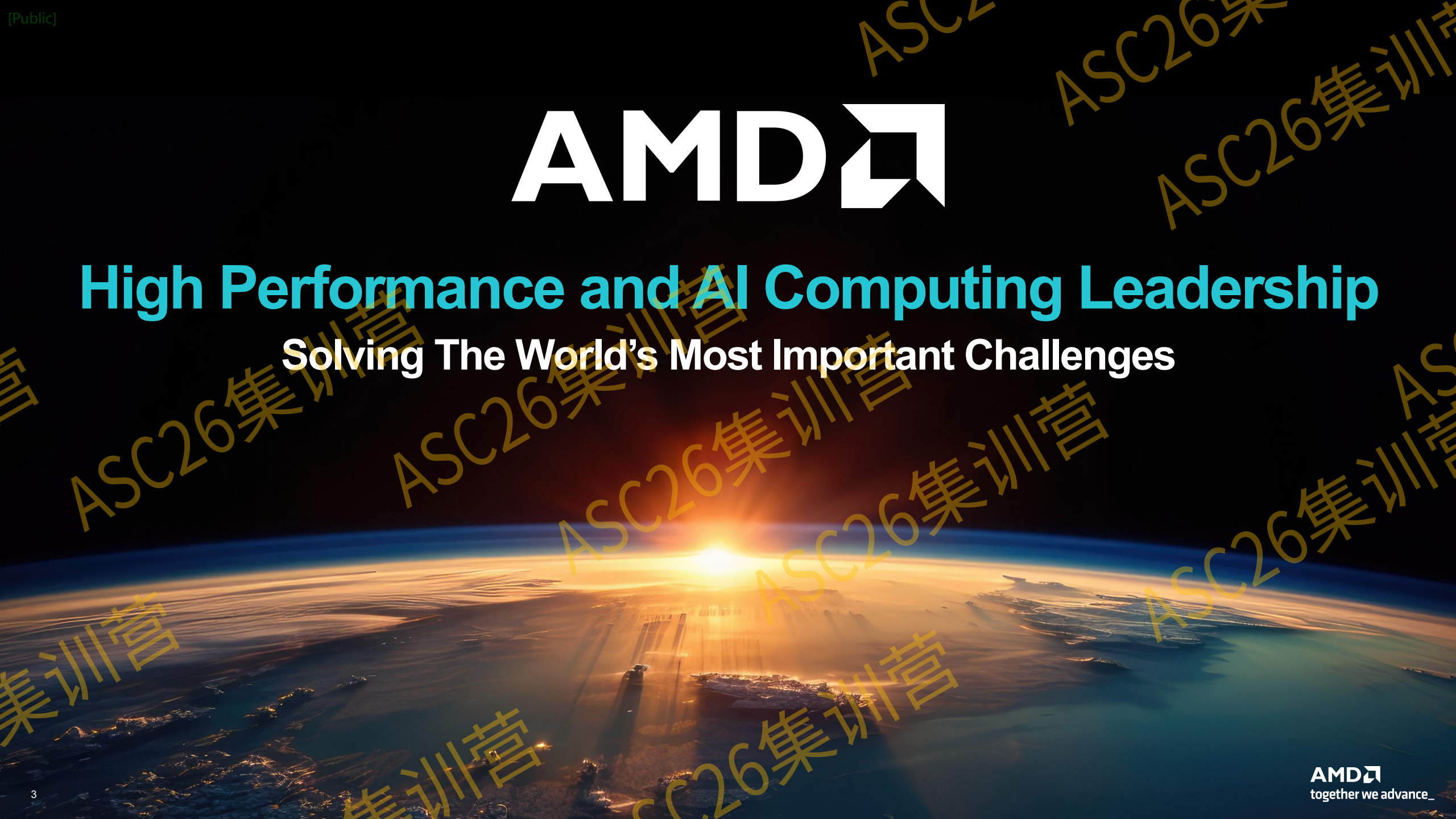
Our Path Forward
RELENTLESS INNOVATION

王强 Frederic Wang
Director, AMD Datacenter Solutions Group



Agenda

01	AMD Corporate Overview
02	EPYC™ 9005 Series CPU Architecture
03	Zen Software Studio
04	EPYC™ System Management Software



56 years

Founded May 1, 1969
Headquartered in Santa Clara, CA

28,000+ employees

Accelerating next-generation computing

$25.8B annual revenue in 2024

Over 25% reinvested towards research and development

3x market cap growth in 5 years

Top 100 most valuable companies in the world

100+ locations

Around the world

AMD
together we advance_

AMD Computing Powers The World

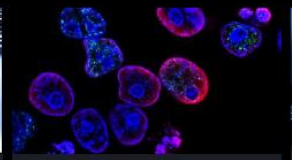
Edge & Intelligent Devices

Cloud and Data Center

HPC

AI Everywhere

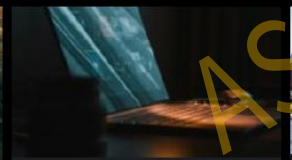
Manufacturing and Automation



Healthcare and Life Sciences

Aerospace

Communications



Consumer and Education

Automotive and Transportation

Financial Services

Science and Discovery

Gaming

Industry's Broadest Solution Portfolio

AMD

Leadership Core IP

CPU  | GPU  | NPU  | DPU  | FPGA 

Common Foundational IP

Infinity Fabric Interconnect | System-on-a-Chip | Chiplets & Packaging | Open-Source Software | End-to-End Product Security

Product Portfolio Driving Revenue Growth

Data Center | Client & Gaming | Embedded

Semi-Custom

Advancing rack-scale energy efficiency

- AMD targets a 20x increase in rack-scale energy efficiency for AI training and inference by 2030, from a 2024 base year.[7]
- Our new 2030 rack-level energy efficiency goal has major implications for equipment consolidation.
- Using training of a typical AI model in 2025 as a benchmark, the gains could enable:[8]
  - Rack consolidation from 275+ racks to <1 fully utilized rack
  - More than a 95% reduction in operational electricity use
  - Carbon emission reduction from nearly 3,000 to 100 metric tCO2 for model training

Learn more at https://www.amd.com/en/corporate-responsibility/data-center-sustainability

Open Development Drives Value & Innovation

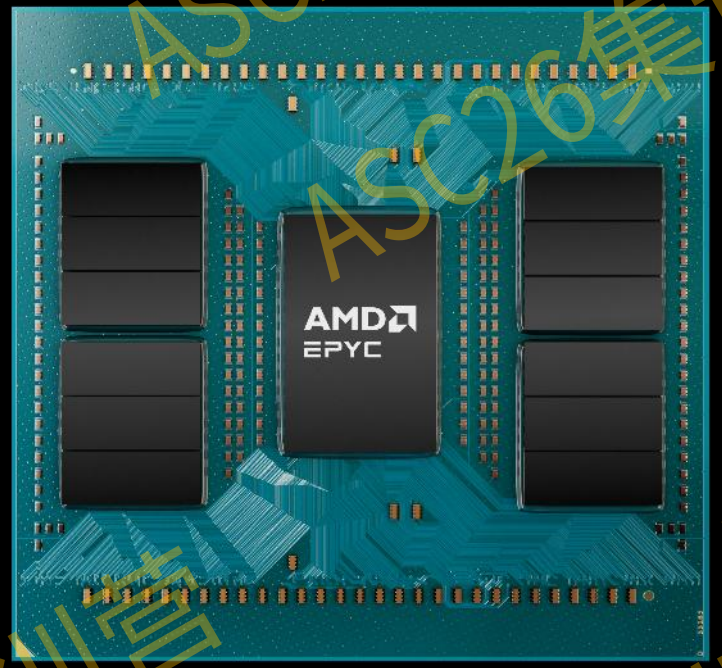
Open Hardware + Open Software + Open Ecosystem

Choice | Flexibility | Rapid Co-Innovation | Portability | Proven

EPYC™ 9005 Series CPU Architecture

5th Gen AMD EPYC™ Processors

Formerly codenamed “Turin”



Choice of CPU for cloud, enterprise & AI

ZEN 5

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

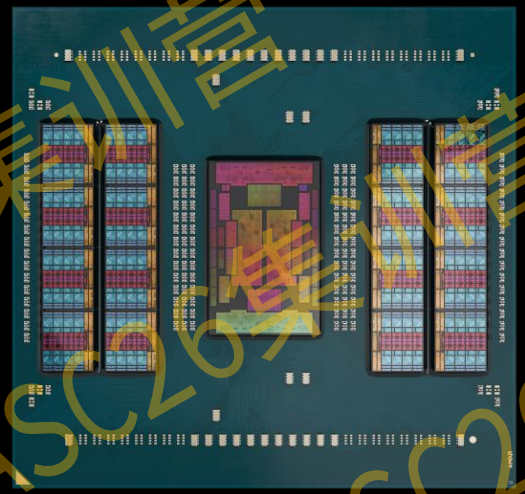
5th Gen AMD EPYC™ Generational Innovations

Compute

- "Zen5" up to **128 cores** / 256 threads
- "Zen5c" up to **192 cores** / 384 threads
- AVX-512 with **full 512b data path**
- New **500W** performance option
- Faster **5GHz** options
- **3/4nm** Zen cores

I/O & Platform

- 2P and 1P Configurations
- Up to 160 lanes of PCIe® Gen5
- **PCIe link encryption**
- SP5 Compatible with **"Genoa"**
- **CXL® 2.0**[1]



Memory

- 12 ch. DDR5 ECC **up to 6400 MT/s**
- Up to 2 DIMMs/channel capacity delivering up to **6TB/socket**
- **Dynamic Post Package Repair** (PPR) for x4 and x8 ECC RDIMMs

Security

- Hardware Root-of-Trust
- **Trusted I/O**
- **FIPS 140-3 in process**

1 - CXL Type 1&2 devices and PCIe link encryption support dependent upon ecosystem readiness
See endnotes 9xx5-048, 072, 083, GD-183A

5th Gen AMD EPYC™ CPU Chiplet Architecture

Up-to-16 CCDs capability

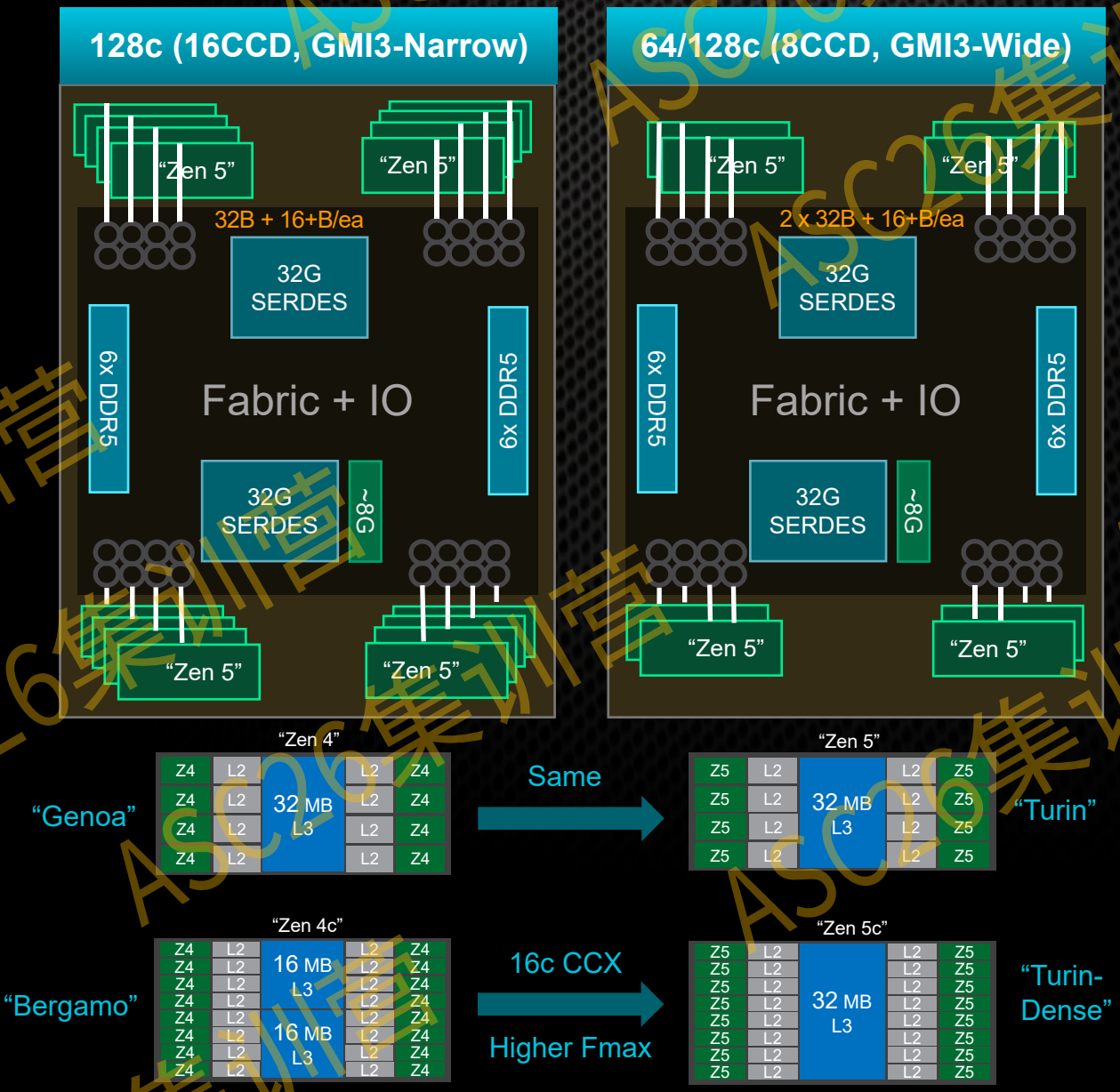
- Refined IOD/CCD and package co-design to enable

Enhanced "Turin"- Dense ("Zen 5c") vs "Bergamo" ("Zen 4c")

- Higher Fmax: 3.7GHz vs 3.1GHz
- 16c Shared L3 (CCX) for cache efficiency

GMI3 Chiplet interface

- Up-to-36Gbps, 20:1 with internal FCLK (1.8GHz max)
- Wide and Narrow connection options
  - >8CCD option use GMI3-Narrow
  - <=8CCD options can use GMI3-Wide
- 2x probe throughput vs GMI2 (3rd Gen AMD EPYC™)
- "16+B" CCD->IOD Datapath: Enhanced probe-response data (32B) and write-heavy traffic (25B) performance
- GMI "Folding" for power management (Half-width)



"Zen 5" and "Zen 5c" SOC

"Zen 5" Optimized for maximum 1T performance

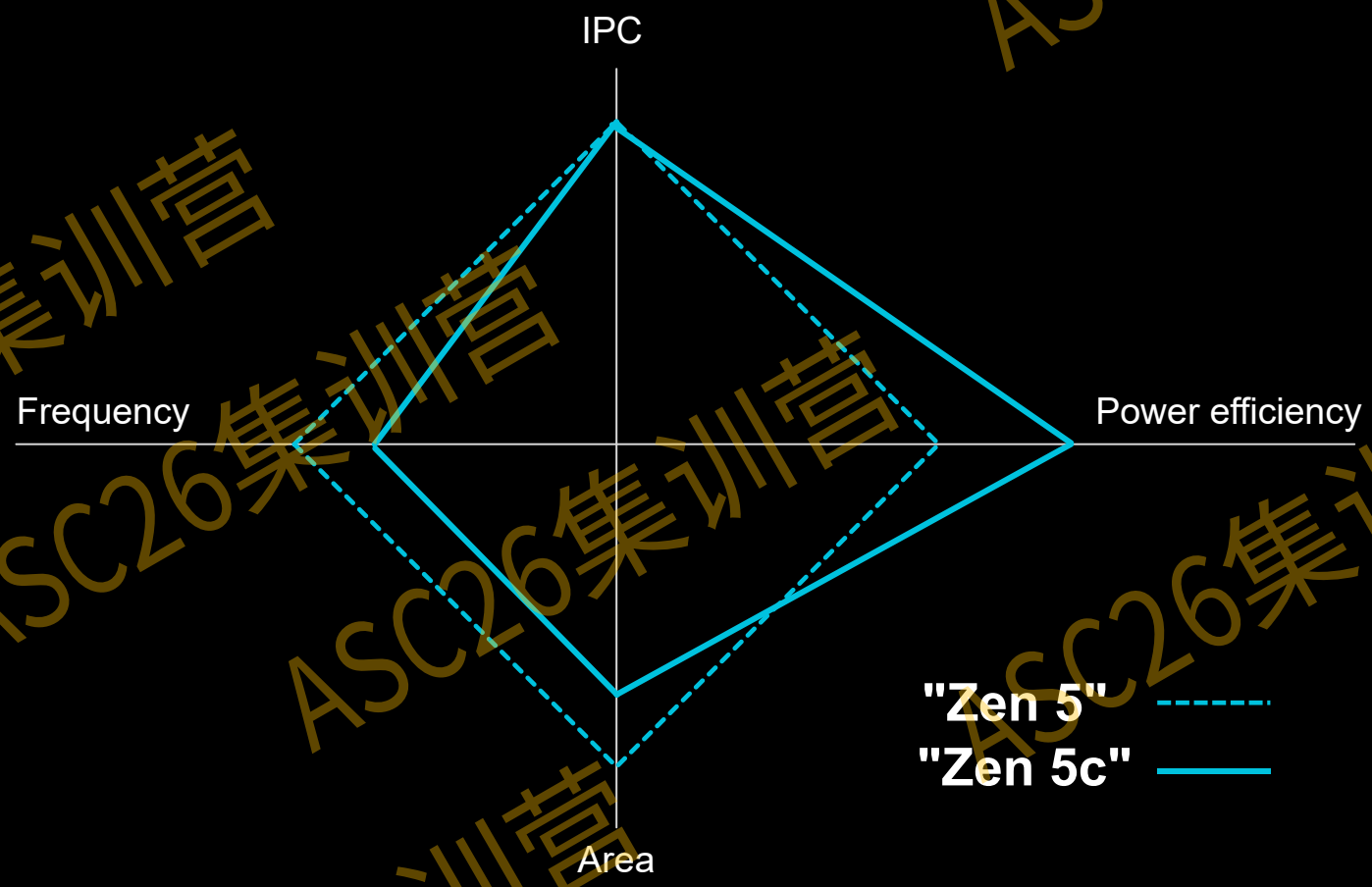
- High max frequency Target
- Large L3 per core

"Zen 5c" Optimized for scalability

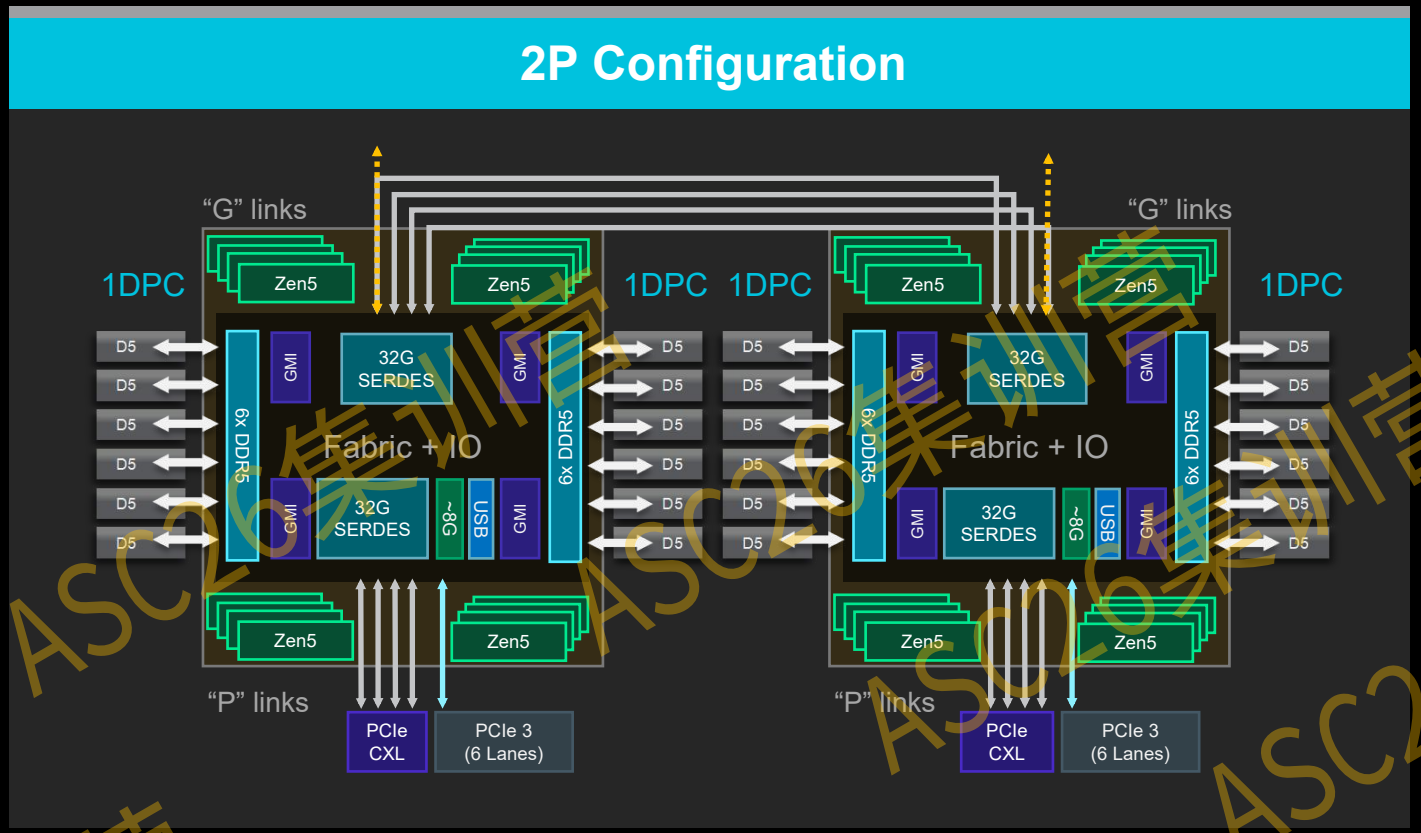
- Same IPC and features
- Lower max frequency
- Increased power efficiency
- Lower L3 per core

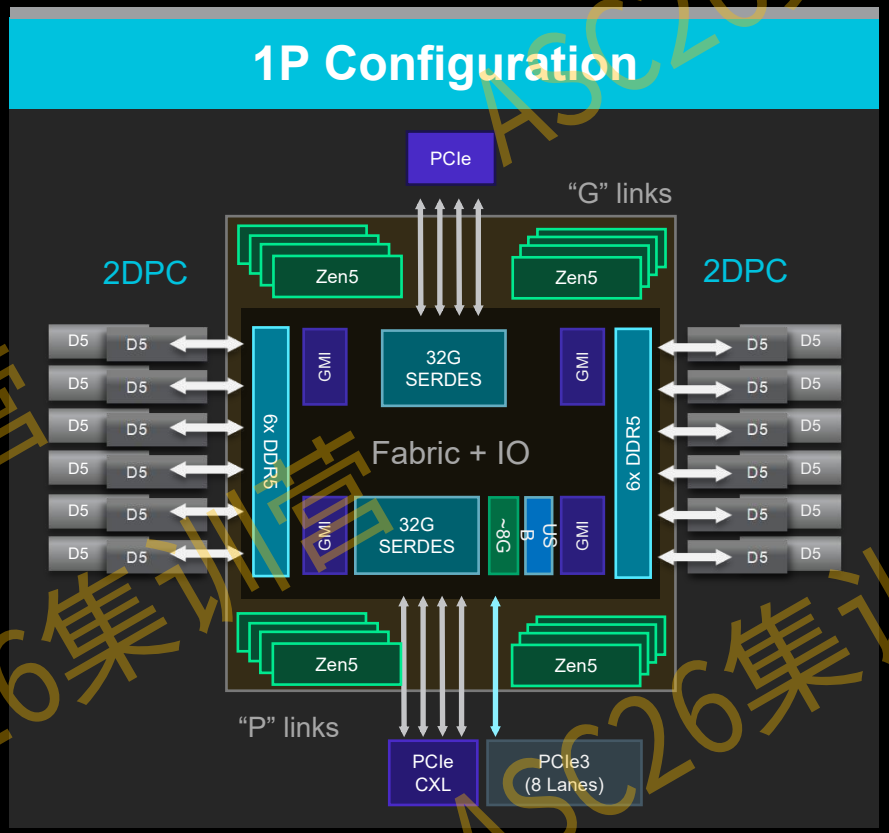
Simplifies software

- Same IPC (less L3 difference) no unique bottlenecks within the core=
- Both full ISA/SMT support



5th Gen AMD EPYC™ SoC Platform Overview



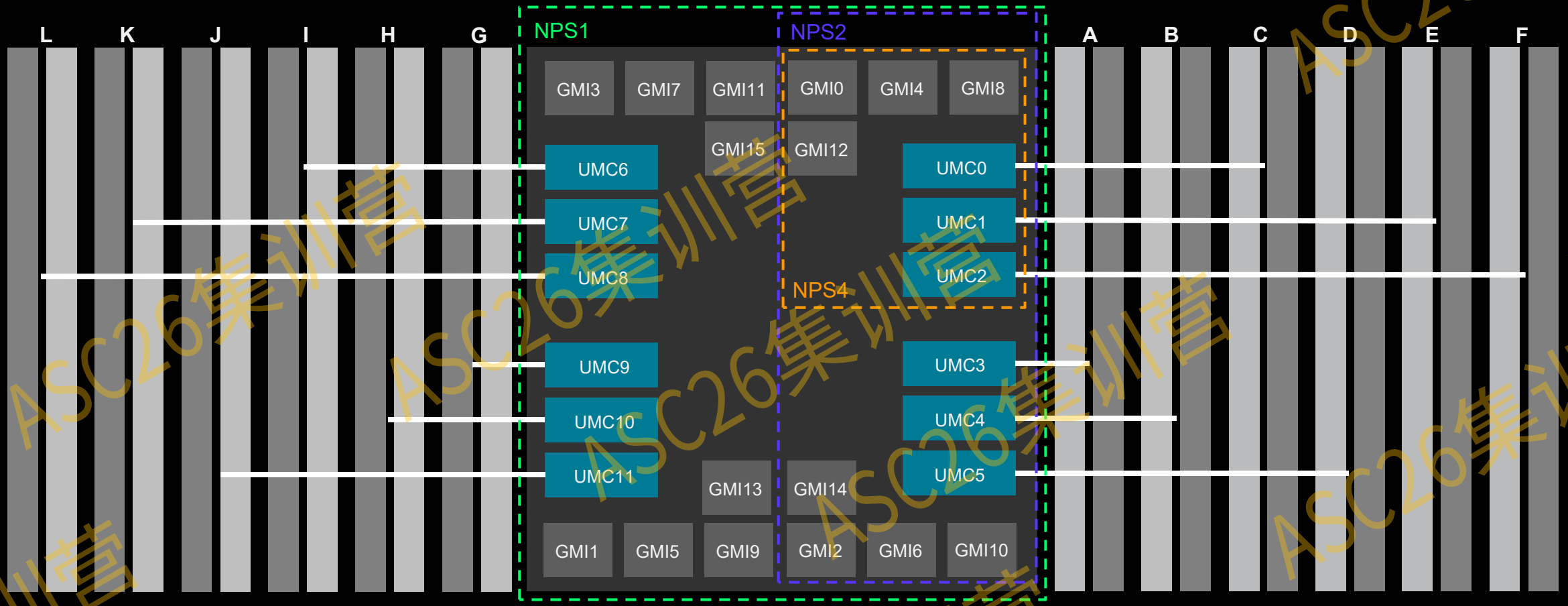


Strong upgrade to existing SP5 (4th Gen AMD EPYC™ CPU) in the same platforms

- **25% improvement in DRAM speed (4800 -> up to 6400*) using JEDEC-standard (non-proprietary) DIMMs**
- **1P PCIe® aggregate bandwidth improvement due to internal SOC topology changes**
- **Enhanced platform option for 500W TDP capability**

*See endnotes: 9xx5-083A

AMD EPYC™ 9005 CPUs

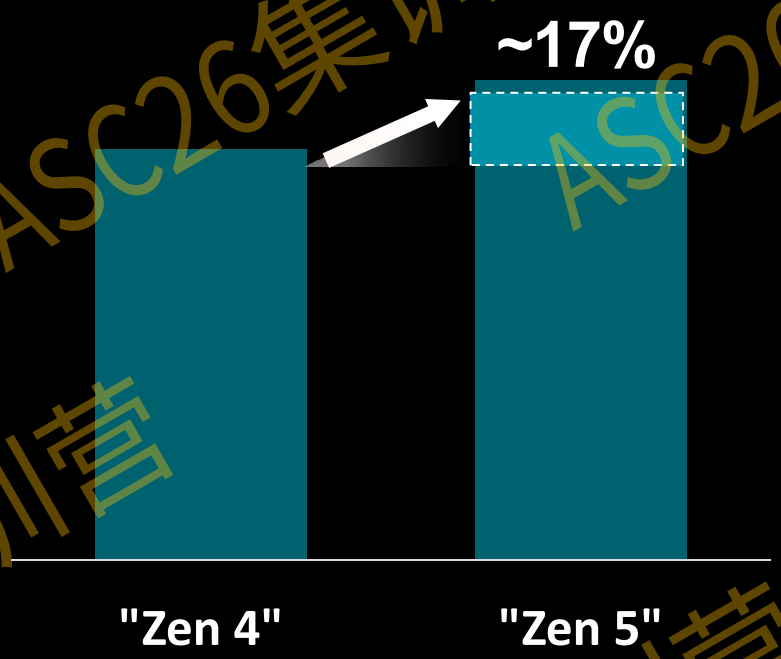


Multiple memory Nodes Per Socket (NPS) configurations for additional optimization

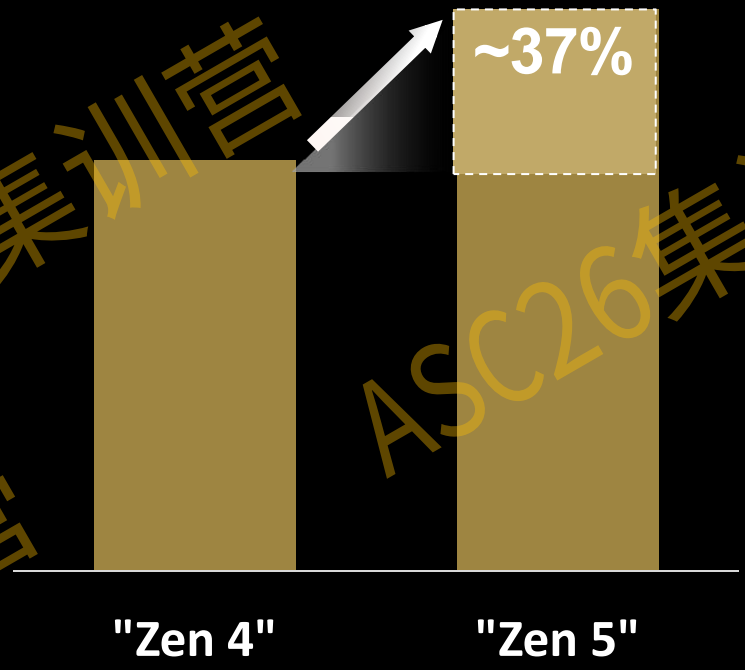
AMD EPYC™ Processors Generational Improvements

"Zen 5" Delivering Exceptional IPC Uplift for Server CPUs

Geomean of 36 Enterprise & Cloud Server Workloads
(Fixed Frequency, 12+1 CCD)



Geomean of 24 AI & HPC Server Workloads
(Fixed Frequency, 12+1 CCD)



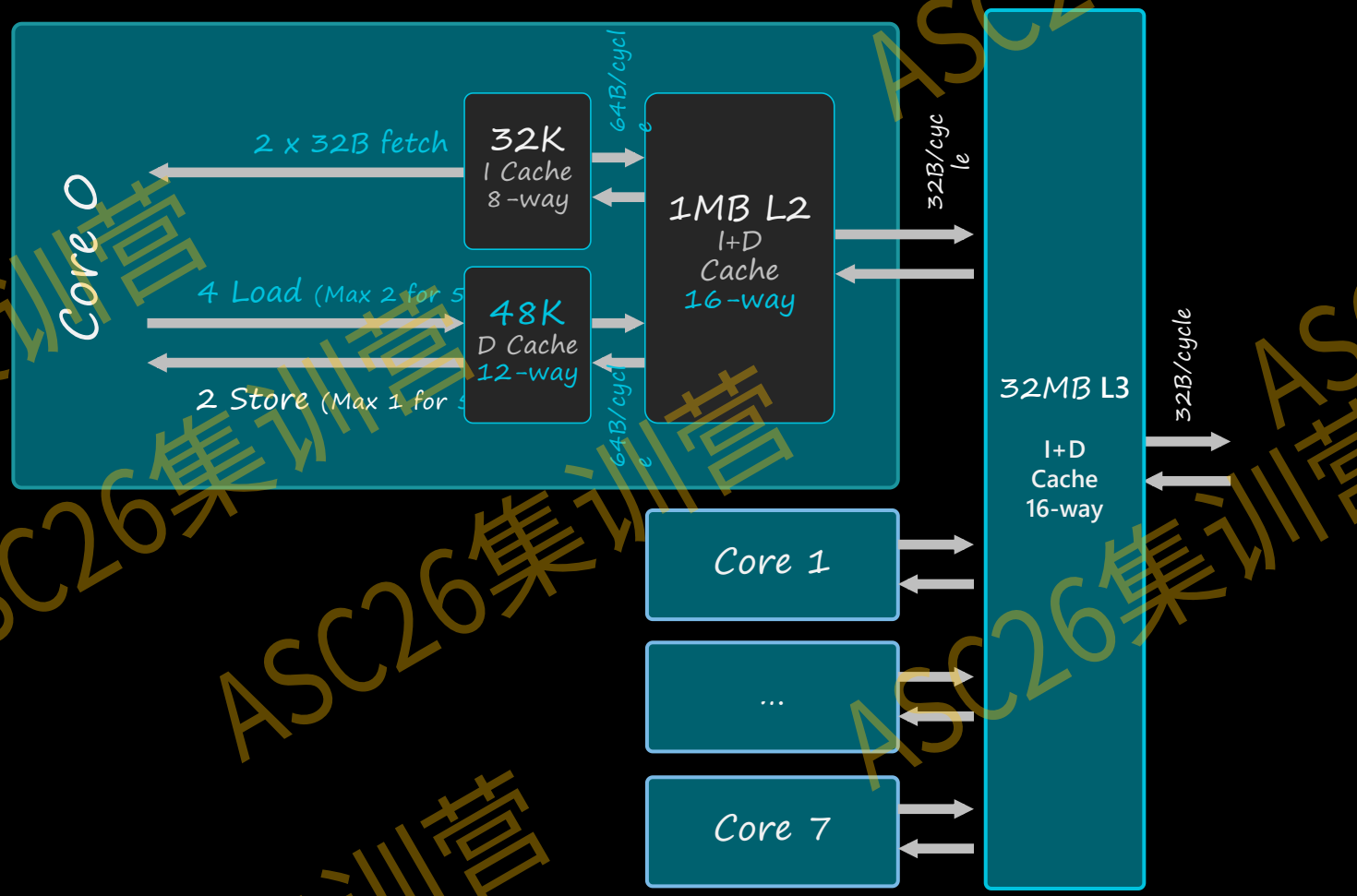
See endnote: 9xx5-001

Key “Zen 5” vs. “Zen 4” Capabilities

Attribute	“Zen 4”	“Zen 5”
L1/L2 BTB	1.5K/7K	1.6K/8K
Return Address Stack	32	52
ITLB L1/L2	64/512	64/2048
Fetched/Decoded Instruction Bytes/cycle	32	64
Op Cache associativity	12-way	16-way
Op Cache bandwidth	9macro-ops	12 inst or fused inst
Dispatch bandwidth (macro-ops/cycle)	6	8
AGU Scheduler	3x24 ALU/AGU	56
ALU Scheduler	1x24 ALU	88
ALU/AGU	4/3	6/4
Int PRF (reg/flag)	224/126	240/192
Vector Reg	192	384
FP Pre-Sched Queue	64	96
FP Scheduler	2x32	3x38
FP Pipes	3	4
Vector Width	256b	256b/512b
ROB/Retire Queue	320	448
LS Mem Pipes support Load/Store	3/1	4/2
DTLB L1/L2	72/3072	96/4096
L1Data Cache	32KB/8-way	48KB/12-way
L2 per core	1MB/8w	1MB/16w
L2 bandwidth	32B/clk	64B/clk

“Zen 5” Core Complex speeds and feeds

- Double the L2 associativity
- Double the L2 bandwidth
- Low latency L3 with 320 L3 in-flight misses
- Baseline from “Zen 4”
  - Fast private 1MB L2 cache
  - L3 shared among all cores in the complex
  - L3 is filled from L2 victims
  - L2 tags duplicated in L3 for probe filtering and fast cache transfer



“Zen 5” Microarchitecture Overview

NextGen Branch Predictor Caches

- I-Cache: 32KB, 8-way; 2x 32B fetch/cycle
- Op-Cache: 6K inst; 2x 6-wide fetch/cycle
- D-Cache: 48KB, 12-way; 4 mem ops/cycle
- L2-Cache: 1MB, 16-way

Dual I-Fetch/decode pipes

- 4 instrucitons per decode pipe
- 8 ops/cycle dispatch

4 inst/pipe 8 ops/cycle dispatched to Integer or FP Execution capabilities

- 6 integer ALU
- 4 AGU, 4 addresses to LS per cycle
- 6 FP ops/cycle; 2-cycle FADD
- Full 512b AVX512 datapaths

Dataflow

- 4 load pipes capable of 2, 512b AVX512 loads
- 2x width L2 cache <-> L1I and L1D caches

2 Threads per core

Optimized Branch Prediction and Fetch

Batch prediction

- Zero-bubble conditional branches
- L2-sized (16K) L1 BTB and larger TAGE
- Larger return address stack (52entry)
- 2 taken predictions/cycle
- Up to 3 prediction windows/cycle

Memory Management

- Aggressive Fetch hides L2 & tablewalk latency
- 2048 entry L2 ITLB

Icache latency and bandwidth

- 64B/cycle fetch
- 2 instruction fetch streams

New Decode Advances

OpCache Storage

- 33% more entry associativity (16-way)
- Dense entries store 6 instructions(fused)
- 2 OC pipes x 6 inst/pipe =>12 inst/cycle

Dual Decode Pipes

- 2 pipes support parallel independent instruction streams
- 4 inst/cycle throughput per pipe
- SMT mode gives each thread a pipe

8-wide dispatch to Int and FP

Wider Dispatch and Execute

8-wide dispatch, rename, retire

Integer scheduler advances

- Unified with age matrix
- More symmetry, simplifying pick

6 ALUs with 3 multipliers, 3 branch units

4 AGUs feed a wider LS with 4 memory addresses per cycle

Execution window growth

- Scheduler growth 88 ALU/56 AGU
- 240 entry physical register file
- ROB 448 entries

Increase Data Bandwidth

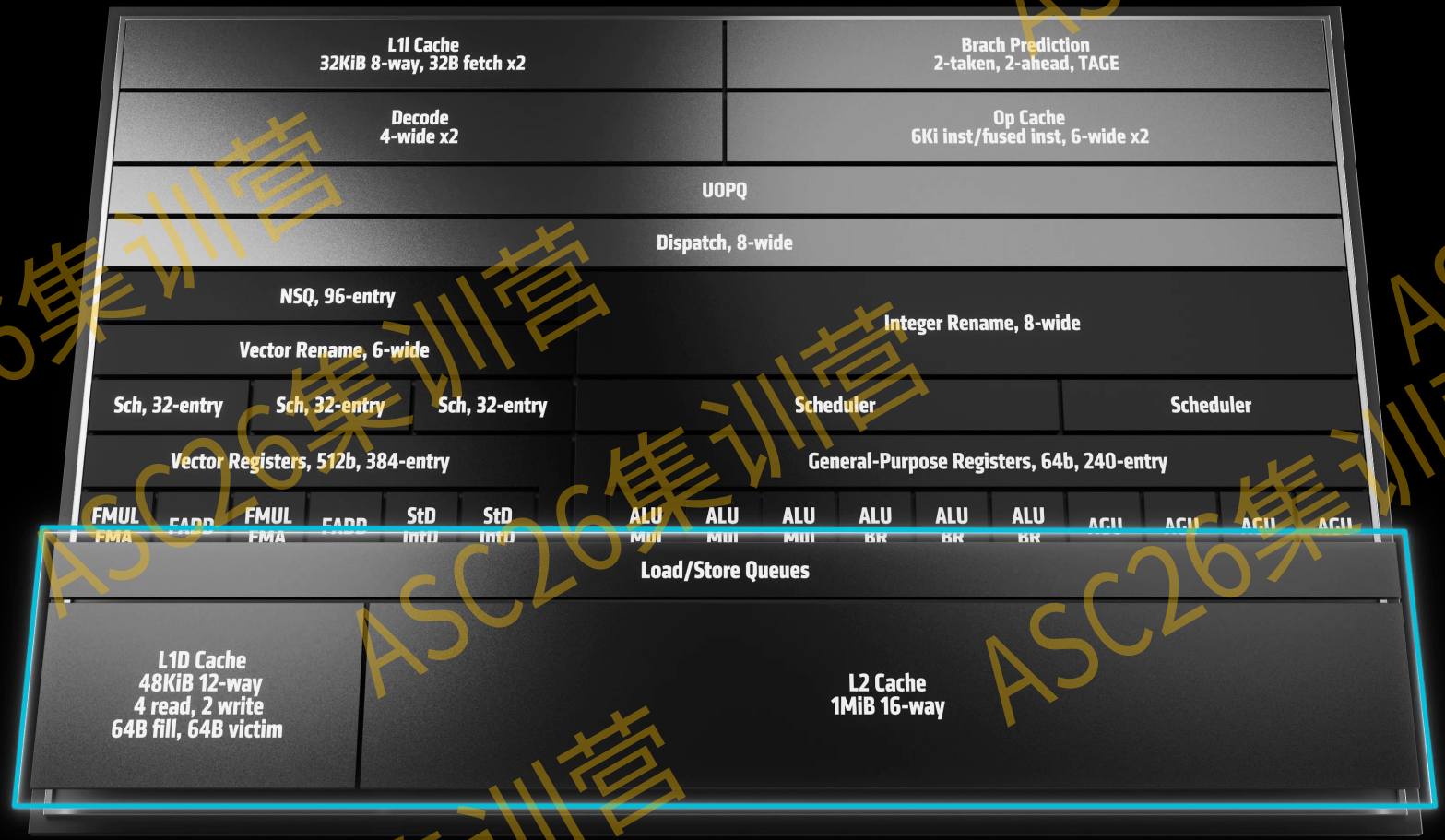
48KB 12-way L1D *keeping 4-cycle load-to-use*

More Bandwidth
- 4 LS pipes for a mix of 4 loads/2 stores per cycle
- 4 Integer load pipes that support 2 FP Pipes
- 2 store commit per cycle
- 64B fill/victim from/to L2 Dcache

TLBs
- L1: 96 entry Fully associative all page size DTLB
- L2: 4K DTLB everything but 1G

Larger In-Flight Window
- Load and store queue growth
- Store coalescing buffer growth
- Scalable load ordering queue

Data prefetching
- New 2D stride prefetcher improves stream and region prefetchers
- Extends workload pattern recognition



Increased FP Capability

FP major features/changes

- AVX512 with full 512b datapath

More bandwidth, less latency

- 4 execution pipelines
- 2 LS/integer register pipelines
- 2 512b loads/cycle, 1 512b store/cycle
- 2-cycle FADD

Execution window growth

- NSQ growth with 8-wide dispatch
- 3 larger schedulers: 1/pipe pair
- Physical register file doubles
- ROB/retire queue growth

Zen Software Studio

AMD Zen Software Studio

AMD Optimizing C/C++ and Fortran Compilers (AOCC)

A high-performance x86 CPU compiler designed for the C, C++, and Fortran programming languages.

AMD Optimizing CPU Libraries (AOCL)

A set of numerical libraries providing routines for various mathematical and scientific computations.

Zen Deep Neural Network (ZenDNN)

A library offering APIs for neural network building blocks to enhance deep learning inference performance.

AMD μProf

Tools that assist developers in profiling and optimizing software for performance and power efficiency.

Python Libraries with AOCL

AMD distributes and supports wheel files for selected Python libraries ensuring optimal performance on AMD Zen processors.

Pre-Built Applications

AMD Zen HPL optimized for AMD EPYC™ processors

AMD Zen HPL-MxP optimized for AMD “Zen4/Zen5”-based processors

AMD Zen HPCG optimized for AMD EPYC™ processors

AMD Zen STREAM for AMD “Zen4/Zen5”-based processors

AMD optimized Spack recipe for HPC workloads

Learn more at https://www.amd.com/en/developer/zen-software-studio.html

AMD Optimizing C/C++ and Fortran Compilers (AOCC)

Platform and Foundation

- Built for 32-bit and 64-bit Linux® platforms, leveraging the robust LLVM™ infrastructure (based on LLVM 17.0.6, Nov 2023) with Clang as the default front-end for C/C++ and Flang for Fortran. For more information, see the AOCC 5.1 Release Notes.

Architecture Optimization

- Tuned for AMD processors across all “Zen” generations (Zen, Zen2, Zen3, Zen4, Zen5).

Language Standards and Compliance

- Supports C17 (default for C), C++17 (default for C++), and Fortran F2008 with Real128 features (coarrays not supported); OpenMP 5.0 for C/C++ and OpenMP 4.5 for Fortran; DWARFv4 debugging by default with DWARFv5 available for C, C++, and Fortran.

Advanced Features and Optimizations

- Includes inter/intraprocedural analysis, SLP and loop vectorization, loop optimizations, OpenMP Debugging Interface (OMPD), and CPU offload support for Fortran OpenMP.

Additional Enhancements

- Optimization level -O2 and -fPIC/-fPIE options are default since 4.1; supports Spack for flexible package management and AMD optimized application recipes.

AMD Optimizing CPU Libraries (AOCL)

AOCL-BLAS

- a portable software framework for performing high-performance Basic Linear Algebra Subprograms (BLAS) functionality.

AOCL-Compression

- a software framework of various lossless data compression and decompression methods tuned and optimized for AMD “Zen”-based CPUs.

AOCL-Cryptography

- AMD’s optimized implementation of cryptographic functions.

AOCL-DA

- a data analytics library providing optimized building blocks for data analysis and classical machine learning problems.

AOCL-DLP

- a high-performance library that provides optimized deep learning primitives for AMD processors.

AOCL-FFTW

a comprehensive collection of fast C routines for computing the Discrete Fourier Transform (DFT) and various special cases.

AOCL-FFTZ

- AMD’s in-house Fast Fourier Transform (FFT) library designed, developed and optimized for AMD “Zen”-based processors. It offers high performant FFT routines for HPC, scientific computing and many such applications.

AOCL-LAPACK

- a portable library for dense matrix computations that provides the functionality present in the Linear Algebra Package (LAPACK).

AOCL-LibM

- a software library containing elementary math functions optimized for x86-64 processor based machines.

AOCL-LibMem

- AMD’s optimized implementation of memory manipulation functions for AMD “Zen”-based CPUs.

AOCL-RNG

- a library that provides a set of pseudo-random number generators, quasi-random number generator and statistical distribution functions optimized for AMD “Zen”-based processors.

AOCL-ScaLAPACK

- a library of high-performance linear algebra routines for parallel distributed memory machines. It depends on external libraries including BLAS and LAPACK for linear algebra computations.

AOCL-SecureRNG

- a library that provides APIs to access the cryptographically secure random numbers generated by the AMD hardware random number generator.

AOCL-Sparse

- a library containing the basic linear algebra subroutines for sparse matrices and vectors optimized for AMD “Zen”-based CPUs.

AOCL-Utils

- a library which provides APIs to check the available CPU features/flags, cache topology, and so on of AMD “Zen”-based CPUs.

AMD ZenDNN (Optimized DNN Acceleration Inference Library for Zen)

Take Performance to the Next Level

Open-Source Ecosystem
License-free and accessible software
Avoid vendor lock-ins

Friction-less Adoption
Purposely designed to leverage existing customer code with zero-to-minimal changes

Performance Boost
On-going SW Roadmap of performance optimization and feature enablement

AMD μProf

CPU Profiling

Profile applications to analyze performance bottlenecks. In-depth analysis at all levels of application: process, thread, module, function, source line, instruction.
Supports call-stack sampling, IMIX reporting. On FreeBSD, only EBP profiling supported.

Profile Types
- TBP – Software timer-based sampling
- EBP – CPU event-based sampling
- IBS – CPU IBS MSR based sampling
- Custom – Choose events from TBP, EBP, IBS to create specialized custom type.

Profile Scopes: Per-process, System-wide

OS: Windows, Linux, FreeBSD

Languages: C, C++, Fortran, Java, .NET, Assembly

CLI option syntax:
```
collect [--config <type> | -e <event-specifier>] [<PROGRAM>] translate -i <session-dir path>
report [--view-config <type>]
profile [--config <type> | -e <event-specifier>] [<PROGRAM>]
diff --baseline <base path> --with <non-base path> -o <out-dir>
info [--help] [options]
```

Supported Predefined Configs list (Refer Appendix-E):
$ AMDuProfCLI info --list collect-configs

Supported Predefined Events list:
$ AMDuProfCLI info --list predefined-events

Supported PMU (EBP) Events list:
$ AMDuProfCLI info --list pmu-events

Supported View Configs list:
$ AMDuProfCLI info --list view-configs

Collect command example:
$ AMDuProfCLI collect --config assess -o /tmp/cpuprof-assess ./classic-app

Report command examples:
$ AMDuProfCLI report -i /tmp/cpuprof-assess/AMDuProf-classic-TBP_Dec-09-2024_12-19-27

Profile command example:
$ AMDuProfCLI profile --config tbp -o /tmp/cpuprof-tbp ./classic-app

Power Profiling

Profile the system for power, thermal, and frequency characteristics.
Report is generated on-the-fly by the Timechart command.

OS: Windows, Linux

CLI option syntax:
AMDuProfCLI timechart [<options>] <program> [<args>]

Supported Counter Categories list:
$ AMDuProfCLI timechart --list

Timechart command example:
$ AMDuProfCLI timechart -e power -t 500 -d 10 -o /tmp/PowerOutput/

AMDuProfPcm (SYSTEM ANALYSIS)

Monitor basic performance metrics. Collects CPU Core, L3, DF perf-counters, and reports of various metrics periodically.

OS: Windows, Linux, FreeBSD

>> Perf mode (rootless) on Linux: Monitor using perf subsystem. Default mode of monitoring.
>> MSR mode (root) on Linux: Monitor using msr module. Select this mode by adding "--msr" to the command.

Prerequisite: Refer Appendix-E

CLI option syntax:
AMDuProfPcm [<COMMANDS>] [<OPTIONS>] -- <PROGRAM> [<ARGS>]

COMMANDS:
roofline [<OPTIONS>] -- <PROGRAM> [<ARGS>] top [<OPTIONS>]
compare <SESSION-PATH1>,<SESSION-PATH2> hreport <SESSION-PATH>
profile [<OPTIONS>] -- <PROGRAM> [<ARGS>]

Help & list of supported metrics:
$ AMDuProfPcm -h

Collect command example:
$ AMDuProfPcm -O /tmp/ -a -d 60
$ AMDuProfPcm -m ipc,l2,l3 -c core=0 -d 60 -O /tmp/
$ AMDuProfPcm -m memory -c core=0 -O /tmp/ -- ./myapp.exe

Collect command example MSR mode:
$ AMDuProfPcm --msr -m ipc -O /tmp/ -- ./myapp.exe

AMDuProfSys (SYSTEM ANALYSIS)

Python-based system analysis tool. Collects data from CPU Core, L3, DF, UMC, and HSMP counters. This tool helps collect hardware events and evaluate simple counter values or complex metrics using collected raw events.

OS: Linux, Windows

Prerequisite (on Linux):
Uses uProf driver / Linux perf subsystem, and basic hardware access primitives. Refer Appendix-D.

Prerequisite (on Windows):
Uses uProf driver to collect PMC events. AMDuProf supplied driver must be installed (Windows installer takes care of the driver installation).

Perf Selection (on Linux):
By default, uses uProf driver for data collection. To use the perf subsystem, command line must include the option --use- linux-perf.

CLI option syntax:
AMDuProfSys collect –config <CONFIG> <WORKLOAD>
AMDuProfSys report -i <SESSION-FILE>

Help command:
$ AMDuProfSys --help-all

Collection + Reporting command example:
$ AMDuProfSys --config core -C 0 -o output taskset -c 0 <application>

Custom metric collection command example:
$ AMDuProfSys --metrics core/BrMisPred="(0x4300C3)/(0x4300C2)",l3/L3CacheAccesses="(0x300C0000040FF04)",core/ratio="BrMisPredExTime/100" -d 20

Python Libraries with AOCL

AMD distributes and supports wheel files for selected Python libraries ensuring optimal performance on AMD Zen processors

Python Libraries

- Enhanced PyTorch performance with AOCL
- Improved NumPy and SciPy performance with AOCL
- New Python Package: NumExpr with AOCL-LibM
- SciPy Sparse Extension with AOCL

Highlights

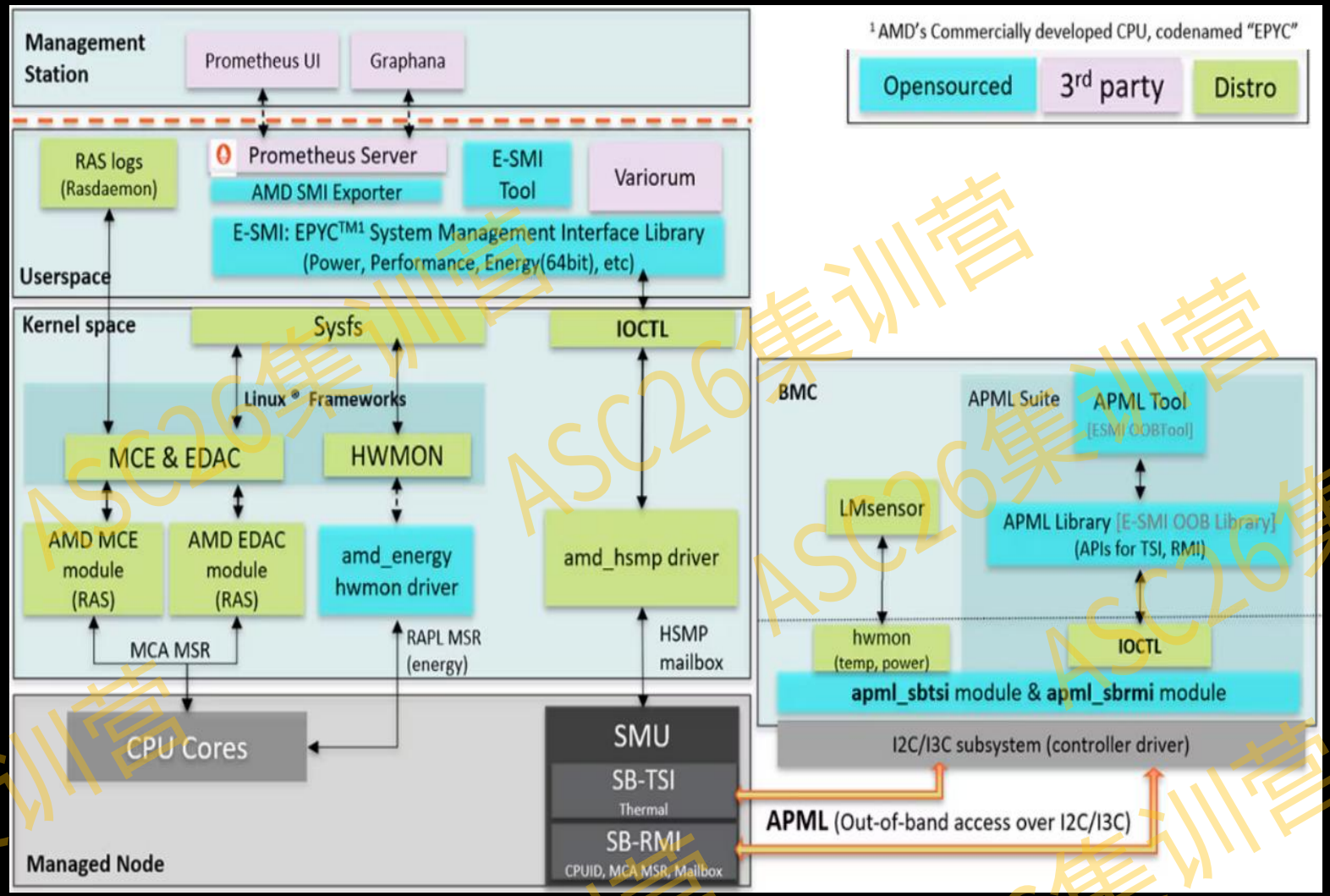
- PyTorch 2.5.0 and 2.4.0
- NumPy 2.1.3 and 1.26.4
- SciPy 1.14.1 and 1.13.1
- NumExpr 2.11.0
- SciPy Sparse Extension with AOCL 0.1.0 verified with above SciPy and NumPy versions

Release Notes

- AOCL-BLAS, AOCL-LAPACK, AOCL-LibM, AOCL-Sparse are compiled with GCC-13.3.1 and AOCC-5.1
- Python Packages are compiled with GCC-13.3.1 and AOCC-5.1
- Compatible for OS with GLIBC 2.28 or later
- Supported Python versions: Python 3.13 (new), Python 3.12, Python 3.11

EPYC™ System Management Software (E-SMS)

EPYC™ System Management Software (E-SMS)



Library lifecycle & support	• Initialization and Shutdown • Auxiliary functions • Test HSMP mailbox
Power & energy (monitor + control)	• Energy Monitor (RAPL MSR) • Power Monitor • Power Control
Performance / frequency (monitor + control)	• Performance (Boost limit) Monitor • Performance (Boost limit) Control
Memory (DDR / DIMM) telemetry	• ddr_bandwidth Monitor • Dimm statistics
Thermal / temperature	• Temperature Query
Fabric / interconnect control	• xGMI bandwidth control • GMI3 width control • APB and LCLK level control
Bandwidth & metrics aggregation	• Bandwidth Monitor • HSMP System Statistics • Metrics Table
Logging Support	• LoopWatch • Logger (CSV format) • JSON/CSV Format summary

Learn more at https://www.amd.com/en/developer/e-sms.html

AMD Technical Resources for Developers

Developer Central

- Public resources for AMD Products
- https://www.amd.com/en/developer.html

Technical Information Portal

- Technical Resources (Docs, Tools, SWs) for AMD Products
- https://docs.amd.com

Zen Software Studio

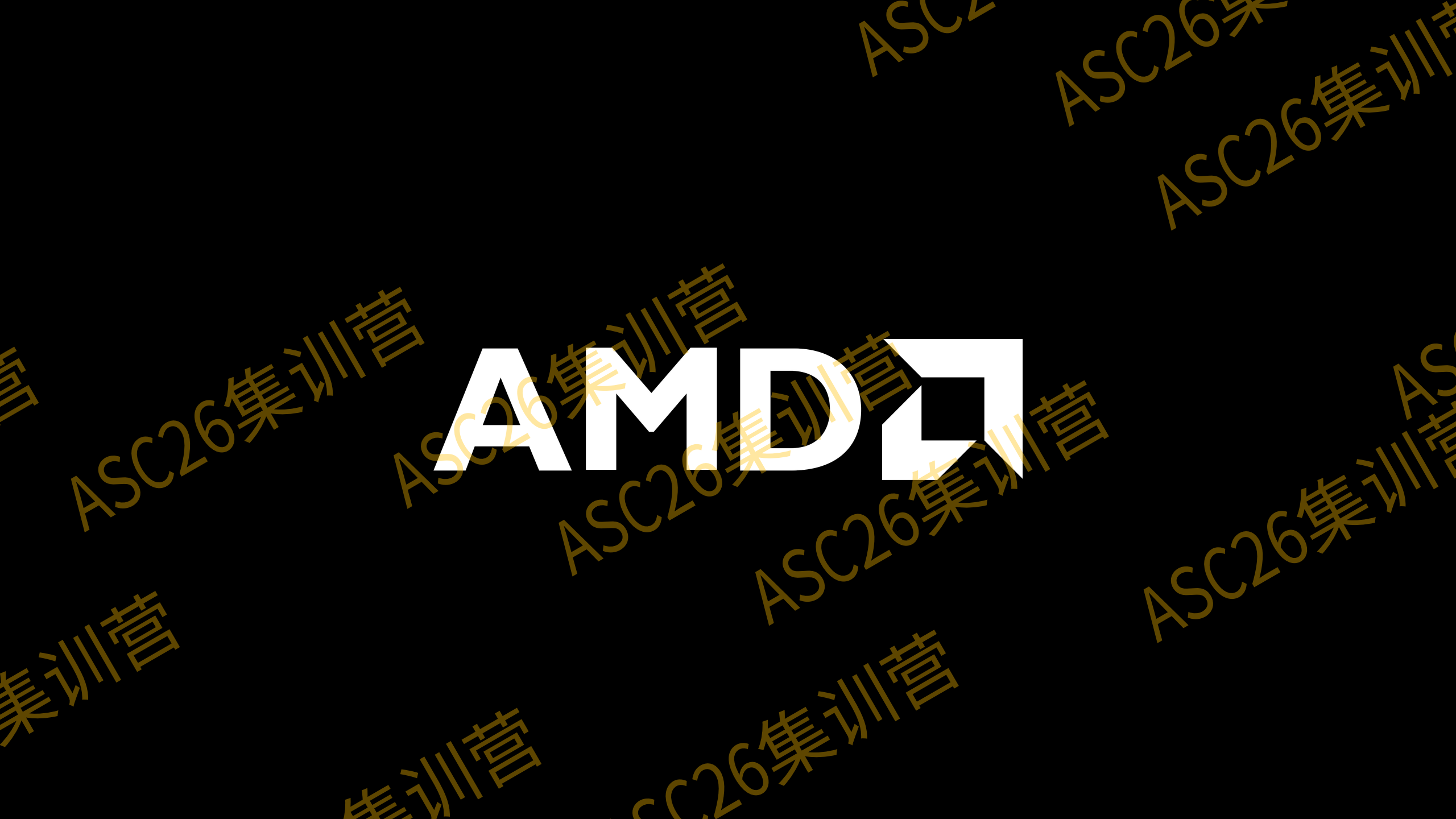
- A comprehensive suite of development tools and libraries designed to optimize software performance on AMD processors utilizing the "Zen" core architecture.
- https://www.amd.com/en/developer/zen-software-studio.html

EPYC™ System Management Software (E-SMS)

- Comprises of kernel modules, user space libraries, and tools to manage power, performance aspects through In-Band and Out-of-Band of the AMD EPYC server CPUs.
- https://www.amd.com/en/developer/e-sms.html

EPYC™ Processors Minimum Operating System (OS) Versions

- https://www.amd.com/en/products/processors/server/epyc/minimum-operating-system.html


AMD
ASC26集训营

Disclaimers and attributions

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. GD-18

Endnotes 1-8

[1] Reported data includes Scope 1 and 2 GHG emissions (base year 2020). Based on AMD calculations that are third-party verified (limited level assurance)

[2] Includes AMD high-performance CPU and GPU accelerators used for AI training and High-Performance Computing in a 4-Accelerator, CPU hosted configuration. Goal calculations are based on performance scores as measured by standard performance metrics (HPC: Linpack DGEMM kernel FLOPS with 4k matrix size. AI training: lower precision training-focused floating-point math GEMM kernels such as FP16 or BF16 FLOPS operating on 4k matrices) divided by the rated power consumption of a representative accelerated compute node including the CPU host + memory, and 4 GPU accelerators.

[3] EPYC-030c: Calculation includes 1) base case kWhr use projections in 2025 conducted with Koomey Analytics based on available research and data that includes segment specific projected 2025 deployment volumes and data center power utilization effectiveness (PUE) including GPU HPC and machine learning (ML) installations and 2) AMD CPU and GPU node power consumptions incorporating segment-specific utilization (active vs. idle) percentages and multiplied by PUE to determine actual total energy use for calculation of the performance per Watt. 38x is calculated using the following formula: (base case HPC node kWhr use projection in 2025 * AMD 2025 perf/Watt improvement using DGEMM and TEC +Base case ML node kWhr use projection in 2025 *AMD 2025 perf/Watt improvement using ML math and TEC) /(Base case projected kWhr usage in 2025). For more information, www.amd.com/en/corporate/corporate-responsibility/data-center-sustainability.html.

[4] “Manufacturing Suppliers” are defined as suppliers that AMD buys from directly and that provide direct materials and/or manufacturing services to AMD

[5] AMD calculations are third-party verified (limited level assurance) based on data supplied by our Manufacturing Suppliers which is not independently verified by AMD.

[6] AMD defines renewable energy as energy from a source that is not depleted when used, such as wind or solar power. AMD does not require a minimum amount of renewable energy to be sourced by Manufacturing Suppliers to be included in the goal. Data is provided by AMD suppliers and has not been independently verified by AMD.

[7] AMD estimated the number of racks to train a typical notable AI model based on EPOCH AI data (https://epoch.ai). For this calculation we assume, based on these data, that a typical model takes 1025 floating point operations to train (based on the median of 2025 data), and that this training takes place over 1 month. FLOPs needed = 10^25 FLOPs/(seconds/month)/Model FLOPs utilization (MFU) = 10^25/(2.6298*10^6)/0.6. Racks = FLOPs needed/(FLOPS/rack in 2024 and 2030). The compute performance estimates from the AMD roadmap suggests that 276 racks would be needed in 2025 to train a typical model over one month using the 2024 MI300X product (assuming 22.656 PFLOPS/rack with 60% MFU) and 276-fold reduction in the number of racks to train the same model over this six-year period. Electricity use for a system to completely train a typical 2025 AI model using a 2024 rack is calculated at ~7GWh, whereas the future 2030 AMD system could train the same model using ~350 MWh, a 95% reduction. AMD then applied carbon intensities per kWh from the International Energy Agency World Energy Outlook 2024 [https://www.iea.org/reports/world-energy-outlook-2024]. IEA’s stated policy case gives carbon intensities for 2023 and 2030. We determined the average annual change in intensity from 2023 to 2030 and applied that to the 2023 intensity to get 2024 intensity (434 CO2 g/kWh) versus the 2030 intensity (312 CO2 g/kWh). Emissions for the 2024 baseline scenario of 7 GWh x 434 CO2 g/kWh ~= 3000 tonnes of CO2, versus the future 2030 scenario of 350 MWh x 312 CO2 g/kWh ~= 109 tonnes of CO2.

[8] AMD based advanced racks for AI training/inference in each year (2024 to 2030) based on AMD roadmaps, also examining historical trends to inform rack design choices and technology improvements to align projected goals and historical trends. The 2024 rack is based on the MI300X node, which is comparable to the Nvidia H100 and reflects current common practice in AI deployments in 2024/2025 timeframe. The 2030 rack is based on an AMD system and silicon design expectations for that time frame. In each case, AMD specified components like GPUs, CPUs, DRAM, storage, cooling, and communications, tracking component and defined rack characteristics for power and performance. Calculations do not include power used for cooling air or water supply outside the racks but do include power for fans and pumps internal to the racks. Performance improvements are estimated based on progress in compute output (delivered, sustained, not peak FLOPS), memory (HBM) bandwidth, and network (scale-up) bandwidth, expressed as indices and weighted by the following factors for training and inference. FLOPS HBM BW Scale-up BW Training 70.0% 10.0% 20.0% Inference 45.0% 32.5% 22.5% Performance and power use per rack together imply trends in performance per watt over time for training and inference, then indices for progress in training and inference are weighted 50:50 to get the final estimate of AMD projected progress by 2030 (20x). The performance number assumes continued AI model progress in exploiting lower precision math formats for both training and inference which results in both an increase in effective FLOPS and a reduction in required bandwidth per FLOP.

Endnotes

9xx5-001: Based on AMD internal testing as of 9/10/2024, geomean performance improvement (IPC) at fixed-frequency. - 5th Gen EPYC CPU Enterprise and Cloud Server Workloads generational IPC Uplift of 1.170x (geomean) using a select set of 36 workloads and is the geomean of estimated scores for total and all subsets of SPECrate®2017_int_base (geomean), estimated scores for total and all subsets of SPECrate®2017_fp_base (geomean), scores for Server Side Java multi instance max ops/sec, representative Cloud Server workloads (geomean), and representative Enterprise server workloads (geomean). “Genoa” Config (all NPS1): EPYC 9654 BIOS TQZ1005D 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-4800 (2Rx4 64GB), 32Gbps xGMI; “Turin” config (all NPS1): EPYC 9V45 BIOS RVOT1000F 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-6000 (2Rx4 64GB), 32Gbps xGMI Utilizing Performance Determinism and the Performance governor on Ubuntu® 22.04 w/ 6.8.0-40-generic kernel OS for all workloads. - 5th Gen EPYC generational ML/HPC Server Workloads IPC Uplift of 1.369x (geomean) using a select set of 24 workloads and is the geomean of representative ML Server Workloads (geomean), and representative HPC Server Workloads (geomean). “Genoa Config (all NPS1) “Genoa” config: EPYC 9654 BIOS TQZ1005D 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-4800 (2Rx4 64GB), 32Gbps xGMI; “Turin” config (all NPS1): EPYC 9V45 BIOS RVOT1000F 12c12t (1c1t/CCD in 12+1), FF 3GHz, 12x DDR5-6000 (2Rx4 64GB), 32Gbps xGMI Utilizing Performance Determinism and the Performance governor on Ubuntu 22.04 w/ 6.8.0-40-generic kernel OS for all workloads except LAMMPS, HPCG, NAMD, OpenFOAM, Gromacs which utilize 24.04 w/ 6.8.0-40-generic kernel. SPEC® and SPECrate® are registered trademarks for Standard Performance Evaluation Corporation. Learn more at spec.org.

9xx5-083A: 5th Gen EPYC processors support up to DDR5-6400 MT/s for targeted customers and configurations. DDR5-6400 support requires OEM enablement and a BIOS update from your server or motherboard manufacturer. Contact your system manufacturer prior to purchase to determine compatibility.