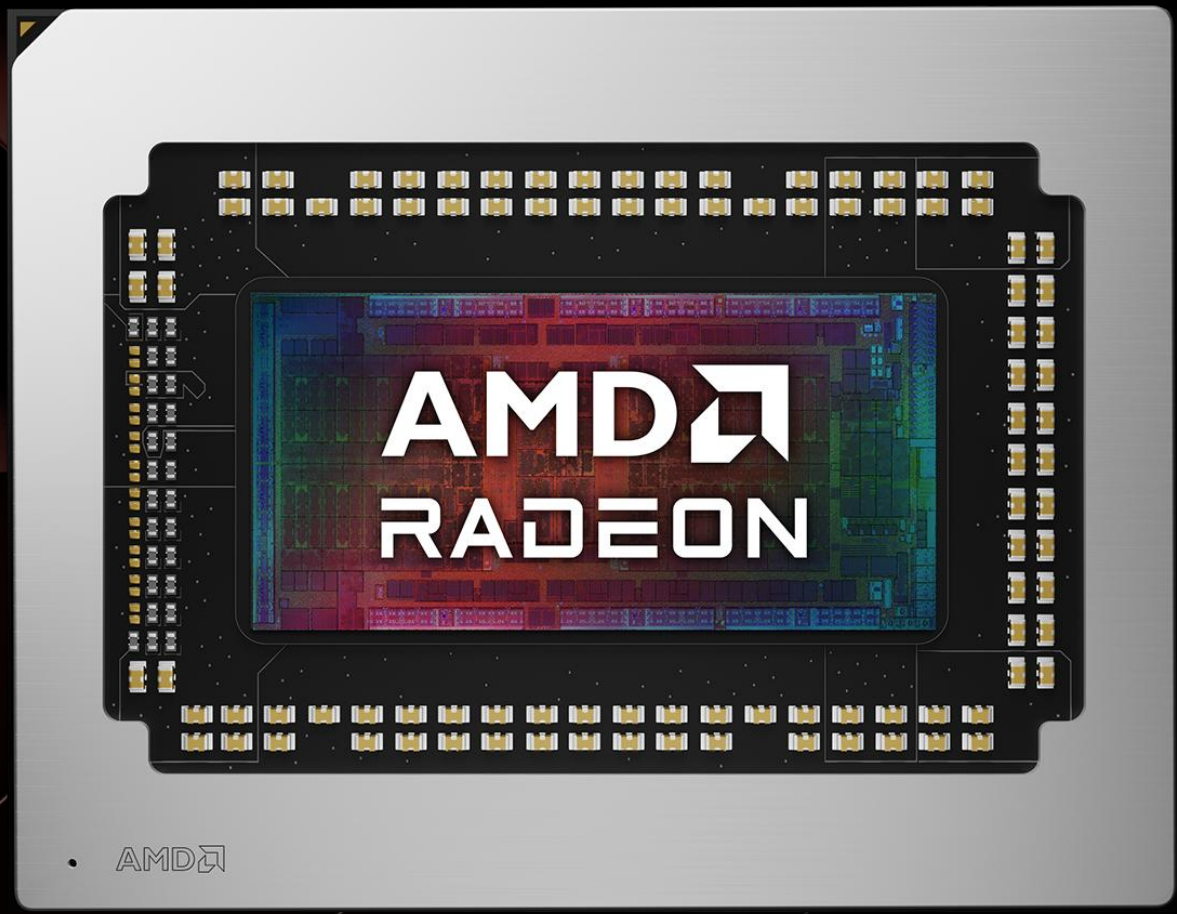
AMD RDNA 4 RADEON 9000 SERIES GPU

Andy Pomianowski

Laks Pappu

HOT CHIPS 2025



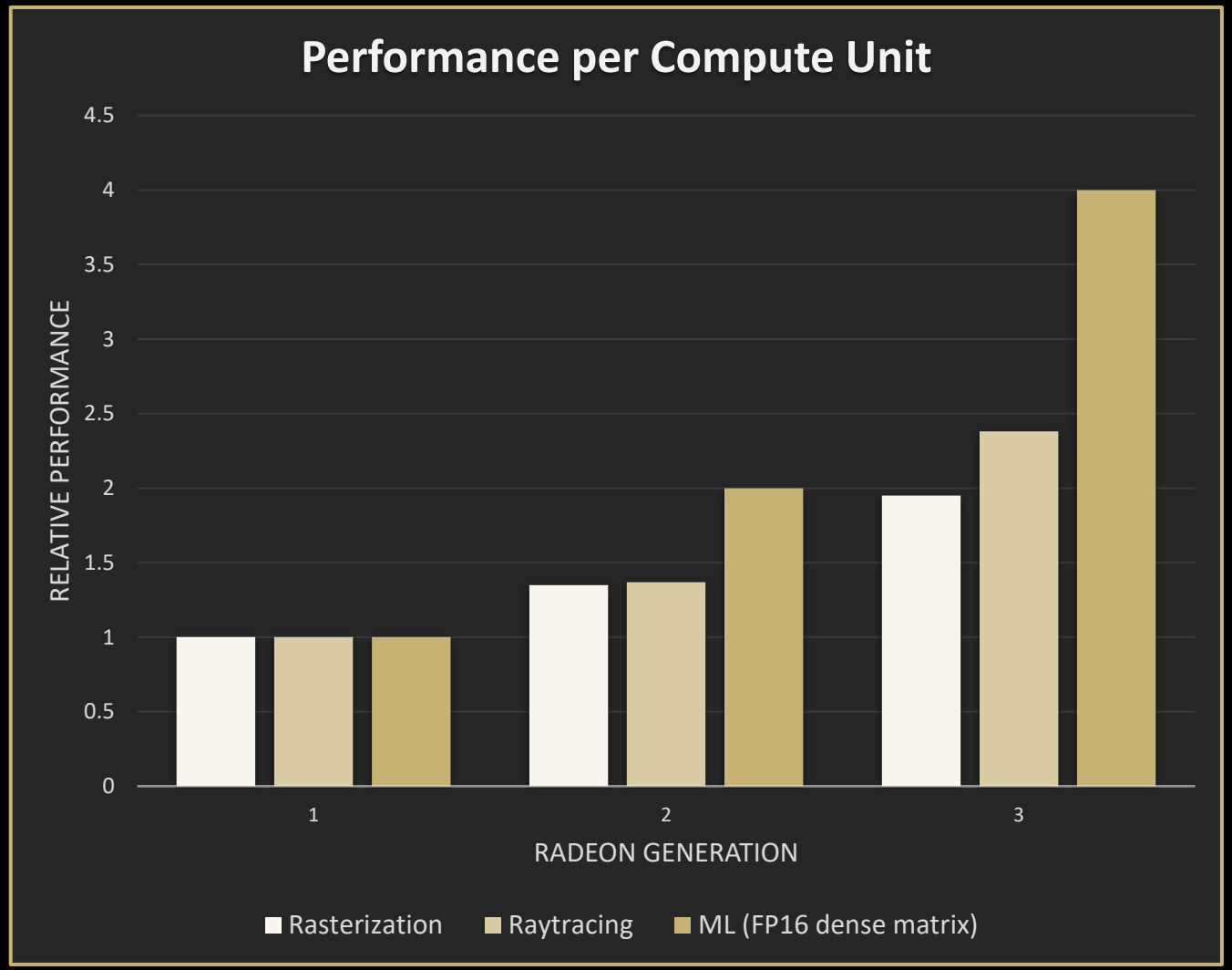
AMD
together we advance_gaming

RDNA 4 VISION

AN ARCHITECTURE BUILT FOR GAMING

WHAT'S NEW

- Heavily optimized for high end gaming workloads
- Improved Rasterization and Compute efficiency
- A step change in Raytracing performance
- Comprehensive high-performance ML support
- Enhanced bandwidth efficiency for all workloads
- Multimedia improvements for gamers and



RASTERIZATION PERFORMANCE BASED ON MEASUREMENTS OF CYBERPUNK ULTRA SETTINGS AT 4K RESOLUTION RUNNING BUILT IN BENCHMARK
RAYTRACING PERFORMANCE BASED ON MEASUREMENTS OF CYBERPUNK ULTRA RT SETTINGS AT 4K RESOLUTION RUNNING BUILT IN BENCHMARK
6800XT (72CU), 7900XT(84CU) AND 9070XT(64CU) USED FOR REFERENCE PERFORMANCE FOR GAMING BENCHMARKS FOR RDNA2, RDNA3, RDNA4
ML PERFORMANCE IS QUOTED AS PEAK OPERATIONS PER COMPUTE UNIT FOR DENSE MATRIX OPERATIONS ASSUMING EQUAL FREQUENCY

RDNA 4 : TARGETING ALL OF RENDERING

3rd Generation Ray Tracing

- Doubled Ray Intersection Rates
- Improved BVH Compression
- Oriented Bounding Boxes
- Accelerated Ray traversal and Shading

Optimized Cache System

- 64 MB 3rd Gen Infinity Cache
- 8MB L2 Cache
- 2MB Aggregate CU Cache

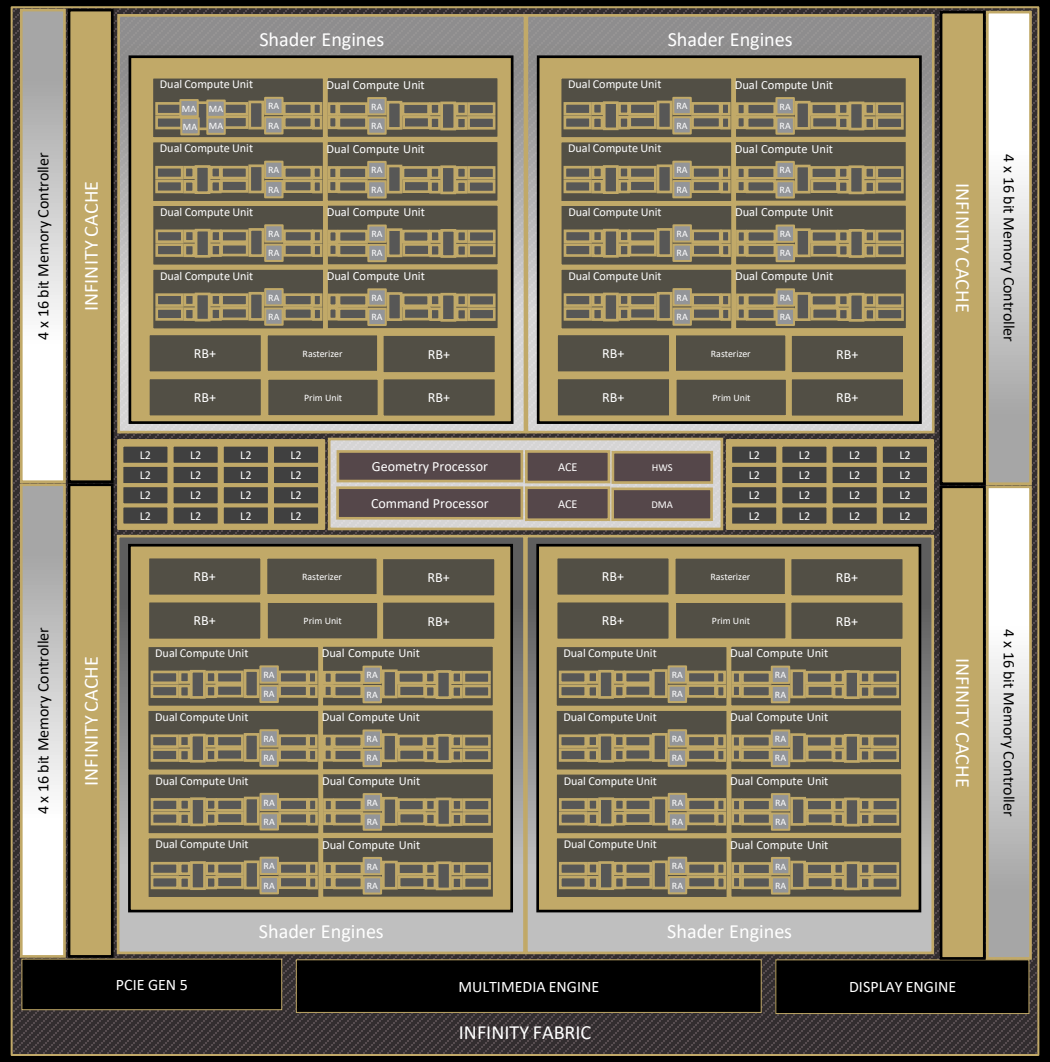
Improved Command Processor

- Enhanced packet accelerators

New Dual Media Engine

- Updated Encode/Decode engine
- Optimized for low-latency streaming
- Up to 25% quality improvement in AVC, H.264, H.265
- Double AV1 throughput

AMD Radeon RX 9070 XT



3rd Generation Matrix Acceleration

- Improved Tensor Dense Rates
- Adds 8b float data types
- Structured Sparsity Support
- ML based Super Resolution

High Speed GDDR6 Memory

- Up to 256b @20 Gbps with 16GB capacity
- Enhanced Memory Compression

AMD Radiance Display™ Engine

- DisplayPort™ 2.1a & HDMI 2.1b
- Updated scaling and sharpening engine

Optimized Monolithic Design

- TSMC 4nm process
- PCIE Gen 5
- 27.5 Billion Transistors
- 356.5 sqmm

RDNA 4

MEDIA AND DISPLAY ENGINES

Enhanced game streaming and recording

- Improved low latency encode quality
- AV1 encoding efficiency improved with B frames
- Increased Encoding Performance

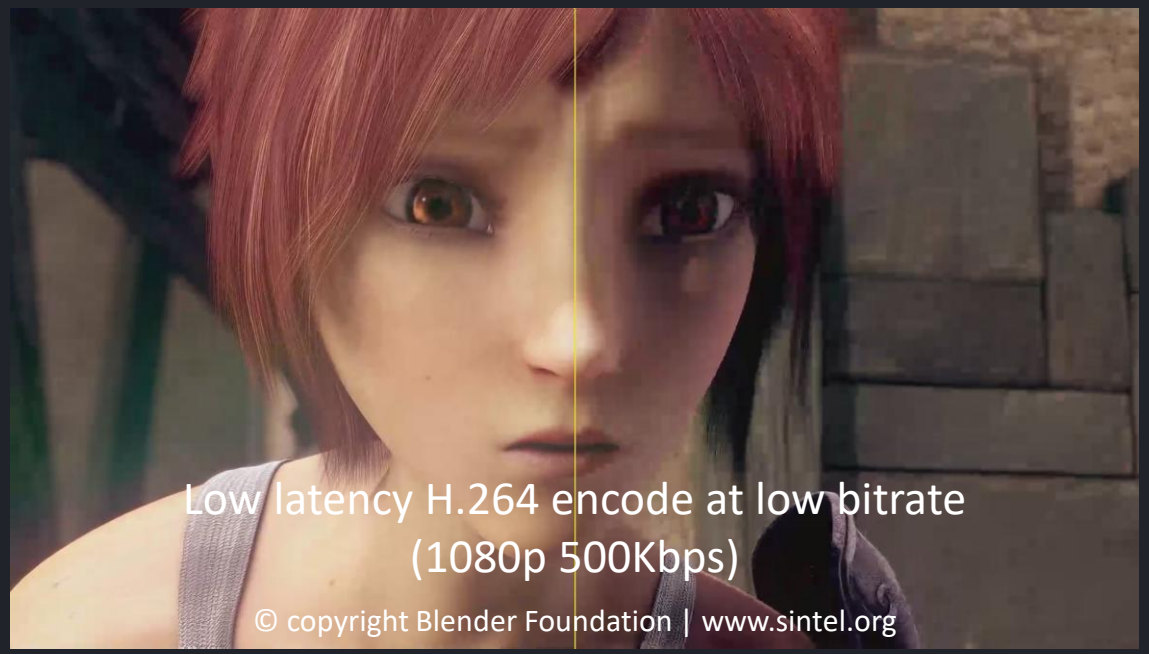
Low power video playback

- More than 50% performance uplift for AV1 and VP9

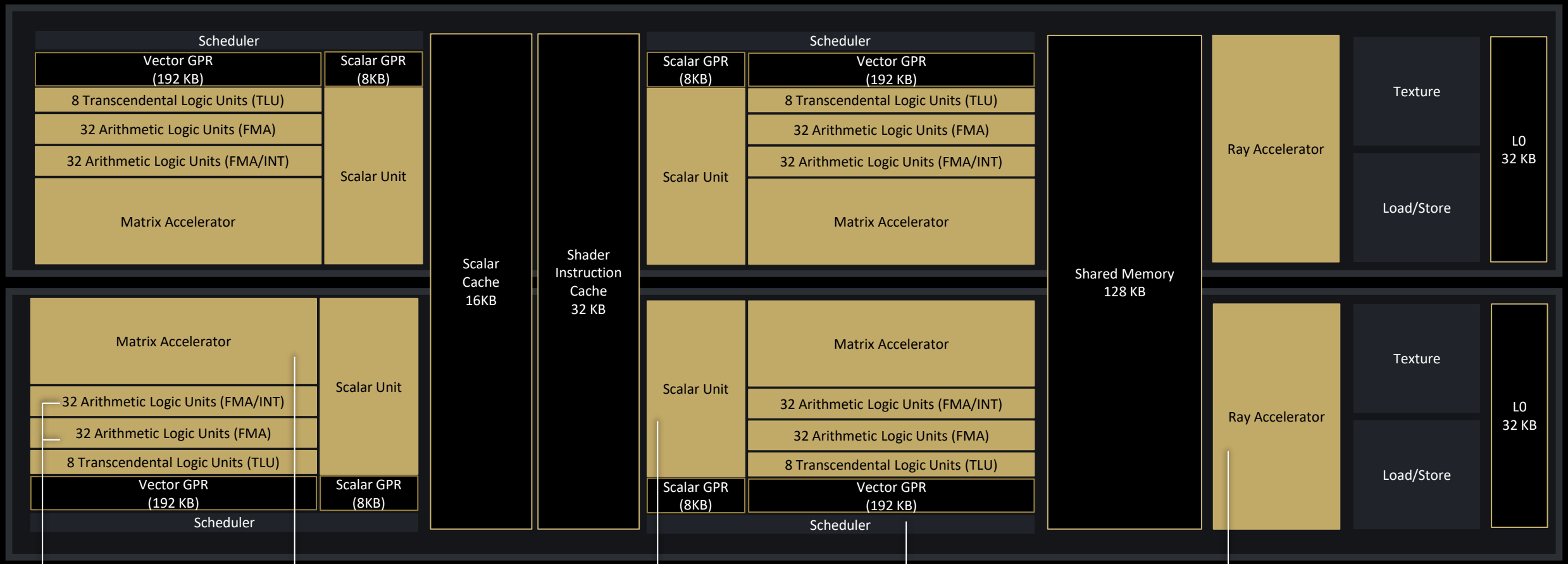
Enhanced FreeSync Power Optimization

- Lower idle power in most 2-display configs

Radeon Image Sharpening 2



COMPUTE ENGINE



Dual SIMD32 Vector Units

Enhanced Matrix Operations

Scalar Unit Updates

Scheduler Updates

Ray Accelerator Improvements

RDNA 4

RAYTRACING ARCHITECTURE

Enhanced Ray Accelerators

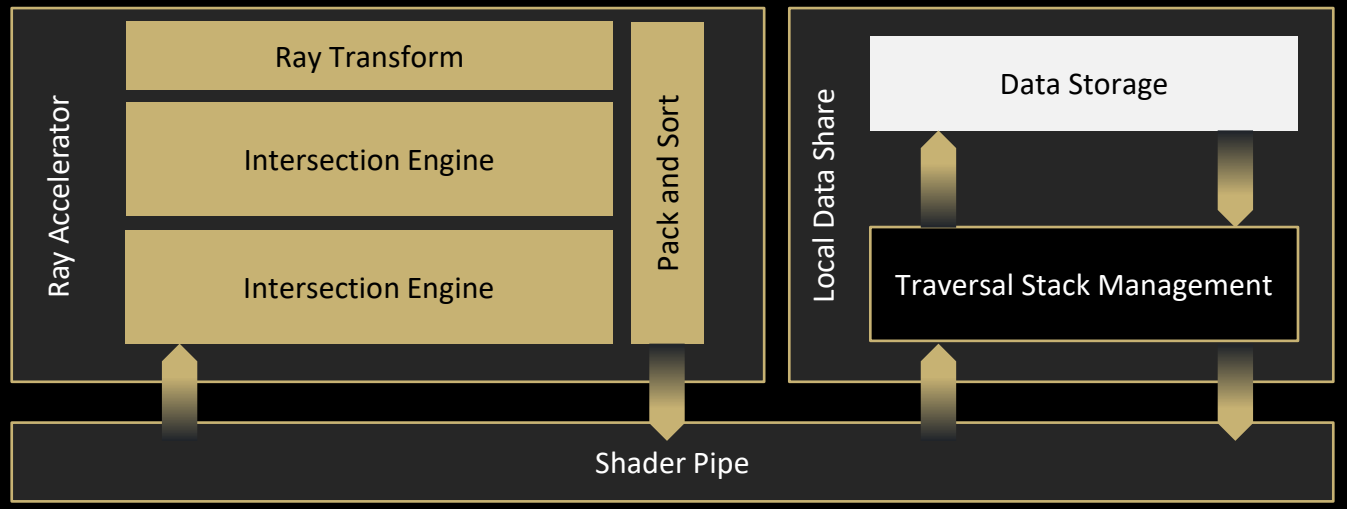
- 8 Ray/Box, 2 Ray/Triangle Units – 2x Increase
- Dedicated Hardware Instance Transform
- Ray Hardware Stack Management acceleration

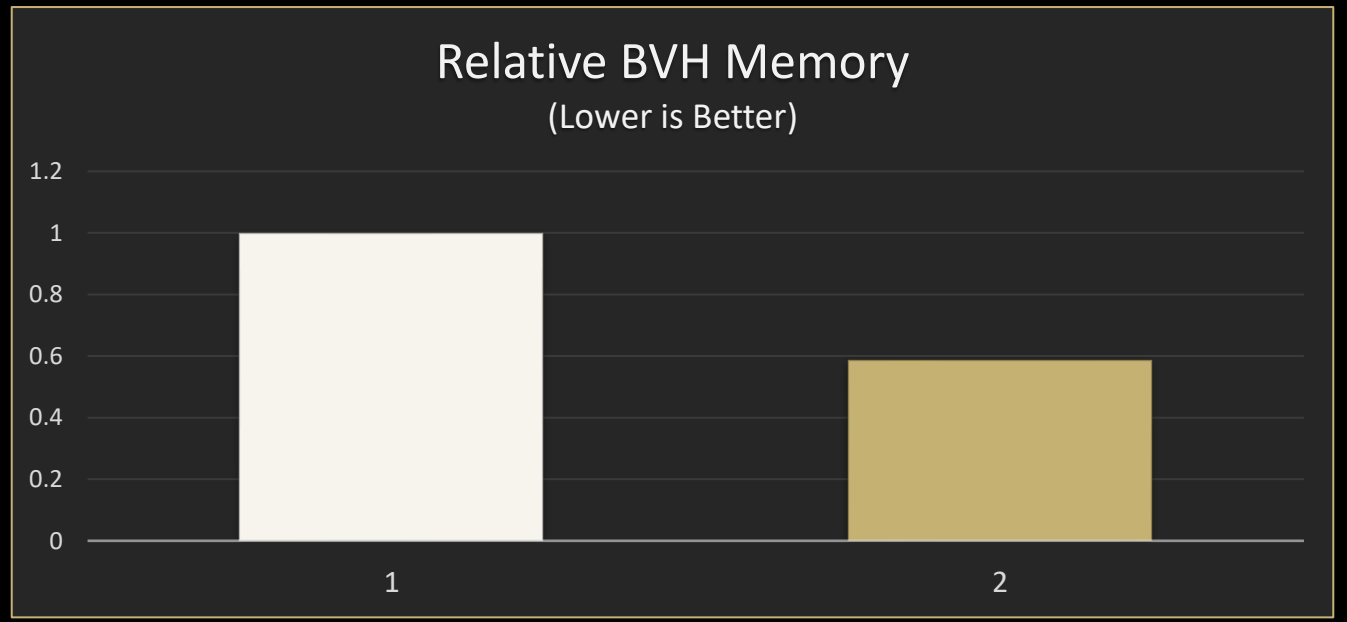
Improved BVH Structure and traversal performance

- BVH8 for reduced traversal steps and latency reduction
- New primitive node compression reducing BVH size
- Oriented bounding boxes reducing false intersections

Accelerated Shading

- Dynamic VGPR management increasing Ray Occupancy
- Out-of-order memory returns reducing latency

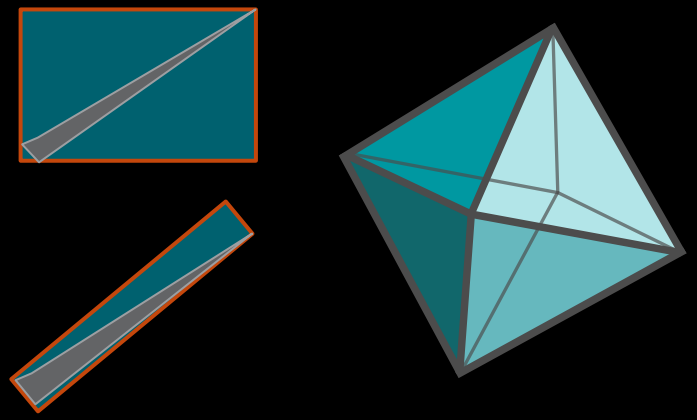




RDNA 4 BVH MEMORY CONSUMPTION BASED ON AVERAGE MEMORY FOOTPRINT ACROSS BROAD SAMPLE OF BVH TREES USING OPTIMIZED REFERENCE BUILDER

DEEPER DIVE

ORIENTED BOUNDING BOXES



World geometry can significantly affect traversal cost

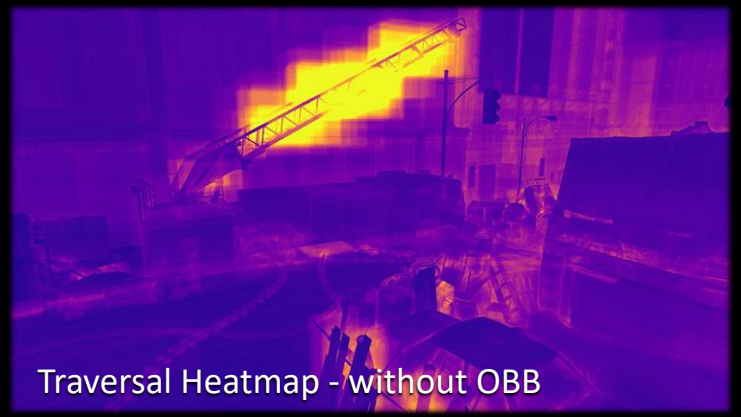
- Traditional BVH bounding boxes are axis-aligned in the world
- Boxes must bound all the geometry contained inside them
- Geometry that is not aligned to the world axes can cause boxes to grow
- This growth triggers false-positive intersections (the box is mostly empty space)

Our solution – allow a unique orientation per box node

- Encode a rotation with each box to more tightly bound the contained geometry
- Aligning the box to the geometry can remove much of the empty space
- Ray direction is transformed on entry to the box to match the encoded rotation

Results

- Off-axis geometry is much more tightly contained in the box nodes
- Number of traversal steps is significantly reduced on average
- Peak cost is reduced, eliminating traversal hotspots
- Performance of traversal improves approximately 10% (geometry



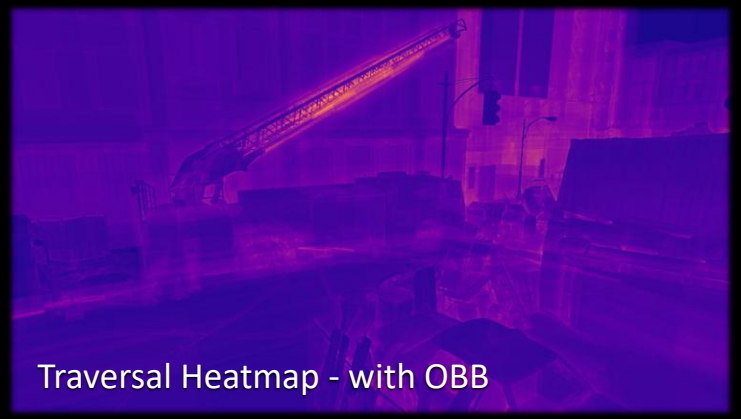


IMAGE EXAMPLE SHOWS TRAVERSAL STEPS IN A FRAME FROM RESIDENT EVIL 3

DEEPER DIVE

OUT-OF-ORDER MEMORY

Latency of memory requests can be critical for performance

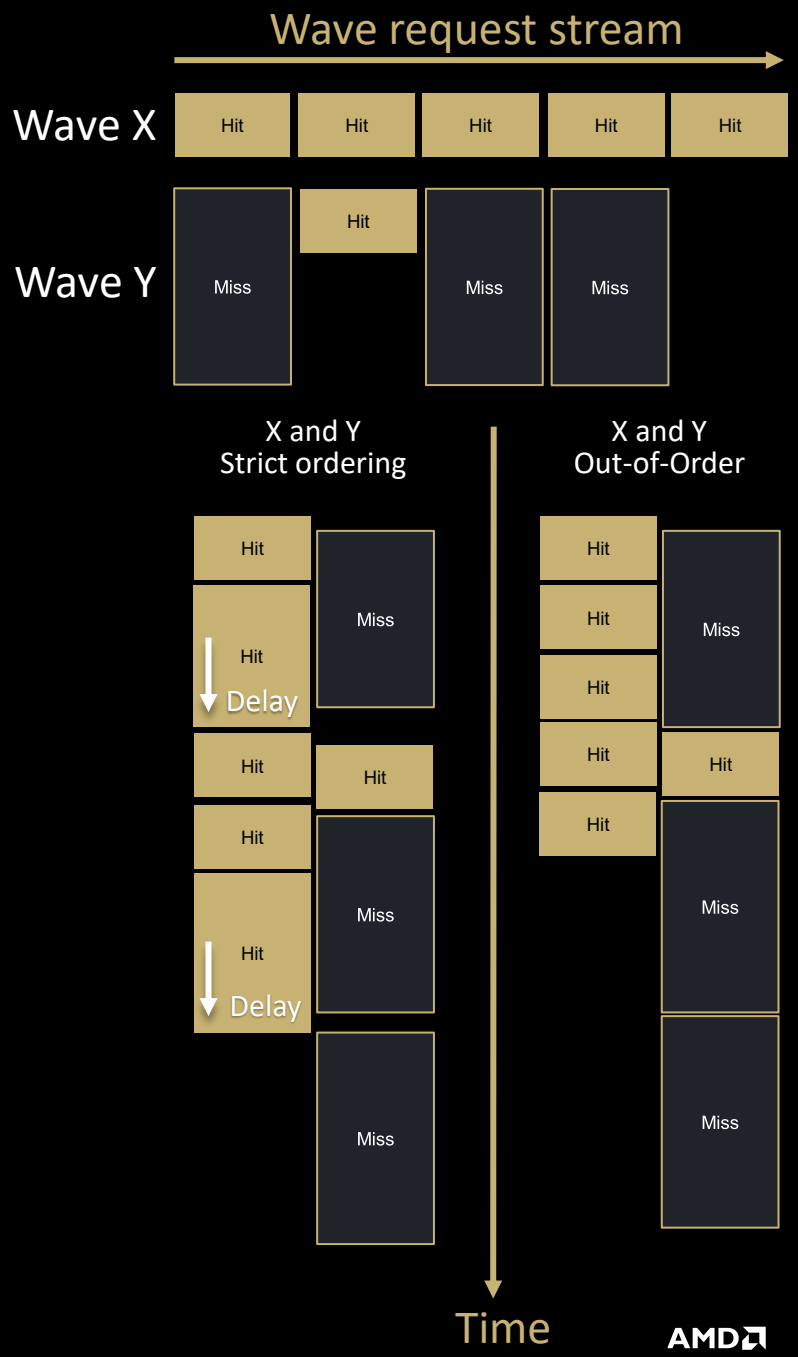
- We see raytracing workloads as highly sensitive to this factor
- Traversal through the BVH structure, texture and buffer requests for shading, etc.

RDNA 4 introduces additional out-of-order queues for memory requests

- RDNA3 data returns were strictly ordered with respect to when the requests were made
- Data items returning quickly could be delayed behind long-latency requests

RDNA 4 allows requests from different shaders to be satisfied out-of-order

- Allows shaders to execute efficiently regardless of some long latency requests (e.g., an access to an un-cached leaf node may no longer hold up surface shading)



This change increases RDNA 4 performance for many workloads

RDNA 4 RAY TRAVERSAL ARCHITECTURE

PUTTING IT ALL TOGETHER

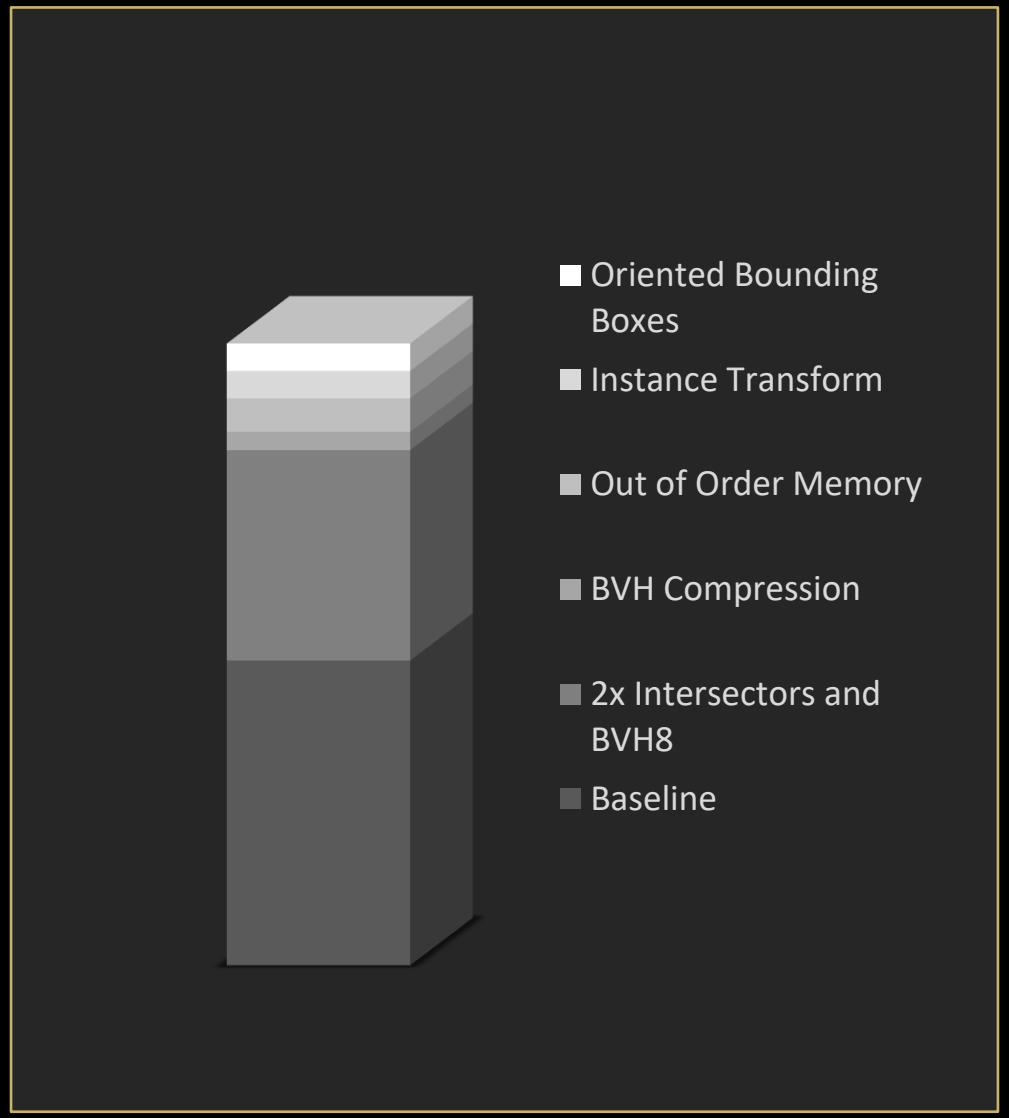
RDNA 4 CUs deliver approximately 2x Ray Traversal performance

- Compared to RDNA 3 at equal clock rates and bandwidth

Final Raytracing performance depends on multiple factors

- OBB impact depends on orientation of geometry in the world
- Traversal and shading efficiency has dependencies on ray coherence

RDNA 4 architecture delivers improvements in all areas providing higher performance for Raytracing for all use cases



RELATIVE RT TRAVERSAL PERFORMANCE BASED ON MODELING DATA FROM AMD'S RAY TRACING ARCHITECTURE MODEL RUNNING RDNA 3 AND RDNA 4 CONFIGURATIONS
DATA COLLECTED BASED ON MODELLED TRAVERSAL PERFORMANCE FROM MULTIPLE GAMING AND BENCHMARK APPLICATIONS

RDNA 4 SHADER

DYNAMIC REGISTERS

Observation: Raytracing shaders show large variance in live registers

- During traversal, a low number of registers is needed
- Result shading often requires a larger number of registers

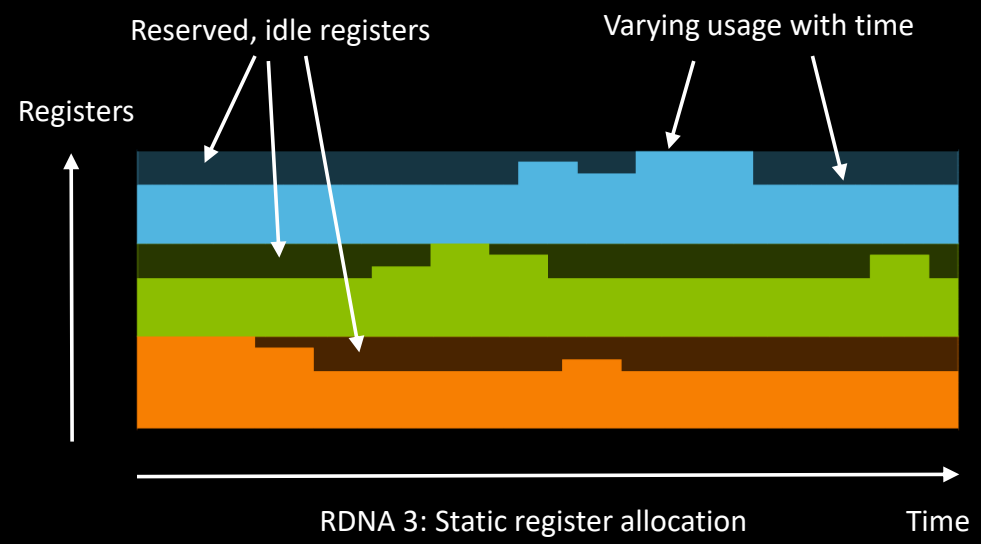
On RDNA 3, shader allocates for the worst case

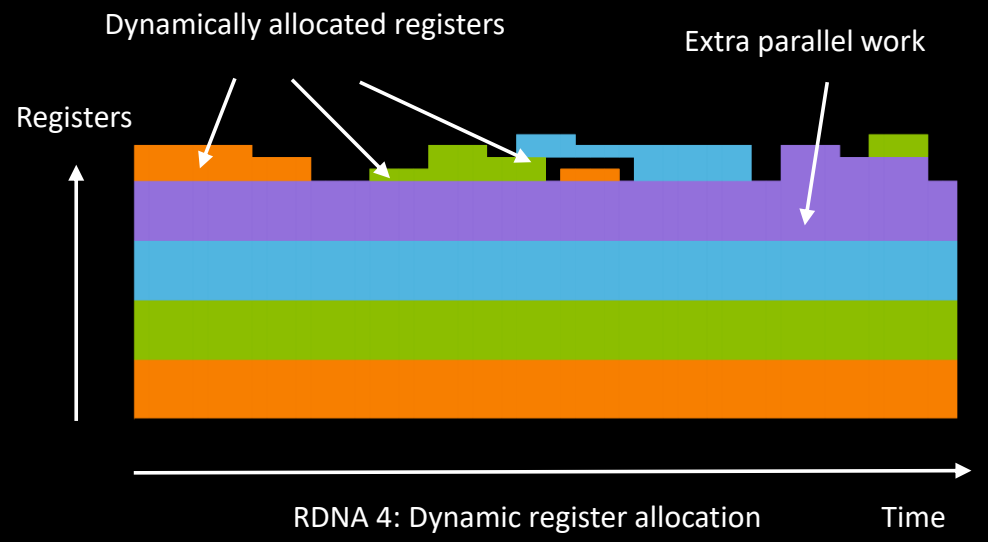
On RDNA 4, shaders have the option to allocate registers dynamically

- Can request registers from the pool when needed
- Can release registers back to the pool when they complete that work
- Software manages the condition where we need to wait for an allocation

By utilizing this capability, RDNA 4 can allow more waves in flight

- Increases occupancy for better handling of memory latency
- Overall efficiency of the shader core can increase significantly

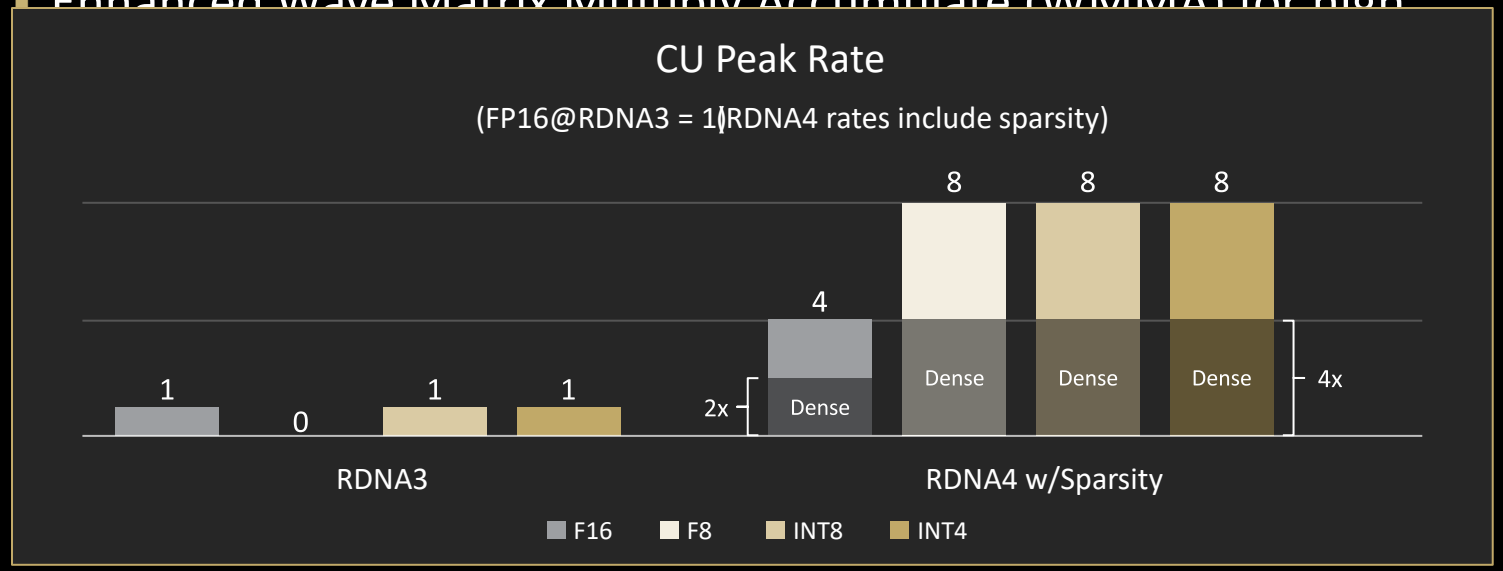




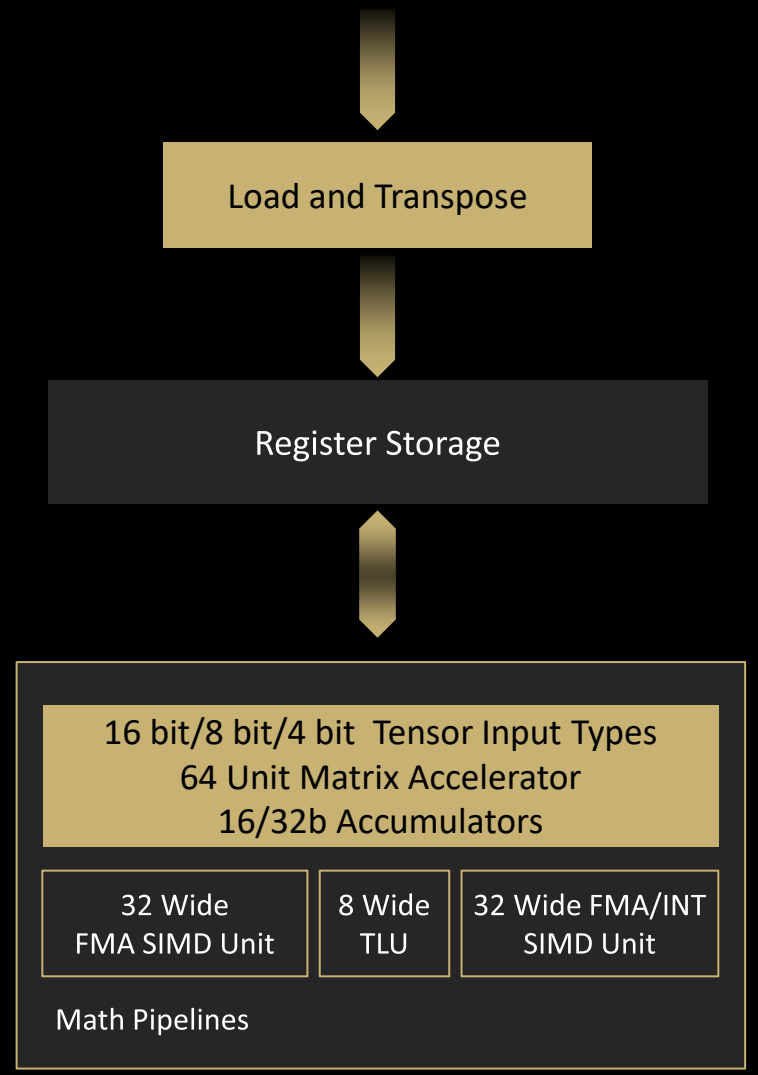
RDNA 4

AI FOR GAMERS AND CREATORS

- RDNA4 focuses on AI capabilities for cutting-edge Gaming and content creation models
- Enhanced Wave Matrix Multiply Accumulate (WMMA) for high

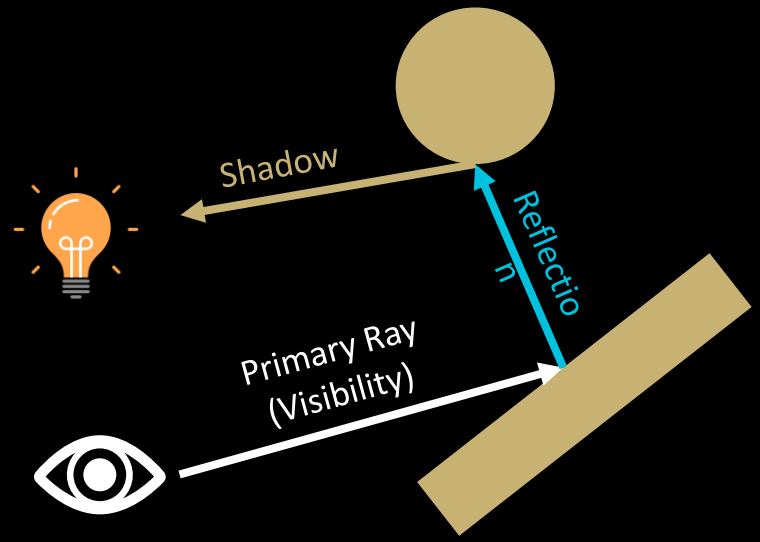


- Support for 8-bit floating point formats enables new and broader capabilities
- 4:2 Structured sparsity enables up to 2x peak performance and improved perf/watt

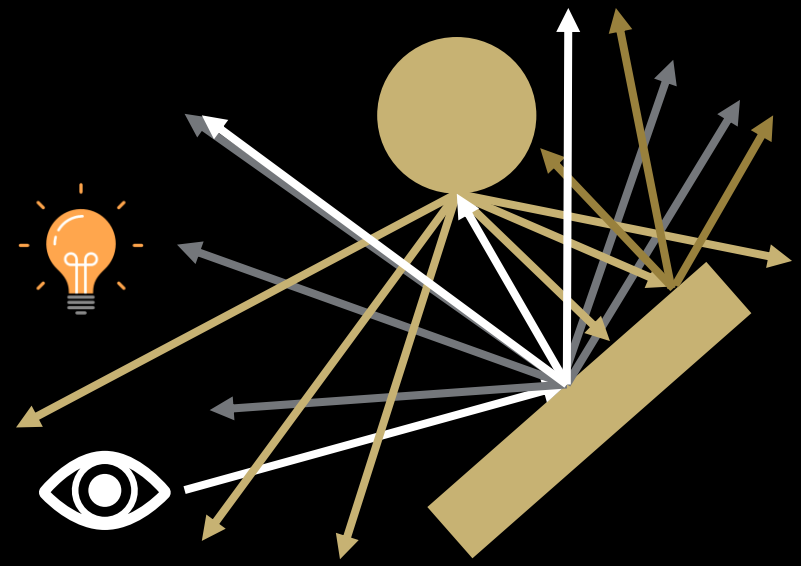


RAYTRACING VERSUS PATHTRACING?

- Traditional Raytracing can take many forms
- Case shown here is just one example
- Can repeat this for every pixel in the scene to create an image
- These techniques can support reflections, refractions, and other effects



- Pathtracing uses raytracing hardware in a different way
- Aims to sample contributions from all possible paths of light
- One way to do this is to shoot very large numbers of rays per pixel
  - Sample a random path for each ray at each intersection
- Is very expensive, but provides a full solution to light transport



RDNA 4 PATH TRACING

PATH TRACED – 1 SAMPLE PER PIXEL

RDNA 4 PATH TRACING

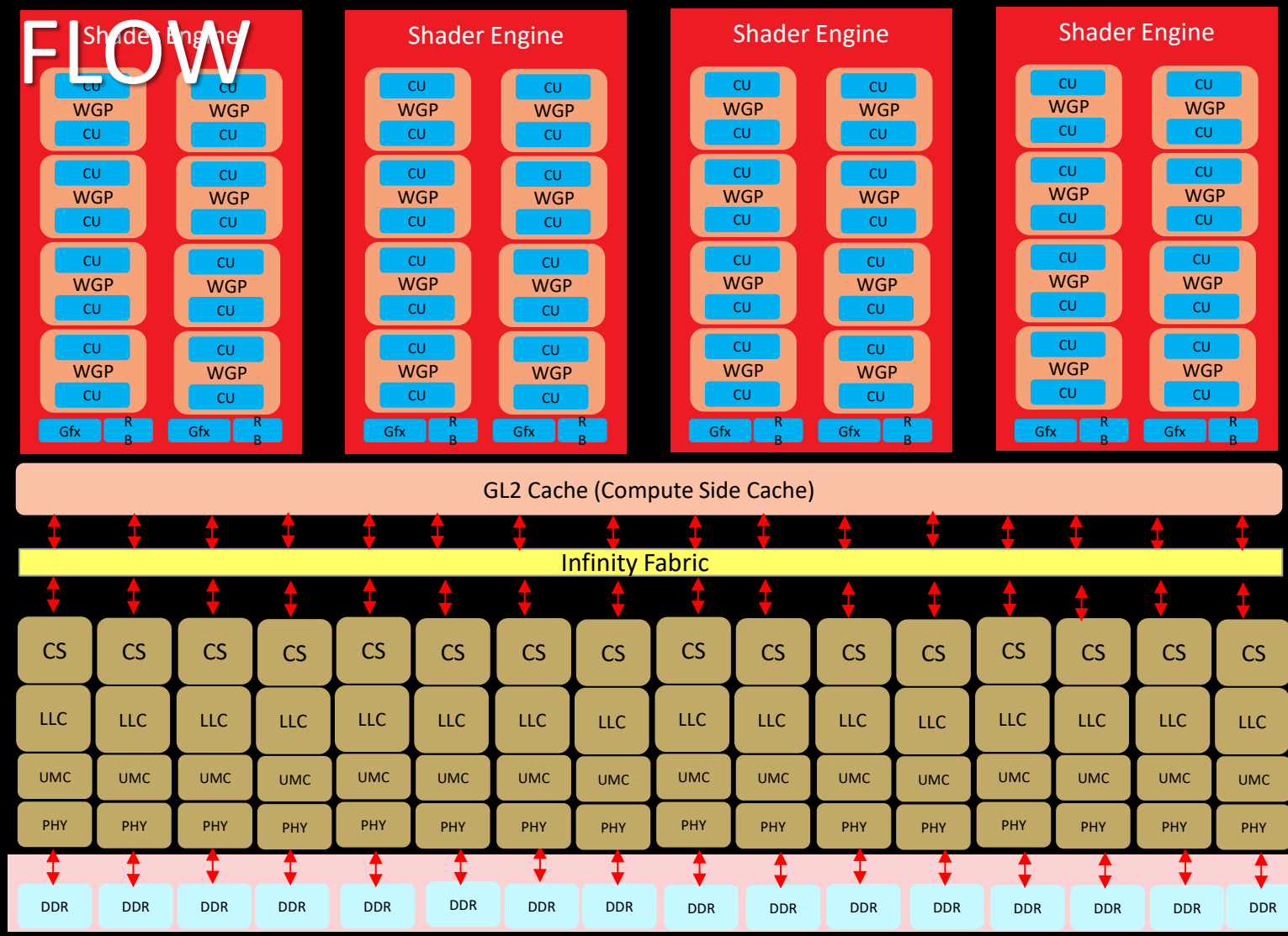
WITH NEURAL SUPERSAMPLING AND DENOISING

AMD
together we advance_gaming

RDNA 4 SOC ARCHITECTURE

Laks Pappu

SOC ARCHITECTURE – DATA FLOW



Work group processors in shader engine(s)

High bandwidth Infinity-Fabric 1KBytes/clock; Freq range: 1.5-2.5 GHZ

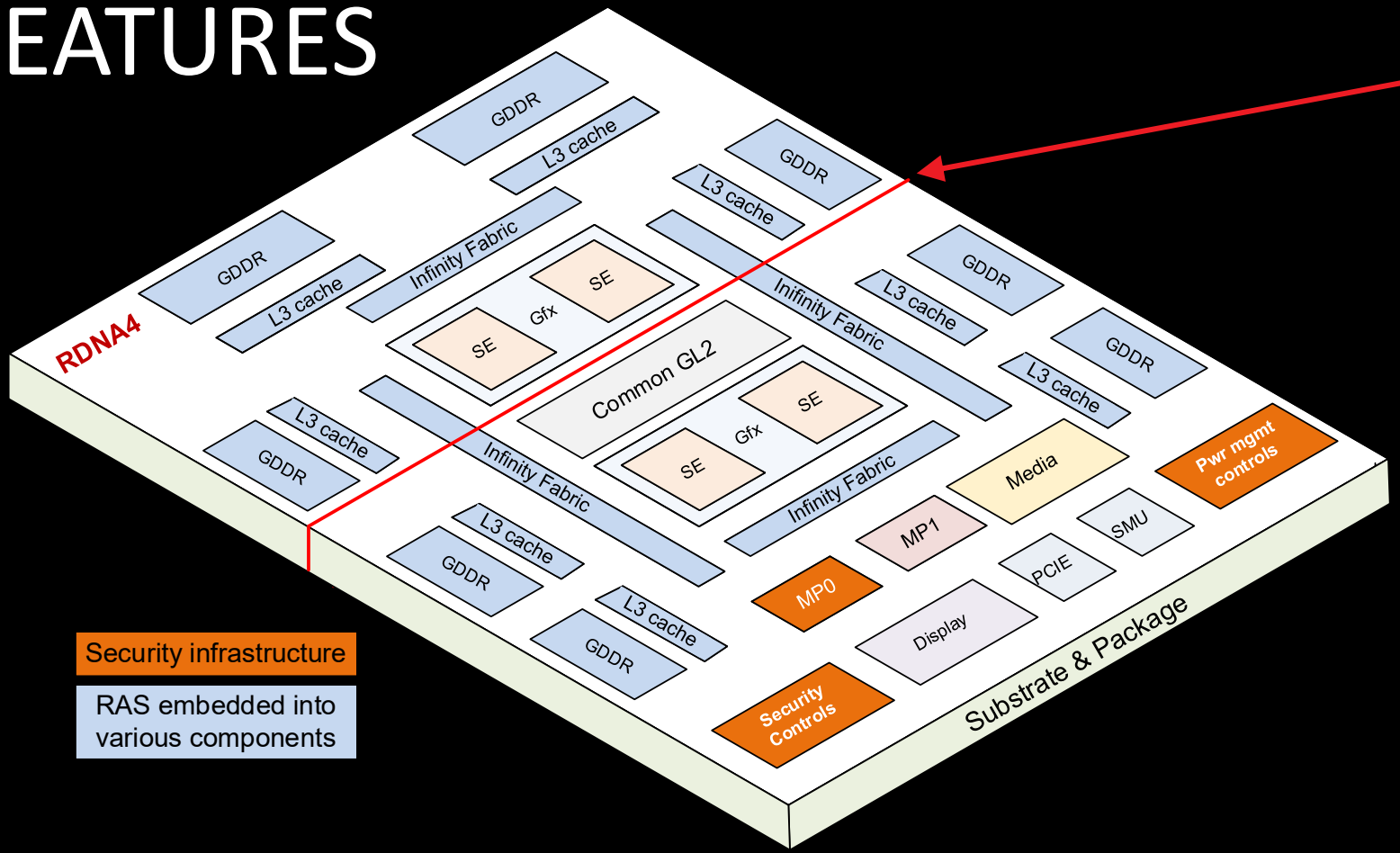
Coherent stations

Last level cache (LLC)

Dual channel Mem controller

DRAM memory

SOC ARCHITECTURE – FEW SALIENT FEATURES



Modular SoC structure

- Ability to spawn smaller SoCs
- On the red line, the bottom portion makes a new SoC

Highly secure architecture

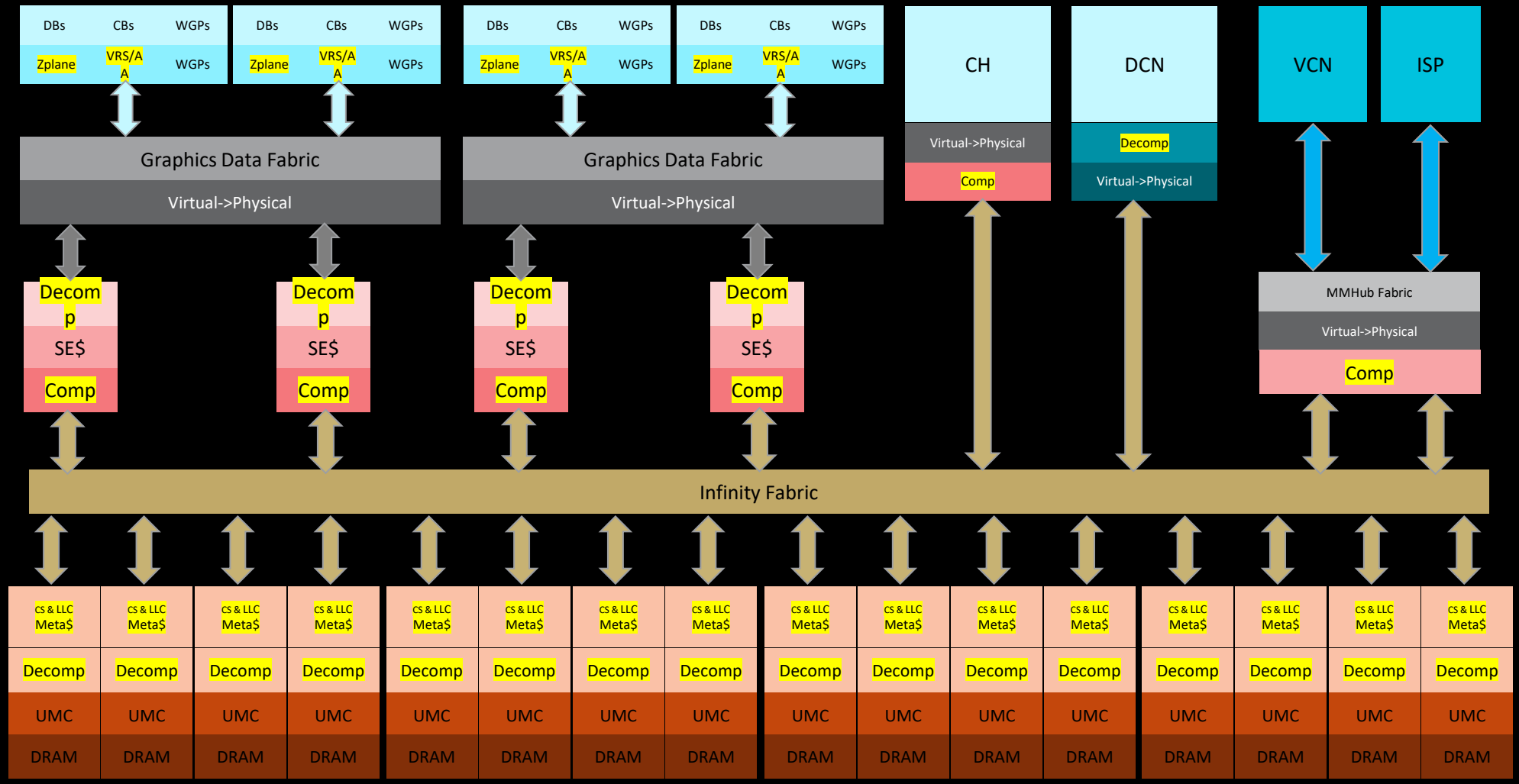
- Access controls and different levels of privileges for components security controls, power management, MP0 (u-controller) supports this mode

RAS and software-controlled re-initialization on poison/uncorrectable parity detection in the SoC

- Infinity-Fabric, L3 cache, GDDR, etc. modules

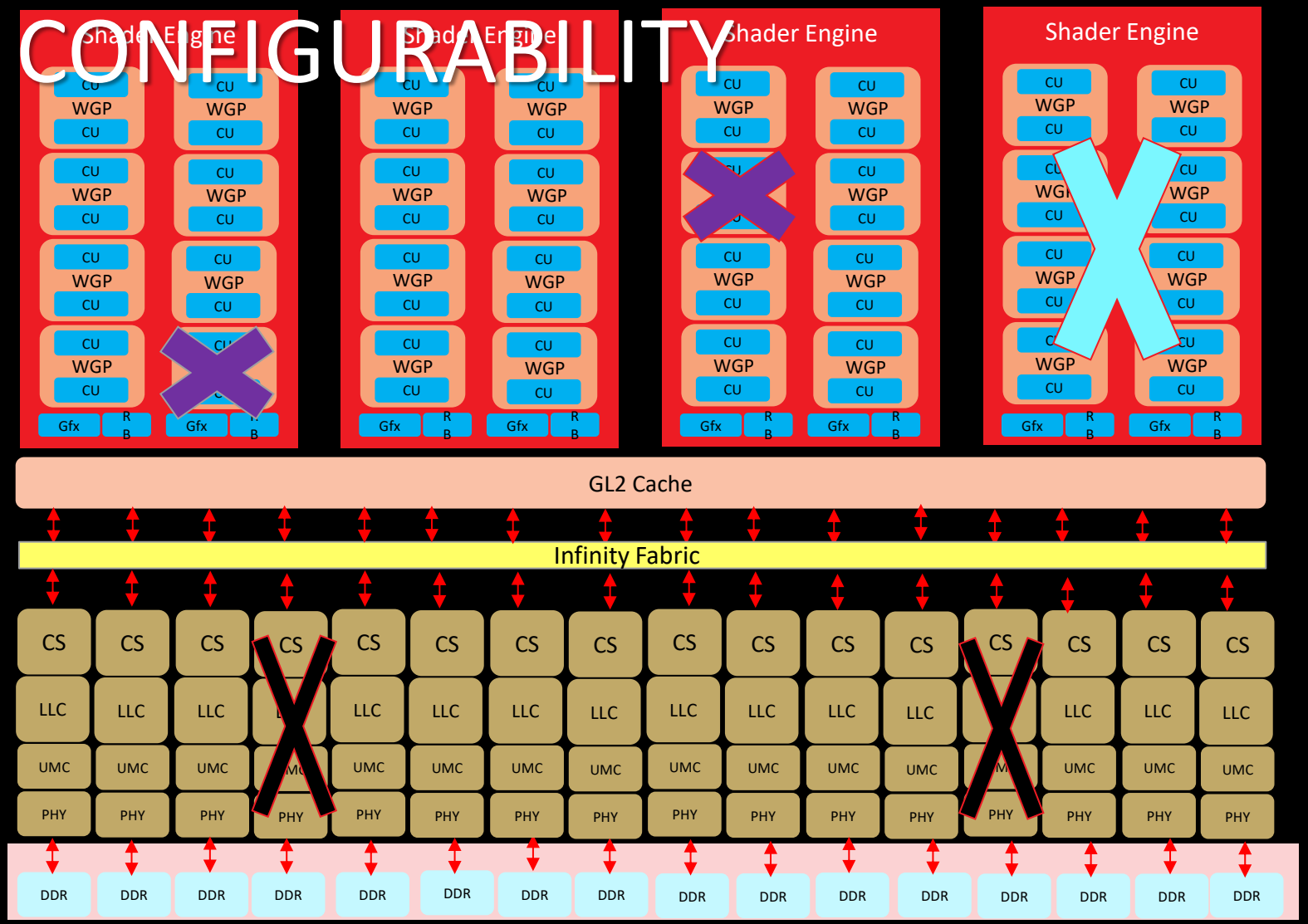
MODULAR, SECURE, RESILIENT

SOC: CENTRAL COMPRESSION/DECOMPRESSION



- ~15% performance improvement for some raster workloads
- ~25% fabric bandwidth reduction (hence lower power) for a few workloads
- Obviates the need for software to be cognizant of compression algorithms (fully contained in HW)

SOC – FLEXIBLE CONFIGURABILITY



- Flexible architecture with fusing options to derive multiple product SKUs
- Market requirements dictate the configuration

SE Harvest

WGP Harvest

Asymmetric harvest

May include weighted PS distro as well as weighted CS distro

Memory device harvest

Single device granularity

64b granularity

SOC RDNA 4 PRODUCT SKUS

SOC ARCHITECTURE FACILITATES MULTIPLE PRODUCT OPTIONS

Product SKUs RDNA 9070	#SEs (Shader Engines)	#CUs (Compute Units)	GDDR size (#bits)	Mem size (GB)	Power (TBP, Watts)
Radeon RX 9070 XT	4	64	256	16	304
Radeon RX 9070	4	56	256	16	220
Radeon RX 9070 GRE	3	48	192	12	220
Radeon AI Pro R9700	4	64	256	32	300

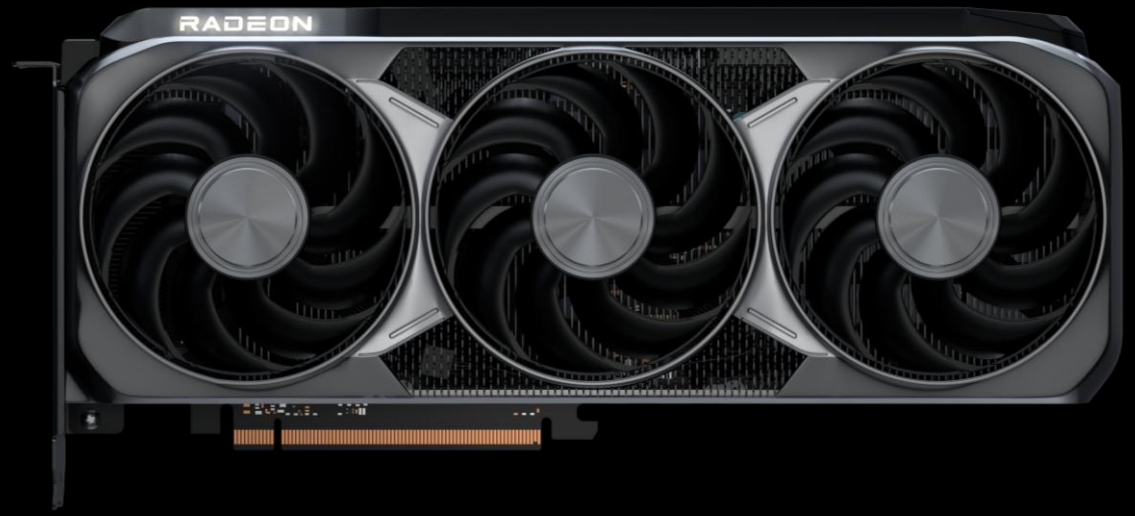
Product SKUs RDNA 9060	#SEs (Shader Engines)	#CUs (Compute Units)	GDDR size (#bits)	Mem size (GB)	Power (TBP, Watts)
Radeon RX 9060 XT 16GB	2	32	128	16	160
Radeon RX 9060	2	28	128	8	132
Radeon RX 9060 XT 8GB	2	32	128	8	150

RDNA 4

WRAPPING UP

- Built for the next generation of gaming and creation with high performance Raytracing and ML
- Advanced compute capabilities, innovative Raytracing features, and ML Enhanced upscaling
- Memory capacity that supports today’s most demanding gaming applications
- Advanced video encoding and streaming capabilities for productivity and media creation
- Highly scalable and secure SoC

AMD RADEON RX 9070 XT



DISCLAIMER

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u.

 Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.


AMD