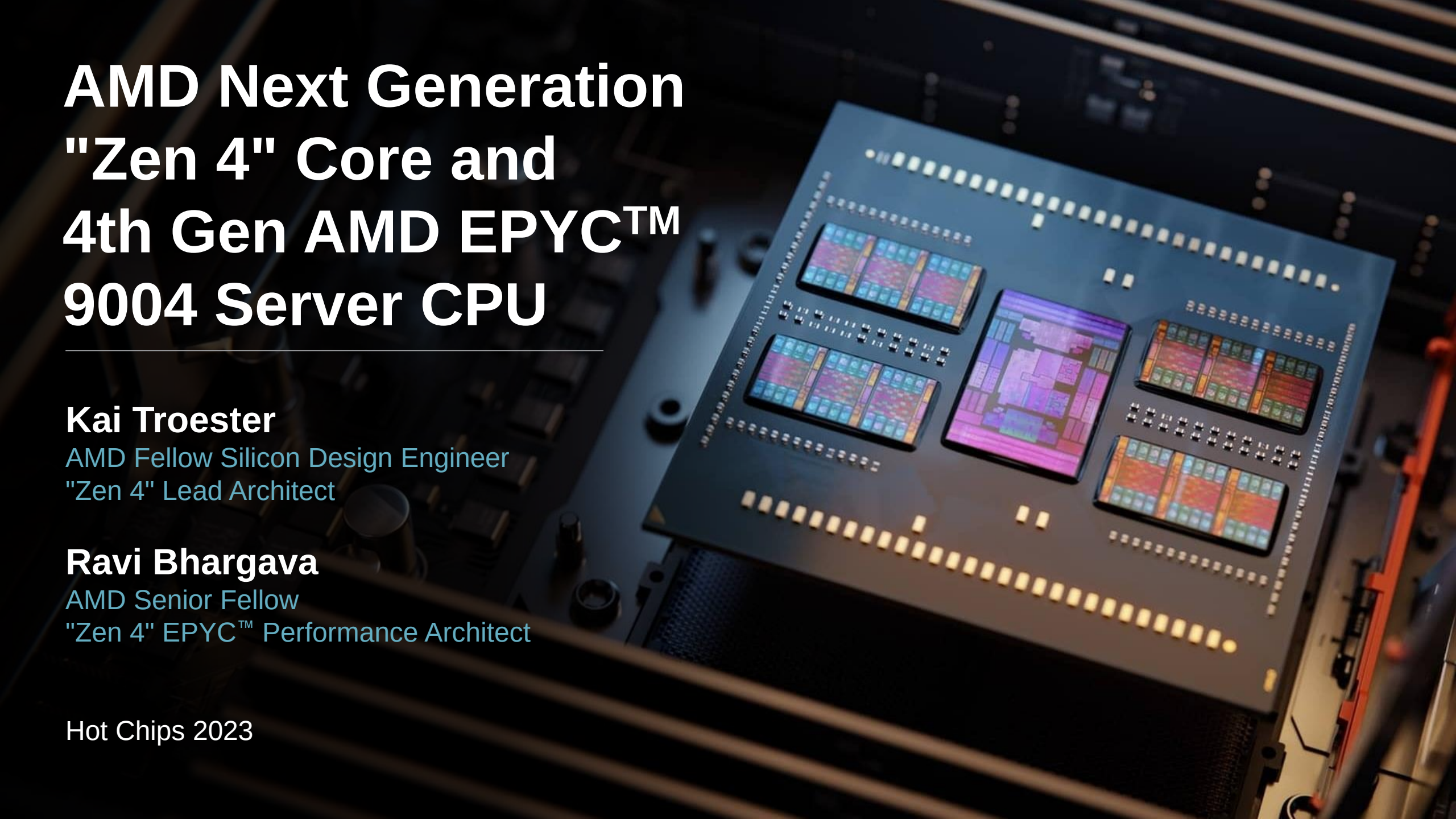
AMD Next Generation "Zen 4" Core and 4th Gen AMD EPYC™ 9004 Server CPU



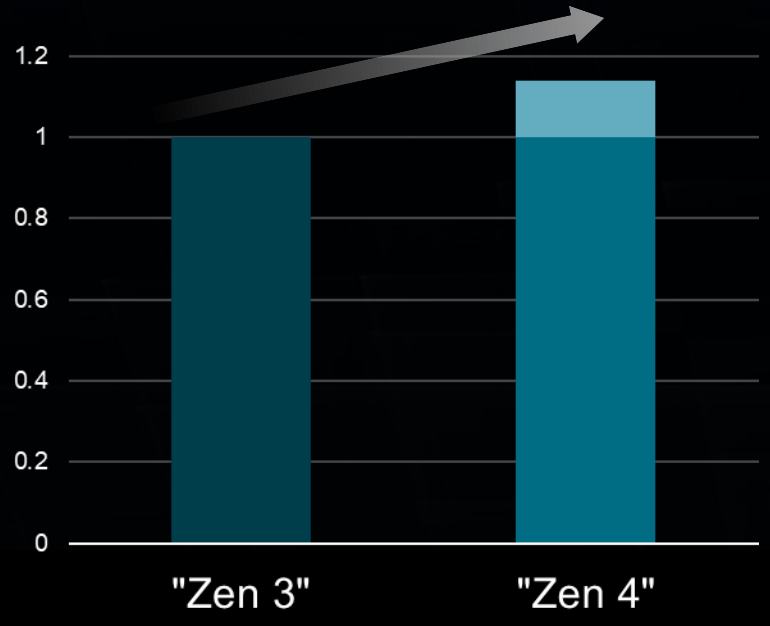
Kai Troester
AMD Fellow Silicon Design Engineer
"Zen 4" Lead Architect

Ravi Bhargava
AMD Senior Fellow
"Zen 4" EPYC™ Performance Architect

Hot Chips 2023

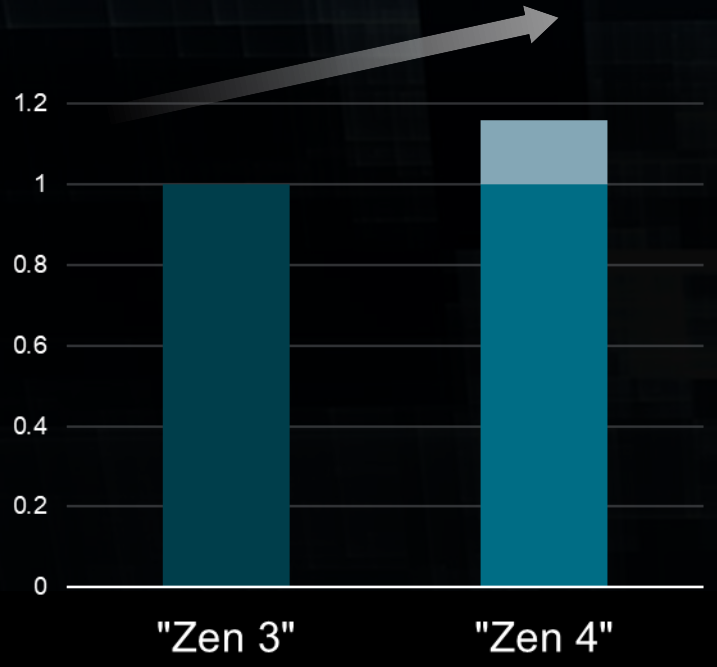
"Zen 4" Improves on Successful "Zen 3" Microarchitecture

~14% higher IPC



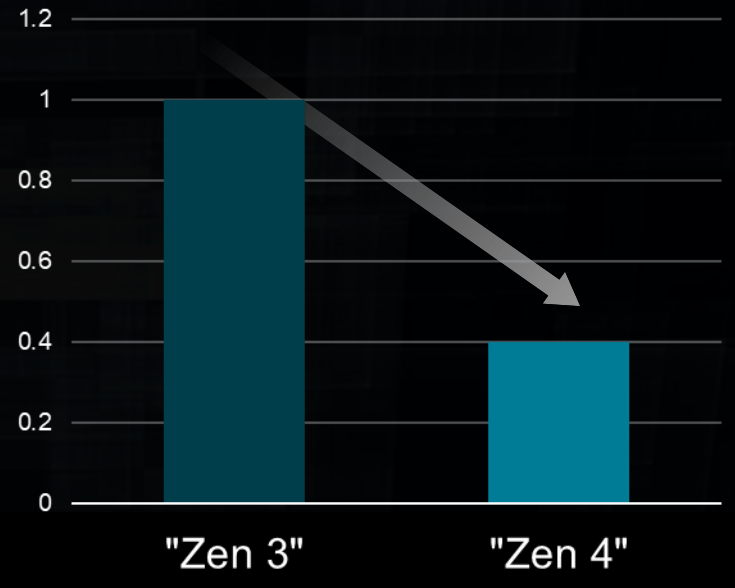
Microarchitectural improvements

+16% higher frequency at same voltage



5nm technology and frequency improvements beyond technology scaling

~60% lower power at same performance



5nm technology and microarchitectural and physical design improvements

See endnotes: EPYC-038, 1, 2

"Zen 4" Fetch Engine

Improved Branch Prediction accuracy

- L1-BTB: +50% more branches
- L2-BTB: +8% more branches
- Larger and improved Conditional Branch Predictor

Improved fetch bandwidth

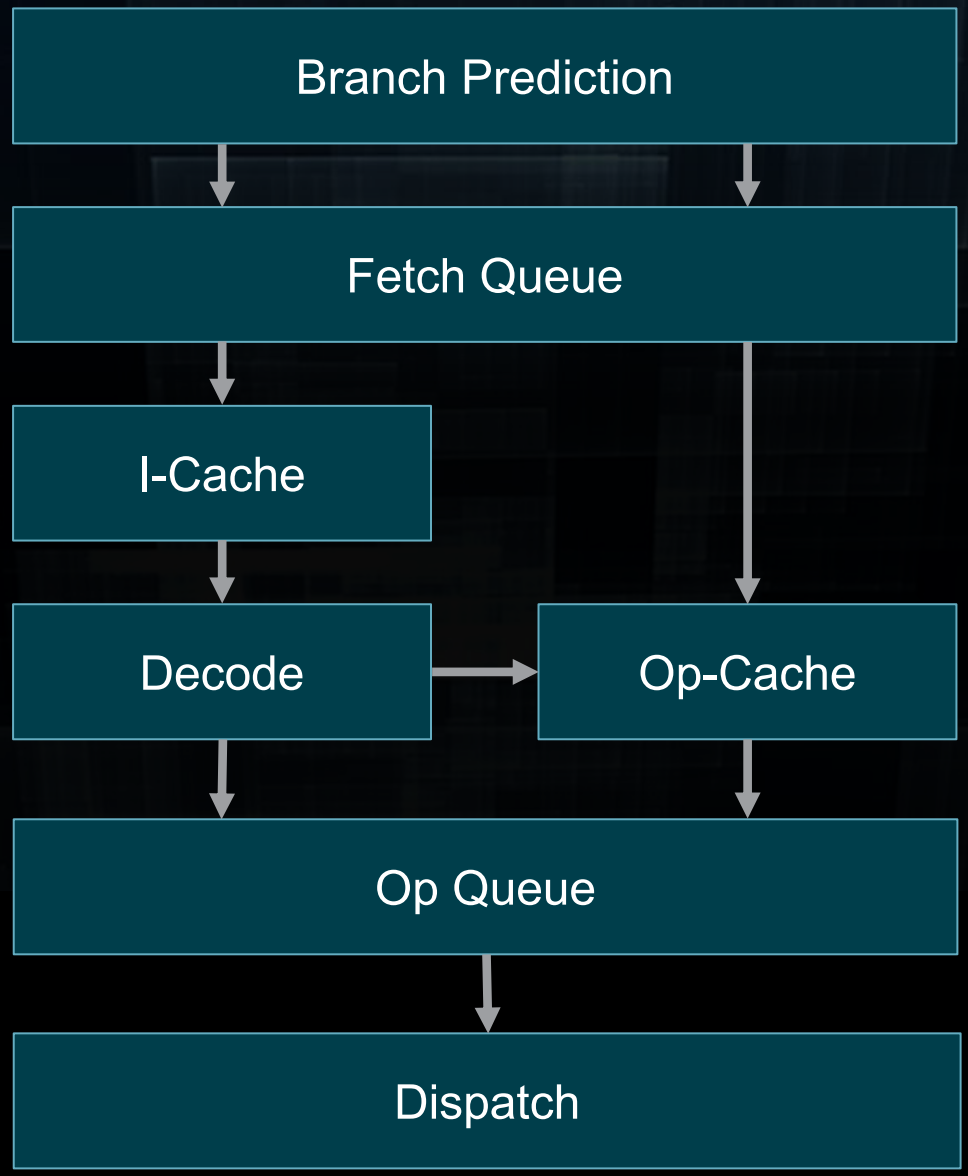
- Up to 2 taken branch predictions per cycle

Improved dispatch bandwidth

- Up to 2 taken branches
- Optimized integer scheduler queue assignment

Larger Op Cache for improved redirect latency and power

- 50% more entries
- max ops per entry increased from 8 to 9



"Zen 4" Execution Engine

Deeper

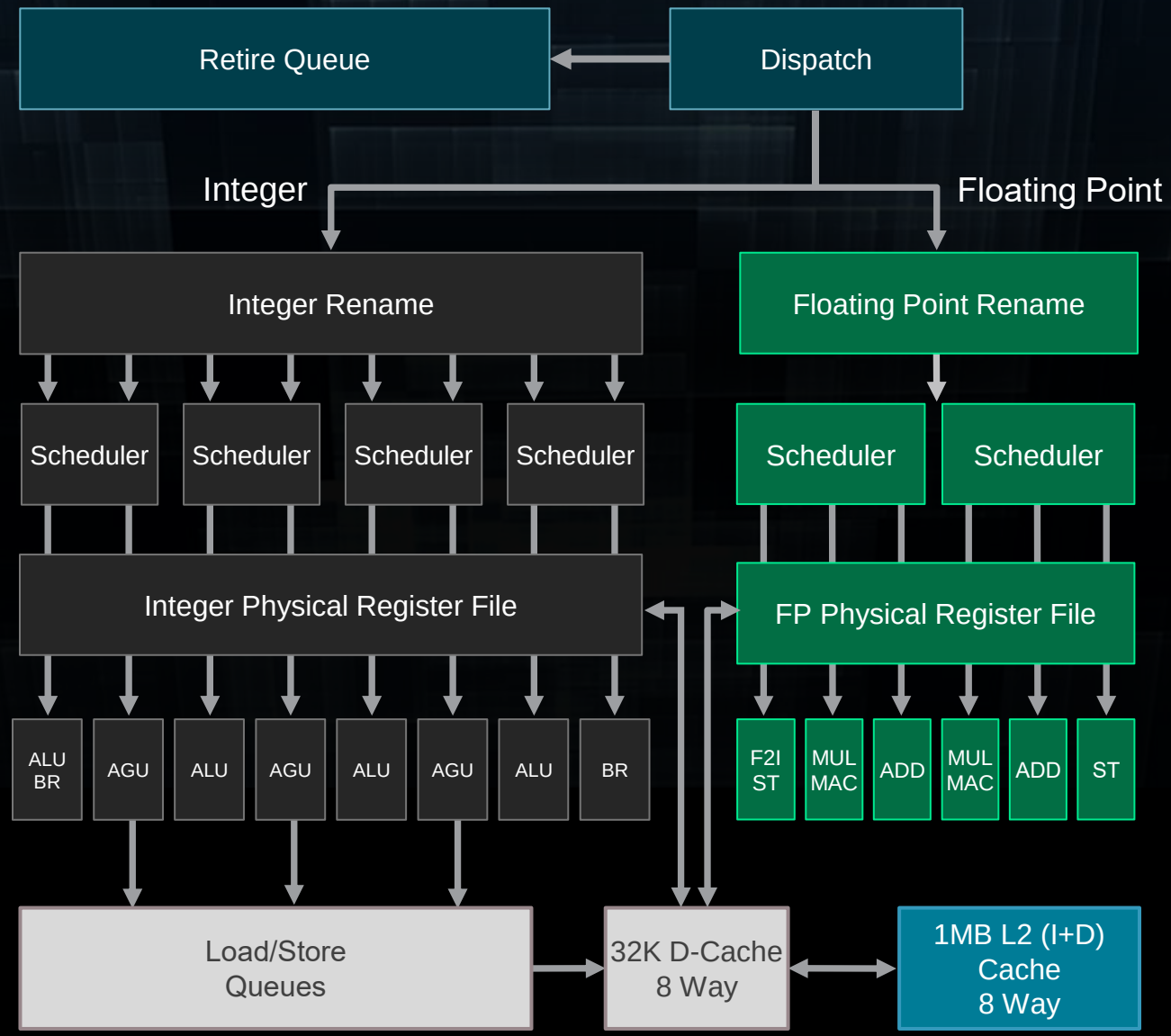
- Larger Retire Queue
- Larger Load and Store Queues
- Larger Integer and FP register files

Faster

- Bit Manipulation Instruction latency improvements
- Byte-writable L1 D-Cache for fewer port conflicts
- Larger L1 and L2 DTLB

New Instructions

- Power efficient AVX-512 support



"Zen 4" Cache

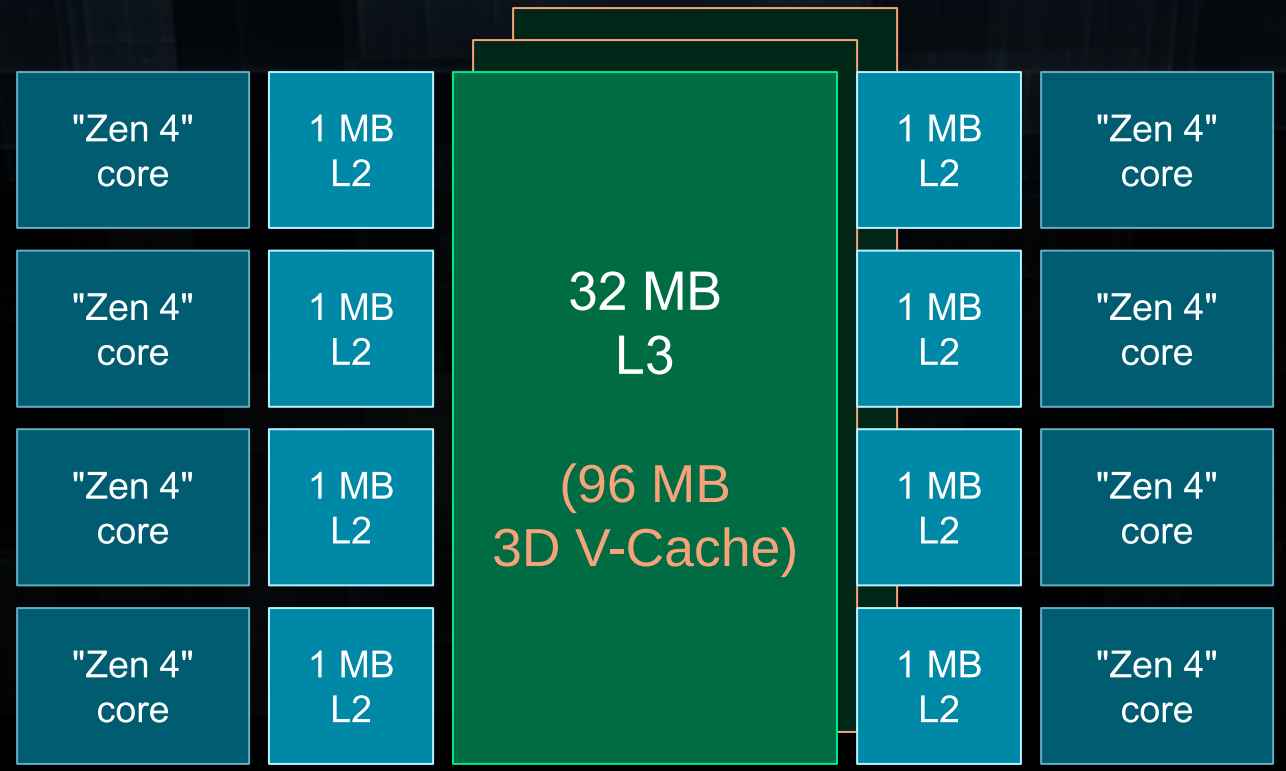
More Cache capacity

- 2x L2 (1 MB) cache capacity over "Zen 3"

More outstanding miss requests

- for increased memory parallelism

AMD 3D V-Cache™ technology allows
up to 96 MB of L3 cache per core complex



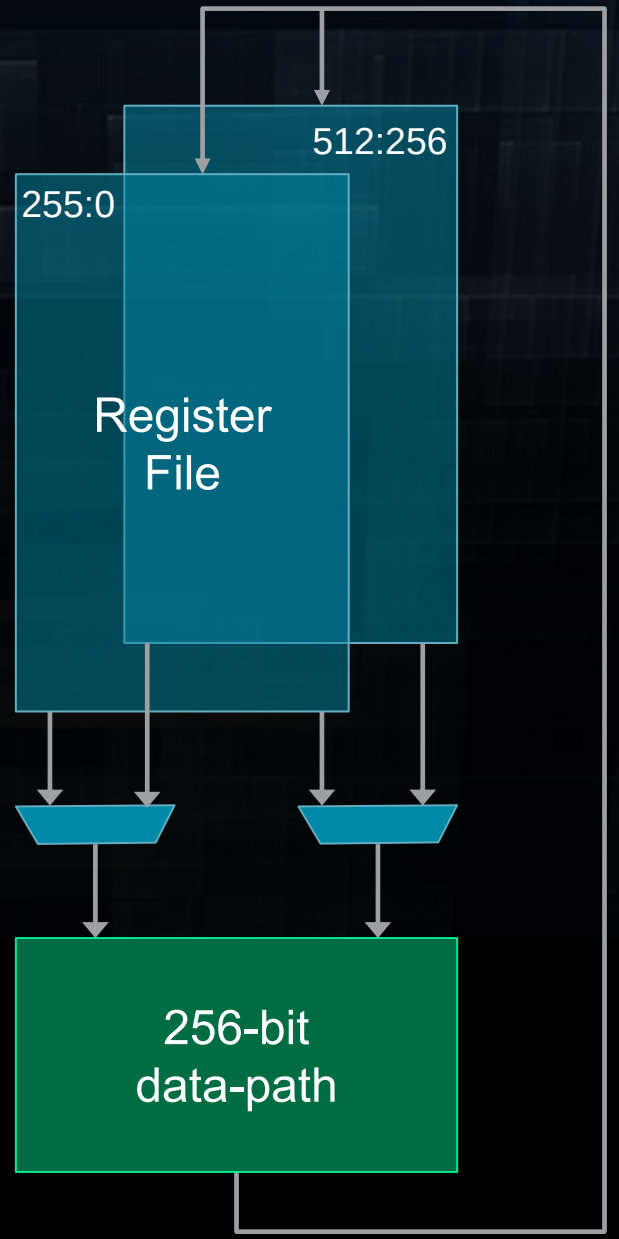
AVX-512

AVX-512 delivers

- Per lane masking capabilities for low overhead code vectorization
- 2x architectural registers for reduced register pressure
- Improved code density
- New instructions for application acceleration
  - VNNI – Vector Neural Network Instructions (Machine Learning)
  - GFNI – Galois Field New instruction (Cryptography)
  - BFLOAT16 – shorter FP data format (Machine Learning)
  - Improved Scatter/Gather instructions (including prefetch)

AVX-512 and "Zen 4"

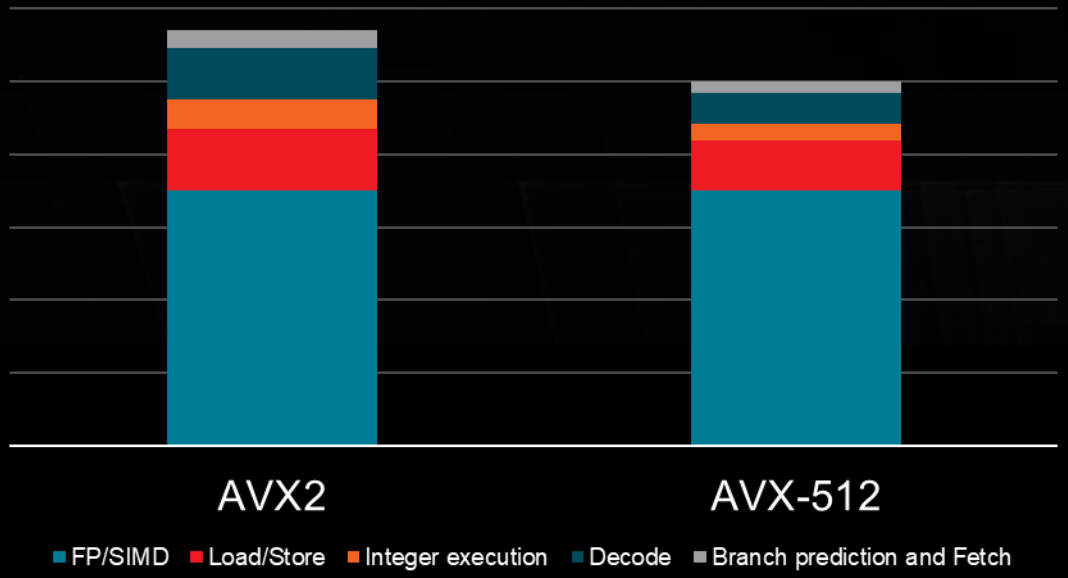
- Utilizes 256-bit data-path
- Most 512-bit ops execute over 2 consecutive cycles
- Minimal physical register file growth
- Power efficient implementation



Power Efficient AVX-512

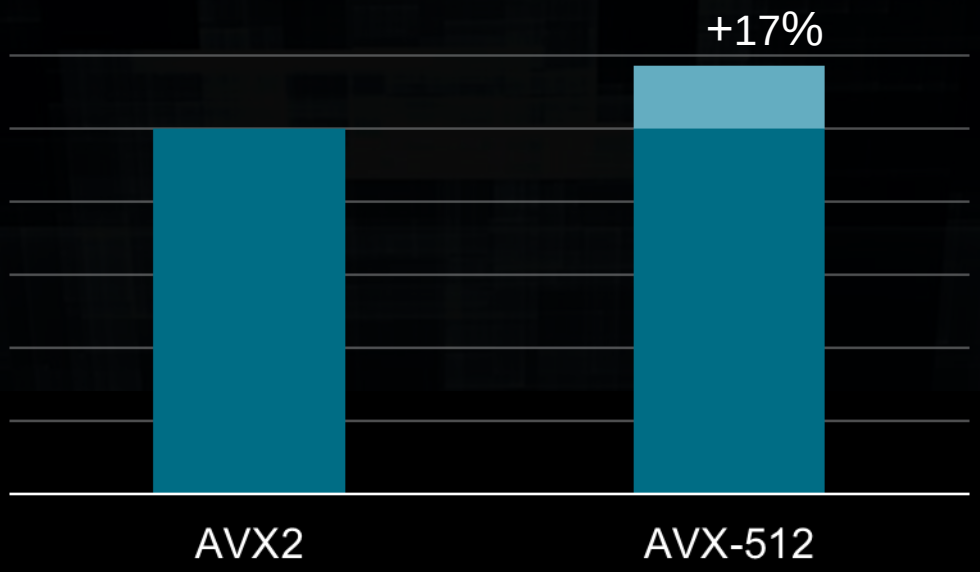
- With AVX-512, the "Zen 4" core processes up to 50% less instructions compared to 256-bit AVX2
- Reduced fetch, decode and out-of-order tracking reduce power consumption of AVX-512 code
- Lower power -> more effective frequency in power limited systems

Relative power consumption of a FP intensive kernel with AVX2 and AVX-512 instructions



for illustrational purposes only

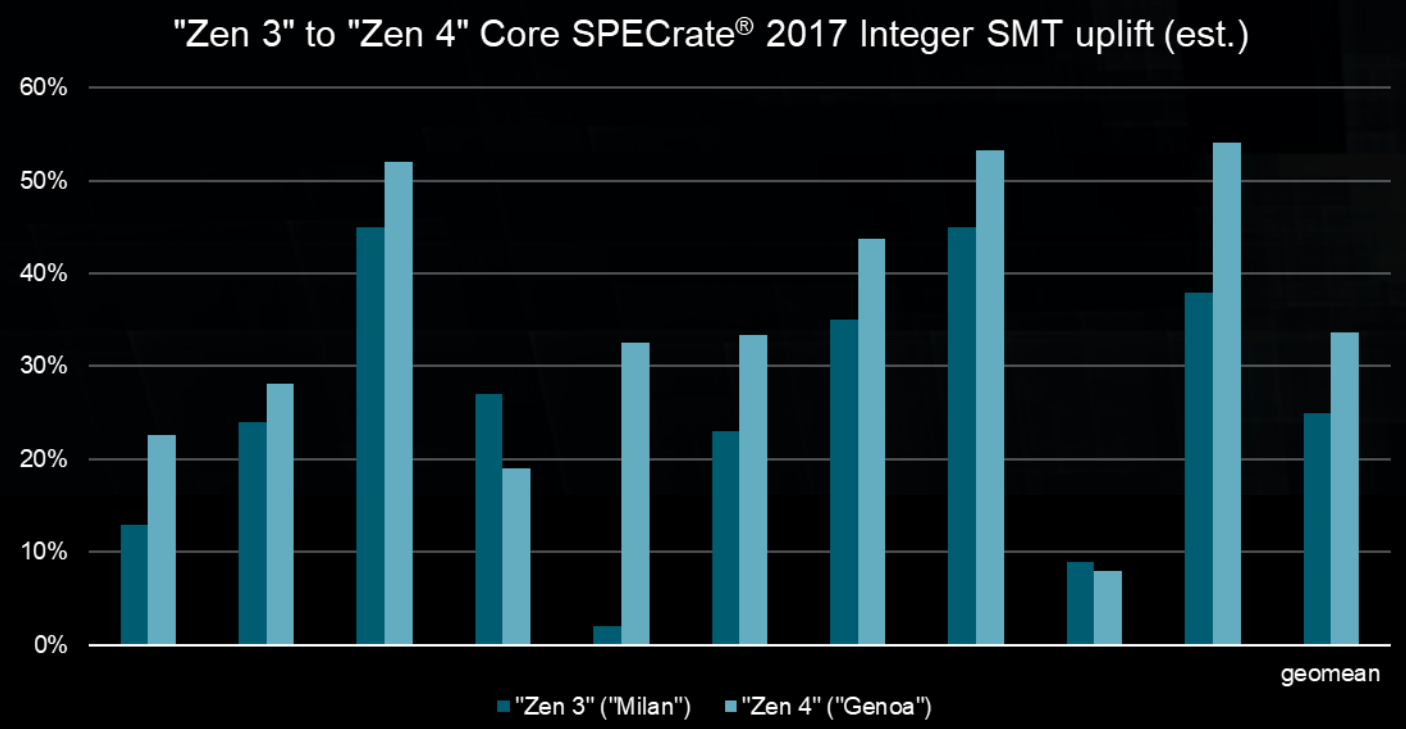
Effective Frequency gain for HPL using AVX-512 instead of AVX2



See endnote: 3

"Zen 4" Delivers More SMT Uplift

- Wider and deeper "Zen 4" micro-architecture delivers higher SMT uplift
- SMT uplift for "Zen 4" increased to 34% compared to 25% for "Zen 3"
- SMT provides flexibility



High Single Thread
performance for compute intensive applications

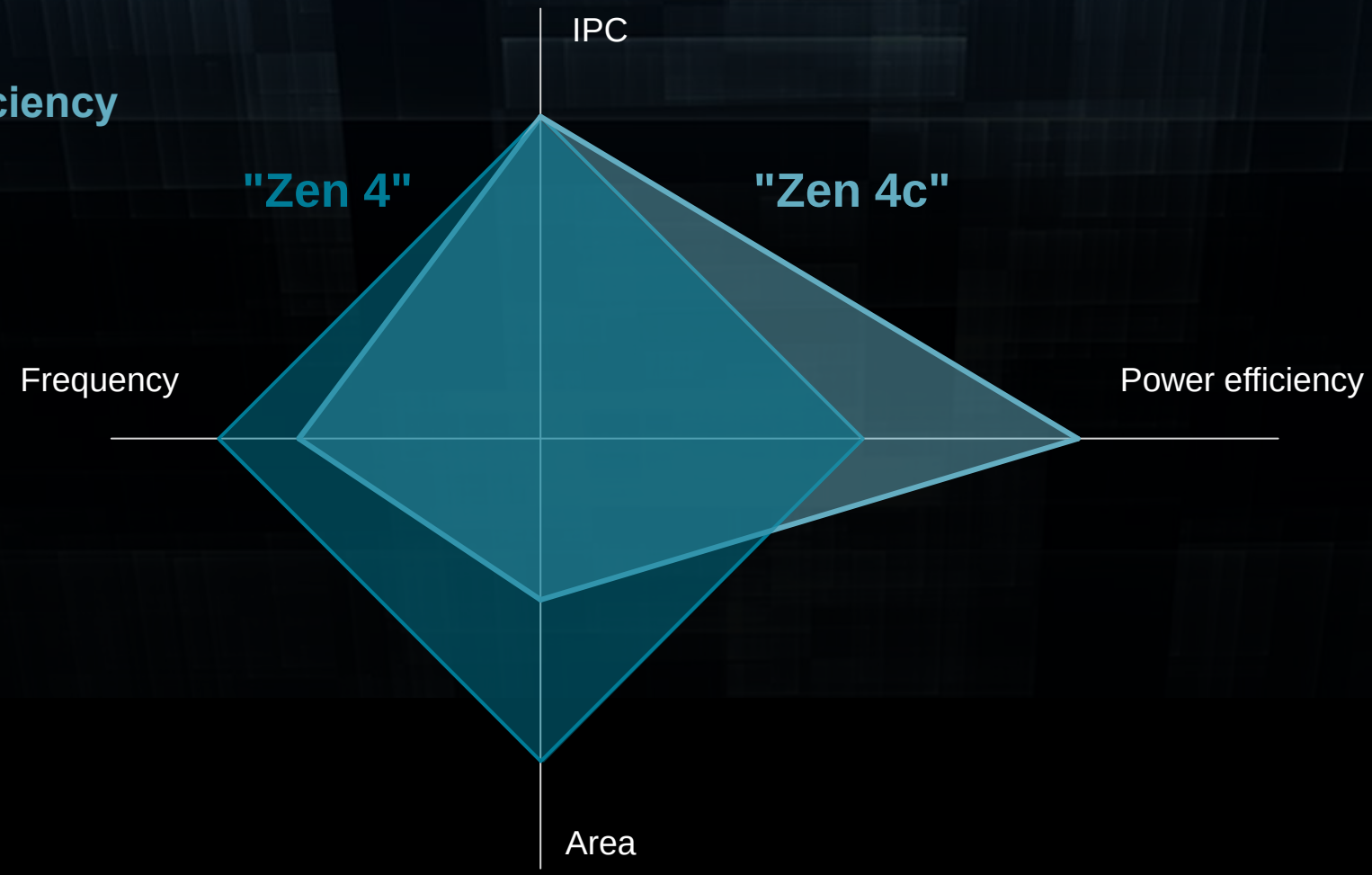
High Throughput
performance for highly threaded applications

See endnote: 4

"Zen 4c" – a Compact "Zen 4"

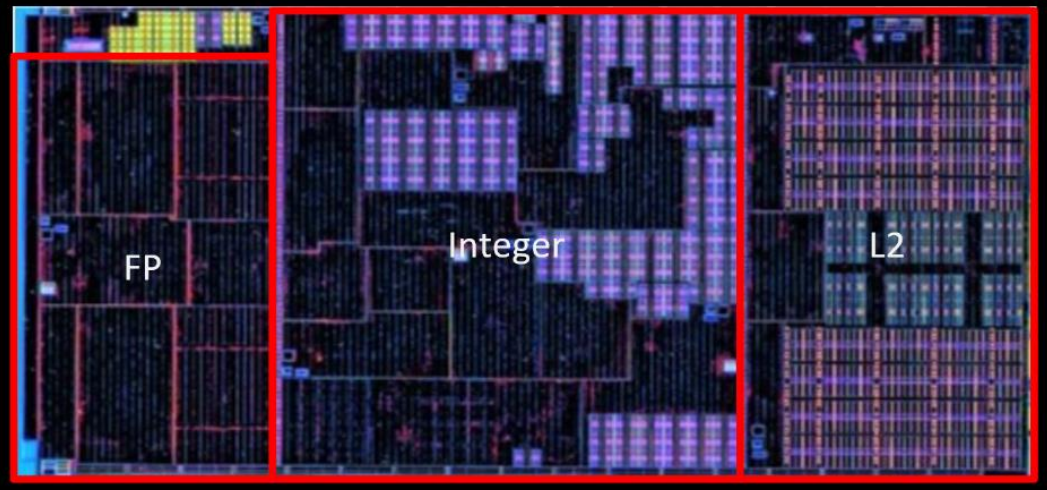
Optimized for density and power efficiency

- Same IPC and features
- Lower max frequency
- Smaller area
- Increased power efficiency

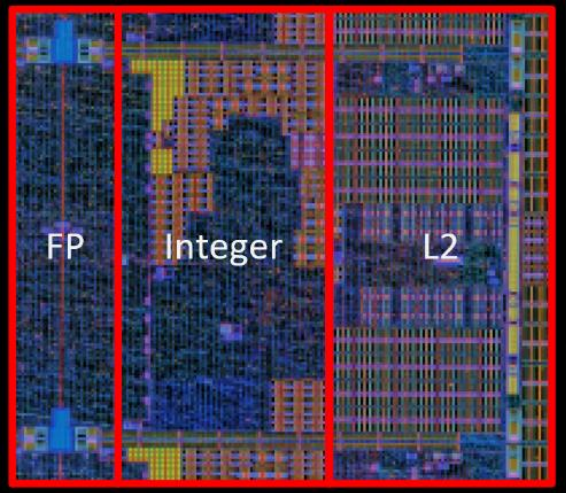


"Zen 4c": ~35% Less Core + L2 Area

"Zen 4" core + L2

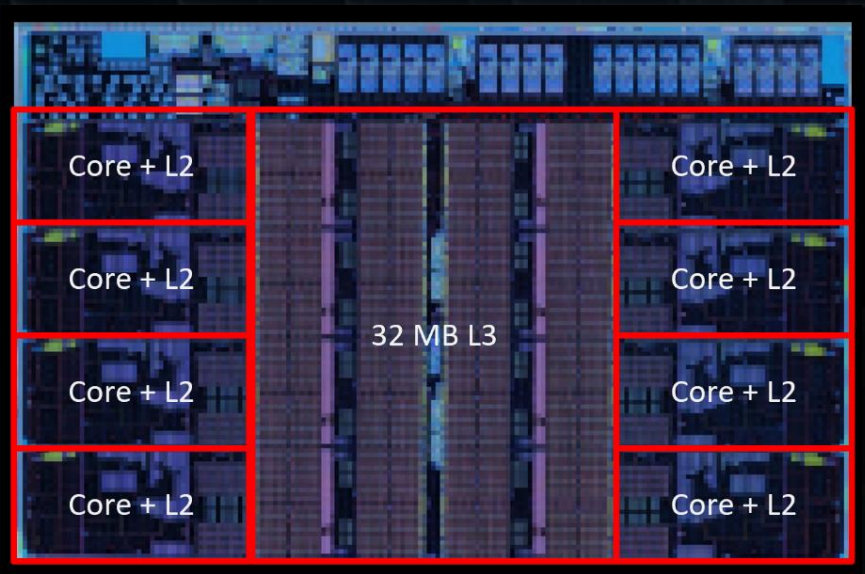


"Zen 4c" core + L2

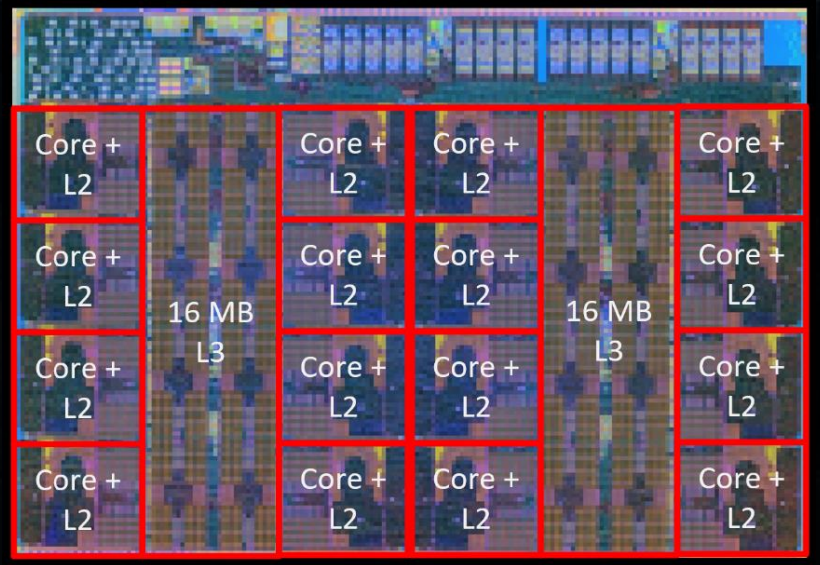


"Zen 4c" Doubles Cores Per Compute Die

"Zen 4" CCD

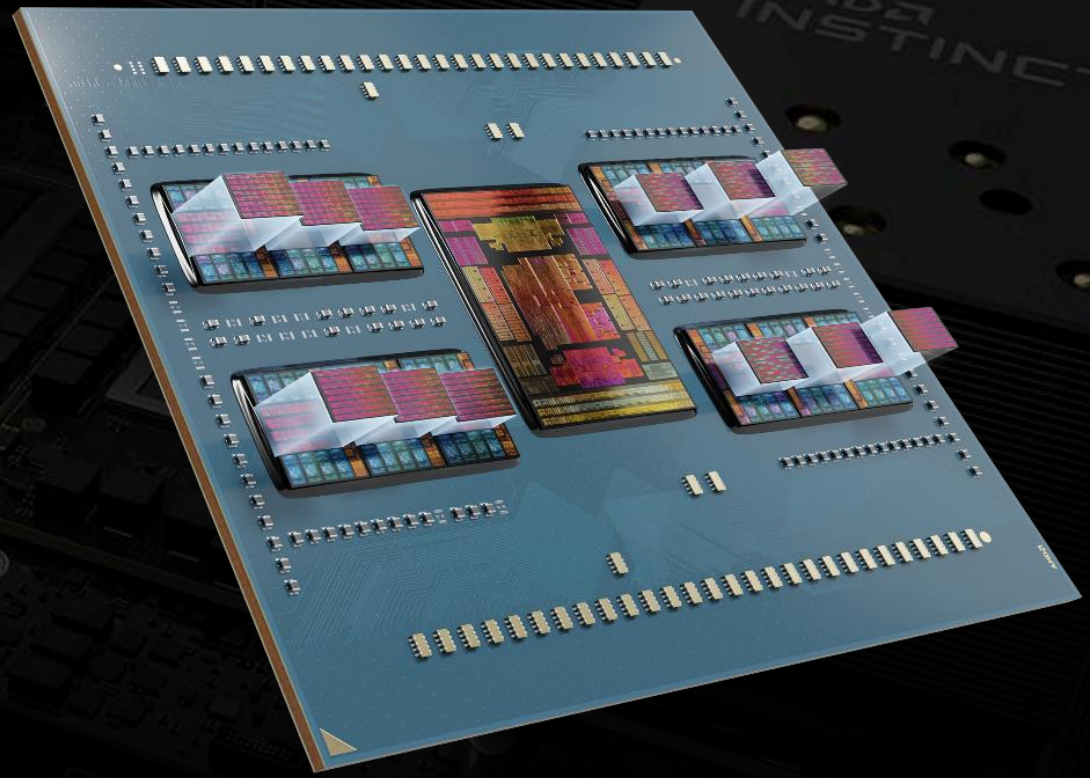


"Zen 4c" CCD



Cores	8	16
L2 cache/core	1 MB	1 MB
L3 cache/core	4 MB	2 MB

AMD "Zen 4" EPYC™ Family Processor Architecture



Ravi Bhargava
AMD Sr. Fellow
"Zen 4" EPYC™ Performance Architect

The "Zen 4" EPYC™ Family of Leadership CPUs

Architected for Datacenter TCO

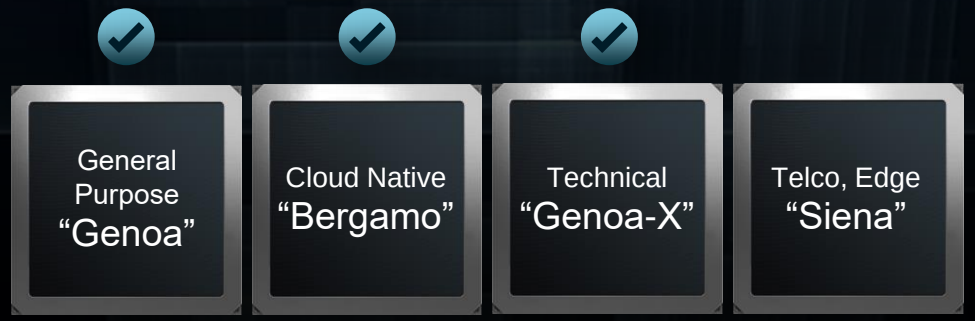
- Foundation of leadership technology
  - “Zen 4” Cores & Infinity Fabric™ 3.0
  - DDR5, PCIe® Gen5, CXL™ 1.1+
- Market-aware SoC variants
- Minimize time to market (TTM)

Chiplets Designed for Flexibility

- Single IOD for Platform Continuity
- Consistent, modular interfaces
- Wide range of performance and power

No Compromise Performance & Power

- All products high capability and high efficiency
- Across entire “Zen 4” EPYC™ family
- Across all configurations within a product stack



4th Gen EPYC™ Processors

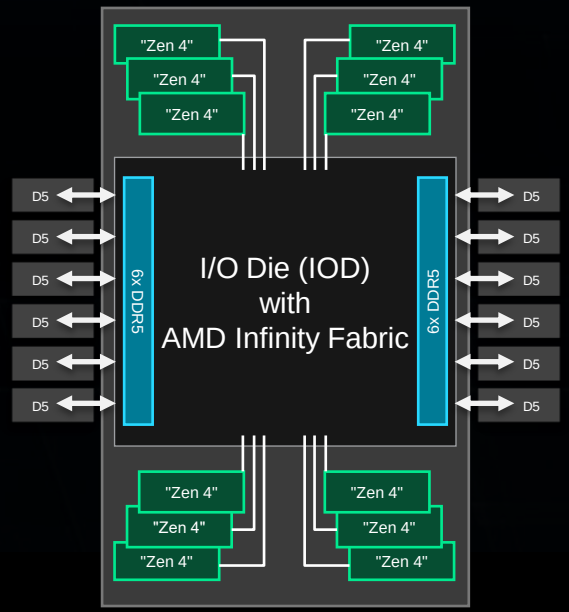
World-class Core Strength + Adaptive Chiplet Architecture
= Market-Tailored Leadership Solutions

KEY DESIGN VECTORS
Core / Thread Density
Performance Per Watt
Throughput Per Socket
Performance Per Core
Form Factor

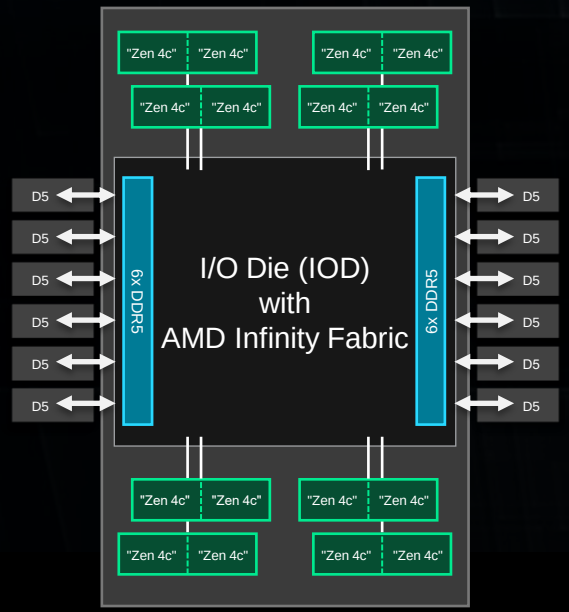
4th Gen EPYC™ Chiplet Portfolio

AMD Infinity Fabric™ 3.0 IOD, "Zen 4" CCD, "Zen 4c" CCD, AMD 3D V-Cache™ technology

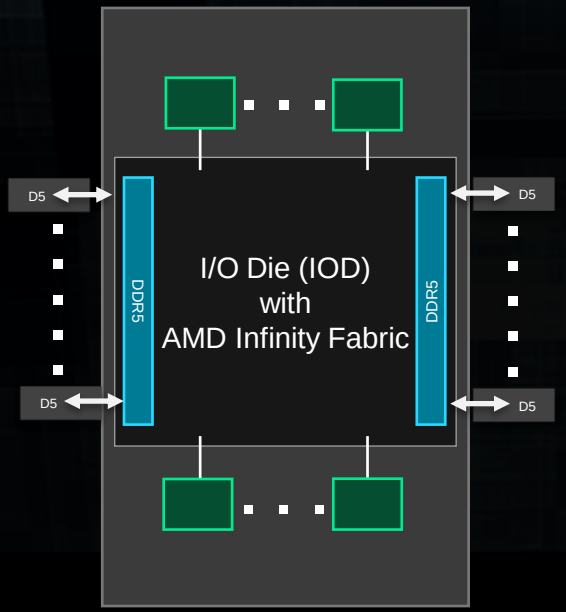
General Purpose
"Genoa"



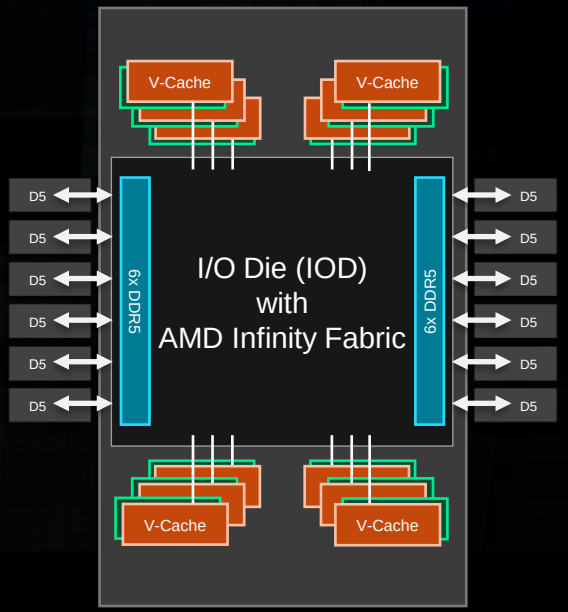
Cloud Native
"Bergamo"



Telco, Edge
"Siena"



Technical
"Genoa-X"

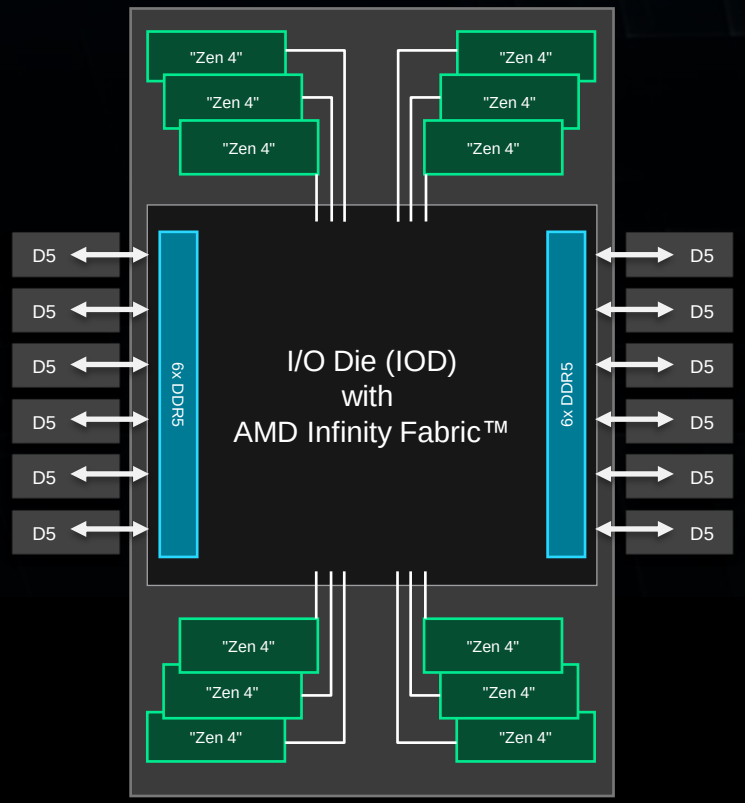


Flexibility by Design: Dynamically & Efficiently Adapt to Industry Needs

“Genoa”: High-Performance, Balanced Computing

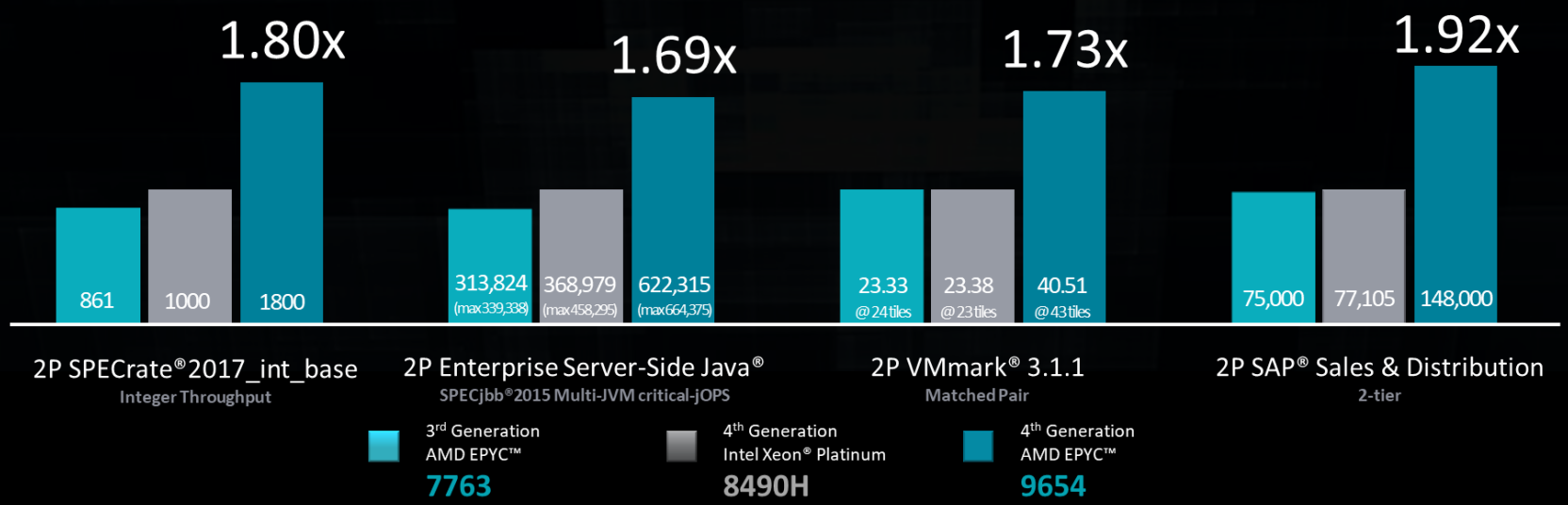
DESIGN VECTORS
Core / Thread Density
Performance Per Watt
Throughput Per Socket
Performance Per Core
Form Factor

4th Gen EPYC™
“Genoa”



Flagship Product for “Zen 4” EPYC™ Processors

- Balanced design optimizes TCO for many customers
- Large generational performance increase
- When launched, leadership performance across many markets



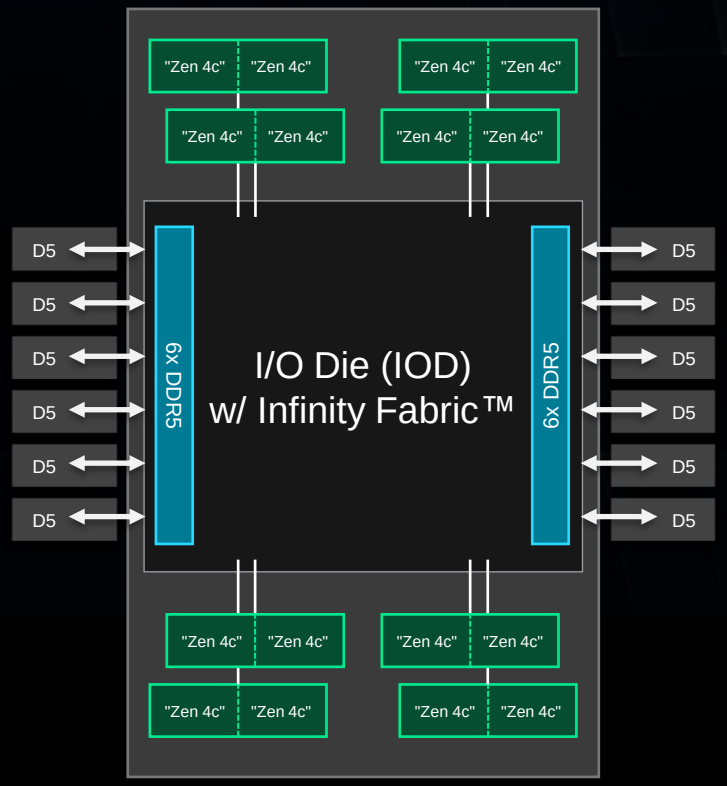
As of 6/13/2023, see SP5-013D,SP5-104A, SP5-049C, SP5-056B

AMD

“Bergamo”: Scaling Out with Power Efficiency

DESIGN VECTORS
Core / Thread Density
Performance Per Watt
Throughput Per Socket
Performance Per Core
Form Factor

4th Gen EPYC™
“Bergamo”

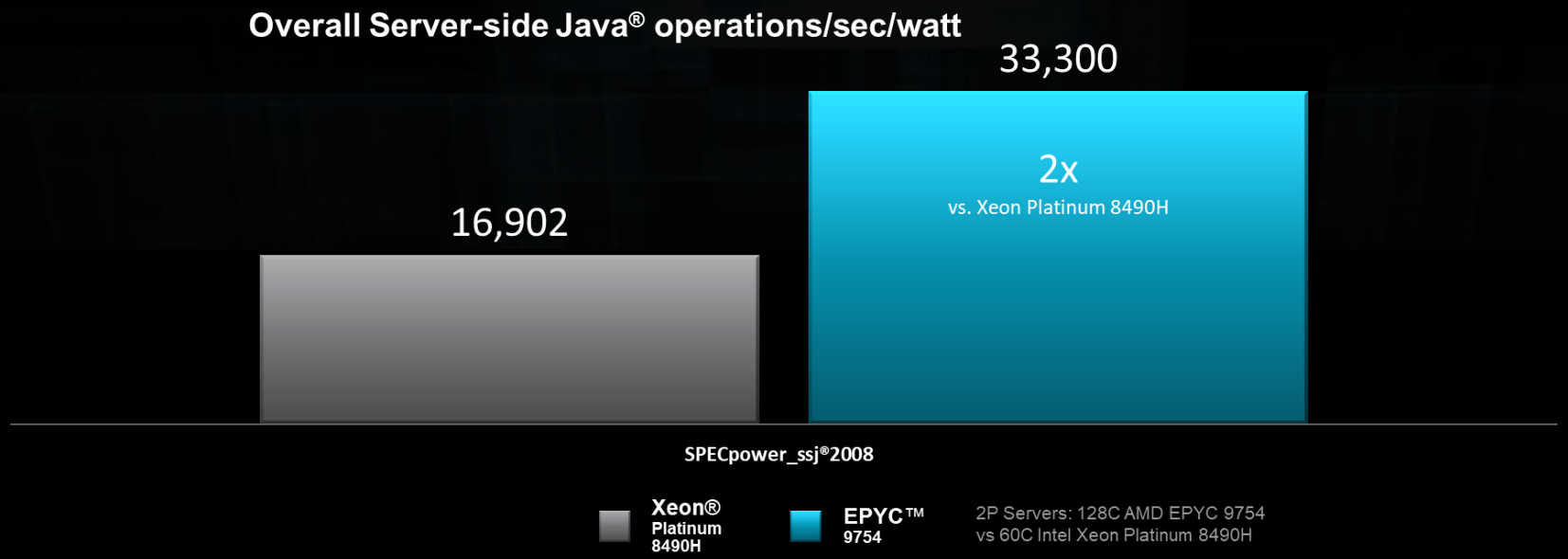


Scaling for 16 Core Complexes (2 per CCD)

- Native support for 128 cores (+33% vs. "Genoa")
- Optimized coherence probe performance for scaling
  - Probe compression, probe reduction, higher probe bandwidth

Natural Transition to “Zen 4c” Core Chiplet

- High-performance IOD/CCD interface unchanged
- Enhanced dynamic and static IOD power management for low TDPs



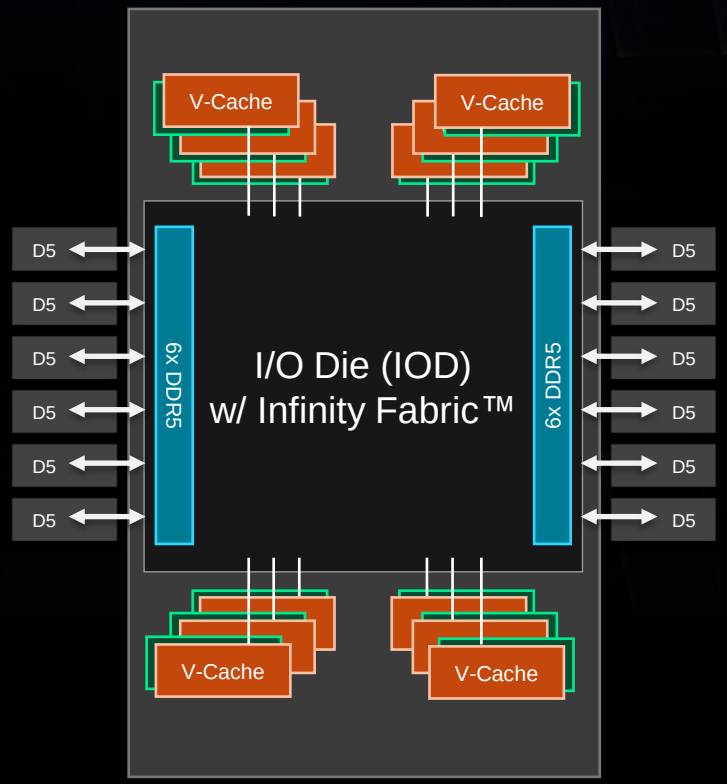
Results may vary due to factors including system configurations, software versions and BIOS settings. See https://www.amd.com/system/files/documents/amd-epyc-9754-pb-spec-power.pdf.

AMD

“Genoa-X”: Maximizing Per-Core Performance

DESIGN VECTORS
Core / Thread Density
Performance Per Watt
Throughput Per Socket
Performance Per Core
Form Factor

4th Gen EPYC™
“Genoa-X”

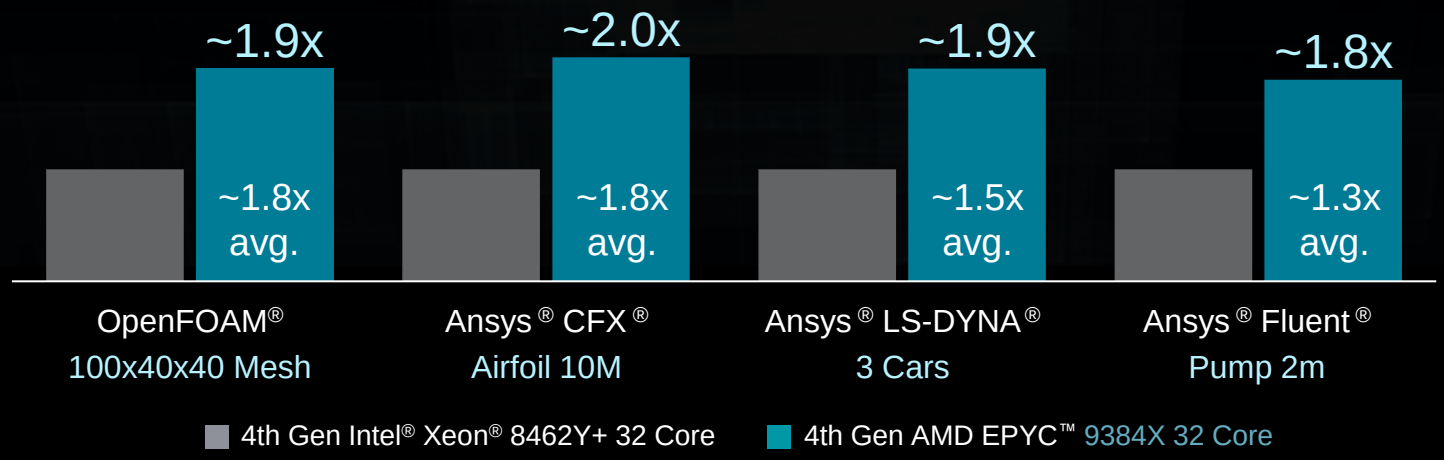


12-CCD config. shown above. 32C uses 8-CCD config.

1.15 GB of L3 cache per socket

- With AMD 3D V-Cache™ technology
- 96MB of L3 cache per CCD!
- Probe filter enhanced to improve coherence coverage
- More effective capacity with more coverage → larger V-Cache uplifts

CFD and FEA 32-Core Comparison



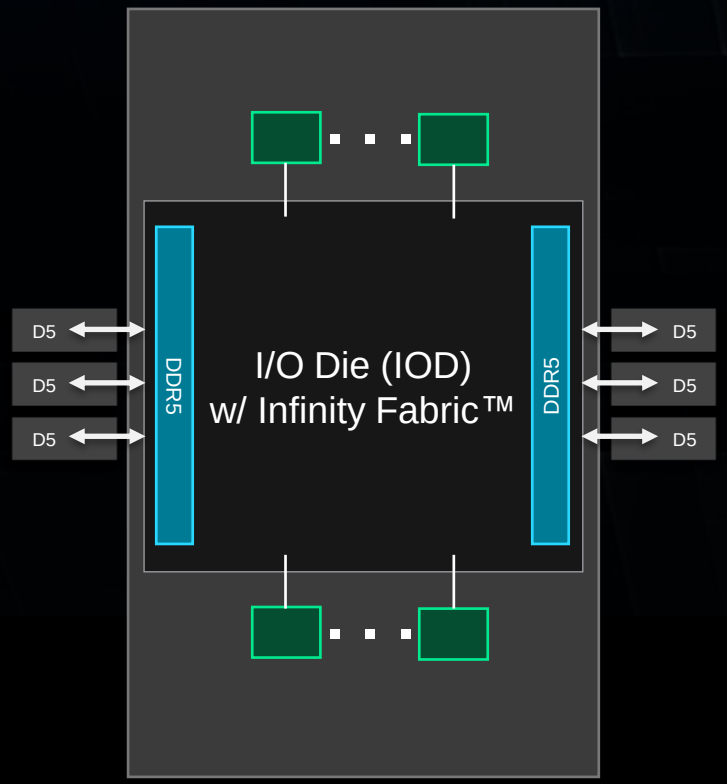
Sources: ANSYS CFX https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-cfx.pdf, ANSYS LS-DYNA https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-ls-dyna.pdf, ANSYS Fluent https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-ansys-fluent.pdf, OpenFOAM https://www.amd.com/system/files/documents/amd-epyc-9004x-pb-openfoam.pdf.

AMD

“Siena”: Resilient, Power-efficient Compute

DESIGN VECTORS
Core / Thread Density
Performance Per Watt
Throughput Per Socket
Performance Per Core
Form Factor

4th Gen EPYC™
“Siena”



New Optimized Package

- Up to 64 cores (128 threads)
- 6 DRAM channels
- Target 70W to 225W

Full-Feature Set at Low Core Counts

- High-performance cores
- Full-featured ISA & RAS
- Support max DDR5, PCIe®, Uncore frequencies

Optimizing CCD Flexibility with GMI3-Wide

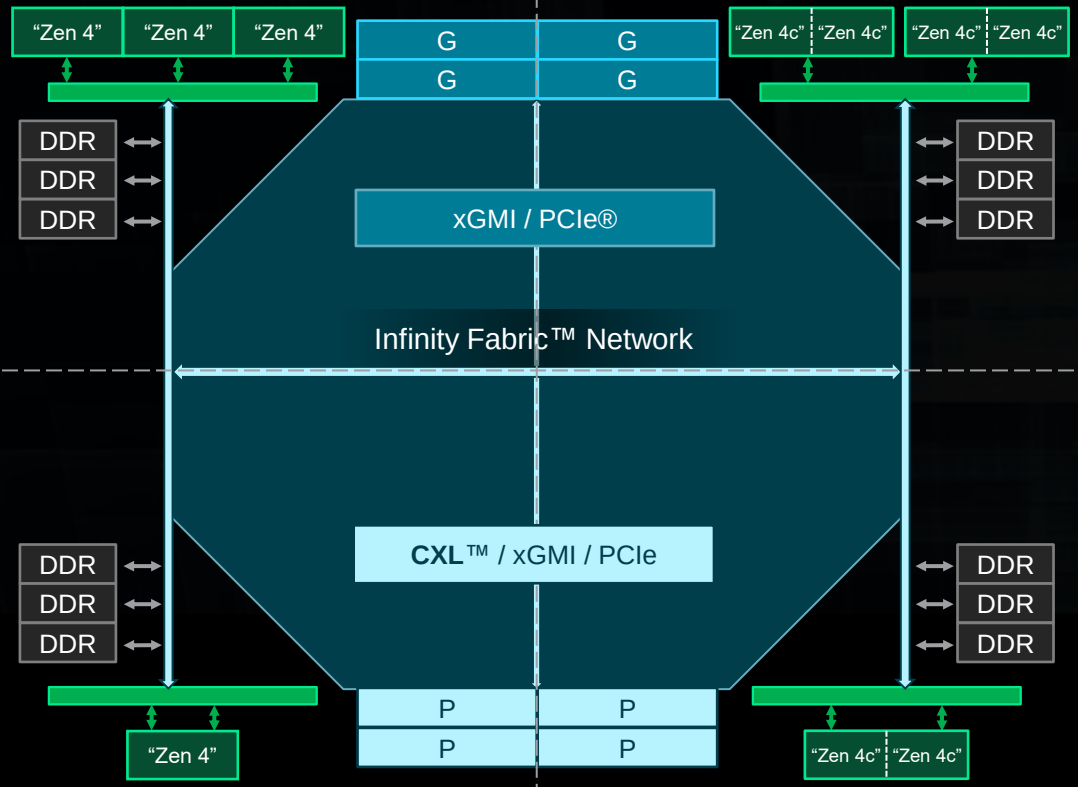
DESIGN VECTORS
Core / Thread Density
Performance Per Watt
Throughput Per Socket
Performance Per Core
Form Factor

3 CCD per Quad

- Ex: 96C “Zen 4” max config
- CCD B/W fairness
- More B/W per CCD vs. Milan

2 CCD Per Quad

- Ex: 128C “Zen 4c” max config
- For "Zen 4" or "Zen 4c" CCD



GMI3-Wide for "Zen 4"

- For "Zen 4" EPYC at <=4 CCD
- 2X GMI3 bandwidth/CCD
- Max perf/core at lower CCD #

GMI3-Wide for "Zen 4c"

- For "Zen 4c" EPYC™ at <=4 CCD
- Efficient B/W at lower CCD #
- Power optimizations for low B/W

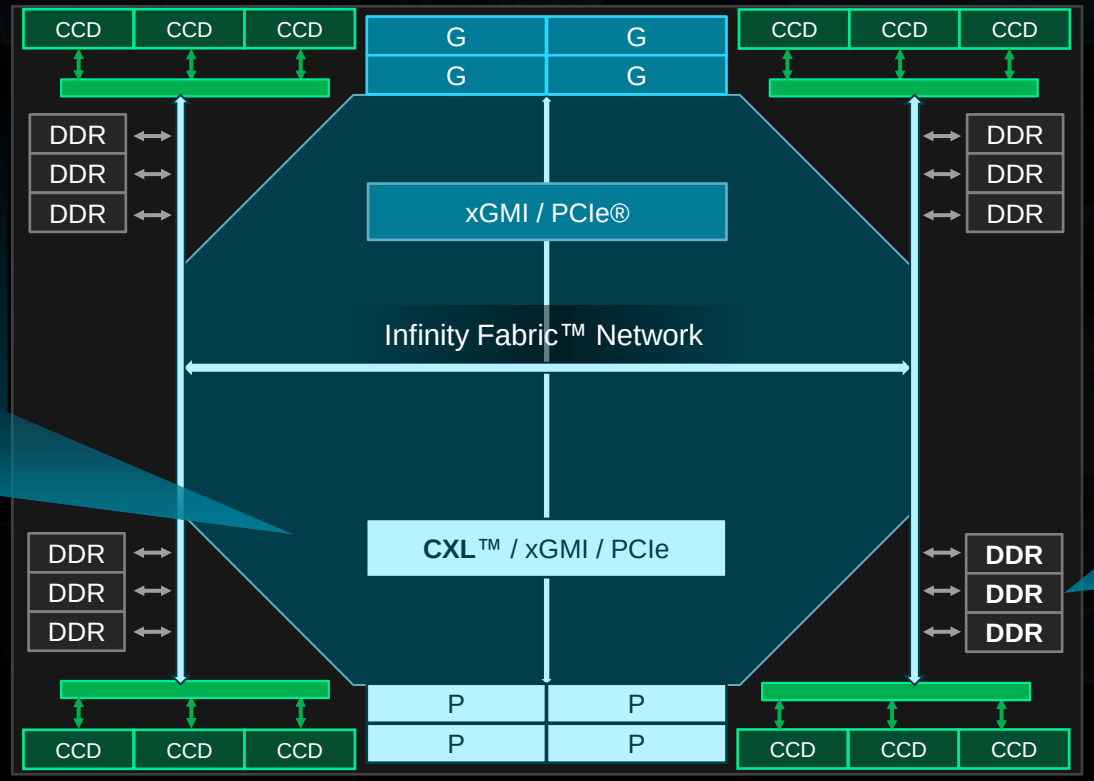
**For illustration: single "Zen 4" EPYC product does not support all options simultaneously

Optimizing Memory Technologies for Capacity

DESIGN VECTORS
Core / Thread Density
Performance Per Watt
Throughput Per Socket
Performance Per Core
Form Factor

CXL Memory Expansion

- 4 x16 Channels at 32Gbps
- CXL™ 1.1+
- Select CXL 2.0 Type3 features (e.g., RAS, persistent memory)
- Targeted for Initial CXL Ecosystem



High Efficiency DDR5

- Enhanced 2 x 40b memory controller
- 75-80% DRAM Util.
  - 2R RDIMM & 3DS
  - 1dpc and 2dpc
- Optimized 1-Rank
  - 1R x8 within 5% of 2R x4
  - Bounded Fault ECC

Cost-Effective and High-Performance
Leaps in Memory Bandwidth and Capacity

"Zen 4" Core and 4th Gen AMD EPYC™ Server CPU's

"Zen 4": New Levels of Performance

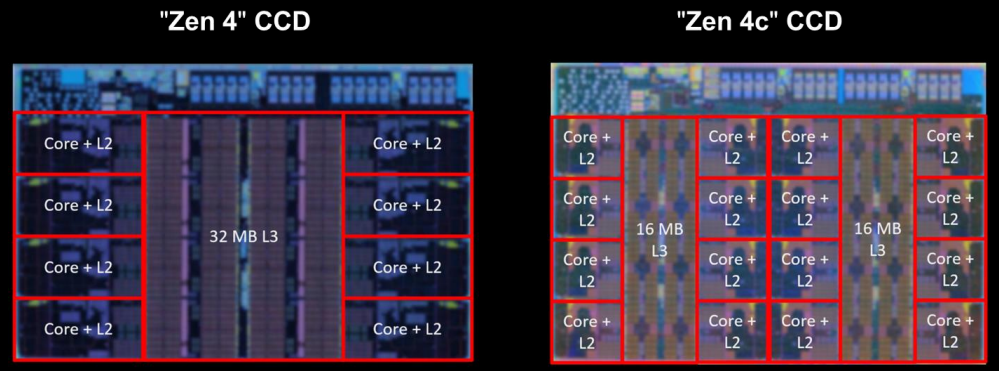
- Generational Single-thread and SMT Uplifts
- Power-efficient AVX-512

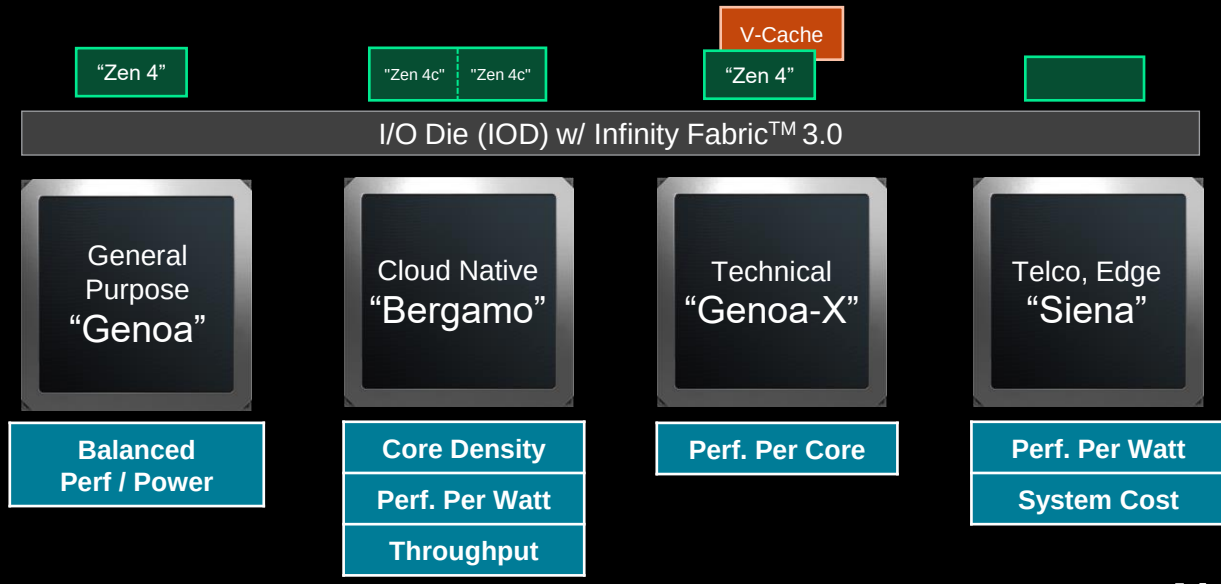
"Zen 4c": Optimized for Cloud

- Optimized for Power Efficiency and Area
- No Compromises on IPC and ISA

Differentiated Family of EPYC CPU's

- Market-aware Design Targets
- Chiplets Architected for Flexibility & TTM
- Performance & Power Efficiency for Every Product





AMD

Endnotes

EPYC-038: Based on AMD internal testing as of 09/19/2022, geomean performance improvement at the same fixed-frequency on a 4th Gen AMD EPYC™ 9554 CPU compared to a 3rd Gen AMD EPYC™ 7763 CPU using a select set of workloads (33) including est. SPECrate®2017_int_base, est. SPECrate®2017_fp_base, and representative server workloads. SPEC®, SPEC CPU®, and SPECrate® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

1: Frequency improvement at same voltage taken from data published in B. Munger et al., ""Zen 4": The AMD 5nm 5.7GHz x86-64 Microprocessor Core," 2023 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 38-39, doi: 10.1109/ISSCC42615.2023.10067540

2: Power reduction improvement derived from data published in B. Munger et al., ""Zen 4": The AMD 5nm 5.7GHz x86-64 Microprocessor Core," 2023 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 38-39, doi: 10.1109/ISSCC42615.2023.10067540

3: AVX2 vs. AVX-512 effective frequency increase based on internal testing as of 09/01/2022 on 2 socket AMD EPYC 9654 96C with SMT off, NPS4, 1.5 TB DDR4-4800 using an internal build of Virginia HPL 2.3 with AVX2 (92.0 % HLP efficiency) against an internal build of Virginia HPL 2.3 with AVX-512 (92.2% HPL efficiency).

4: "Zen 3" vs. "Zen 4" estimated SPECrate® 2017 Integer SMT uplift determined using top-of-stack AMD EPYC 7763 ("Milan") at 2.5 GHz fixed frequency, Memory: 3200-DDR4, 32GB x 8 channels, NPS1 and top of stack AMD EPYC 9654 ("Genoa") at 2.5 GHz fixed frequency, Memory: 4800-DDR5, 64GB x12 Channels, NPS1. SPECint® 2017 benchmark compiled with GCC 10.2 -O3. 2 copies run on the same core with SMT off and SMT on.

SP5TCO-032: 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 96 core AMD EPYC_9654 powered server with a SPECrate2017_int_base score of 1790, https://spec.org/cpu2017/results/res2022q4/cpu2017-20221024-32607.pdf; 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-032.

Endnotes (2)

- 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.

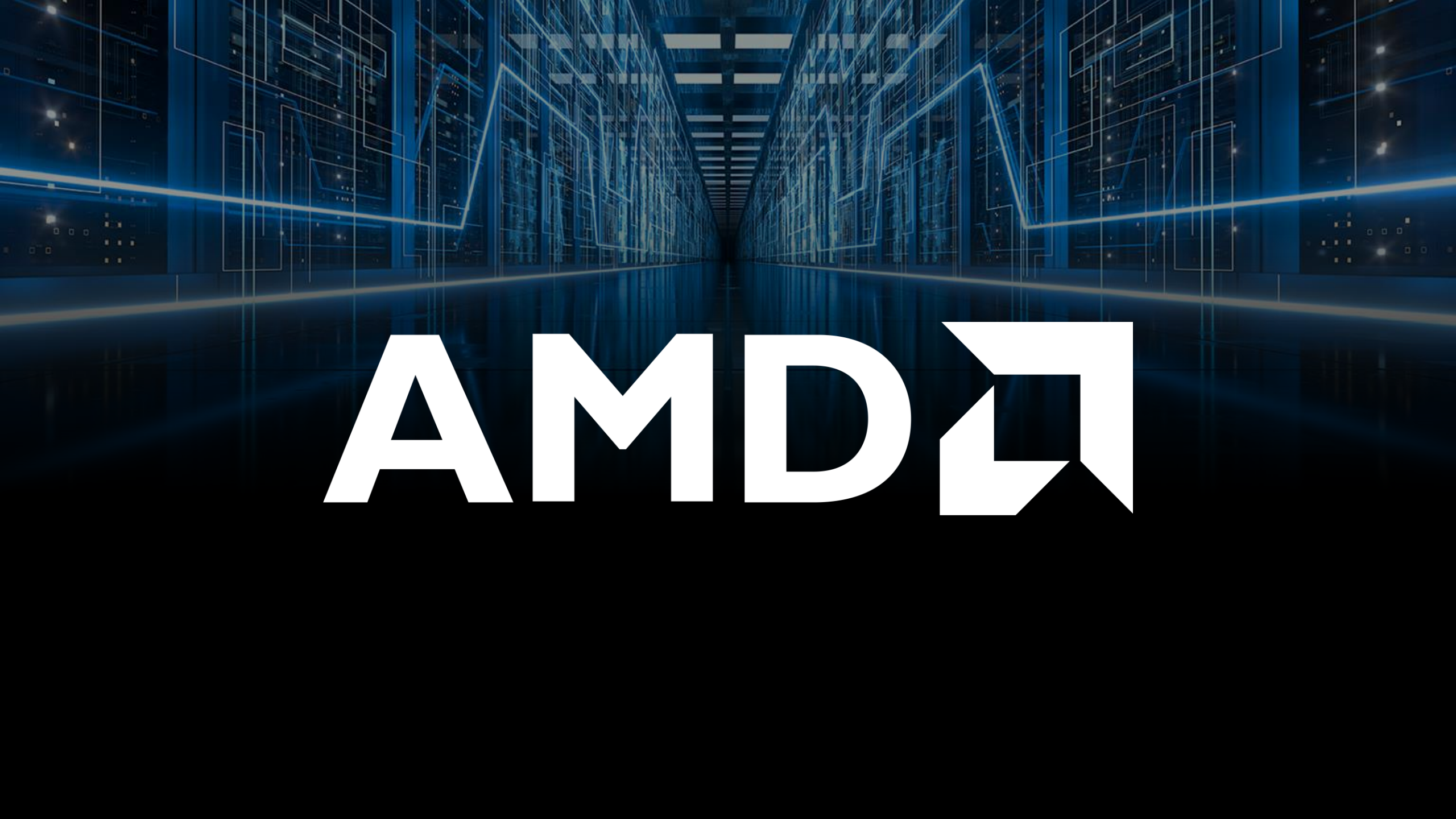
- SP5-056B: SAP® SD 2-tier comparison based on published results as of 6/13/2023. Configurations: 2P 96-core EPYC 9654 powered server (148,000 benchmark users, https://www.sap.com/dmc/benchmark/2022/Cert22029.pdf) versus 2P 60-core Xeon Platinum 8480+ (77,105 benchmark users, https://www.sap.com/dmc/benchmark/2023/Cert23021.pdf) 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.

Disclaimer and copyright

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.


AMD