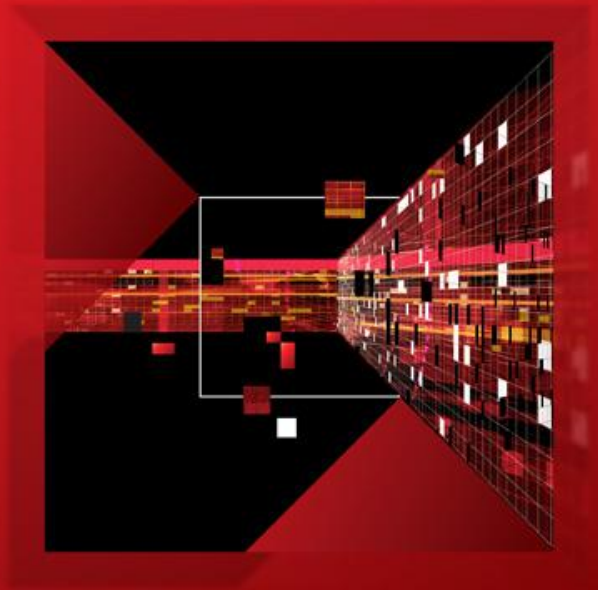
EVOLUTION OF AMD GRAPHICS

Eric Demers, Graphics CTO, CVP

EARLY GPU ERAS

FROM FIXED FUNCTION HARDWARE TO PRESENT



1ST ERA: FIXED FUNCTION

Prior to 2002

- Graphics specific hardware
- No general purpose compute capability

Example: ATI Rage

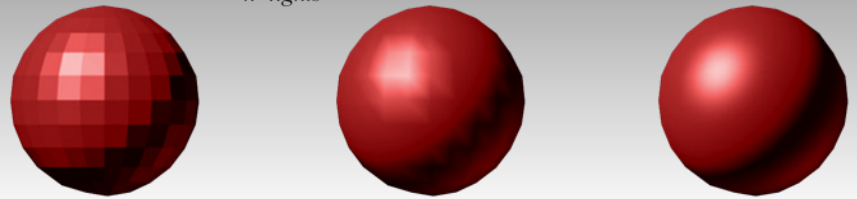
Images used with permission of Creative Assembly & Codemasters

A DOT PRODUCT AND A SCALAR HANDLES IT ALL!

$$V_{eye}\begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix} = MVP \begin{bmatrix} m_0 & m_4 & m_8 & m_{12} \\ m_1 & m_5 & m_9 & m_{13} \\ m_2 & m_6 & m_{10} & m_{14} \\ m_3 & m_7 & m_{11} & m_{15} \end{bmatrix} \bullet V_{obj} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

$$V_{tex}\begin{bmatrix} s \\ t \\ r \\ q \end{bmatrix} = M_{proj} \bullet V_{in}$$

$$C_p = k_a L_a + \sum_{n-lights} Att_n (k_d (\hat{L}_n \bullet \hat{N}) + k_s (\hat{R}_n \bullet \hat{V})^{\alpha})$$



- Geometry transformation is based on dot products as part of matrix multiply
- 32b SPFP Multiply-accumulate basic hardware
- Lighting requires various scalar and dot products
- 8b-12b limited precision required, including transcendental functions

2ND ERA: SIMPLE SHADER

2002 - 2006

- Graphics focused
- Floating point processing
- Limited shaders

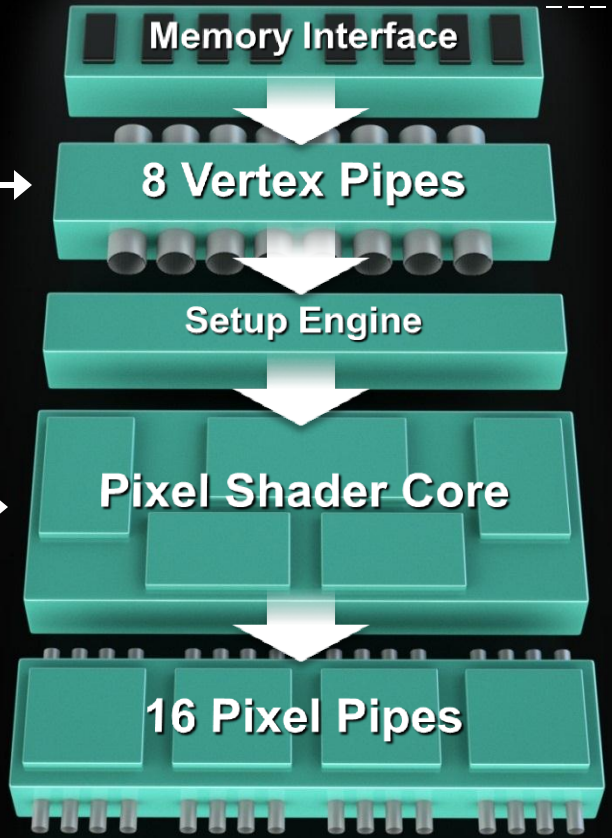
Example:
ATI Radeon™ 9700 Pro

Images used with permission of Creative Assembly & Codemasters

FIRST GENERATION SHADERS

128-bit Vector ALU
32-bit Scalar ALU
(DX8)

96b SPFP
Allows application specified shading
(DX9)



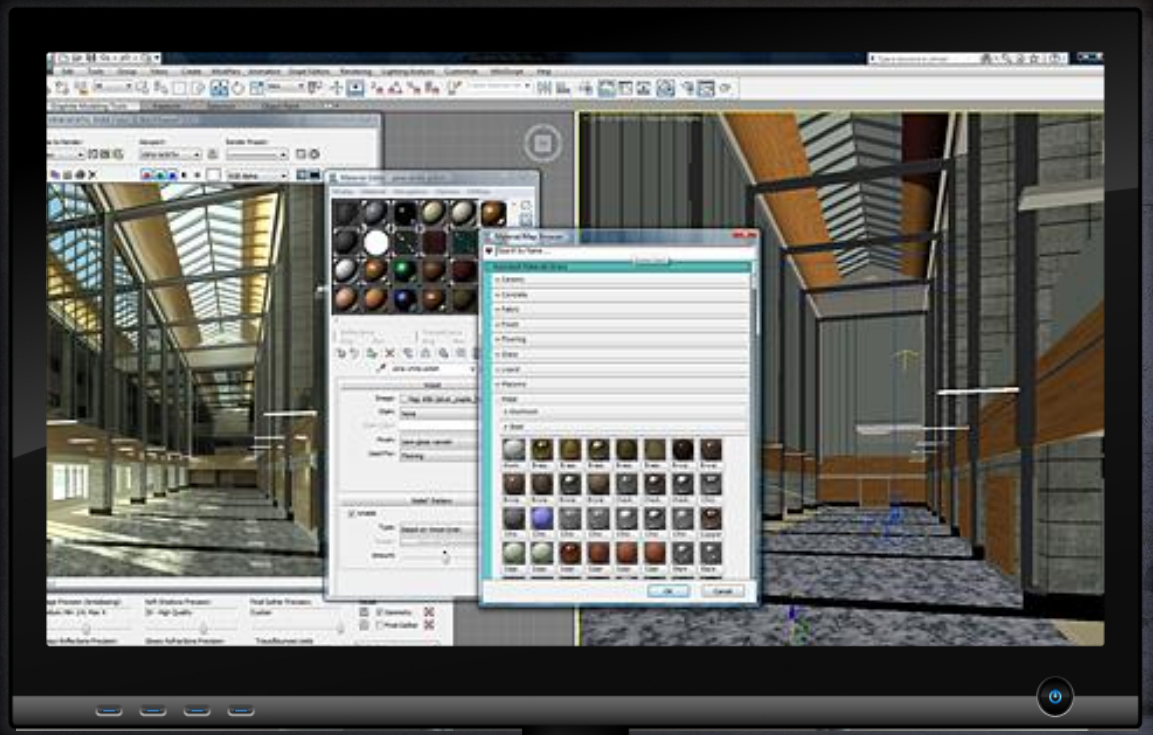
Not required to be IEEE compliant

3RD ERA: GRAPHICS PARALLEL CORE

2007 to Present

- Graphics is key
- Unified shader architecture
- Basic general purpose compute
  - CAL, Brook, OpenCL™

Example:
ATI Radeon™ HD 2900



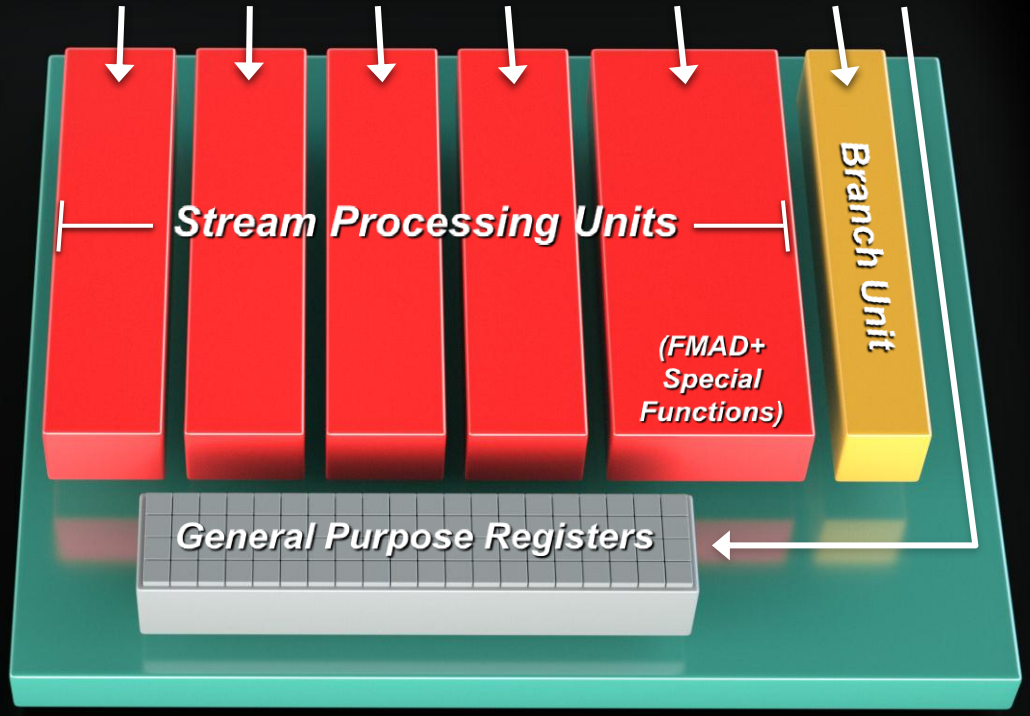
Images used with permission of Dassault Systemes, Codemasters & Autodesk

GRAPHICS HISTORY AND VLIW5

Centralized core all shading

- Multiple engines with VLIW5 core
- Focus on IEEE math
- Texture cache as primary I/O

Graphics performance still primary objective



3RD ERA EVOLVES: GPU COMPUTE

2010+

- Graphics important
- But also optimized for compute
- Enabling high performance computing

Example:
Radeon™ HD 6970

Image used with permission of Codemasters, EXA & OPTIS

EMERGING GENERAL COMPUTE

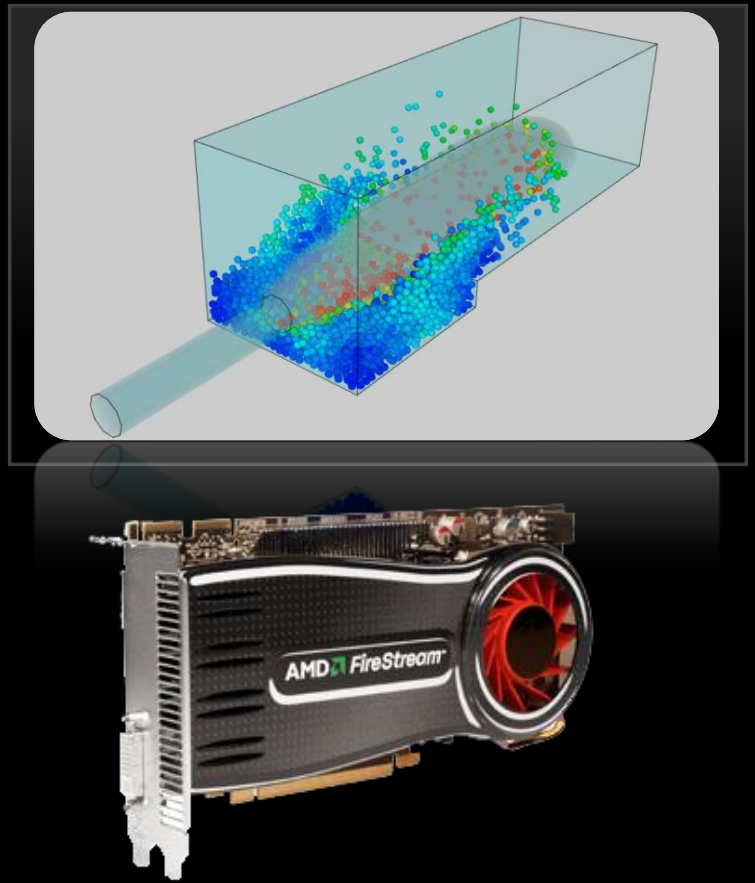


Image used with permission of DEM Solutions

HPC/Server Applications

Cloud-based Computing

- Commercial cloud, cloud-based gaming, and virtual desktop

Massive Data Mining

- Image, video, audio processing
- Pattern analytics and search

Research

- Research clusters with mixed workloads

Production HPC

- Seismic, financial analysis, Pharmaceutical

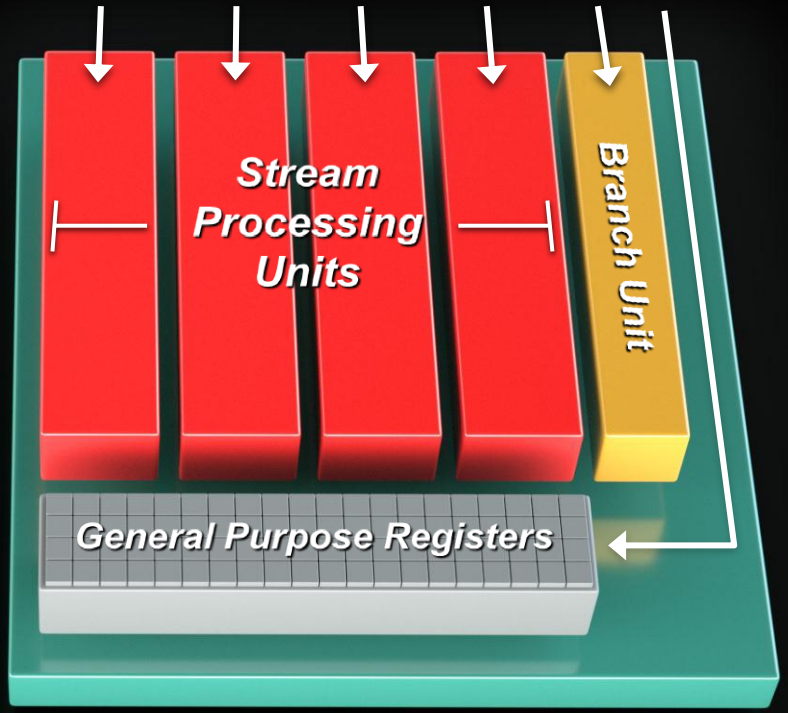
SYMMETRICAL VLIW4 ARCHITECTURE

Better balance for today's workloads

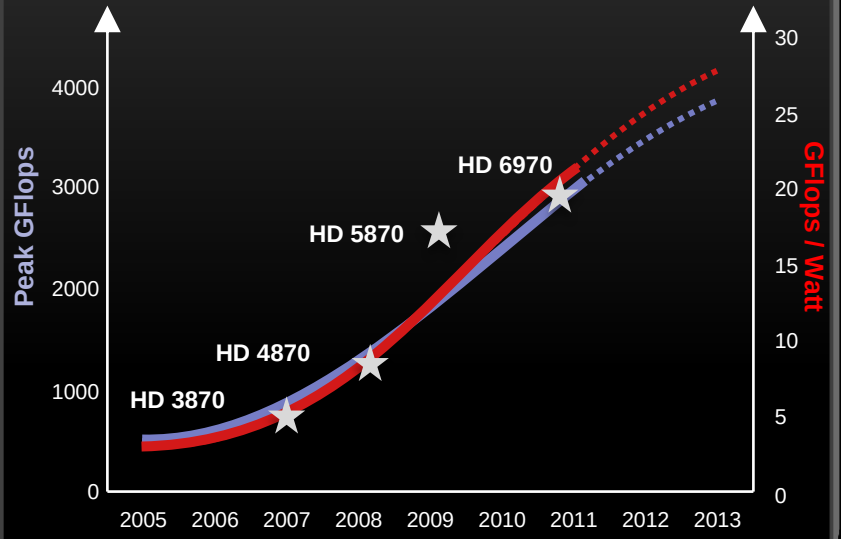
- Symmetrical and less lanes
- More efficient for generalized algorithms and compiler

Simplified & Optimized

- Single replicated FMAD design



Performance has continued to get better



ISVs have embraced industry standards such as OpenCL™ or DirectCompute

Engaged

Houdini 3D ANIMATION TOOLS

esi get it right®

ANSYS

Altair

Autodesk

Solution Partner PLM SIEMENS

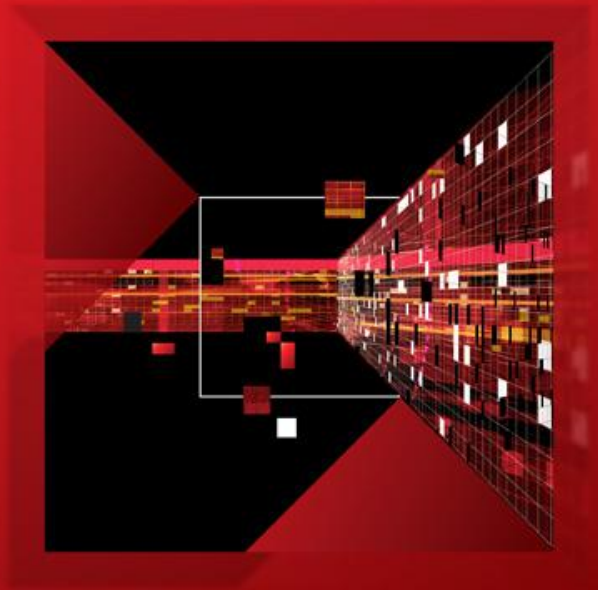
MSC Software

ISV list is qualified – not comprehensive

For the next era, we want to unlock the full potential…

NEXT GENERATION GPU

A NEW ARCHITECTURE

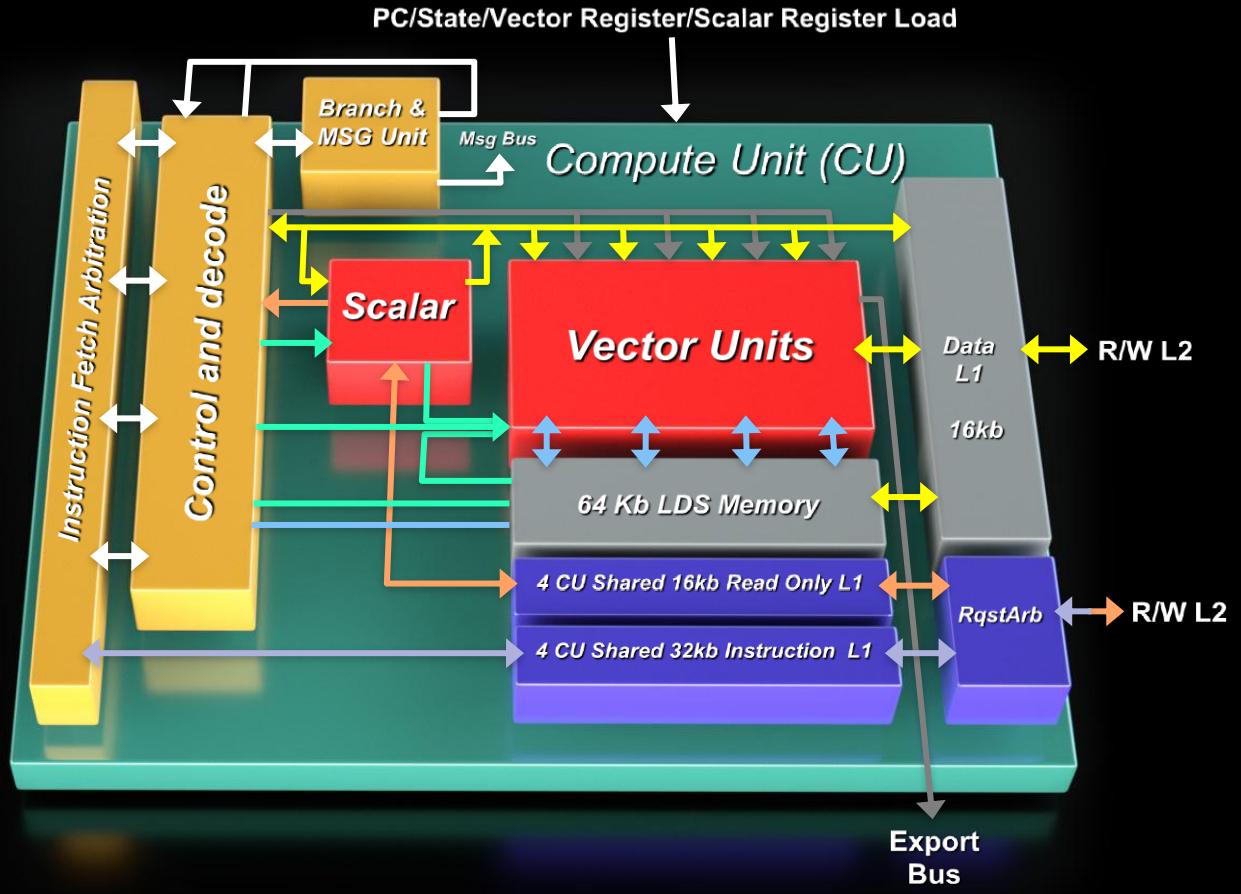


ROADMAP FOR AMD FUSION SYSTEM ARCHITECTURE (FSA)

Physical Integration	Optimized Platforms	Architectural Integration	System Integration
Integrate CPU & GPU in silicon	GPU Compute C++ support	Unified Address Space for CPU and GPU	GPU compute context switch
Unified Memory Controller	User mode scheduling	GPU uses pageable system memory via CPU pointers	GPU graphics pre-emption
Common Manufacturing Technology	Bi-Directional Power Mgmt between CPU and GPU	Fully coherent memory between CPU & GPU	Quality of Service
			Extend to Discrete GPU

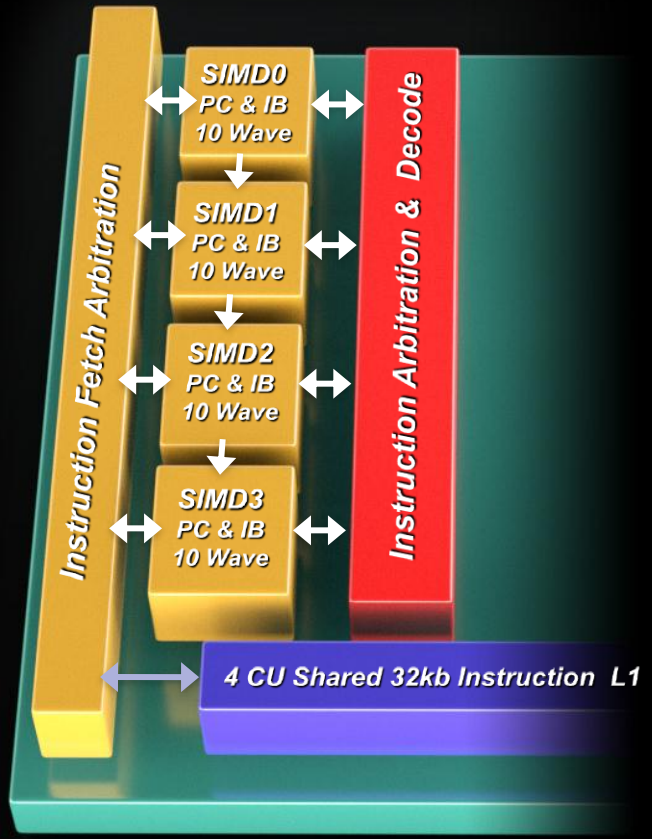
COMPUTE UNIT: NEW ARCHITECTURE

- Observable and controllable processing
  - Context switching
- Single lane programming
- Supports x86 virtual memory
- Supports C++ constructs
  - Virtual functions
  - DLLs…

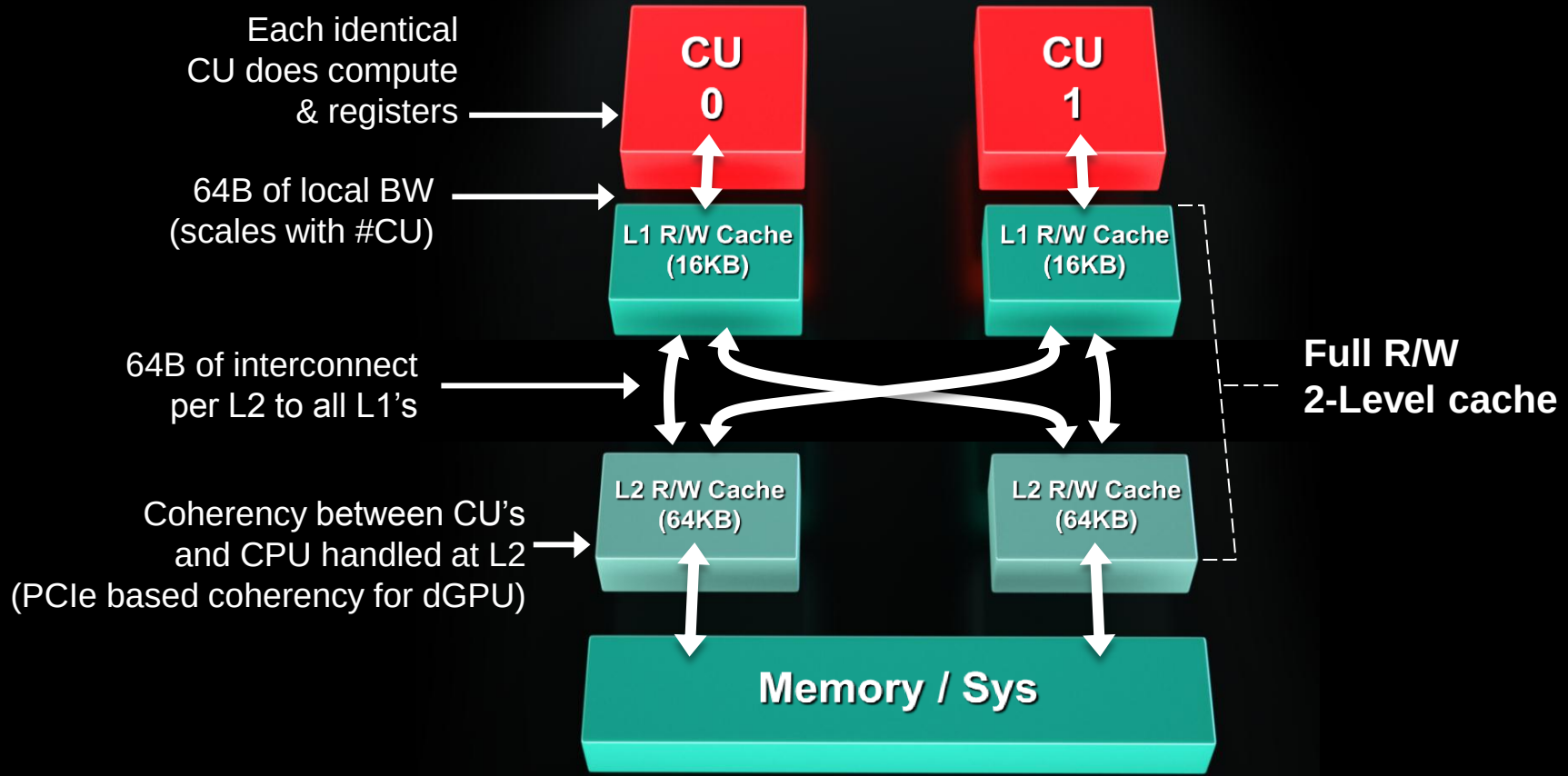


A NEW ARCHITECTURE FOR A NEW ERA

- A multitude of elements
  - MIMD: 4 threads per cycle per vector, from different apps, per CU
  - SIMD: 64 FMAD vector for 4 waves per cycle
  - SMT: 40 waves per CU active
  - Vector unit & Scalar unit coprocessor
- Powered by multiple command streams
  - Support of multiple asynchronous and independent command streams

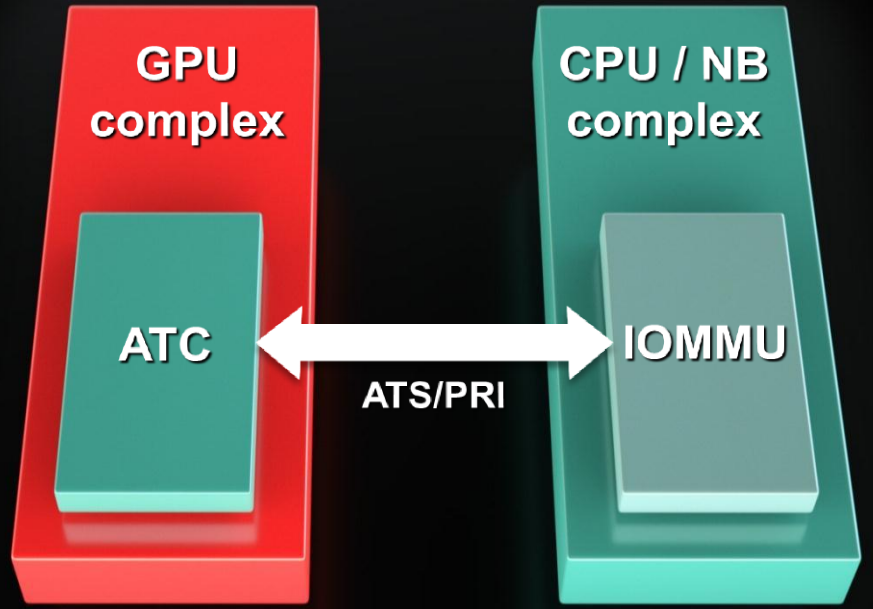


GENERALIZED READ / WRITE CACHE DESIGN



SWITCH THE COMPUTE – DON'T MOVE THE DATA

- X86 Virtual memory is the foundation
  - IOMMU for GPU, MMU for CPU
  - 64b x86 pointer will work for GPU
  - GPU will page fault
  - GPU will have address translation caches
  - Can over allocate memory
- OS Will service both MMU and IOMMU
- Under a unified address space
  - CPU and GPU will use the same 64b pointers



RECAP OF FSA FEATURES

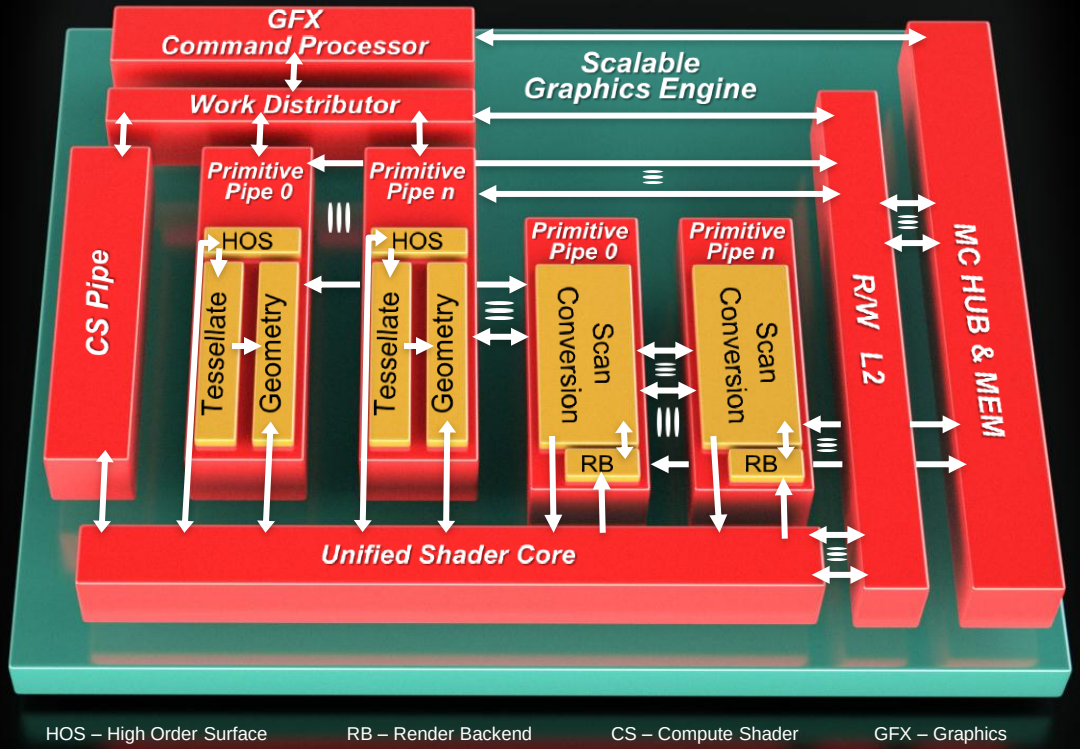
- Full GPU support for C, C++ and other high-level languages
- Unified virtual address space between CPU and GPU
- GPU can access all system memory, and handle page faults
- Memory coherence between CPU and GPU
- Pre-emptive scheduling and context switching of GPU
- FSA enabled on discrete GPUs as well as AMD Fusion

Features added incrementally each year

Each bundle of features brings incremental benefit

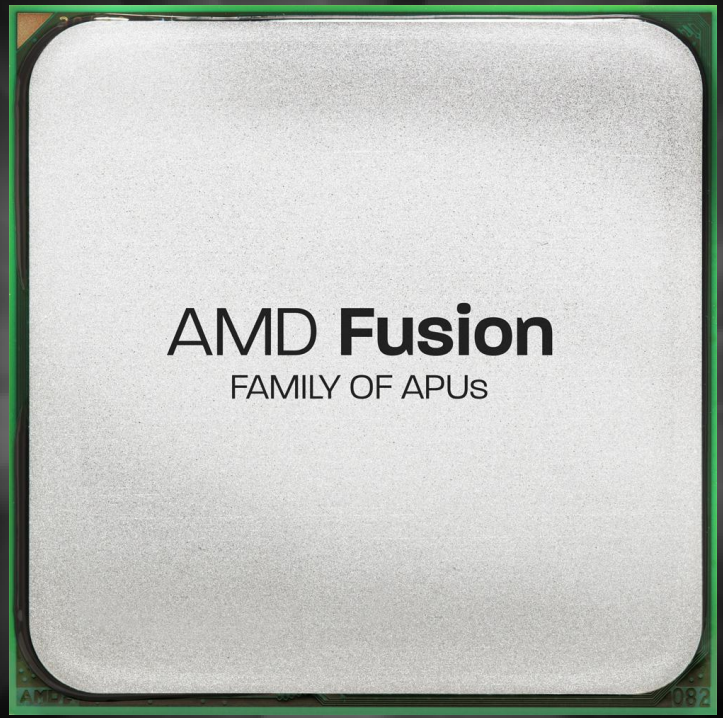
WHAT ABOUT 3D?

- Graphics continue to drive forward
  - PRTs to drive virtual texturing
  - Supporting next iteration of graphics APIs
- Still have fixed function hardware
  - Raster Ops and Z units still independent
  - R/W cache supports all texture ops
- Going forward, FSA – 3D
  - Leveraging compute model for graphics
  - Enable FSA back into graphics APIs



Platform will deliver

- True Virtual Memory
  - 3D Volumes, massive data sets
  - Streaming and on demand memory
- Simpler & powerful programming model
  - Task graph
  - Braided and nested parallelism
- Today's Demands
  - Scalable and Open
  - Emerging General Compute
  - High density computing
  - Essential on our road to the Holodeck!



WHERE ARE WE TAKING YOU?


AMD
Fusion 11
DEVELOPER SUMMIT