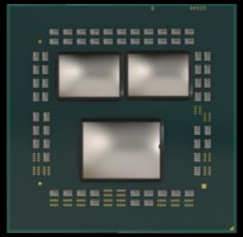
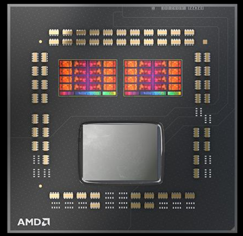
Next Generation “Zen 5” Core

Brad Cohen and Mahesh Subramony
Co-author: Mike Clark

Hot Chips 2024

The “ZEN” Path





2020 — 2024

“ZEN 3”

- +19% IPC[3] over “Zen 2”
- 8-core complex
- 32/16MB L3 per complex
- CET shadow stack
- 7nm/6nm

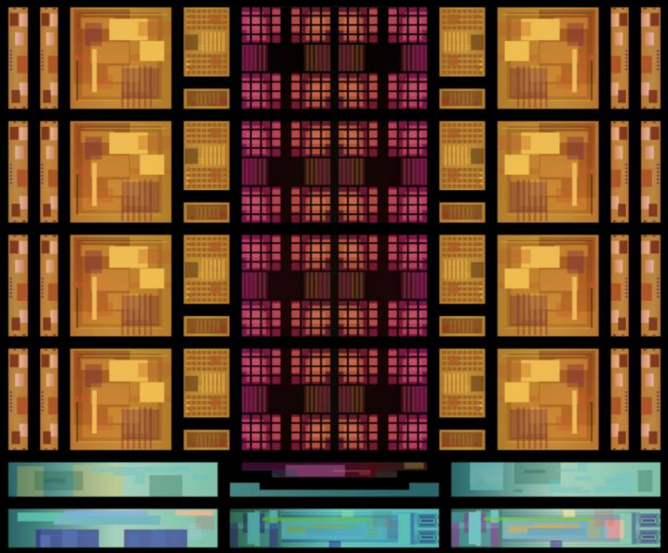
“ZEN 4”

- +14% IPC over “Zen 3”
- AVX-512 on FP-256
- 1M L2
- VNNI and BFLOAT16
- 5nm/4nm

“ZEN 5”

- +16% IPC[3] over “Zen 4”
- AVX-512 variants, FP-512
- 8-wide dispatch, 6 ALU
- Dual pipe fetch/decode
- 4nm/3nm

*1. R5K-003;2.: EPYC-038; 3. all results are ‘up to’. See endnote GNR-03



Design Objectives

Performance

- Deliver another major 1T and 2T performance increase
- Balanced cross-core 1T and 2T instruction and data throughput
- Create front end parallelism
- Increased execution parallelism
- High throughput, efficient data movement and prefetching
- AVX512 with 512bit FP datapath for throughput and AI uplift

New Capabilities

- Additional ISA extensions
- New security features

Platform Support

- Deliver “Zen 5” and “Zen 5c” core variants
- Support configurable FP512/FP256 datapath
- Support scaling and energy efficiency

“Zen 5” Microarchitecture Overview

2 Threads/Core

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 inst/pipe

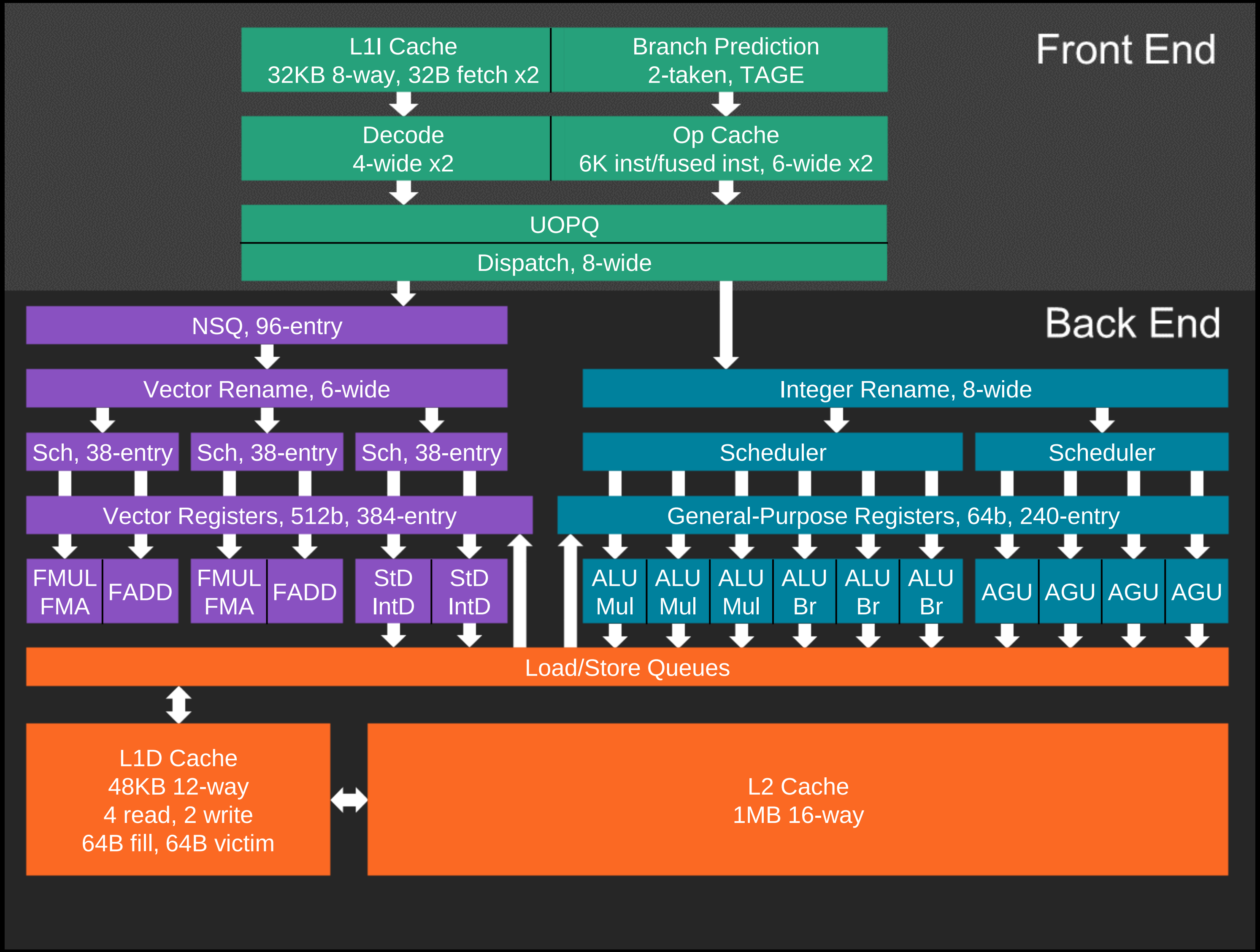
8 ops/cycle dispatched to Integer or FP

Execution capabilities

- 6 integer ALU
- 4 AGU, 4 addresses to LS per cycle
- 4 FP ops/cycle; 2cycle FADD

TLBs

- L1: 64entry ITLB, 96entry DTLB
- L2: 2K ITLB; 4K DTLB everything but 1G



AMD

"Zen 5" Delivers a New Level of Balanced Throughput

Front end parallelism

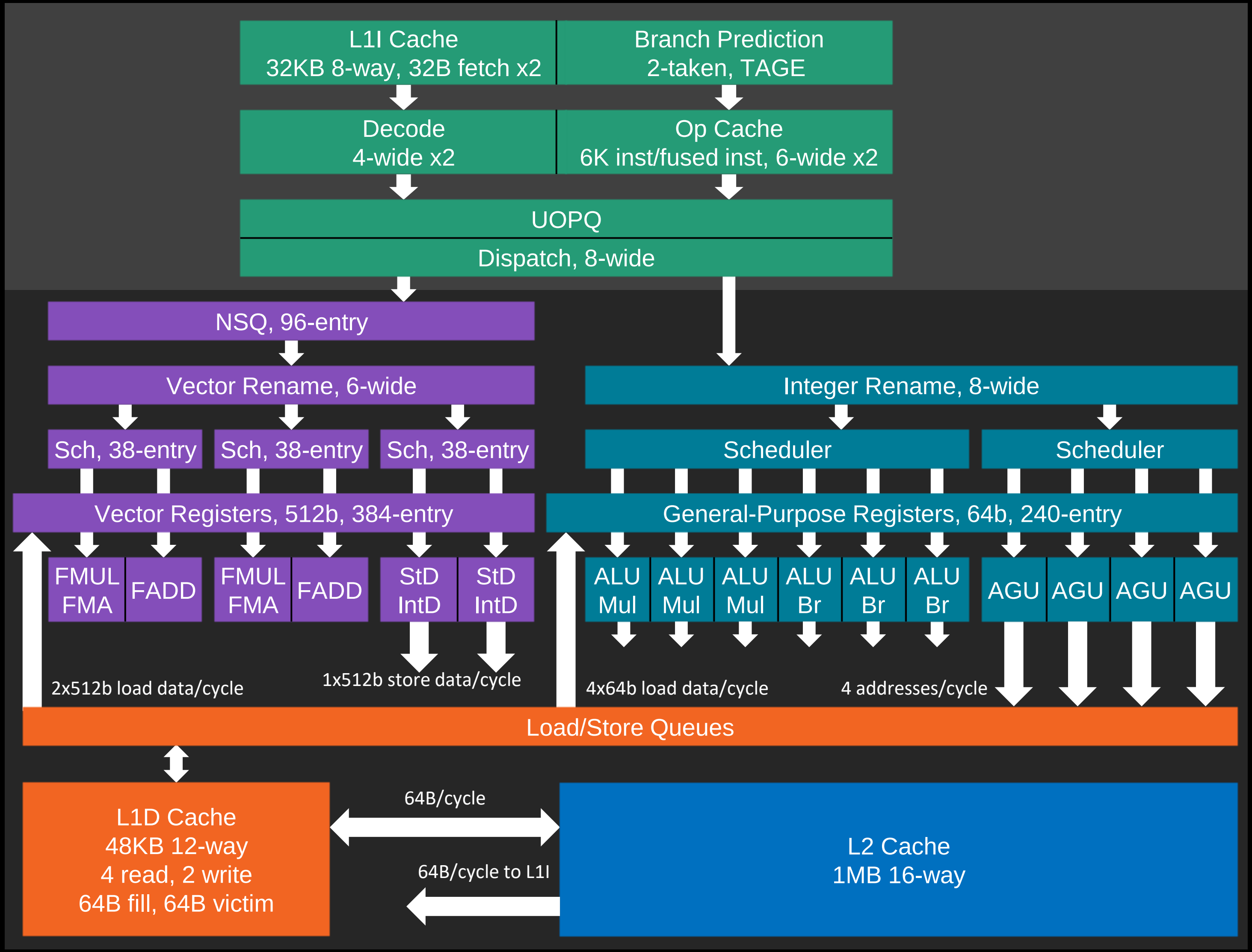
- 2 predicted taken branches per cycle
- 2x Op Cache pipes
- 2x Instruction fetch/decode pipes
- 8 wide dispatch

Execution

- Integer: 6ALU, 4AGU addresses->LS
- FPU: full 512b AVX512 datapaths
- FPU: 4 execution pipes

Dataflow

- 4 load pipes support 2, 512b AVX512 pipes
- 48K,12-way L1D cache delivers 4 memops/cycle
- 2x width L2 cache <-> L1I and L1D caches



“Zen 5” Instruction Fetch Advances

Branch prediction: fewer bubbles, more accuracy, and throughput

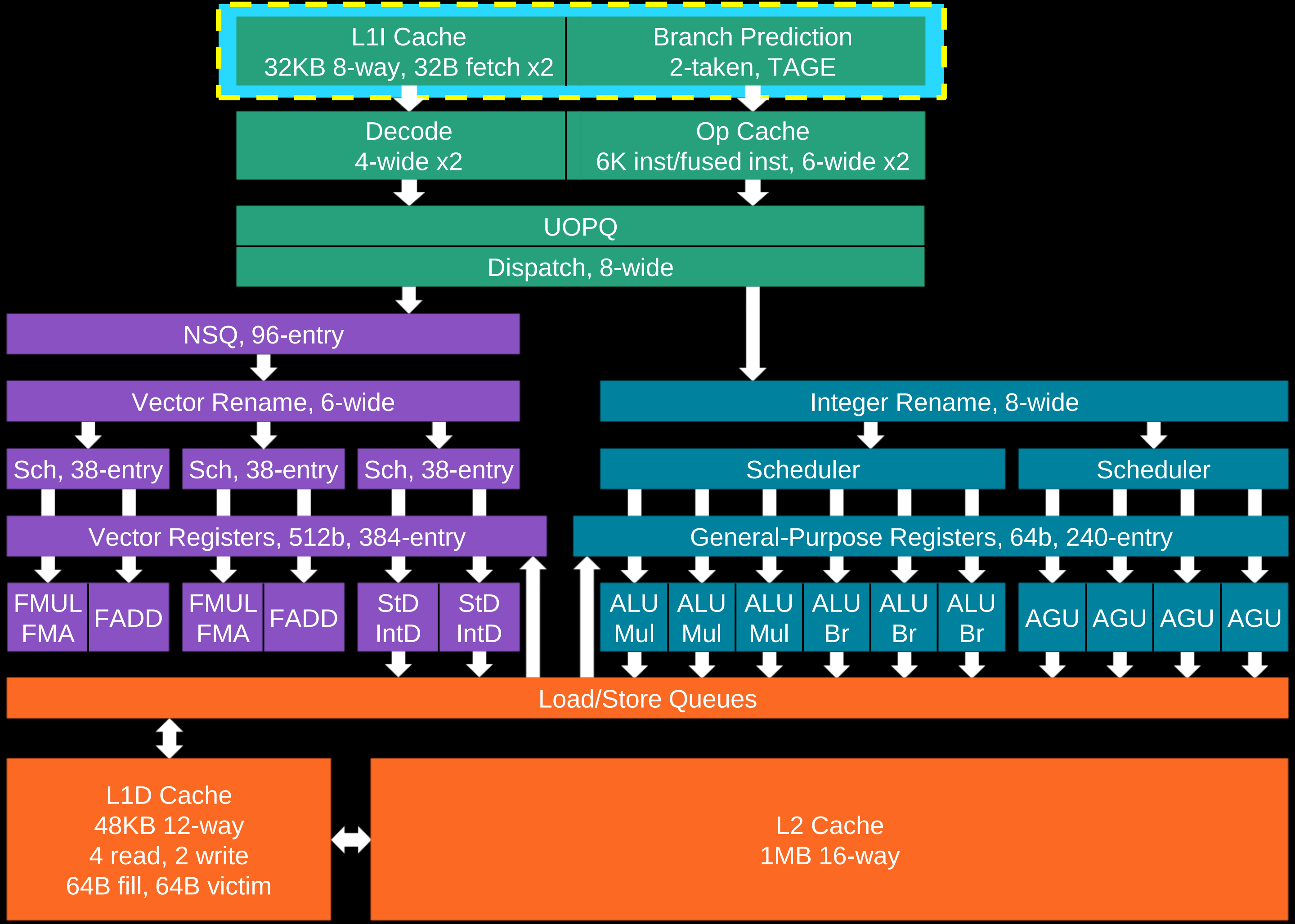
- Zero-bubble conditional branches
- L2-sized (16K) L1 BTB and larger TAGE
- Larger return address stack (52entry)
- 2 taken predictions/cycle

Memory Management

- Aggressive fetch hides L2 & tablewalk latencies
- 4x the L2 ITLB (2048 entry)

Icache latency and bandwidth

- 64B/cycle fetch
- 2 instruction fetch streams



“Zen 5” Instruction Decode Advances

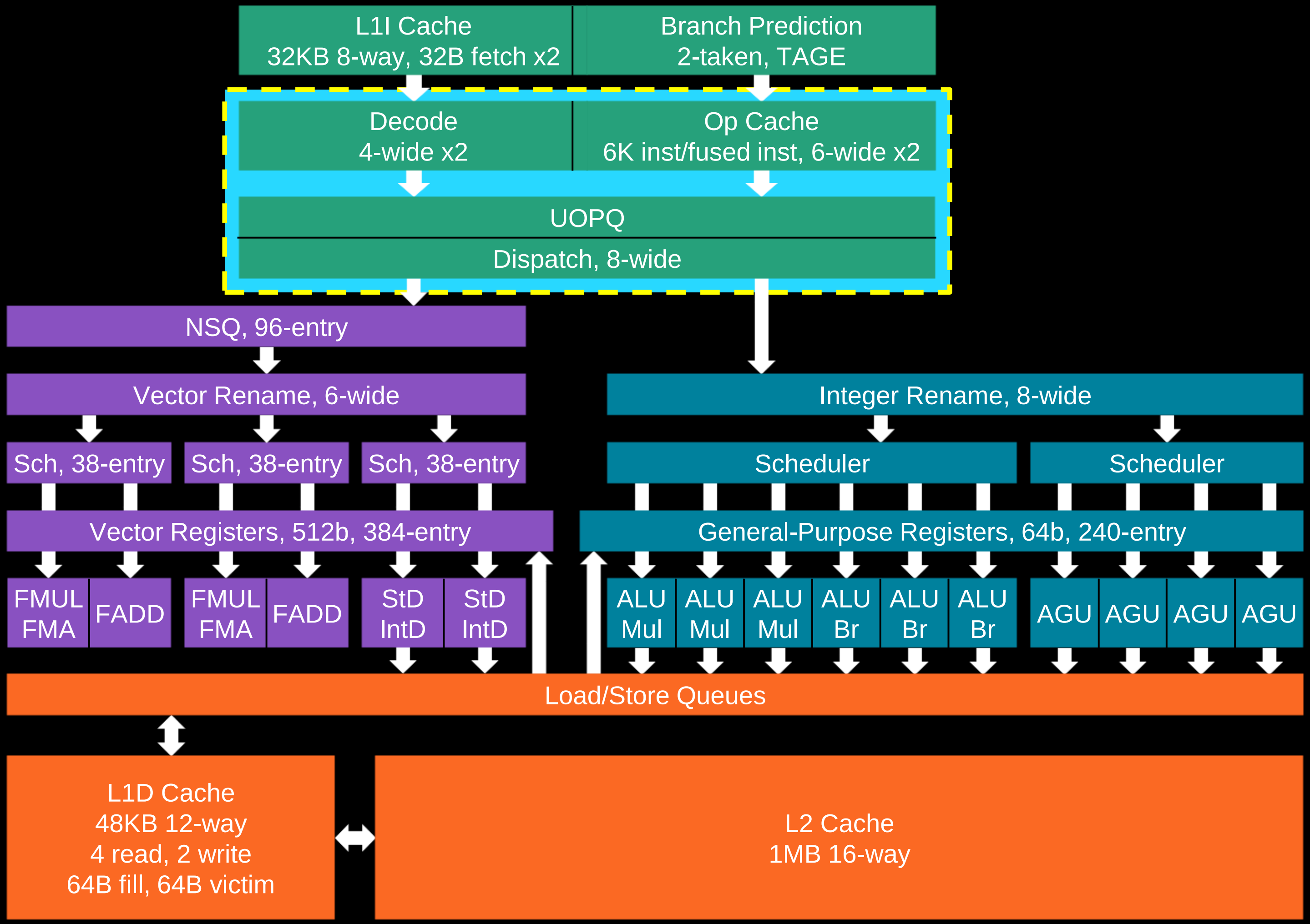
OpCache: higher density with greater coverage and throughput

- 33% more entry associativity (16-way)
- Dense entries store 6 instructions or fused instruction, not ops
- 2 OC pipes x 6 inst/pipe =>12 inst/cycle

Dual Decode Pipes

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

8-wide dispatch to Int and FP execution



“Zen 5” Integer Execution Advances

8-wide dispatch, rename, retire

Integer scheduler advances

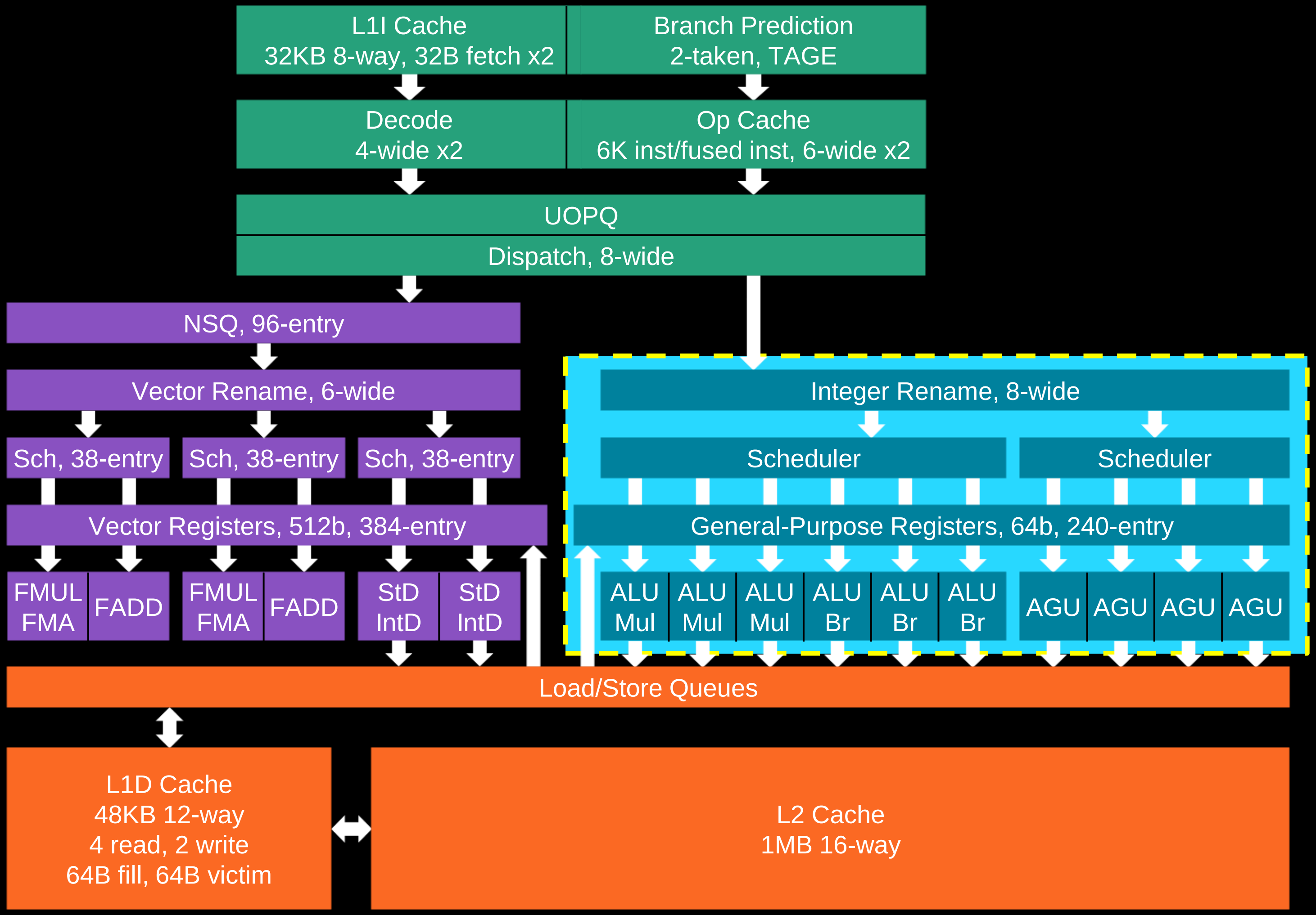
- Unified with age matrix
- More symmetry, simplifying pick

6 ALU with 3 multipliers, 3 branch units

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

Execution window growth

- Scheduler growth
- 240 entry physical register file
- ROB/retire queue 448/224 1T/2T entries



“Zen 5” FP/Vector Execution

FP major features/changes

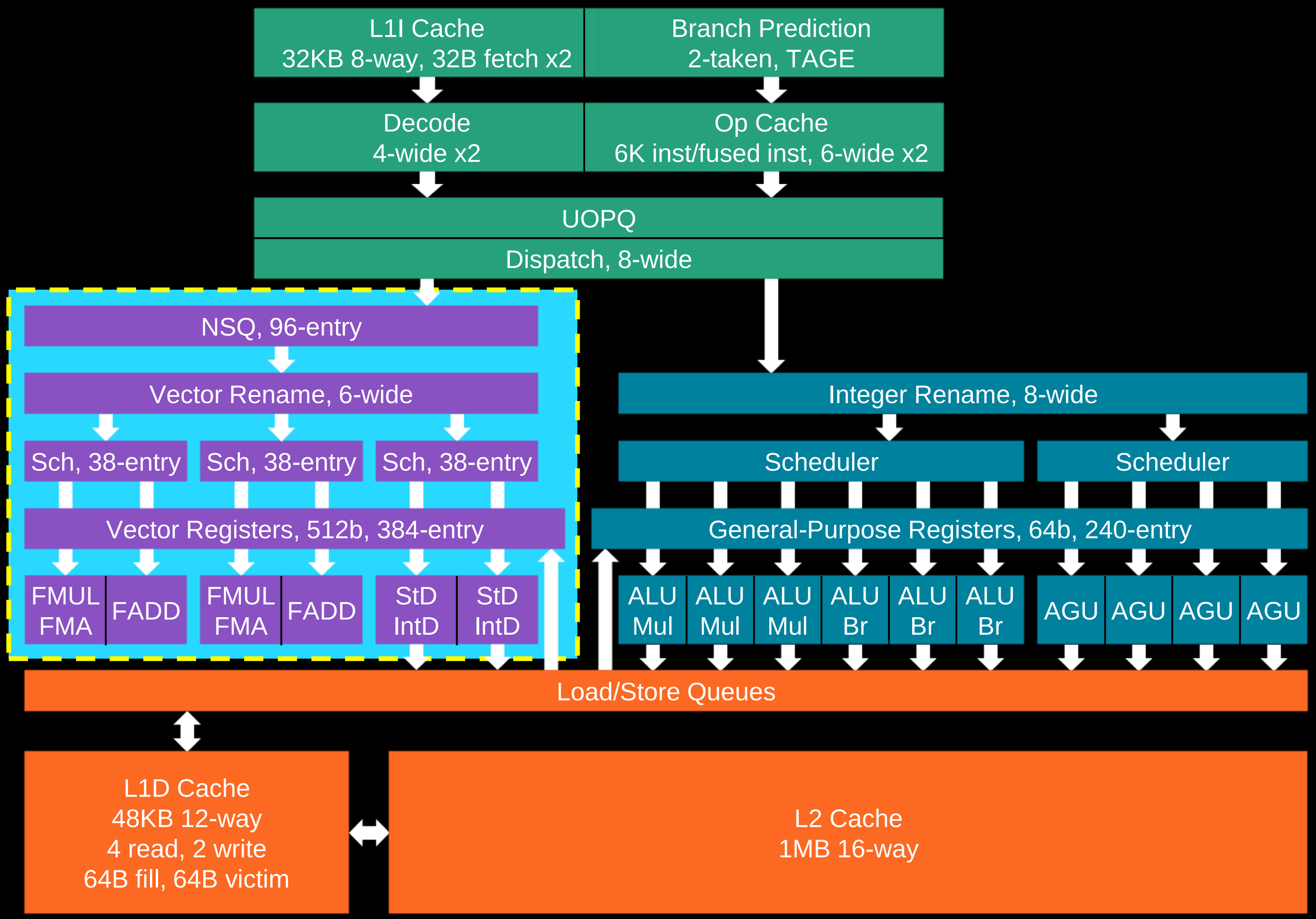
- AVX512 with full 512b datapath

More bandwidth, less latency

- Full FP512 execution datapath
- 4 1op/cycle execution pipelines
- 2 LS/integer register pipelines
- 2 512b loads/cycle, 1 512b store/cycle
- 2cycle FADD

Execution window growth

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



Notes: NSQ = non-scheduling queue

“Zen 5” Load/Store Advances

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

More Bandwidth

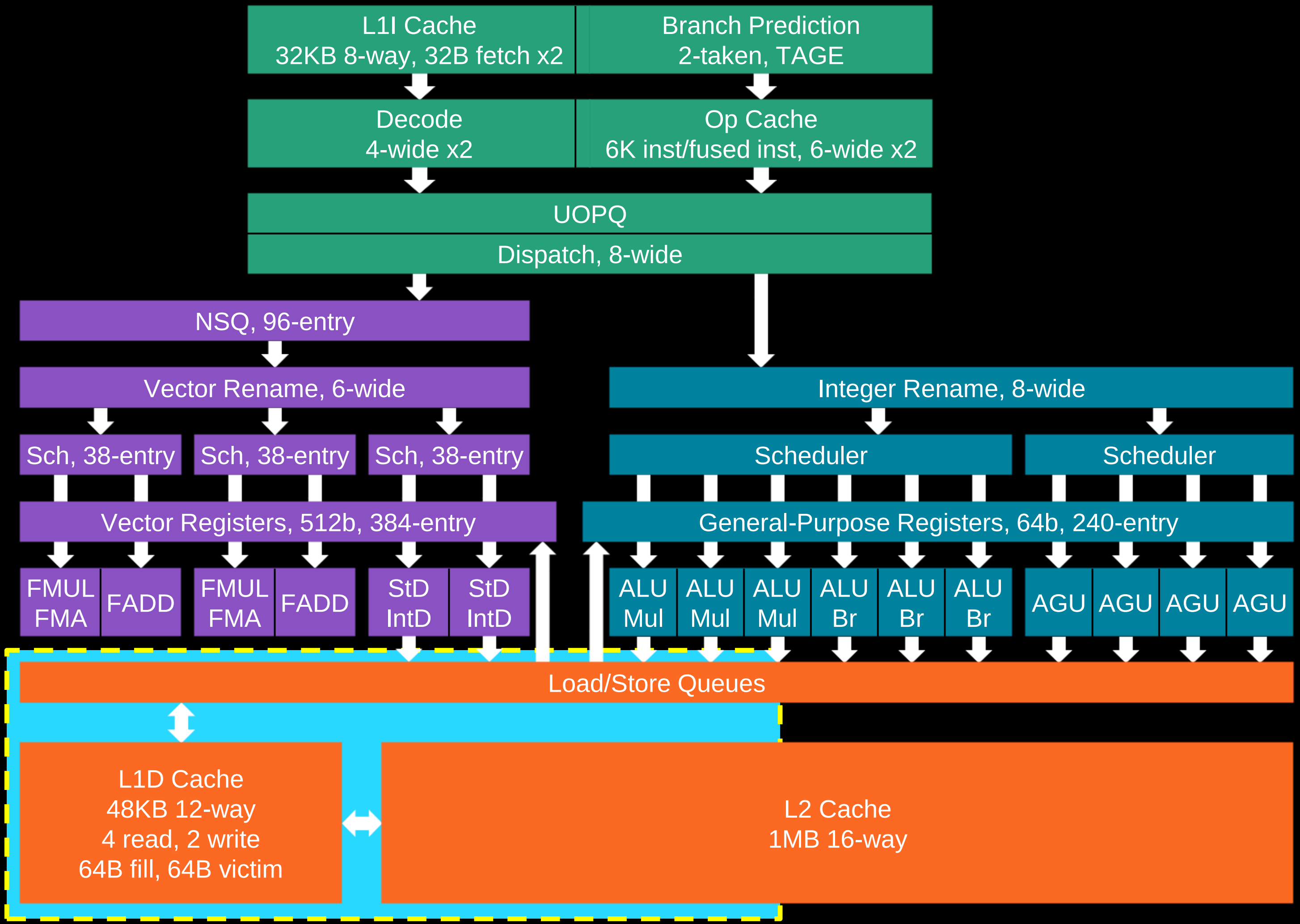
- 4 LS pipes for a mix of 4 loads/2 stores per cycle
- 4 Integer load pipes support 2, 512b AVX512 pipes
- 2 store commit per cycle
- 64B fill/victim from/to L2 DCache
- String store optimizations - eliminates dest mem read, frees bandwidth if the Store will overwrite the block

Larger In-Flight Window

- Load and store queue growth
- Store coalescing buffer growth
- More in-flight misses (scalable tracking)
- Scalable load ordering queue

Data prefetching

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



“Zen 5” Cache Hierarchy

2x L2/core interface bandwidth

- 64B/clk -> L1I
- 64B/clk -> L1D
- 64B/clk <- L1D

2x L2 associativity, now 16-way

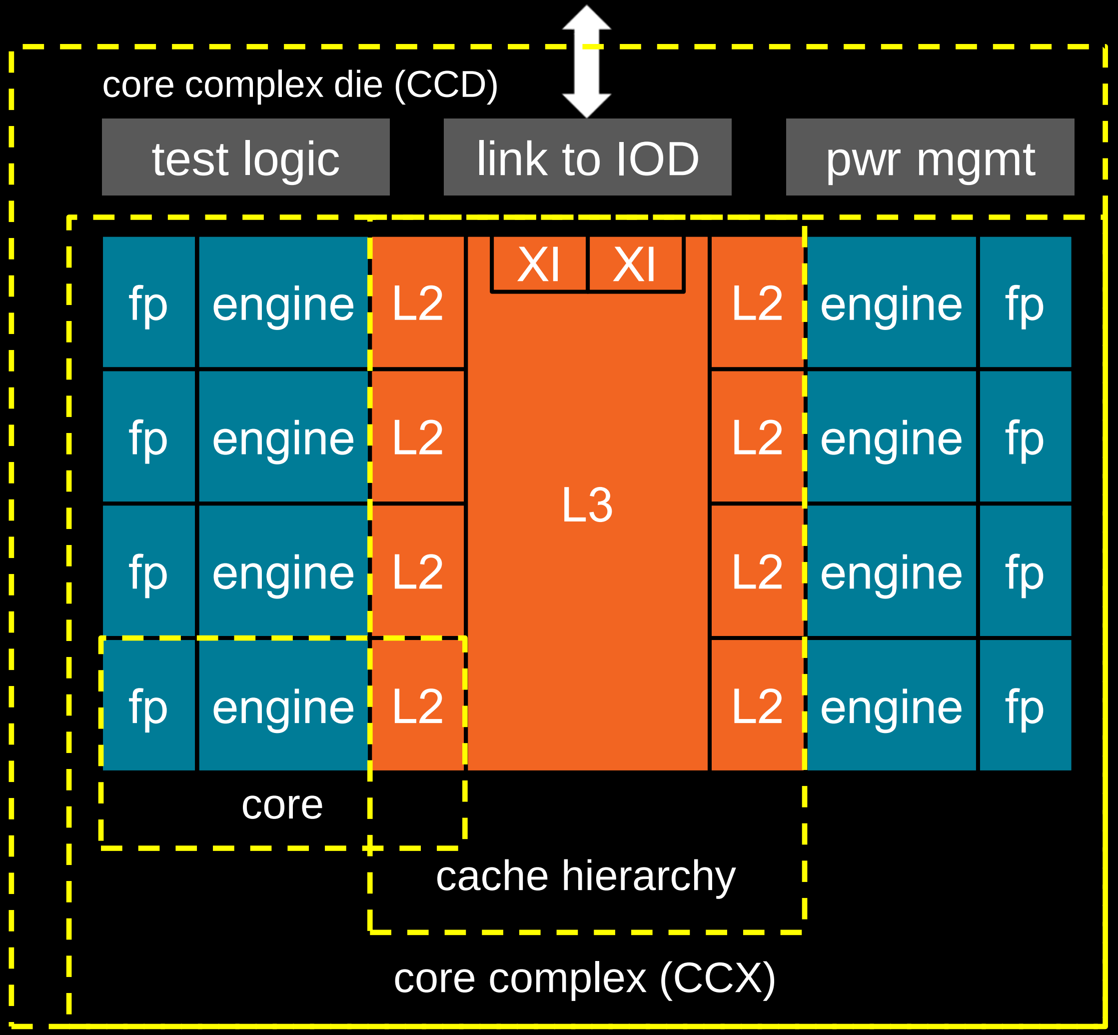
Improved L3 latency

- "Zen 5" 8c32m - 3.5 cycles less

Supports more L3 in-flight misses

Configurability

- L3 per core example (32M/16M L3)
- 4MB/core (8c32M)
- 2MB/core (8c16M)



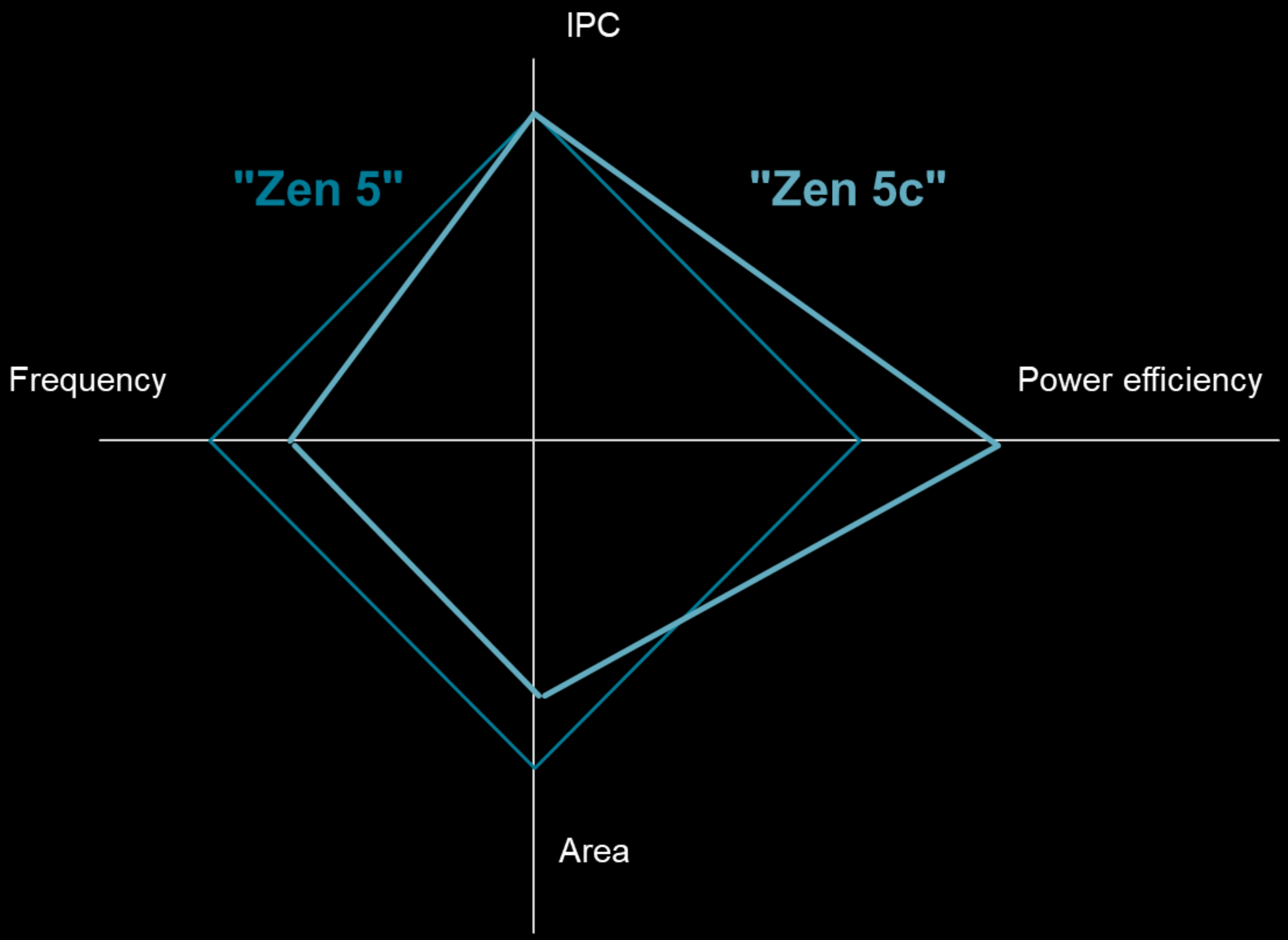
AMD

“Zen 5” Family

"Zen 5" family members
- "Zen 5" – max 1T performance
- "Zen 5c" – perf/w & perf/area optimized

"Zen 5" family members bring
- Better optimization/targeted solutions -> end use
- Software consistency and ease of use

"Zen 5" family enables platform breadth
- "Zen 5" and "Zen 5c" provide scalability/performance options



images not to scale
AMD

New “Zen 5” ISA

Attribute	ISA Feature
Instructions	▪ MOVDIRI/MOVD64B – move 4,8 or 64 bytes as a direct store, bypassing caches ▪ VP2INTERSECT[DQ] – AVX512 vector pair intersection to a pair of mask registers ▪ VNNI/VEX – extends AVX512 instruction to VEX encoding ▪ PREFETCH[I*] – software prefetch of instruction lines into cache hierarchy
Kernel/Virtualization/QoS	▪ PMC virtualization – provides security for a guest vs. hypervisor; isolates PMC/guest ▪ QoS for CDMA
Security	▪ SEV – secure AVIC ▪ SEV – enhanced perf and HV controls ▪ SEV-SNP segmented RMP – reverse map table
Perf Analysis/Trace	▪ L3X PMC – stacked L3 PMCs ▪ General enhancements

“Zen 5” Power Efficiencies

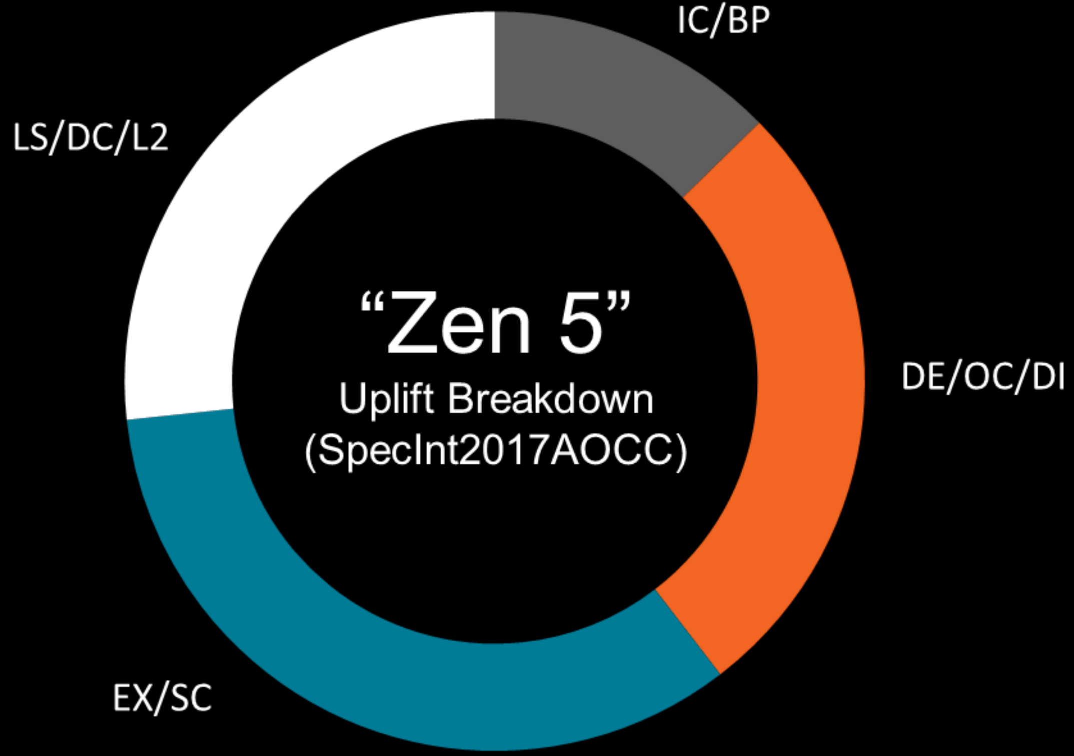
"Zen 5" builds upon "Zen" family Perf/Watt advantages

- Continued Power Gating Improvements
- Continued 2T support – major perf/watt benefits
- Reduced power state entry/exit times
- Better branch prediction – eliminate wasted work
- Optimizing operations - eliminating operations and bus, cache hierarchy, inter-core traffic
  - Optimizing string operations
  - Prefetcher effectiveness and efficiencies
- Power efficiencies => System performance increase

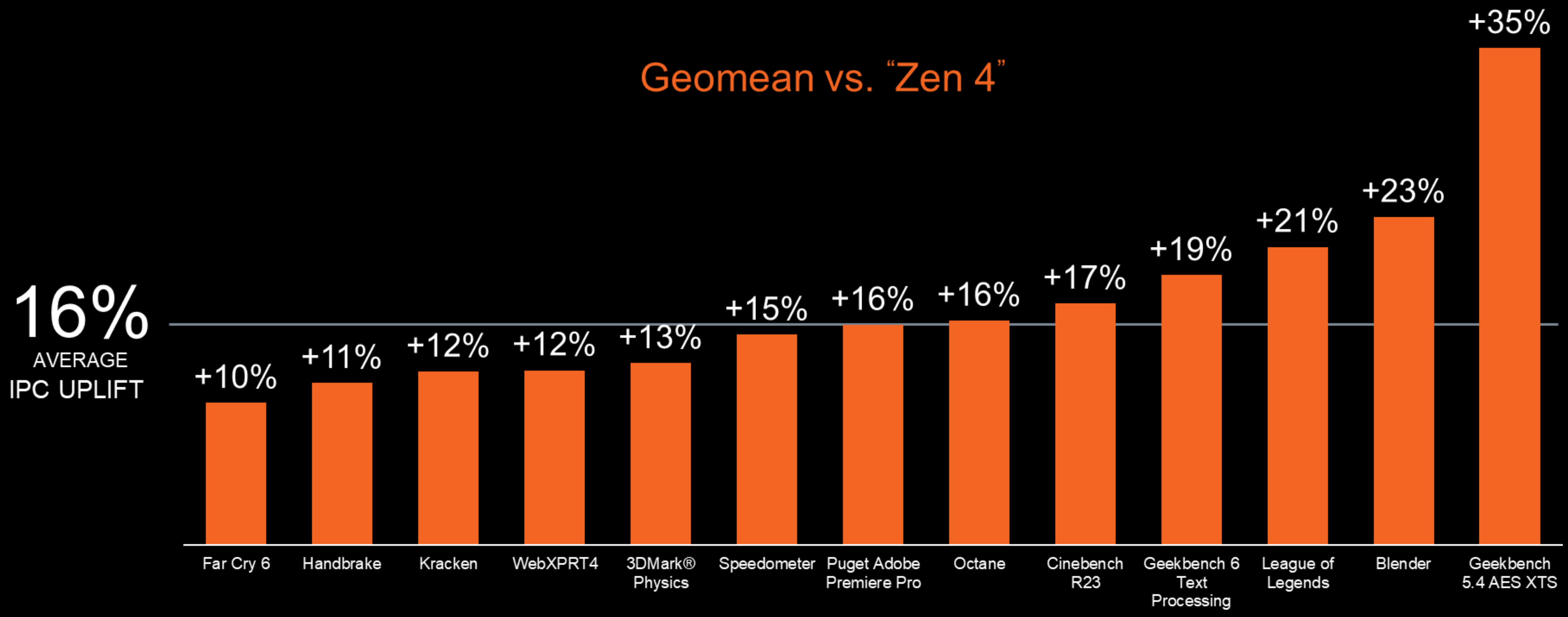
AMD

Key "Zen 5" vs. "Zen 4" Advances

Attribute	Zen 4	Zen 5
L1/L2 BTB	1.5K/7K	16K/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
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



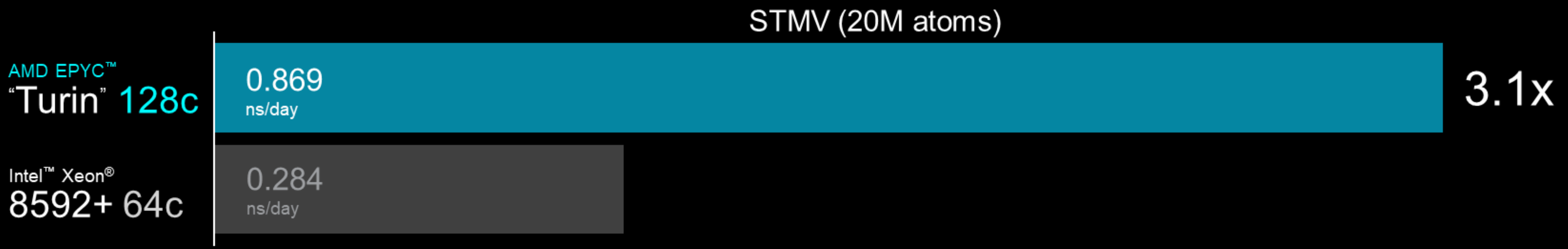
“Zen 5” IPC Uplift for PCs

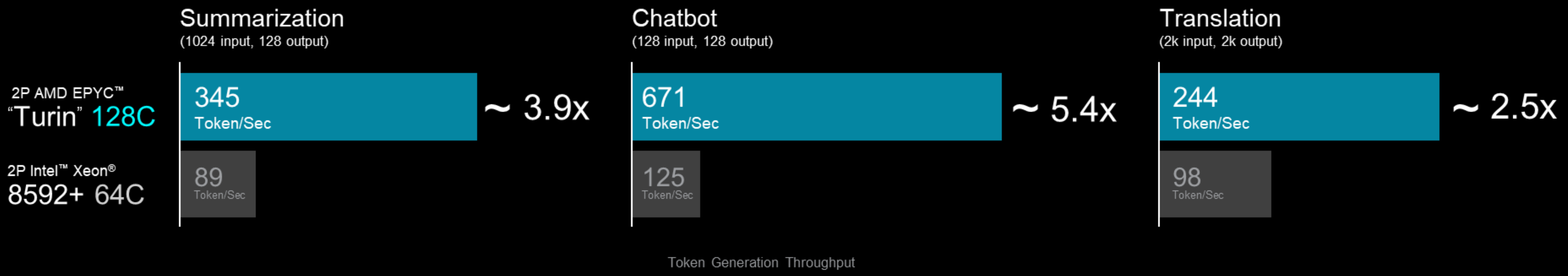


*all results are ‘up to’. See endnote GNR-03

5th Gen AMD EPYC™ “Turin” Processors

Molecular Dynamics with NAMD

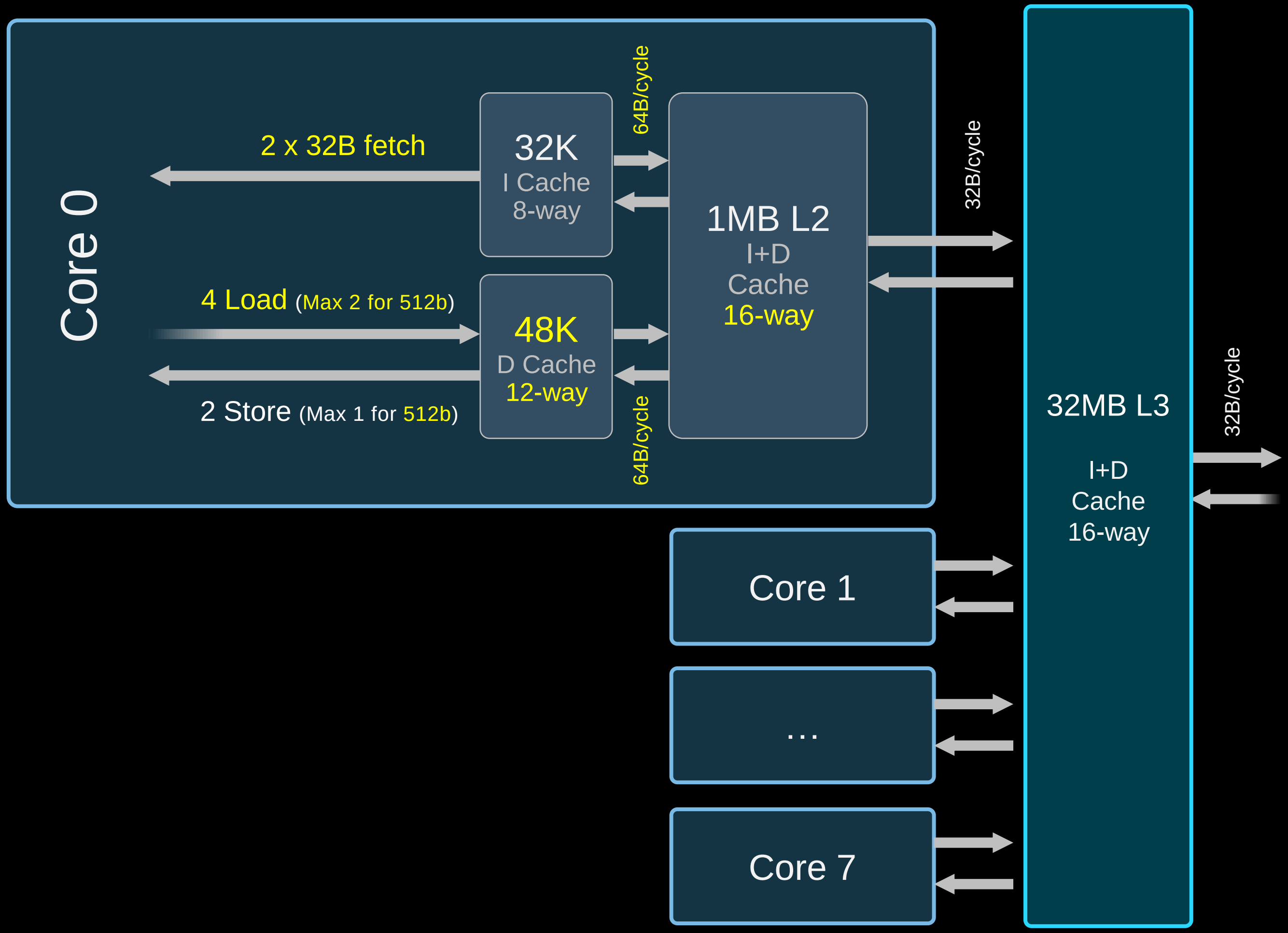




See endnote CTT-001 Llama 2, -7B, INT4, Token Generation Latency at 50 ms, See endnote CTT-002

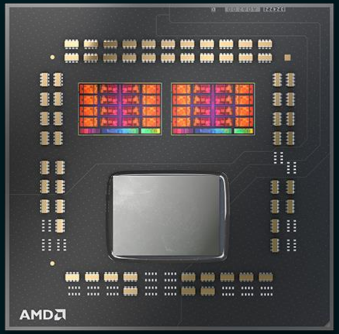
"Zen 5" Core Complex Speeds and Feeds

- 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” Core Complexes Across SoCs

8 "Zen 5"
32MB L3

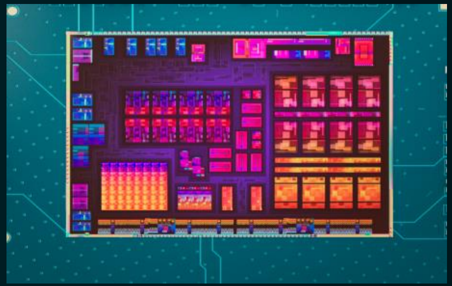


“Granite Ridge”

- Homogenous architecture
- Up to 2 core complex die (CCD)
- Each CCD: 8 "Zen 5" – 32MB L3 core complex (CCX)

4 "Zen 5"
16MB L3

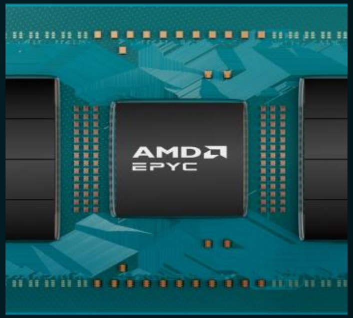
8 "Zen 5c"
8MB L3



“Strix Point”

- Heterogenous architecture
- Dual CCXs
- 4 "Zen 5" – 16MB L3, 8 "Zen 5c" – 8MB L3

Zen 5, Zen5c
More..



Futures

- Smaller, larger CCXs
- Homogenous or heterogenous
- Data center to embedded

AMD "Strix Point" SoC

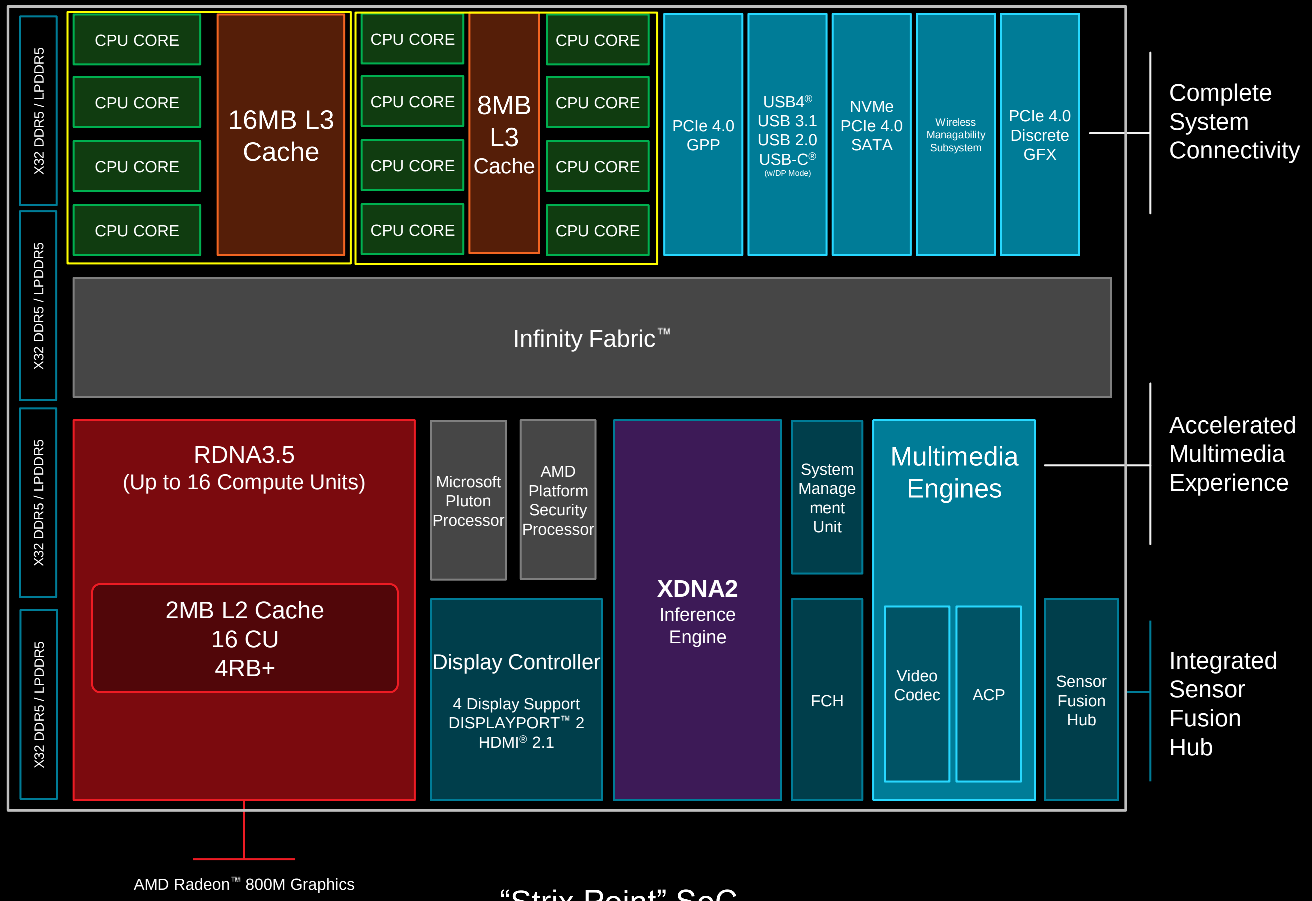
CPU
- 4C8T "Zen 5" – 1MB L2/core, 16MB L3 CCX
- 8C16T "Zen 5c" – 1MB L2/core, 8MB L3 CCX
- Datapath – 32B/cycle port each

GPU
- 8 WGP (16 CU) RDNA 3.5
- Datapath – 4 x 32B/cycle ports

NPU
- 4 x 8 Array XDNA 2 Inference Engine
- Datapath – 32B/cycle

Accelerators / uControllers
- Video encode/decode
- Audio co-processor
- Display controller
- System management, security, wireless manageability

IO
- 128b LPDDR5/DDR5 (7500/5600 MT/s)
- 16L PCIe Gen4
- 4 simultaneous display streams
- 8 USB ports
  - 2 USB4 v1
  - 1 USB3 Type-C
  - 2 USB3.2 Gen2
  - 3 USB2
- I2c, SPI/eSPI, GPIO



"Strix Point" SoC

AMD

AMD RDNA™ 3.5

Mobile Optimized RDNA™ 3 Architecture

Architectural Improvements

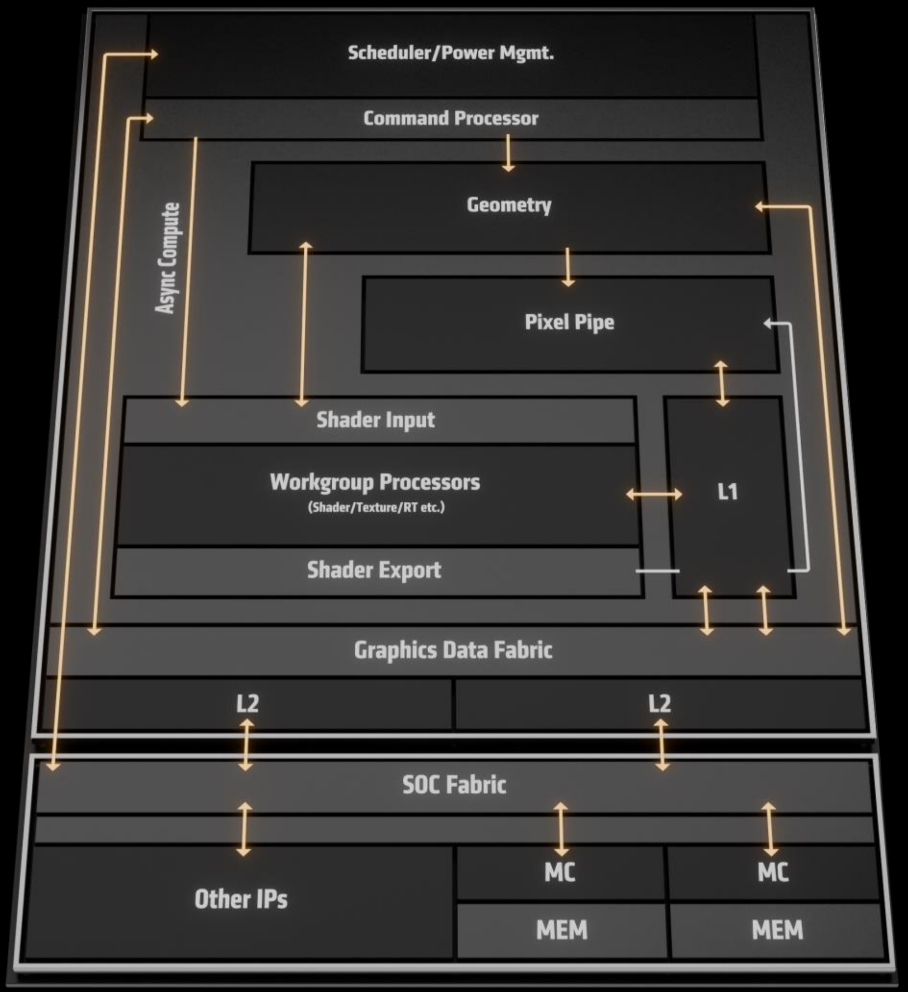
Larger Engine
- 1SE, 2SA, 8 WGP, 4 RB+, 2MB GL2 Engine
- 2.9G Fmax results in >11 TFLOPs (~30% higher)

Texture Subsystem
- 2x Sampler Rate, Point sampling acceleration

Shader Subsystem
- 2x Interpolation and Comparison rates
- Floating point in SALU
- Skip single-use VGPR writes

Raster Subsystem
- Sub-batching allows hardware to be efficient
- Programmable bin order

Memory Subsystem Improvements
- LPDDR5 awareness
- Improved compression



AMD XDNA™ 2 Architectural Innovations

World's First "Win24 Ready" NPU on x86 Processor

AMD XDNA™ 2 Architecture

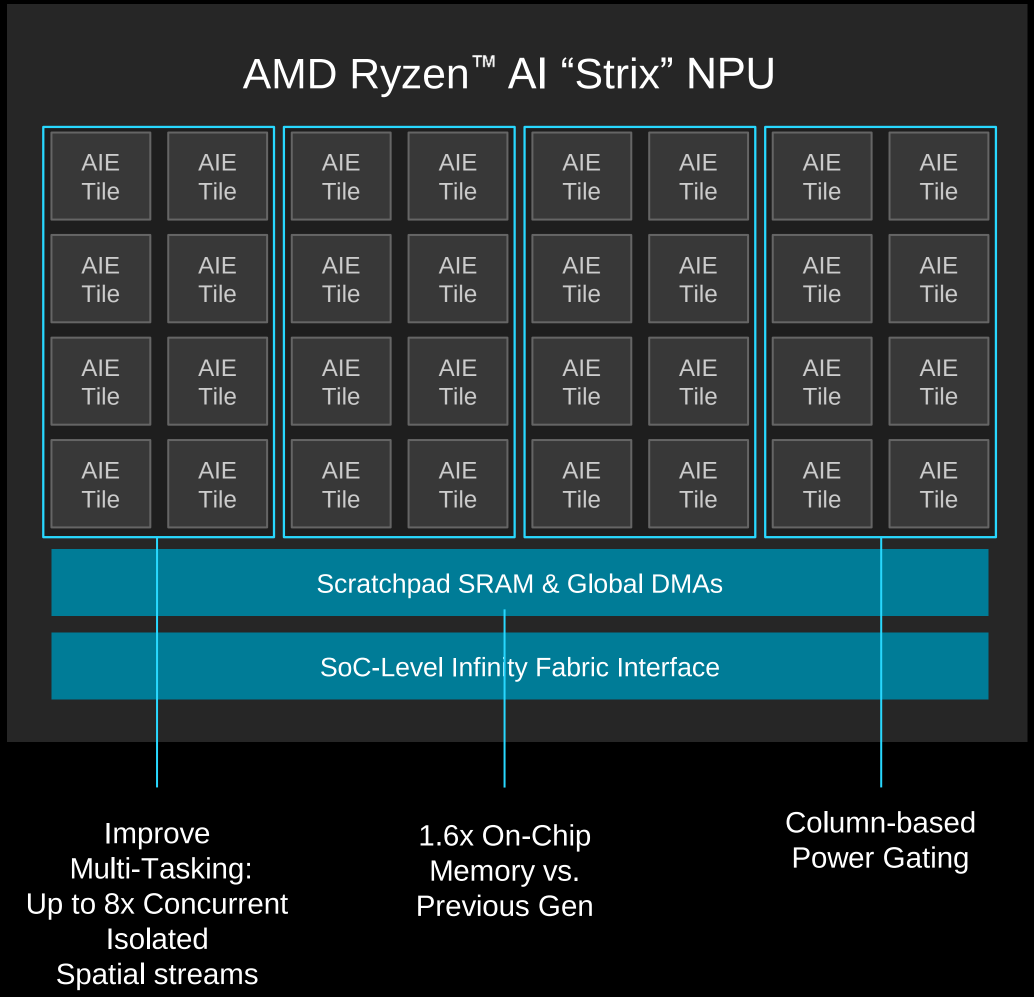
Broad AI Model Support
- Generative AI, Unlocking new AI PC Experiences

Peak Performance
- 50 INT8 TOPS
- 50 Block FP16 TFLOPS

Gen-on-Gen Improvements from Phoenix
- 2x more concurrent spatial streams
- 1.6x on-chip memory capacity

Advanced Features
- Block floating point support
- Enhanced support for non-linear functions (tanh, exp)
- 50% weight sparsity

Improved Power Efficiency
- Per column power gating
- Up to 2x Perf/W improvement over prior generation (8040 Series)



AMD “Granite Ridge” SoC

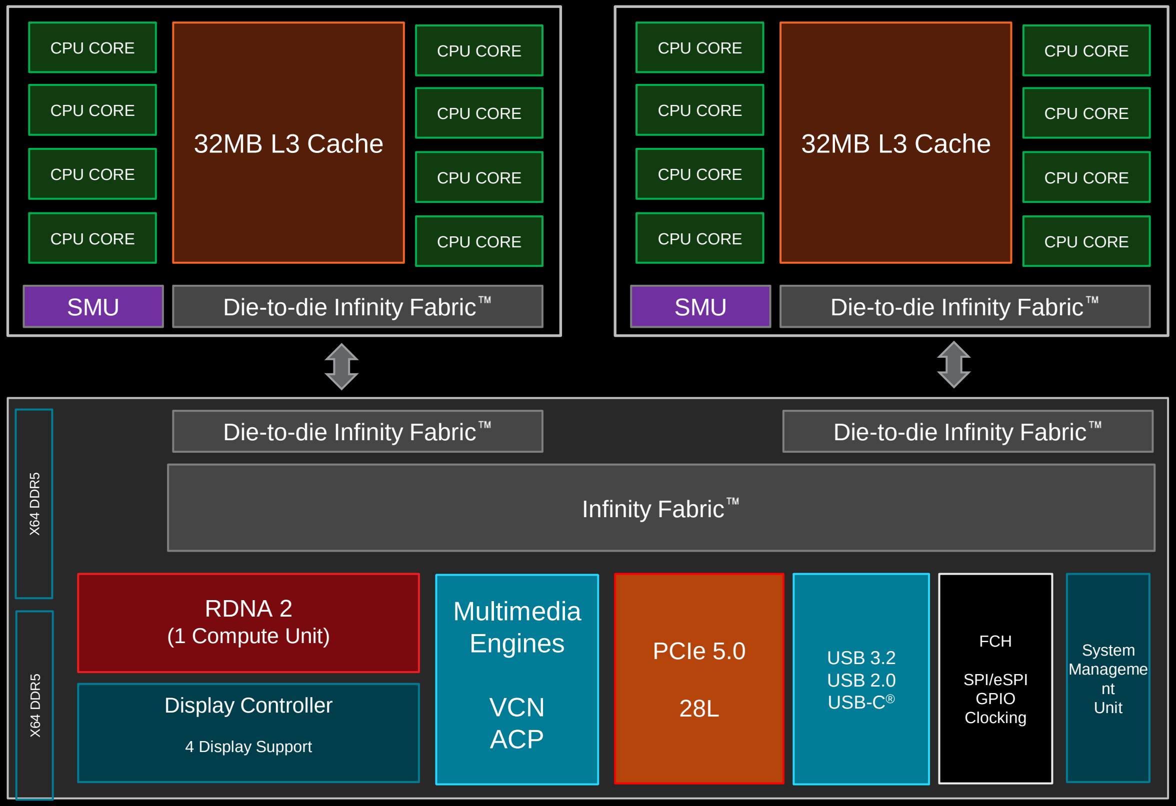
“Zen 5” based Desktop CPU

CPU
- Upto 2 x 8C16T – 1MB L2/Core, 32MB L3 CCDs
- 512b datapath for FP, optimized for high frequency
- Die-to-die Infinity Fabric™
- Datapath – 32B/cycle port each

GPU
- 1 WGP (2 CU) RDNA 2
- Datapath – 2 x 32B/cycle ports

IO
- 128b DDR5 5600 MT/s
- 28L PCIe® Gen5
- 5 USB ports
  - 3 USB3.3 Type-C
  - 1 USB3.2 Gen2 Type-A
  - 1 USB2
- 4 Simultaneous display streams
- I2C, SPI/eSPI, GPIO

Package
- Fully Compatible with AM5



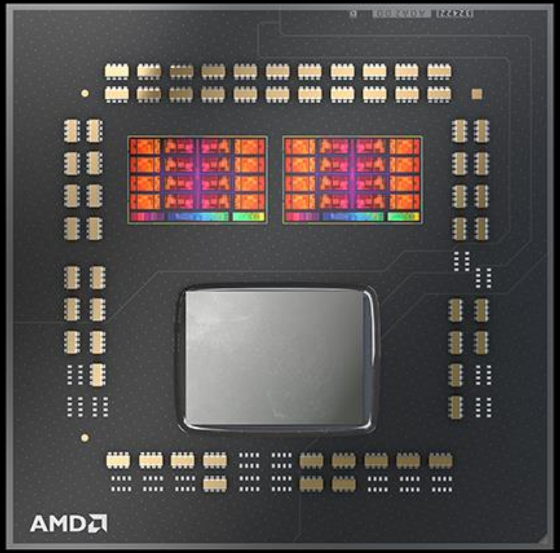
“Granite Ridge” SOC

AMD

Summary: AMD Delivers Again!

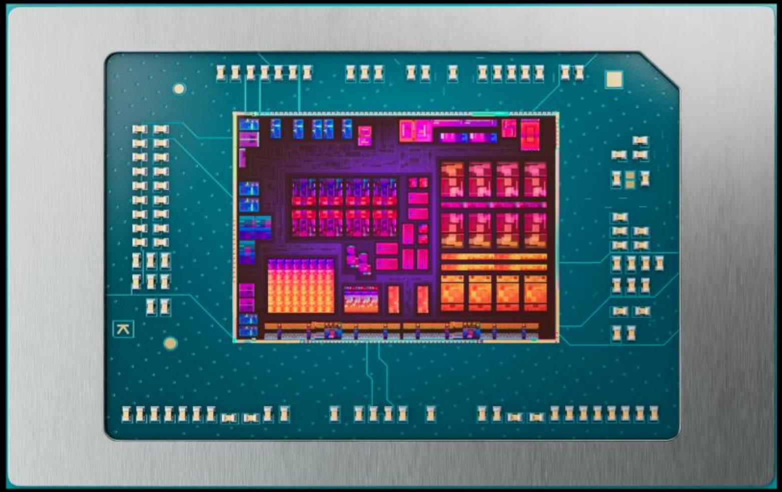
"Zen 5"

- On-cadence with major performance increase
- Balanced cross-core 1T/2T instruction and data throughput
- AVX512 with 512bit FP data-paths for throughput and AI uplift
- Efficient, performant, configurable solutions which scale:
  - "Zen 5" - peak performance / "Zen 5c" - efficient throughput
  - Configurable FP and cache hierarchy
  - Supports 4nm and 3nm



"Strix Point", "Granite Ridge"

- Commanding Performance and Gaming Leadership with Granite Ridge
  - Continuing our support of the AM5 infrastructure
- With increased compute and efficiency across the entire chip, Strix Point delivers a complete, uncompromising AI PC solution
  - Continuing support in the FP8 infrastructure



AMD continues to drive Leadership Performance and Efficiency

Endnotes

GD-122: “Zen” is a codename for AMD architecture and is not a product name.

R5K-003: Testing by AMD performance labs as of 09/01/2020. IPC evaluated with a selection of 25 workloads running at a locked 4GHz frequency on 8-core "Zen 2" Ryzen 7 3800XT and "Zen 3" Ryzen 7 5800X desktop processors configured with Windows® 10, NVIDIA GeForce RTX 2080 Ti (451.77), Samsung 860 Pro SSD, and 2x8GB DDR4-3600. Results may vary.

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.

GNR-03: Testing as of May 2024 by AMD Performance labs. "Zen 5" system configured with: Ryzen 9 9950X GIGABYTE X670E AORUS MASTER motherboard, Balanced, DDR5-6000, Radeon RX 7900 XTX, VBS=ON, SAM=ON, KRACKENX63 vs. "Zen 4" system configured with: Ryzen 7 7700X, ASUS ROG Crosshair X670E motherboard, Balanced, DDR5-6000, Radeon RX 7900 XTX, VBS=ON, SAM=ON, KRAKENX62 {FixedFrequency=4.0 GHz}. Applications tested include: Handbrake, League of Legends, FarCry 6, Puget Adobe Premiere Pro, 3DMark Physics, Kraken, Blender, Cinebench (n-thread), Geekbench, Octane, Speedometer, and WebXPRT. System manufacturers may vary configurations, yielding different results. GNR-03.

CTT-001 - AMD testing as of 05/30/2024. The detailed results show average ns/day (nanoseconds per day) for the 2P Intel Xeon 8592+ system and the AMD 5th Gen EPYC (pre-production silicon) system running the namd-stmv20m test of the NAMD 2.15alpha2 benchmark. EPYC test run results followed by Xeon test run results in parenthesis. • namd-stmv20m: EPYC Normalized to Xeon 3.085x, 3.049x, 3.059x for an average of ~3.06x the performance/~206% higher performance System configurations: AMD: 2 x 128-core AMD 5th Gen EPYC on AMD reference platform; Memory: 1.5 TB RAM; BIOS: Pre-production; BIOS options: SMT=OFF, NPS=4, OS: RHEL 9.4 kernel 5.14.0-427.16.1.el9_4.x86_64; Kernel options: amd_iommu=on iommu=pt mitigations=off; Runtime options: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor - Performance, Disable C2 States Intel: 2 x 64-core Intel Xeon 8592+ production system; Memory 1.0 TB RAM; Hyperthreading=OFF, Profile=Maximum Performance; OS: RHEL 9.4 kernel 5.14.0-427.16.1.el9_4.x86_64; Kernel options: processor.max_cstate=1 intel_idle.max_cstate=0 iommu=pt mitigations=off; Runtime options: Clear caches, NUMA Balancing 0, randomize_va_space 0, THP ON, CPU Governor=Performance Results may vary based on factors including but not limited to production silicon, system configurations, software versions and final BIOS version and settings.

CTT-002: AMD internal testing as of 5/31/2024 on Llama2-7B-CHAT-HF comparing 2P 5th Gen EPYC 128C “Turin” (pre-production) powered server to 2P 5th Gen Intel Xeon 8592+ powered server. All testing with weights quantized to INT4 and with latency under msec. System Configurations: 2P 5th Gen EPYC 128-C pre-production silicon (128C/256T) on AMD reference system, BIOS: pre-production (Determinism=Power, NPS=1), Memory: 1.5TB, OS: Ubuntu® 22.04.3 LTS | 5.15.0-105-generic. 2P Xeon Platinum 8952+ (64C/128T) production system, (SNC=Disabled), Memory: 1TB ; OS: Ubuntu 22.04.3 LTS | 5.15.0-94-generic Results may vary due to factors including but not limited to production silicon, system configurations, software versions and BIOS settings.

Disclaimer & Attribution

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 SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2024 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names are for informational purposes only and may be trademarks of their respective owners.

AMD