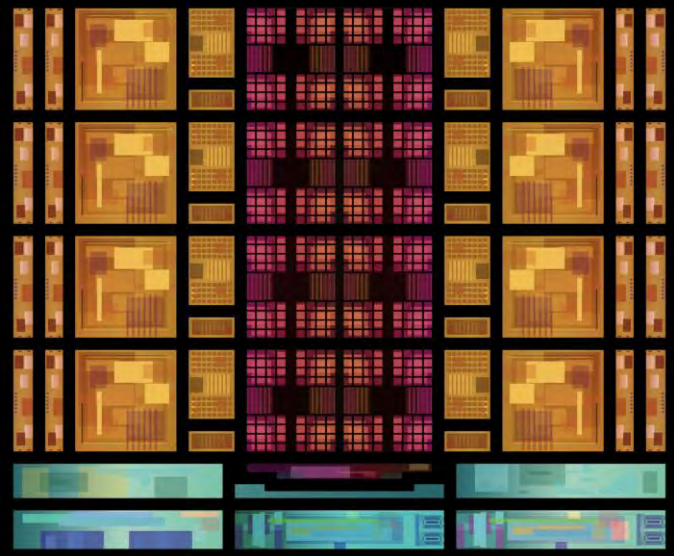
# Design Objectives

**Performance**

- Deliver another major 1T and 2T performance increase
- New Foundation of Compute for the Future
- AVX512 on 512-bit datapath to increase throughput and AI uplift

**Platform Support**

- Deliver “Zen 5” and “Zen 5c” (compact) core variants
- Support configurable FP512/FP256 datapath
- Support scaling and energy efficiency
- Support 4 and 3 nm
- Enhanced ISA capabilities



# “Zen5” 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 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**




AMD

# Optimized Branch Prediction and Fetch

## Branch 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

# Increased 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 can pair into 2, FP Pipes
- 2 store commit per cycle
- 64B fill/victim from/to L2 Dcache

## TLBs

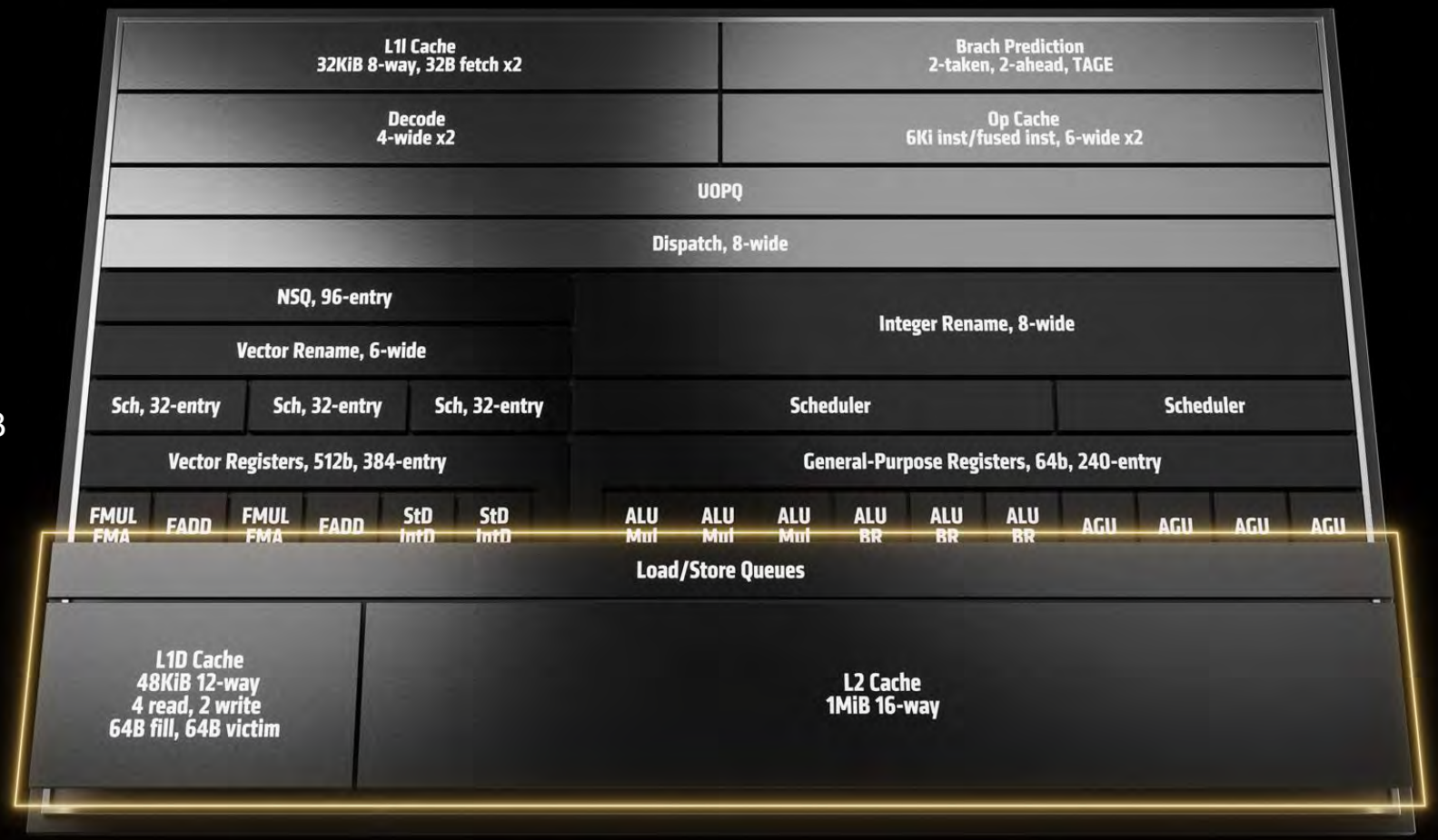
- L1: 96entry Fully associative all page size DTLB
- L2: 4K DTLB everything but 1G

## Larger In-Flight Window

- Load and store queue growth
- Store coalescing buffer growth
- Scalable load ordering queue

## Data prefetching

- New 2D stride prefetcher also 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 5" and "Zen 5c" in Heterogeneous SOC

**"Zen 5" and "Zen 5c" in separate core complexes**

**"Zen 5" Optimized for maximum 1T performance**

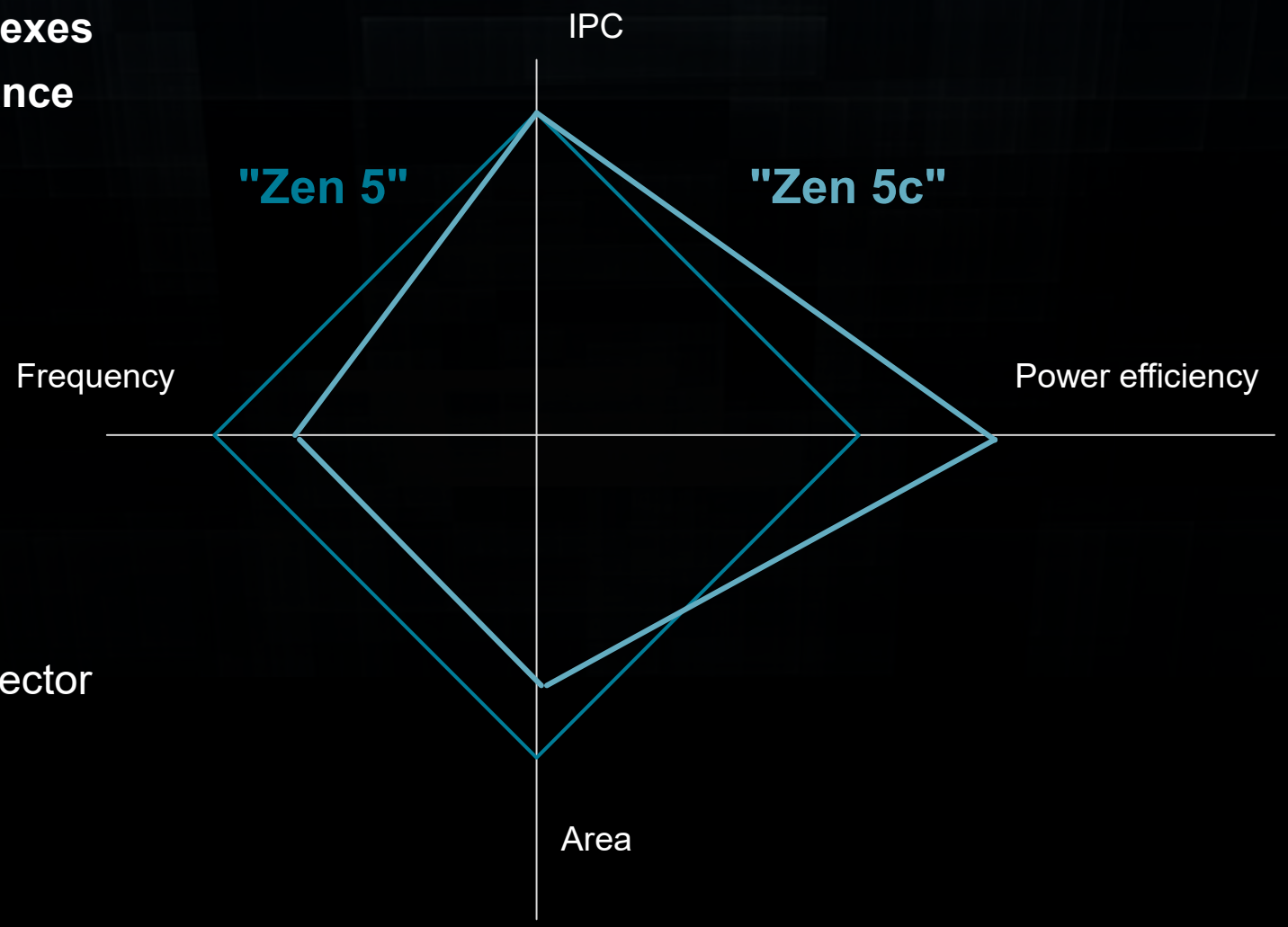
- High max frequency Target
- Large L3 per core

**"Zen 5c" Optimized for scalability**

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

**Simplifies software scheduling**

- Same IPC means no unique bottlenecks like vector performance
- Both support SMT
- Modulate between ultimate performance vs. efficiency
- Scheduling "mistakes" minimized over time

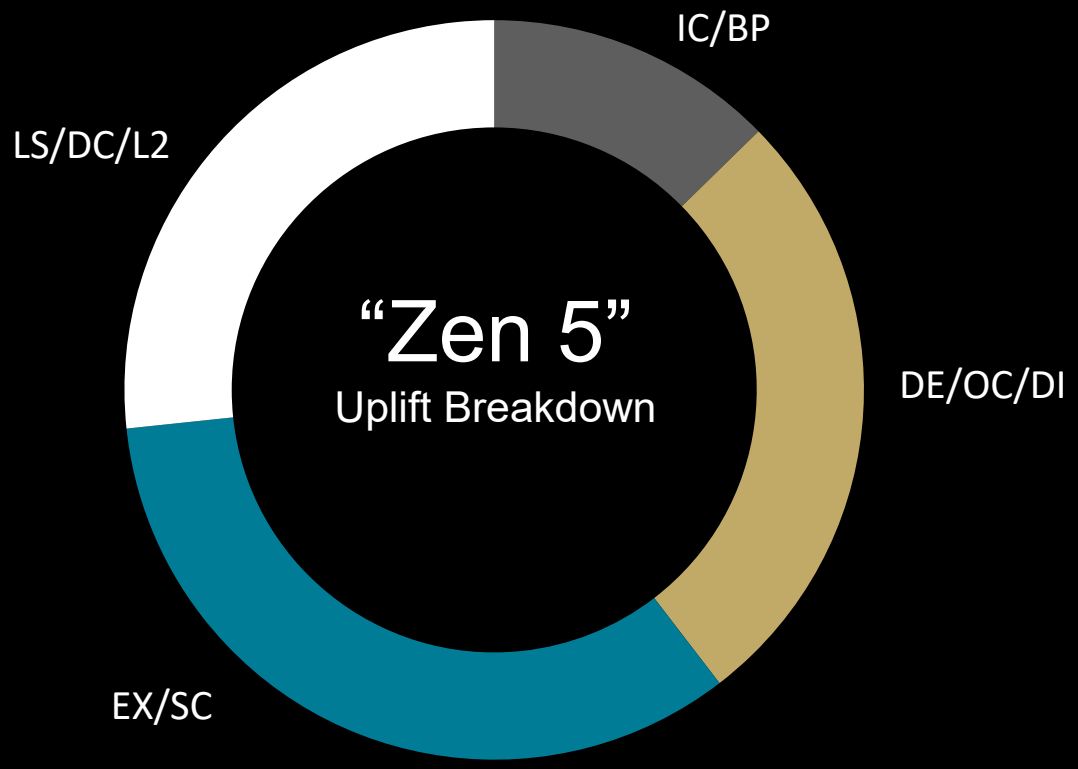


# New "Zen 5" ISA

| Attribute | ISA Feature |
| --- | --- |
| **Instructions** | ▪ MOVDIRI/MOVD64B – move 4,8 or 64 bytes as a direct store, bypassing caches<br>▪ VP2INTERSECT[DQ] – AVX512 vector pair intersection to a pair of mask registers<br>▪ VNNI/VEX – extends AVX512 instruction to VEX encoding<br>▪ 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<br>▪ Heterogeneous Topology |

# Key “Zen 5” vs. “Zen 4” Capabilities

| 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 |
| 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 |

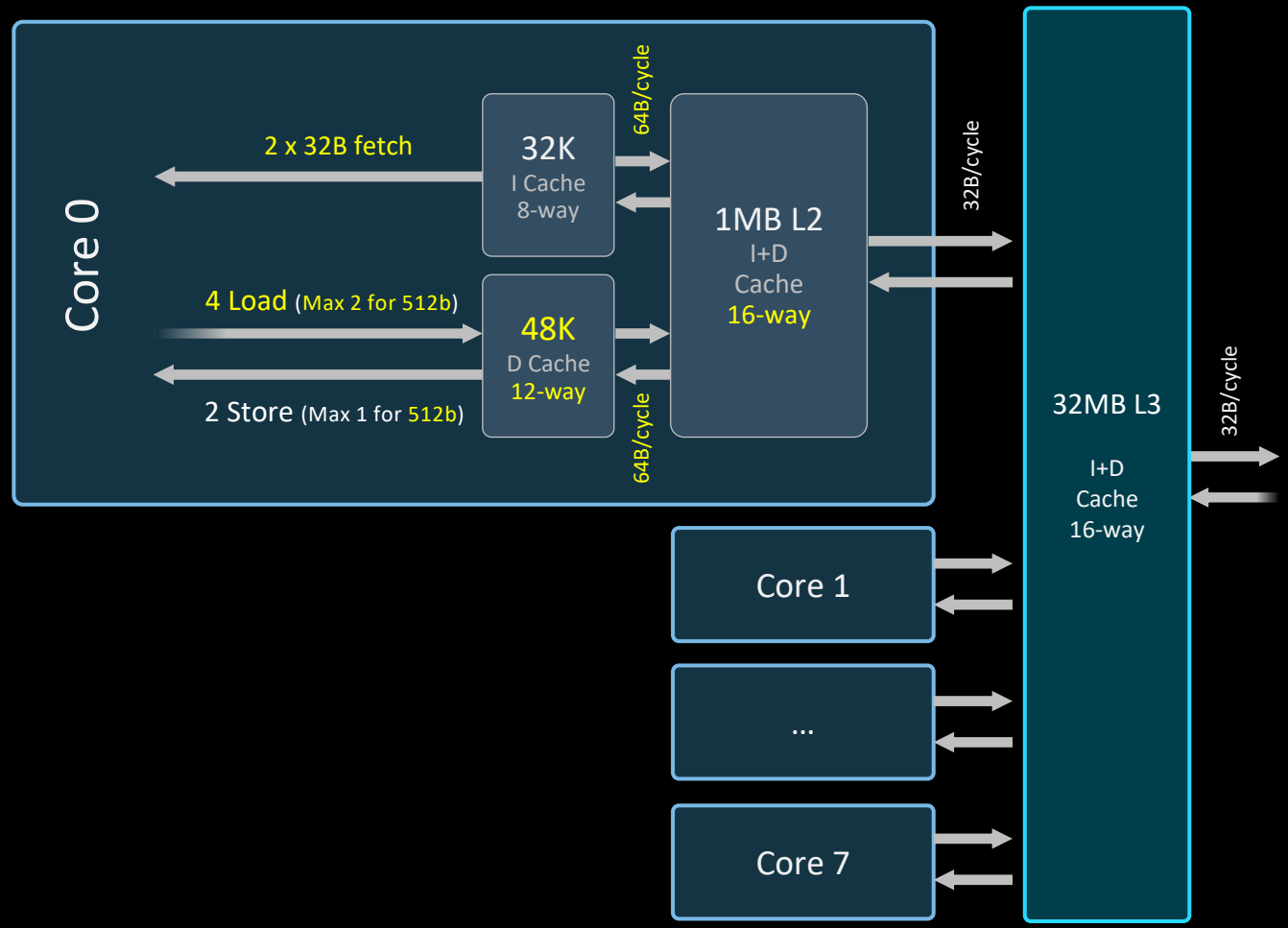





AMD

# “Zen 5” Core Complex speeds and feeds

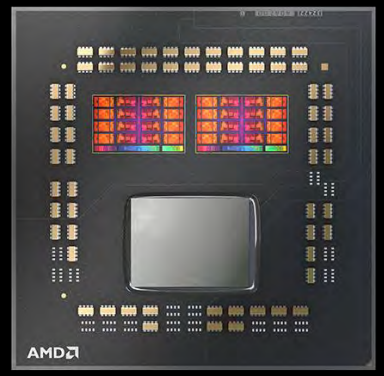
- Double the L2 associativity
- Double the L2 Bandwidth
- Low latency L3 with 320 L3 in-flight misses
- Baseline from “Zen 4”:
  - Fast private 1MB L2 cache
  - L3 shared among all cores in the complex
  - L3 is filled from L2 victims
  - L2 tags duplicated in L3 for probe filtering and fast cache transfer



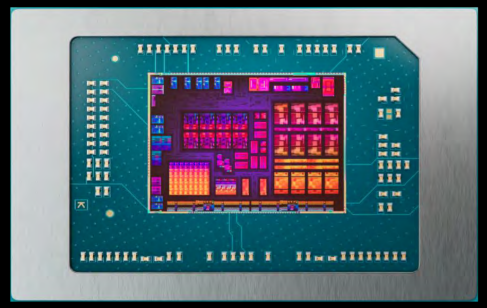


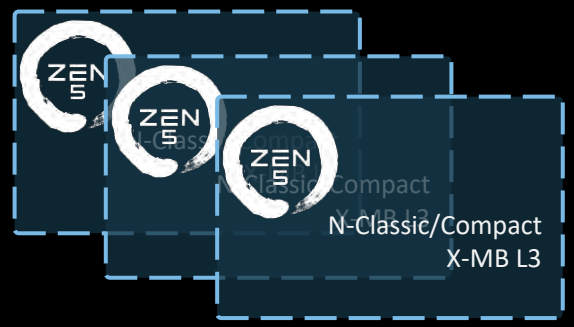

AMD

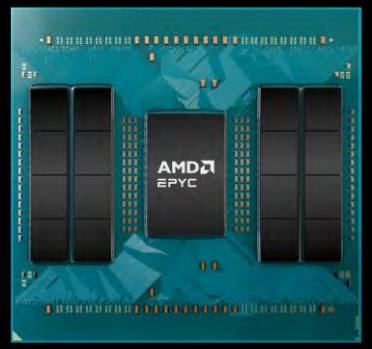
# Zen5 Core Complexes across SOCs

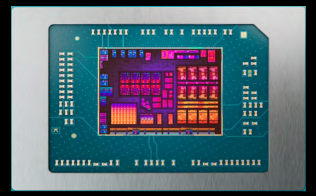









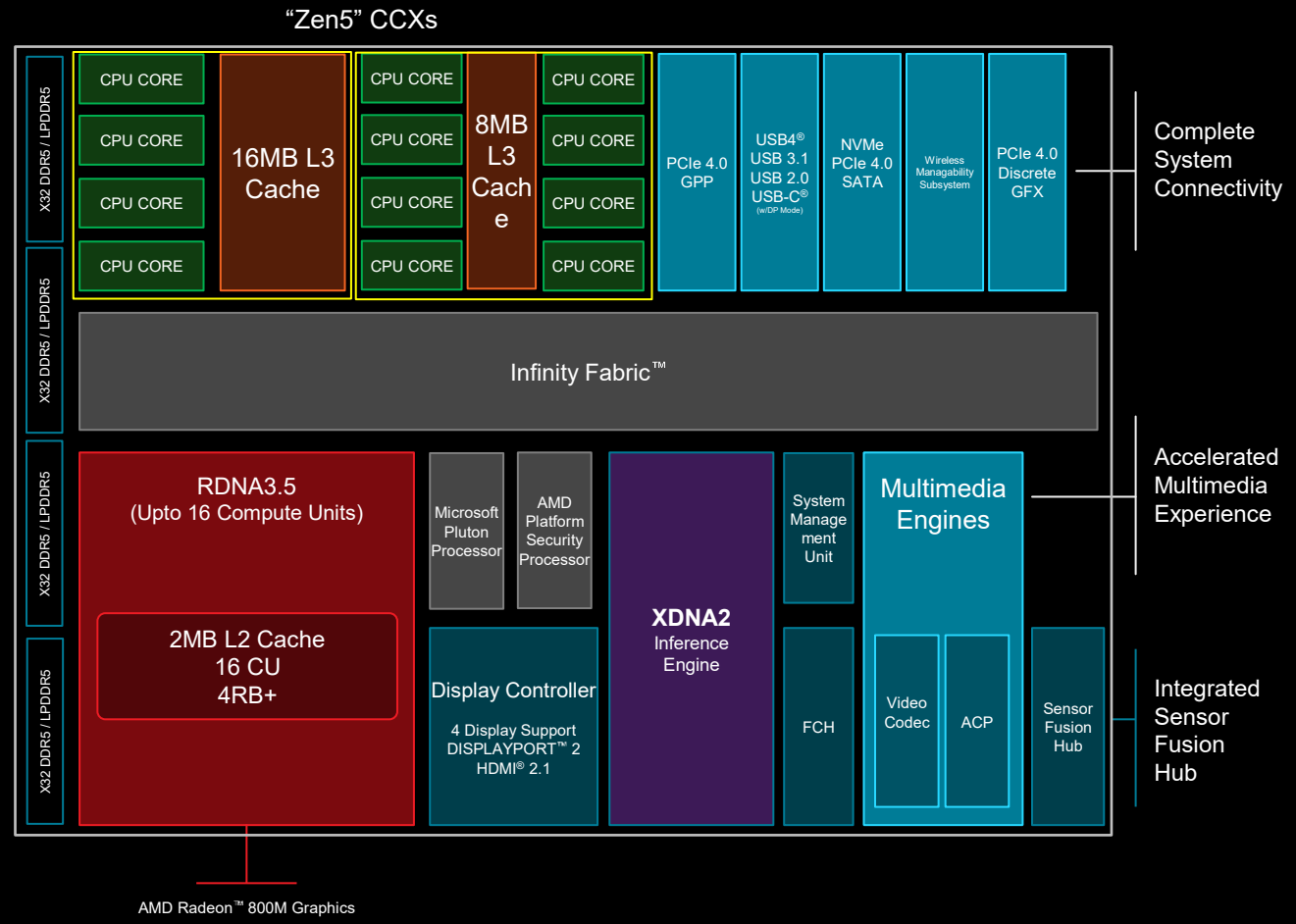
- **"Granite Ridge"**
  - Homogenous Architecture
  - Upto Dual CCDs
  - 8-Classic – 32MB L3
- **"Strix Point"**
  - Heterogenous Architecture
  - Dual CCXs
  - 4-Classic -16MB L3, 8-Compact – 8MB L3
- **Futures**
  - Smaller, Larger CCXs
  - Homogenous or Heterogenous
  - Data Center to Embedded



AMD

# AMD “Strix Point” SOC

- **CPU**
  - 4C8T Zen5 – 1MB L2/core, 16MB L3 CCX
  - 8C16T Zen5c – 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

## AMD RDNA™ 3.5 Improvements

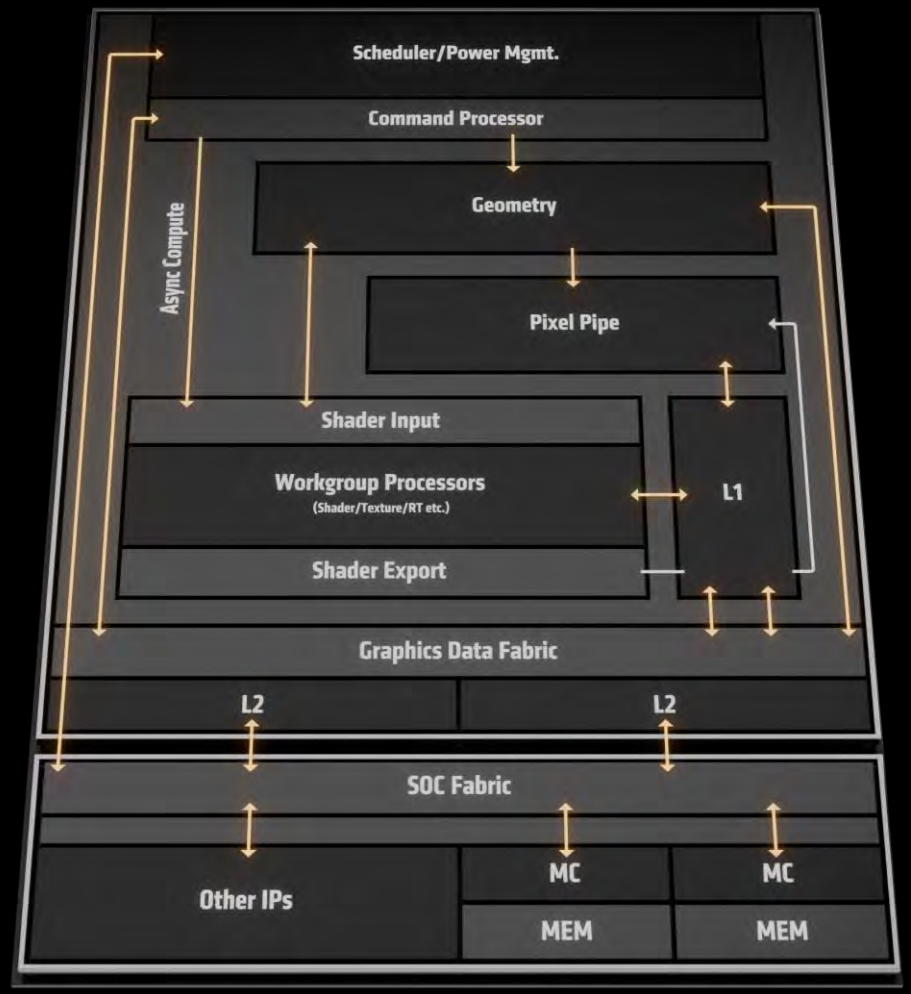
### Larger Engine
1SE, 2SA, 8 WGP, 4 RB+, 2MB GL2 Engine
2.9G Fmax results in >11 TFLOPs (~30% higher)

### Texture Subsystem
2x Sampler Rate, Point sampling acceleration

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

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

### Memory Subsystem Improvements
LPDDR5 awareness
Improved compression






AMD

# AMD XDNA™ 2 Architectural Innovations

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

## AMD XDNA™ 2 Architecture

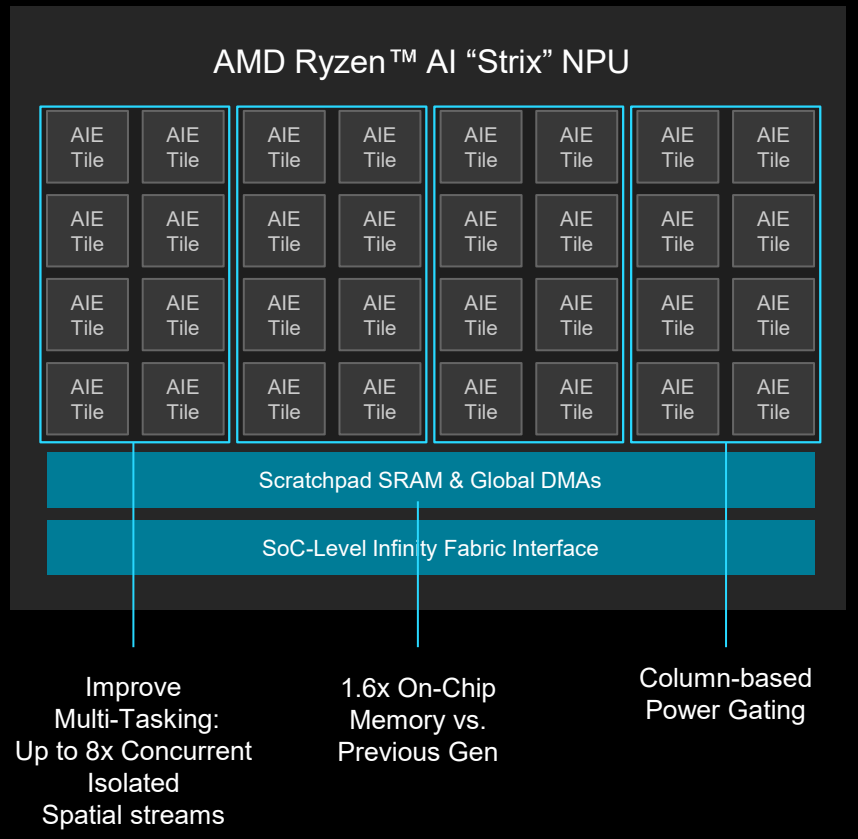
**Broad AI Model Support**
Generative AI, Unlocking new AI PC Experiences

**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

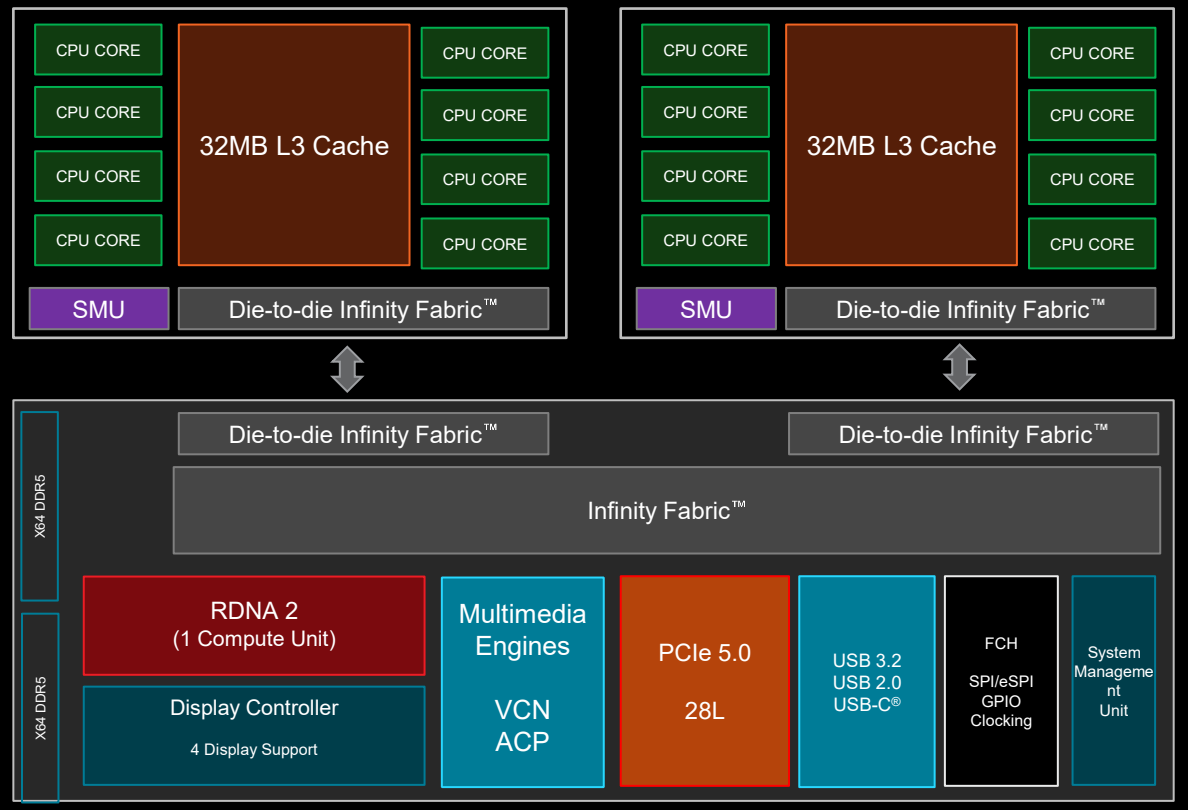
**Advanced Features**
Block floating point support
Enhanced support for non-linear functions (tanh, exp)
50% weight sparsity

**Improved Power Efficiency**
Per column power gating
Up to 2x Perf/W improvement



# AMD “Granite Ridge” SOC

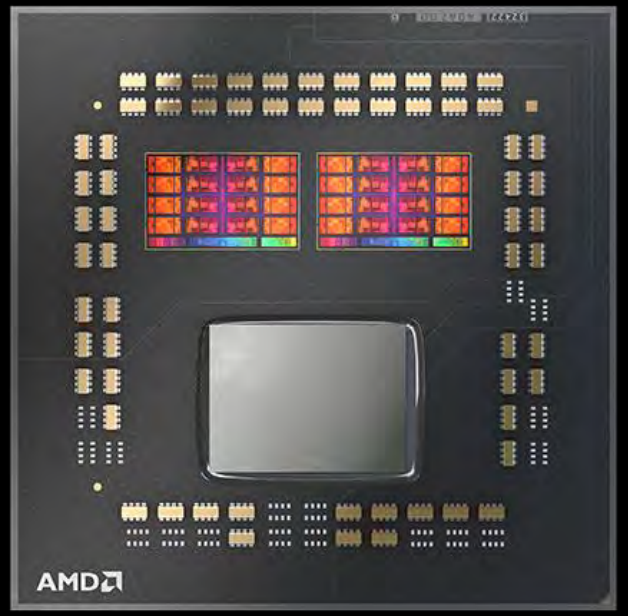
- **CPU**
  - Upto 2 x 8C16T – 1MB L2/Core, 32MB L3 CCDs
  - 512b datapath for FP, optimized for high frequency
  - Datapath – 32B/cycle port each
- **GPU**
  - 1 WGP (2 CU) RDNA 2
  - Datapath – 2 x 32B/cycle ports
- **Accelerators / uControllers**
  - Video Encode/Decode
  - Audio Co-processor
  - Display Controller
  - System Management, Security
- **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



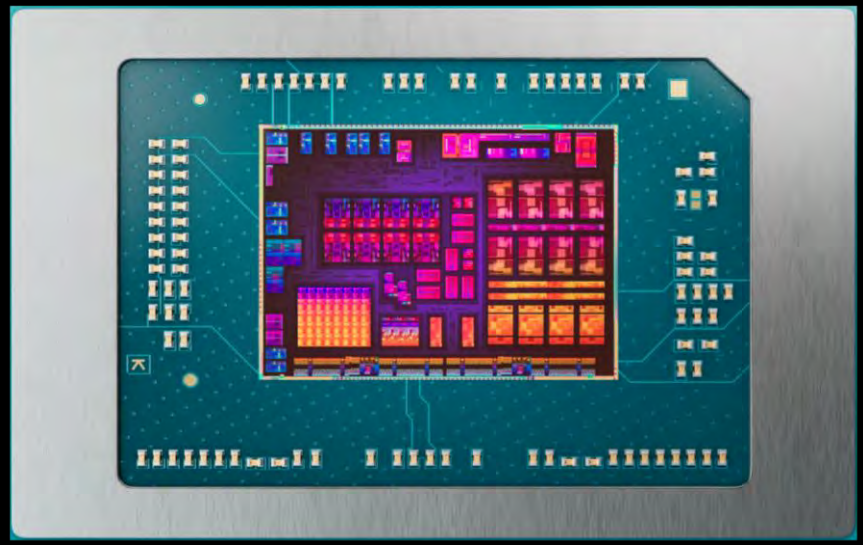

“Granite Ridge” SOC

# Summary: AMD Delivers Again!

- "Zen 5" :
- Yet another on-cadence 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:
  - Variants: Peak performance (“Zen 5” and “Zen 5c”)
  - Configurable FP and cache hierarchy
  - Multiple processes across the product line




- “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 no-compromise AI PC solution
  - Continuing support in the FP8 infrastructure



- **AMD** continues to drive Leadership Performance and Efficiency