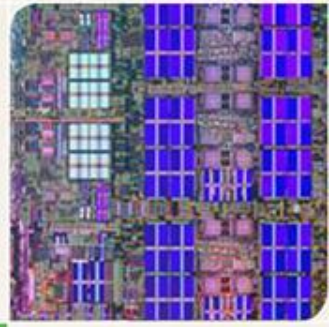
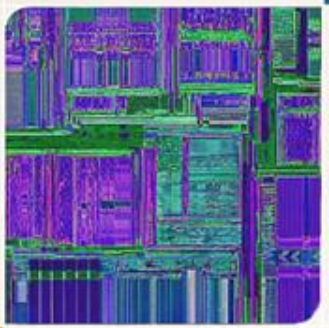
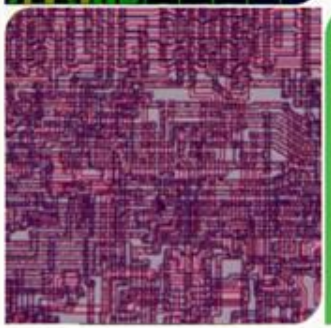




CS4803DGC Design Game Console

Spring 2010

Prof. Hyesoon Kim

Georgia Tech College of Computing

AMD presentations from Richard Huddy and Michael Doggett



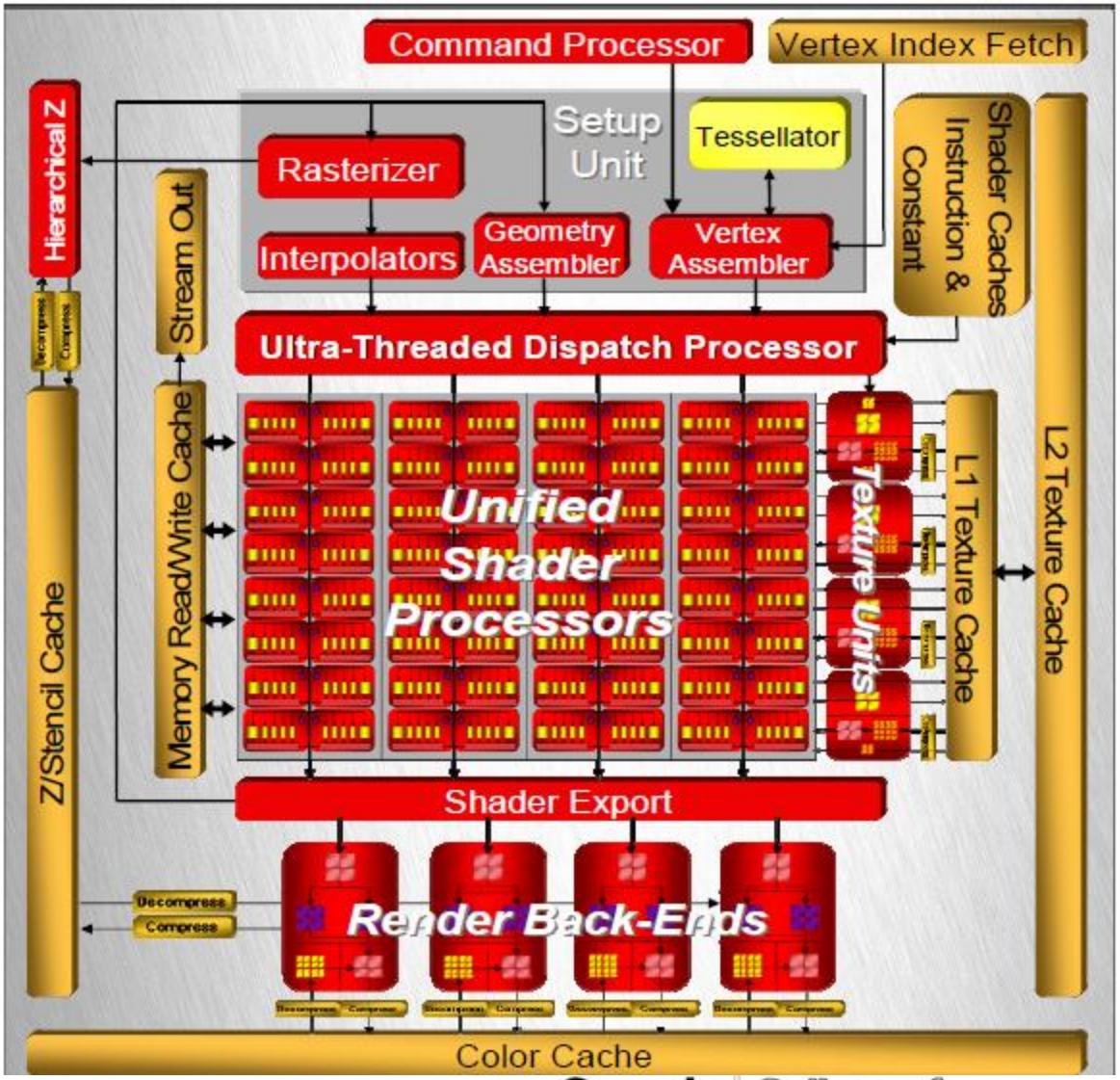
Radeon HD 2000 Series

Radeon	2900	2600	2400
Stream Processors	320	120	40
SIMDs	4	3	2
Pipelines	16	8	4
Texture Units	16	8	4
Render Backens	16	4	4
L2 texture cache (KB)	256	128	0
Technology (nm)	80	65	65
Area (mm2)	420	153	82
Transistors (millions)	720	390	180
Memory bandwidth	512	128	64
Optimized for	High clock speed	Power efficiency	Power efficiency

AMD presentations from Richard Huddy and Michael Doggett

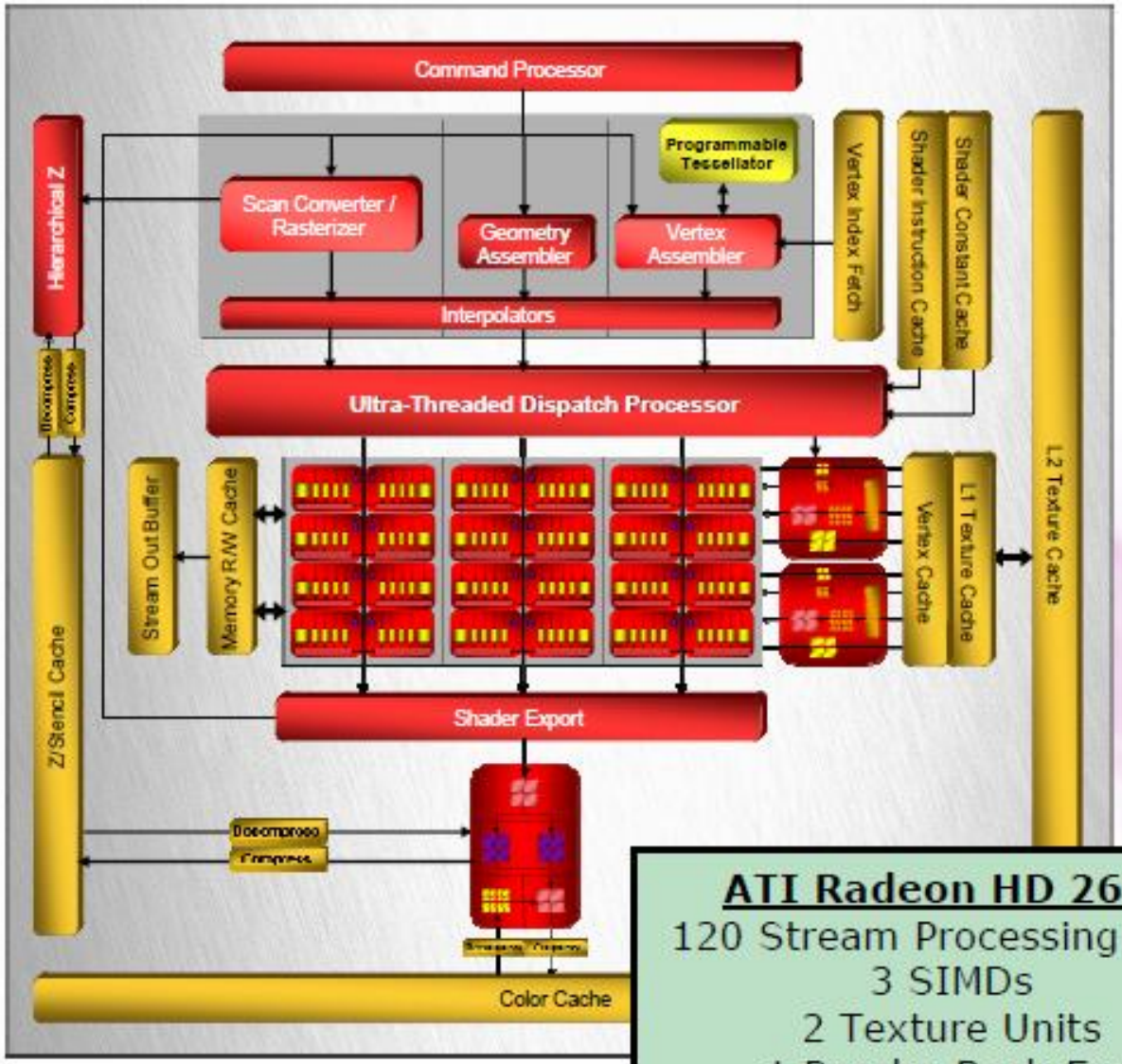
Radeon 2900 Top Level

- 320 Stream processing units
- 4 SMIDs
- 4 Texture Units
- 4 Render Back-end



AMD presentations from Richard Huddy and Michael Doggett

Radeon HD 2600 Top Level



ATI Radeon HD 2600

120 Stream Processing Units

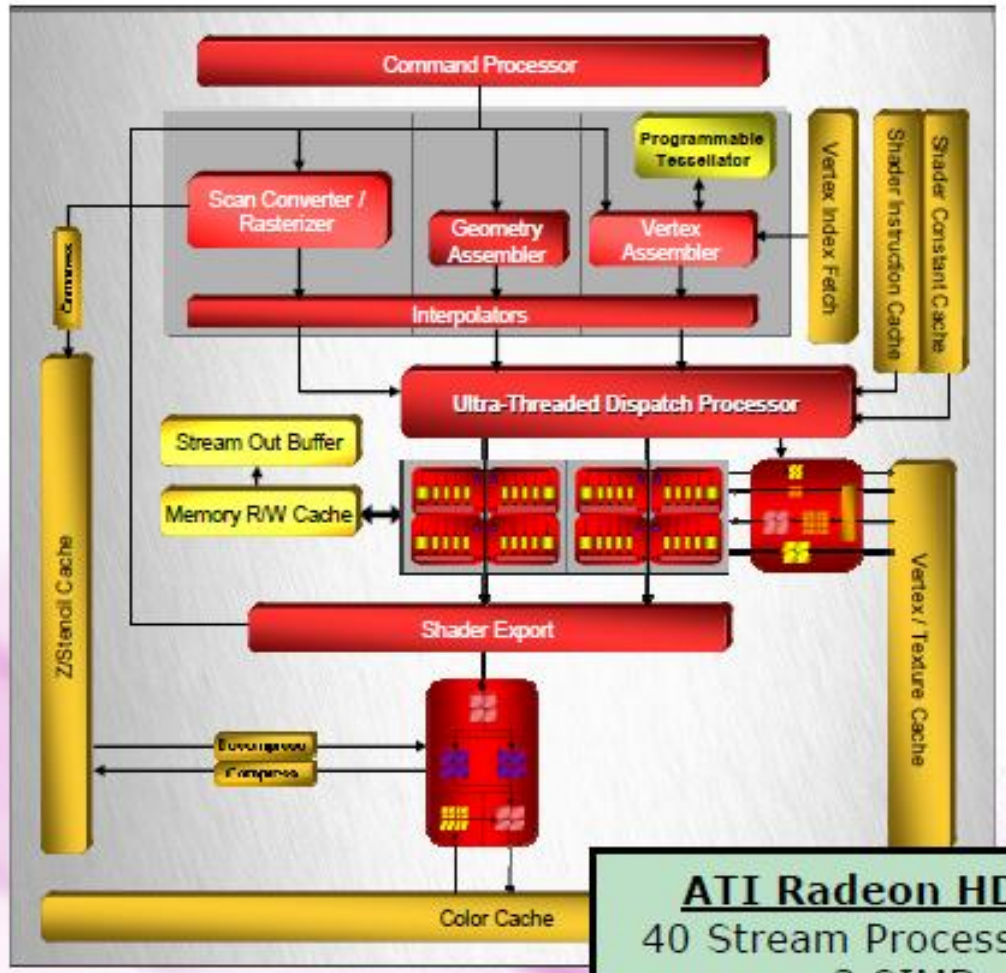
3 SIMDs

2 Texture Units

1 Render Back-End

AMD presentations from Richard Huddy and Michael Doggett

Radeon HD 2400 Top Level



ATI Radeon HD 2400

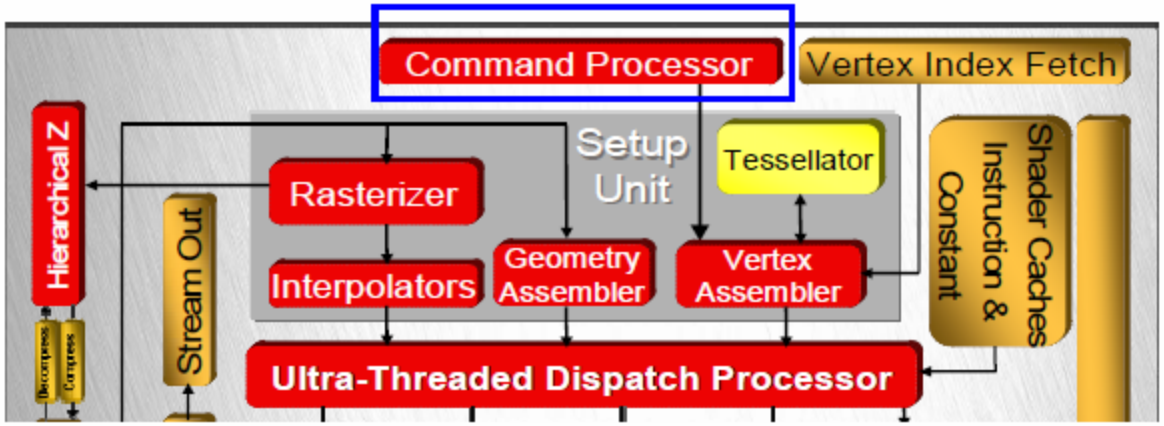
40 Stream Processing Units
2 SIMDs
1 Texture Unit
1 Render Back-End
Shared vertex/texture cache

AMD presentations from Richard Huddy and Michael Doggett

College of Computing

Command Processor

- GPU interface with host
  - Processes command stream from graphics driver
- A custom RISC based Micro-Coded engine
- First class memory client with Read/Write access
- State management



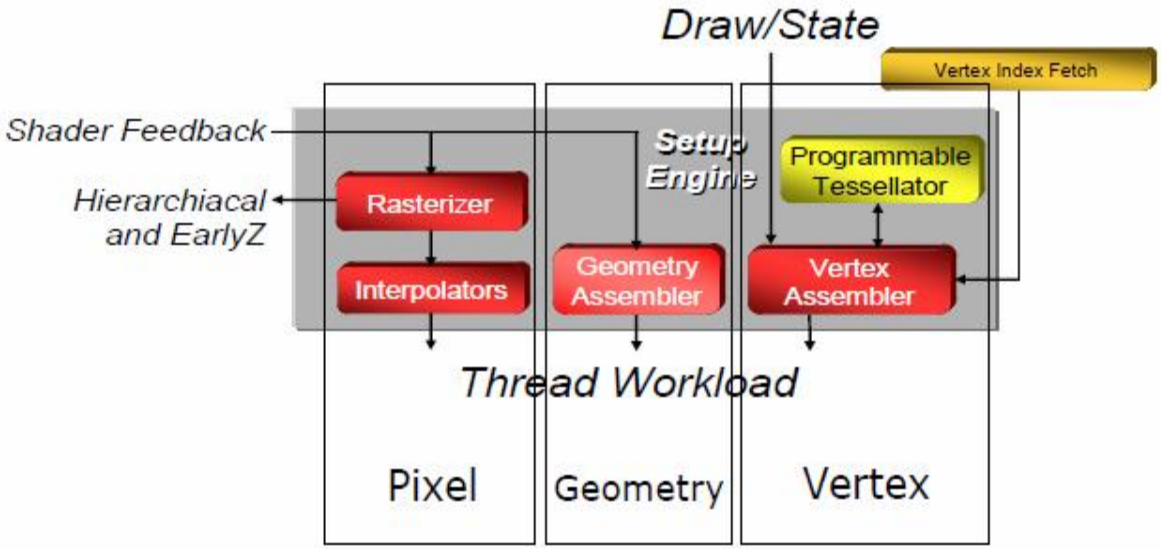
AMD presentations from Richard Huddy and Michael Doggett

Setup engine

Prepares data for processing by the stream processing units

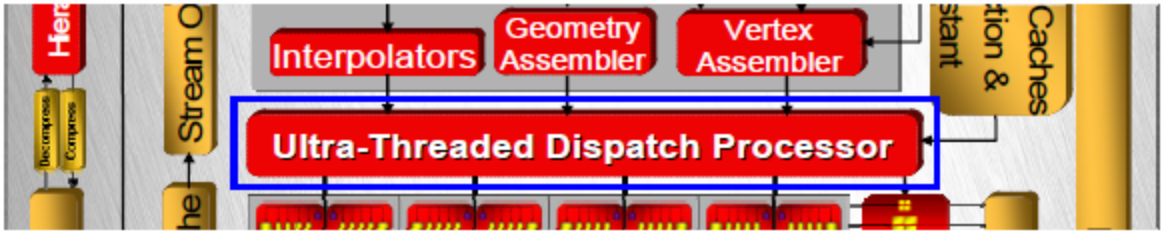
3 groups of blocks feeding 3 data streams

Each group feeding 16 elements



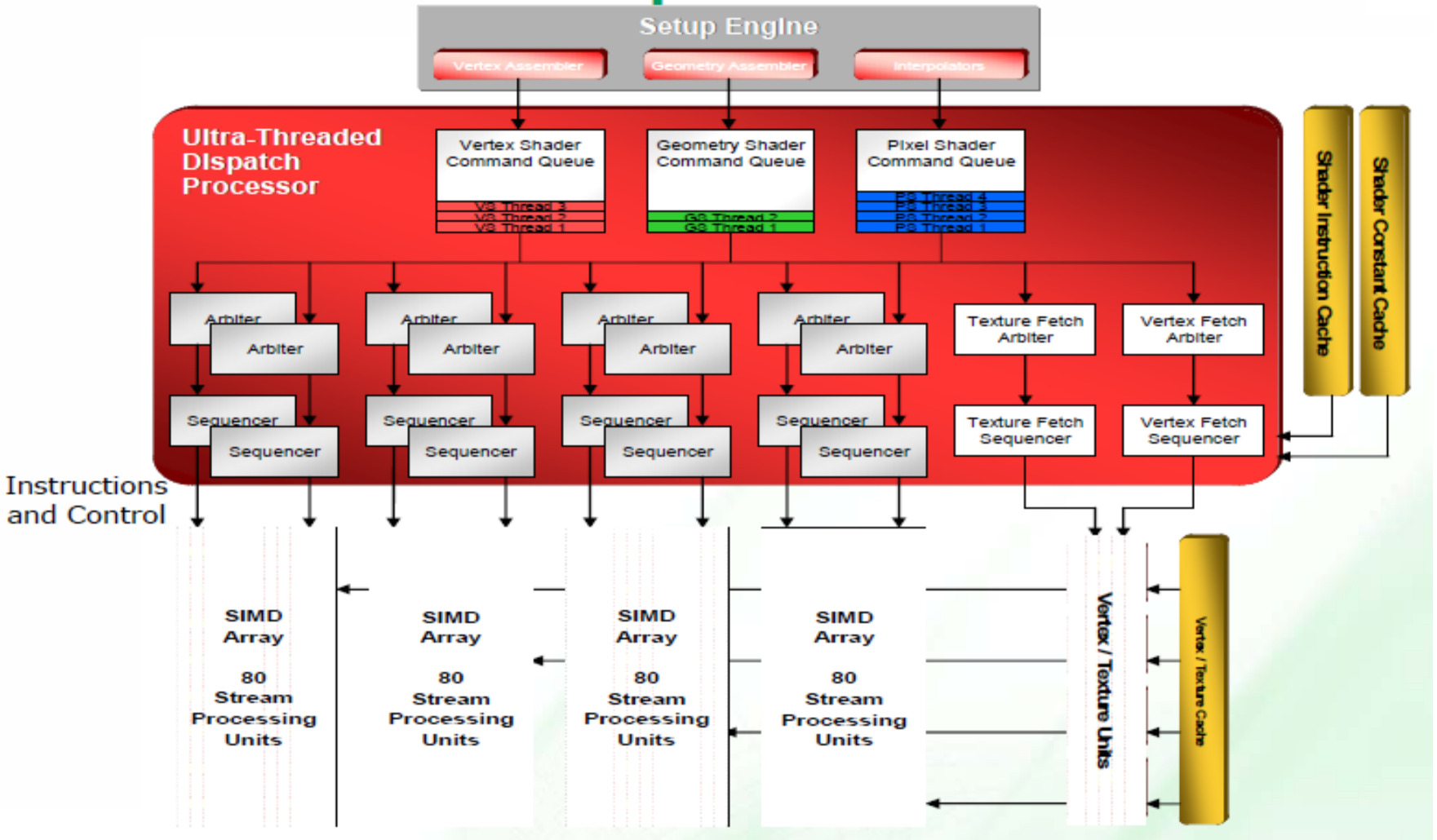
- Vertex blocks: Primitive tessellation, Inputs-index & instancing
  - Sends vertex addresses to shader core
- Geometry blocks : Uses on/off chip staging
  - Sends processed vertex addresses, near neighbor addresses and topological information
- Pixel blocks : Scan conversion, Triangle setup, Rasterizations, and interpolation
  - Interfaces to depth to perform Hiz/EarlyZ checks

Ultra-Thread Dispatch Processor



- Main control for the shader core
- Separate command queues for each shader type
  - Each thread consists of a number of instructions that will operate on a block of input data
  - All workloads have threads of 64 elements
  - 100's of threads in flight
  - Threads are put to sleep when they request a slow responding resource

Ultra-Threaded Dispatch Processor



AMD presentations from Richard Huddy and Michael Doggett

Arbiter in Ultra-threaded Dispatch Processor

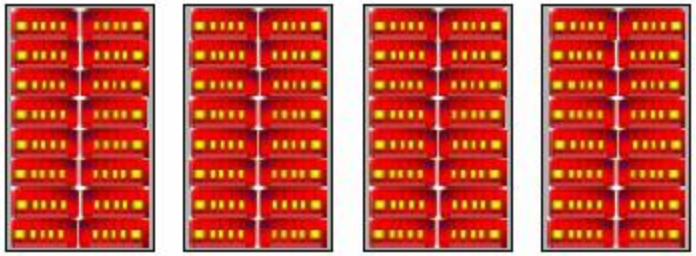
- Initial arbiter to select with thread to submit
- Two arbiter units per SIMD array
  - Allows each SIMD to be pipelined, with two operations at a time in process
- Dedicated arbiter units for texture and vertex fetches
  - Can be scheduled independently from math operations
- Executing threads can be bumped at any time if a higher priority thread is pulled from the command queues
  - Temporary data saved so thread can resume later
- Arbitration policy
  - Age/need/availability
  - When in doubt favor pixels
  - Programmable



Ultra-threaded Dispatch Processor

- Dedicated shader caches
  - Instruction cache allows unlimited shader length
  - Constant cache allows unlimited number of constants
  - Both caches take advantage of data re-use to improve state change overhead and efficiency
- Latency hiding
  - Cache miss, switches to another thread
  - Suspended threads remain in the command queues until their requested data arrives
  - Ultra-threaded dispatch processor can queue up hundreds of threads

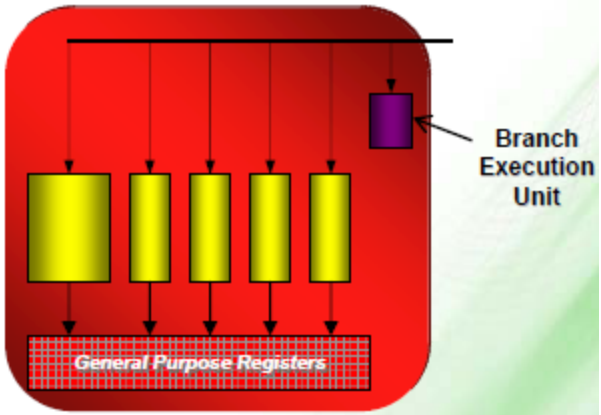
Shader Core



- 4 parallel SIMD units
- Each unit receives independent ALU instruction
- Very Long Instruction Word (VLIW)
  - Each instruction word can include up to 6 independent, co-issued operations (5 math + 1 flow control)
  - All operations are performed in parallel on each data element in the current thread
- Texture fetch and vertex fetch instructions are issued and executed separately
  - Allows fetches to begin executing before the requested data is required by the shader
- ALU Instruction (1 to 7 64-bit words)
  - 5 scalar ops- 64 bits for src/dst/controls/op
  - 2 additional for literal constants

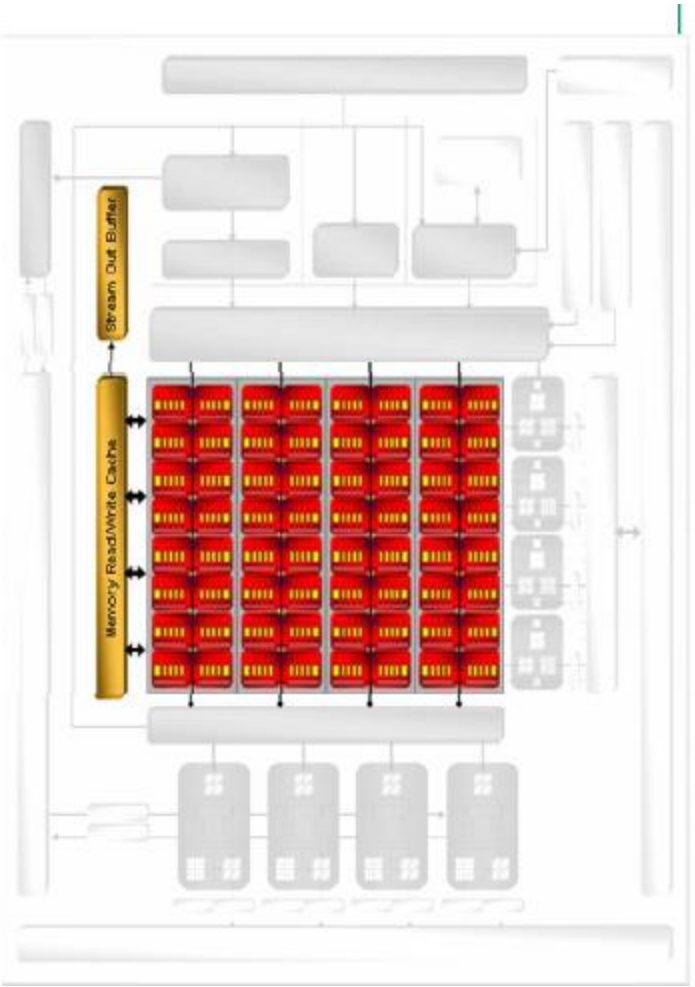
Stream Processing Units

- 5 Scalar Units
  - Each scalar unit does FP Multiply-Add (MAD) and integer operations
  - One also handles transcendental instructions (SIN, COS, LOG, EXP, etc.)
  - IEEE 32-bit floating point precision
  - Integer and bitwise operation support
- Branch Execution unit
- Up to 6 operations co-issued

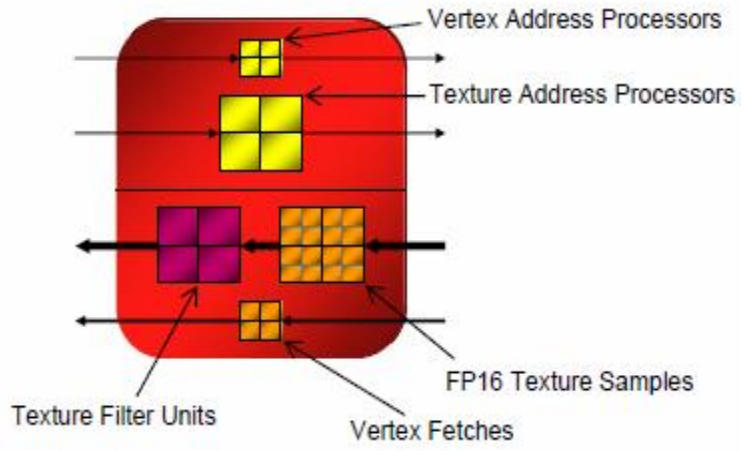


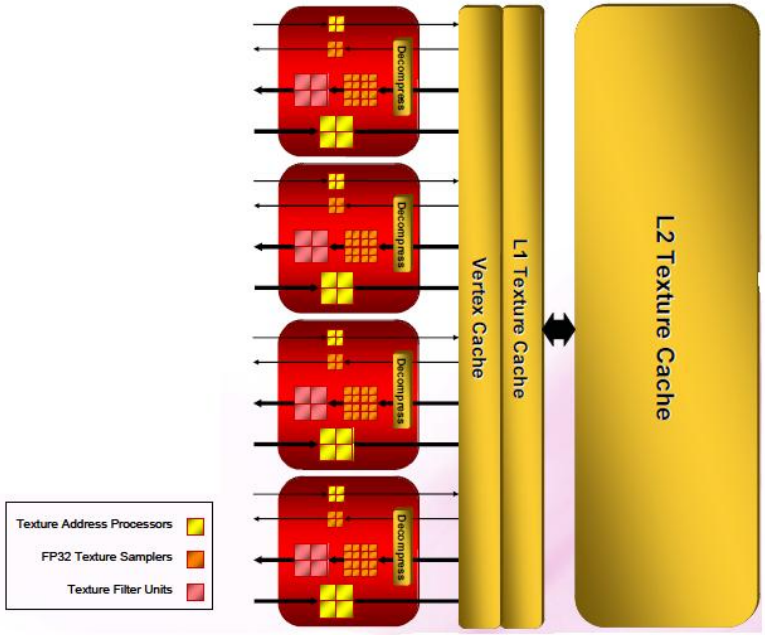
Memory Read/Write Cache

- Virtualizes register space
  - Allow overflow to graphics memory
  - Can be read from or written to by and SIMD (texture & vertex caches are read-only)
  - 8KB Fully associative cache, write combining
- Stream out
  - Allows shader output to bypass render back-ends and color buffer
  - Render to vertex buffer
  - Outputs sequential stream of data instead of bitmaps
- Uses: Used for inter-thread communication



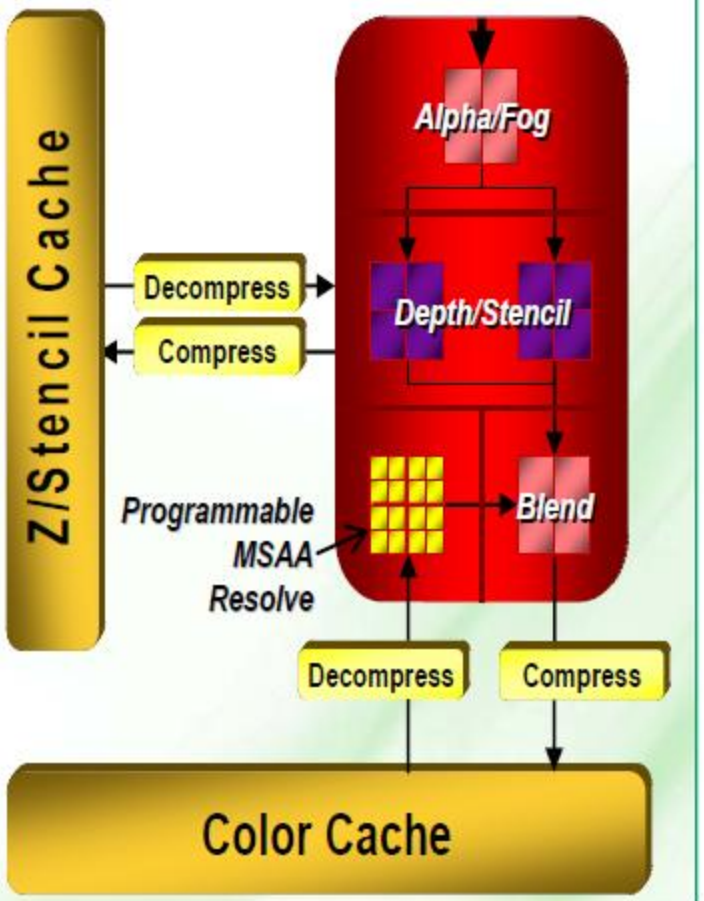
Texture Units





- Fetch Units
  - 8 fetch address processor each (32 total)
    - 4 filtered and unfiltered
  - 20 texture samplers each (80 total)
    - Can fetch a single data value per clock
  - 4 filtered texels (with BW) (16 total)
    - Bilinear filter one 64-bit FP color value per clocks for each pixel
    - 128-bit FP textures filtered at half speed
    - Trilinear and anisotropic filtering
- Fetch caches
  - Unified caches across all SIMDs
  - Vertex/Unfiltered cache
    - 4kB L1, 32 Kb L2
  - Texture cache
    - 32KB L1, 256 KB L2 (128KB for HD 2600, HD2400 uses single level vertex/texture cache)

Render Back-Ends



- Double rate depth/stencil test
  - 32 pixels per clock for HD 2900
  - 8 pixels per clock for HD2600&HD2400
- Programmable MSAA (multi-sample anti-aliasing) resolve
  - Allows custom AA filters
- New blend-able DX10 surface formats
  - 128-bit and 11:11:10 floating point format
- Up to 8 Multiple Render Targets (MRT) with MSAA support

Depth, Stencil, and Compression Improvements



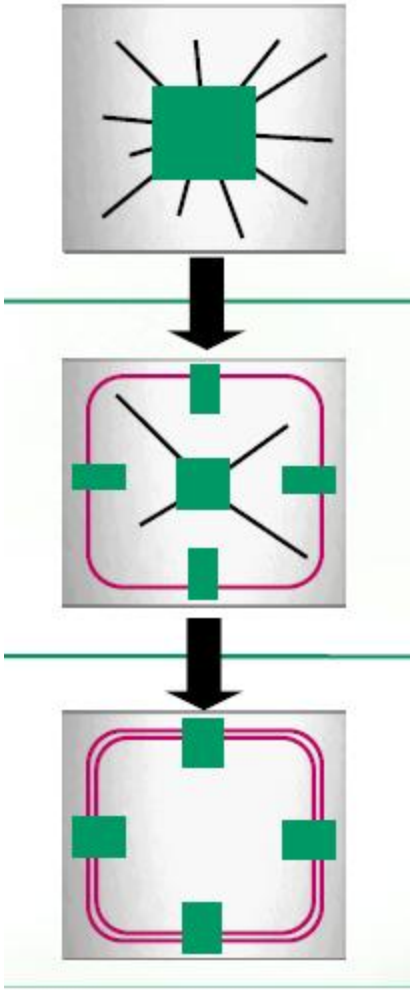
- Improved Z & Stencil compression
  - Up to 16:1 in standard mode
  - Z & stencil now compressed separately with each other for better efficiency
- Z Range optimization
  - Limit depth test operations to a programmable depth range (useful for speeding up stencil shadowing)
- Re-Z
  - Can check Z buffer twice – once before pixel shader, and again after
  - Allows early Z before shading in all cases
- Improved Hierarchical Z buffer
  - Adds hierarchical stencil (HiS) for better stencil shadow performance
  - Handles most situations where it had to be disabled in the past
- 32-bit floating point z-buffer support

Memory Controller Progression

Centralized

Partially distributed

Fully distributed



Crossbar

ATI Radeon X850& earlier +
All computing GPUs

Hybrid Ring Bus

ATI Radeon X1000 Series

Ring Bus

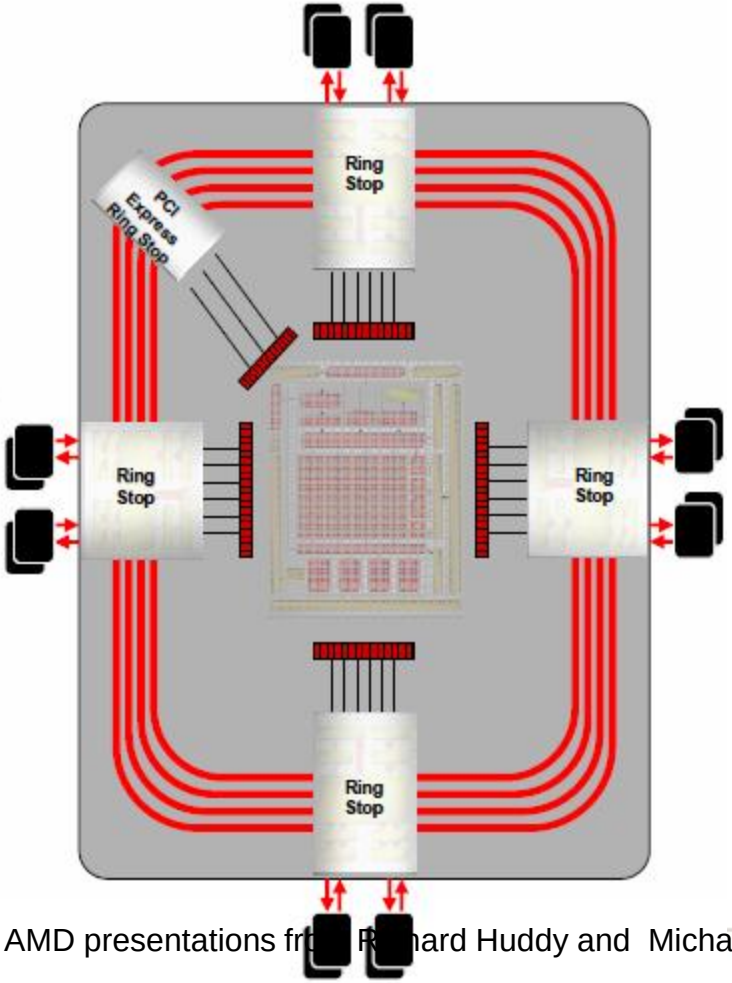
ATI Radeon HD 2000 series

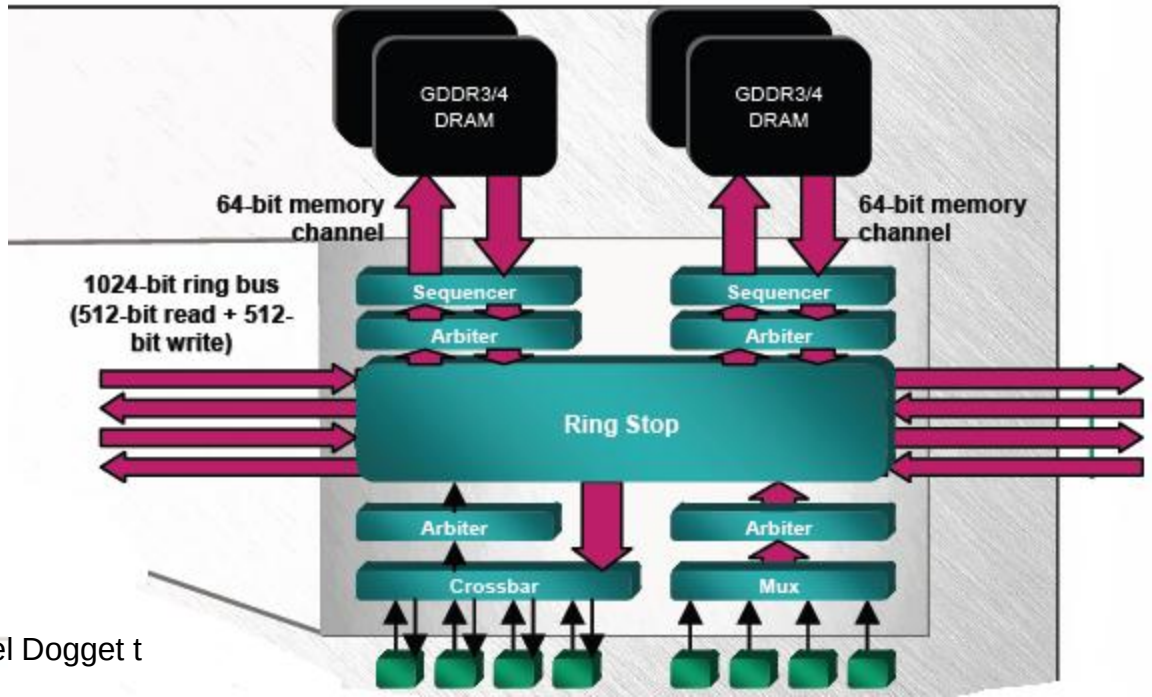
AMD presentations from Richard Huddy and Michael Doggett

Memory Interface and Controller

Over 100GB/s memory bandwidth
Fully distributed design
Highly scalable

- 512-bit interface
  - Compacts, stacked I/O pad design
  - More bandwidth with existing memory technology
  - Improved cost: bandwidth ratio
  - 8x64 bit memory channels
- Double ring bus
  - 512 bit read and write





AMD presentations from Richard Huddy and Michael Doggett

Memory controller

- Benefits of a 512-bit interface
  - More bandwidth with existing memory technology
  - Lower memory clock required to achieve target bandwidth
- Benefits of the ring bus
  - Simplifies routing to improve scalability
  - Reduces wire delay
  - Reduces number of repeaters required

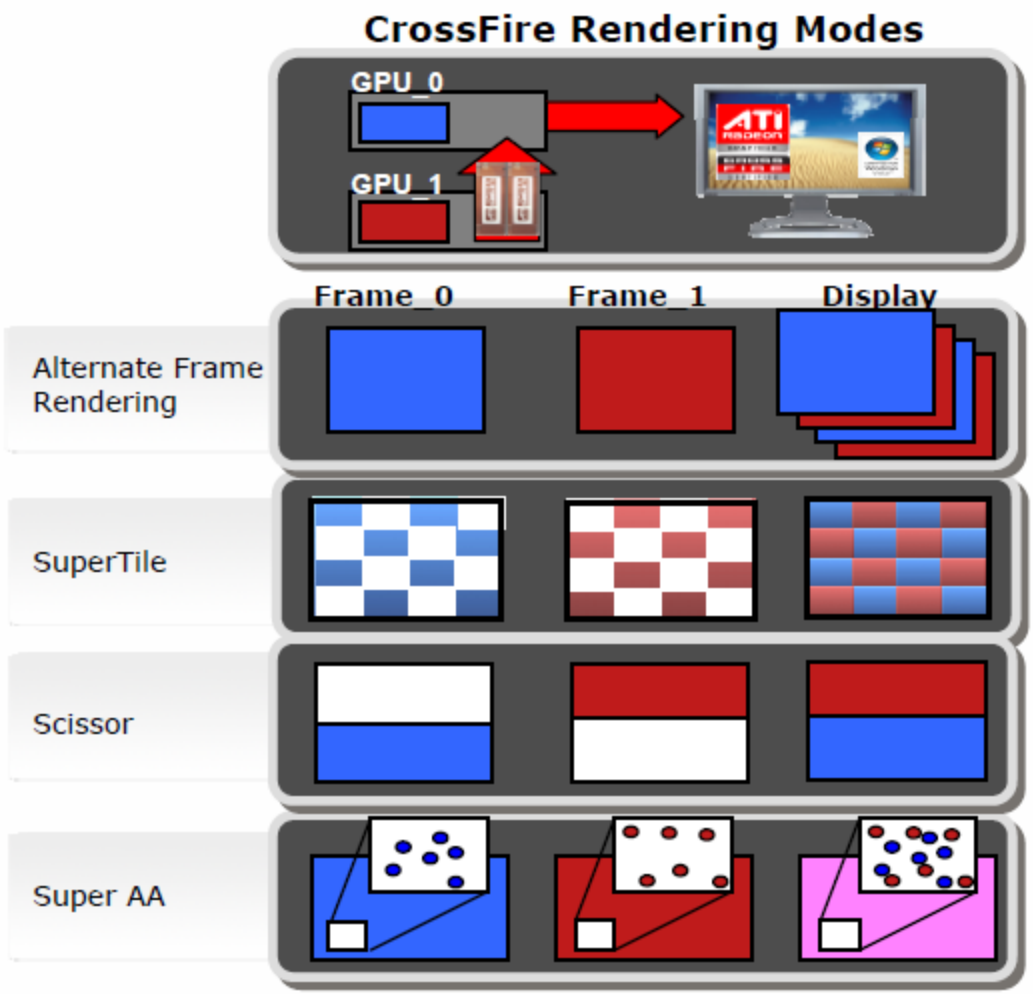


Tessellation

- Programmable tessellation unit
  - Based on xbox 360 technology
  - Provides highly effective geometry data compression
  - Orders of magnitude faster than CPU-based or geometry shader-based tessellation
- Enables:
  - More detailed animation
  - More realistic characters
  - Complex terrain
  - More sophisticated shader effect

CrossFire

- All ATIs Radeon HD 2000 series GPUs feature
- High bandwidth dual-link GPU interconnect
- Supports display resolutions up to 2560x2048 @ 60Hz
- Built for future scalability





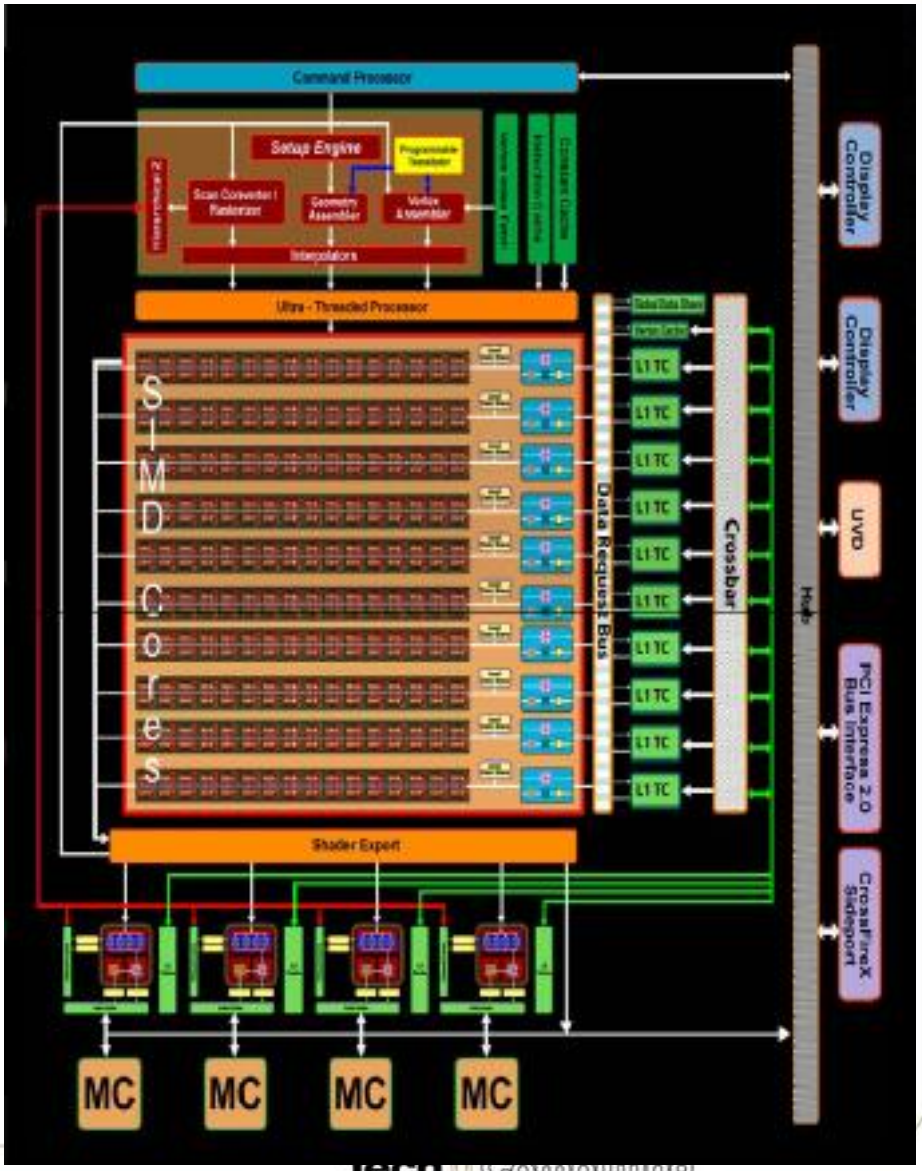
TERASCALE ARCHITECTURE

Design Goals

- Focus on efficiency
  - Old equation: architecture advances = f (performance, features)
  - New equation, architecture advances = f(perf/watt, perf/$, features)
- Scale up processing power & AA performance
- Enhance stream computing capability
  - Faster and more flexible
- DirectX 10.1, tessllation, CFAA, GDDR5, PCI-E 2.0

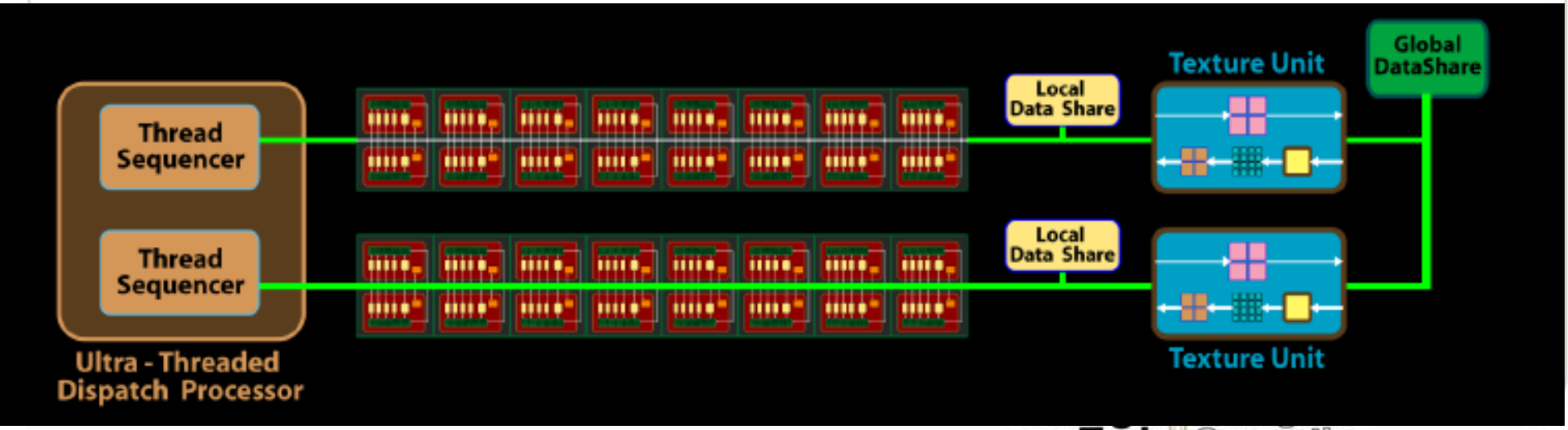
Terascale Graphics Engine

- 800 stream processing units
- Texture
- New texture cache design
- New memory architecture
- Optimized render back-ends for faster anti-aliasing
- Enhanced geometry shader and tessellator performance



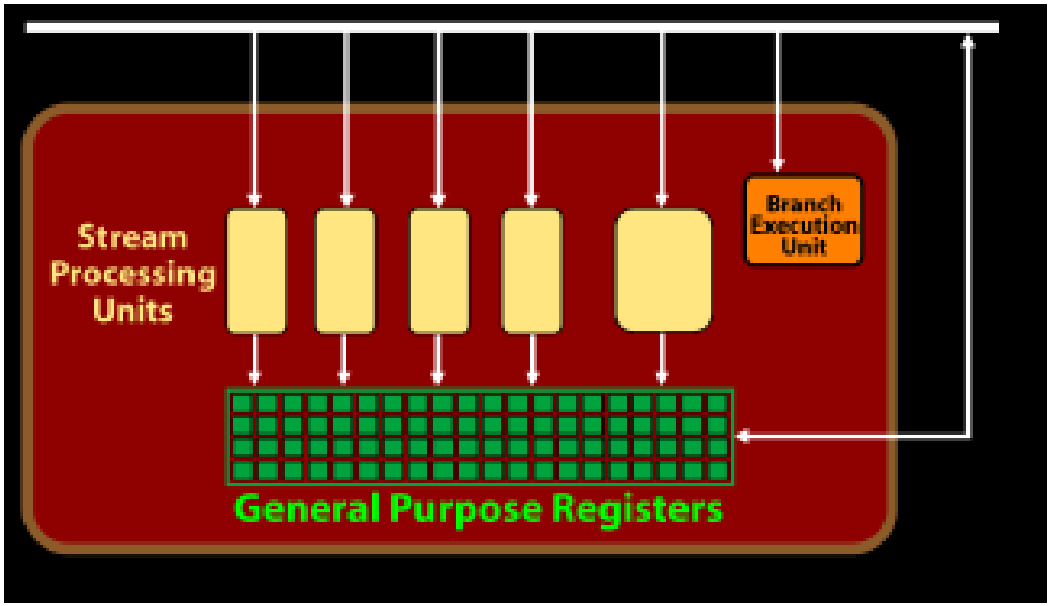
SIMD Cores

- Each core
  - 80 scalar stream processing units + 16KB local store
  - Has its own control logic
  - 4 dedicated texture units + L1 cache
  - 16 global data share
- 4:1 ALU:TEX ratio



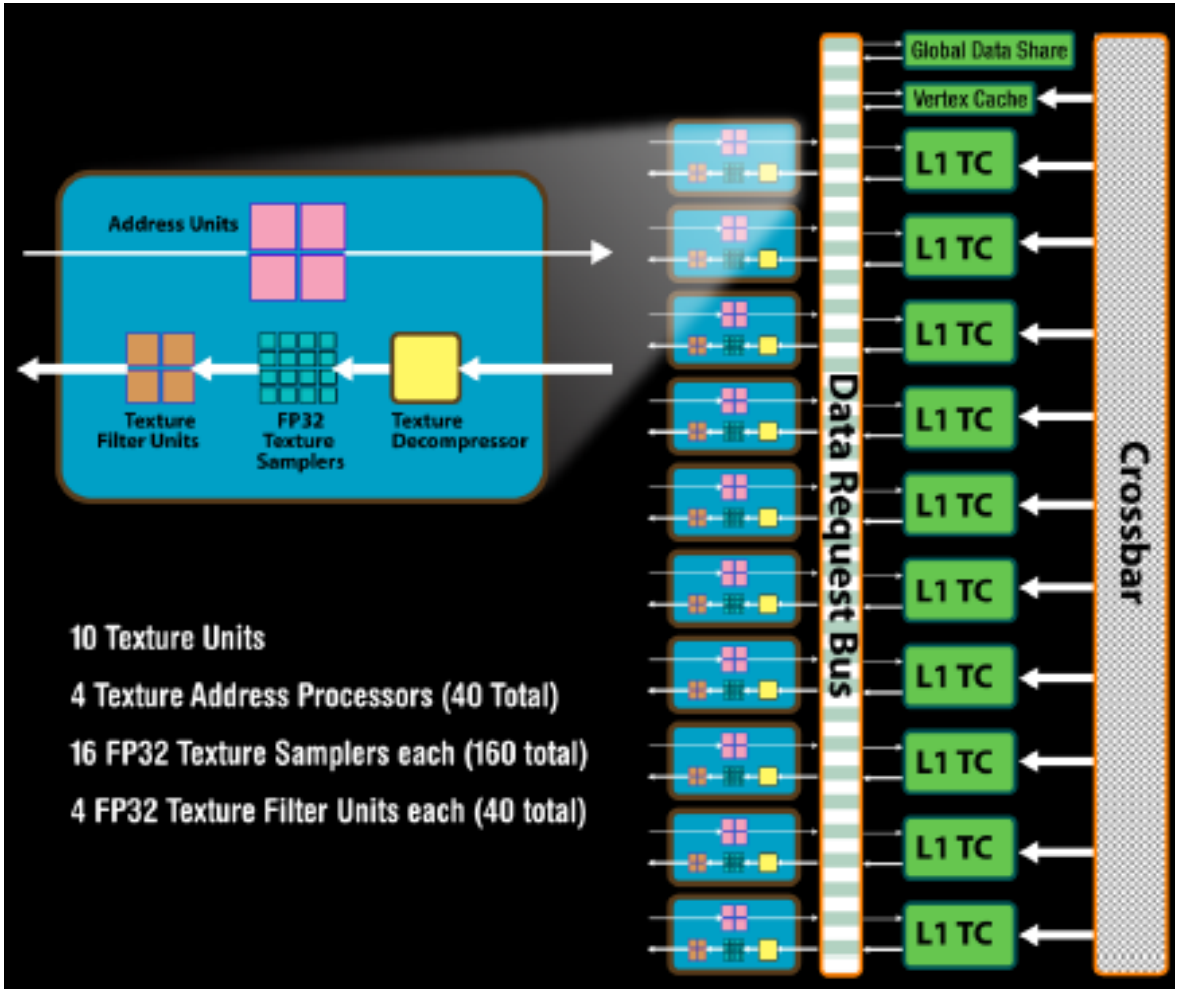
Stream Processing Units

- 40% increase in performance per mm^2
- More aggressive clock gating for improved performance per watt
- Fast double precision processing units



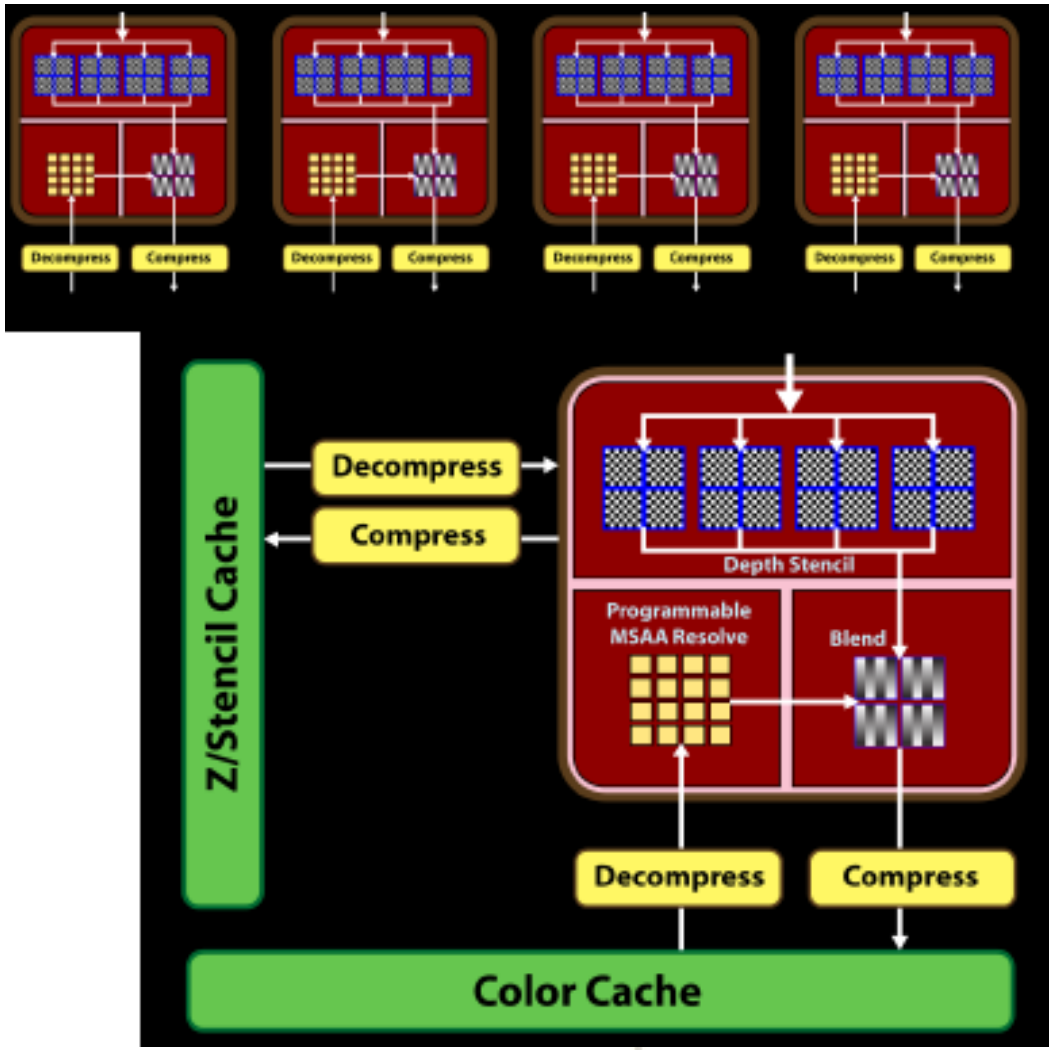
Texture Units

- Streamlined design
  - 70% increase in performance/mm^2
- More performance
  - Double the texture cache bandwidth
  - 2.5x increase in 32-bit filter rate
  - 1.25x increase in 64-bit filter rate
- New cache design
  - L2s aligned with memory channels
  - Separate vertex cache
  - Increased bandwidth



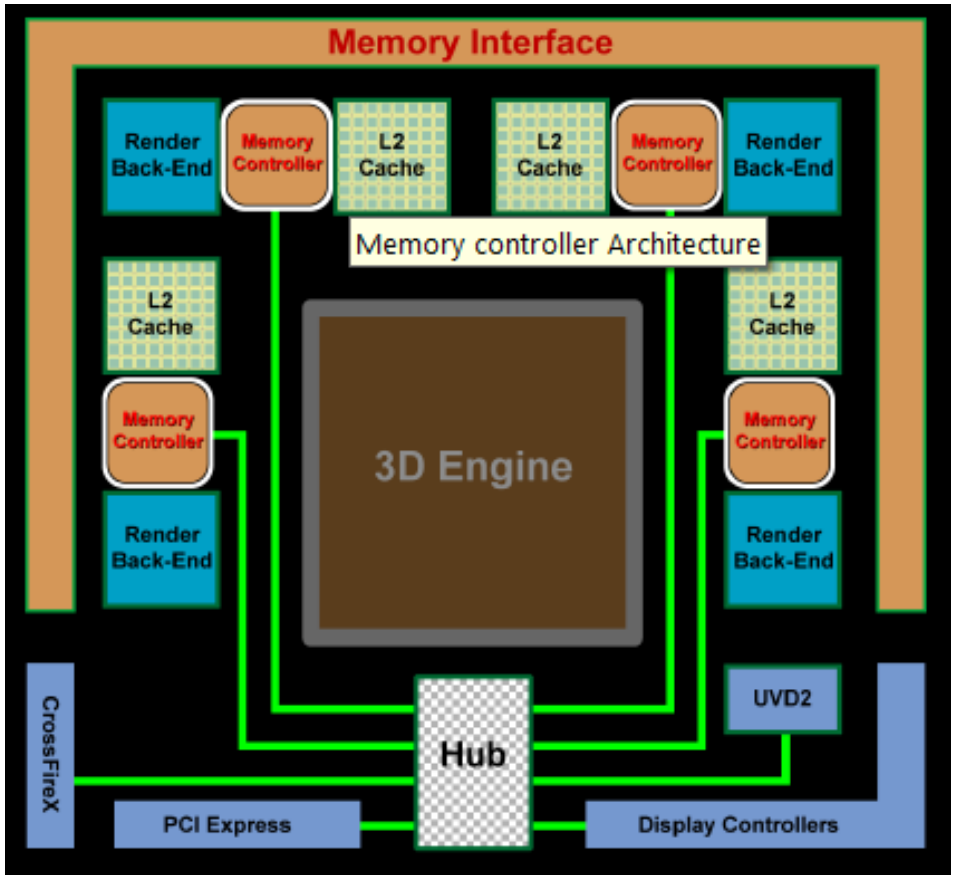
Render Back-Ends

- Focus on improving AA performance per mm^2
- Doubled peak rate for depth/stencil ops to 64 per clock



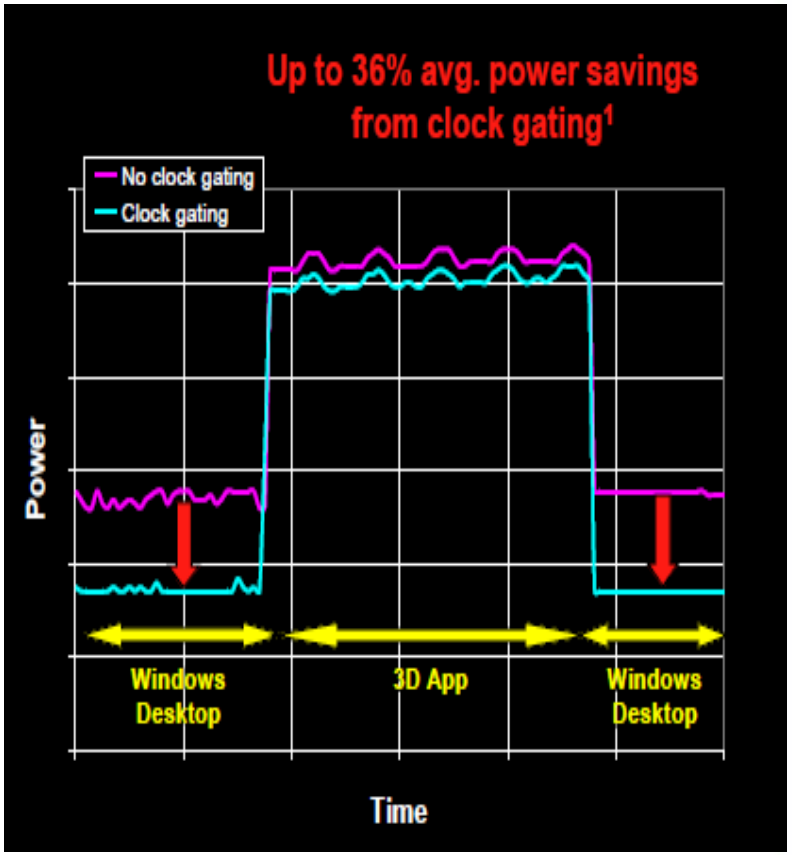
Memory Controllers

- New distributed design with hub
- Controllers distributed around periphery of chip, adjacent to primary bandwidth consumers
- 256-bit interface allows reduced latency
- Hub handles relatively low bandwidth traffic
  - PCI Express, CrossFireX interconnect, UVD2, display controllers



Dynamic Power Management

- On-chip microcontroller
- Controls clock gating, engine/memory clock speeds, voltages, and fan controllers



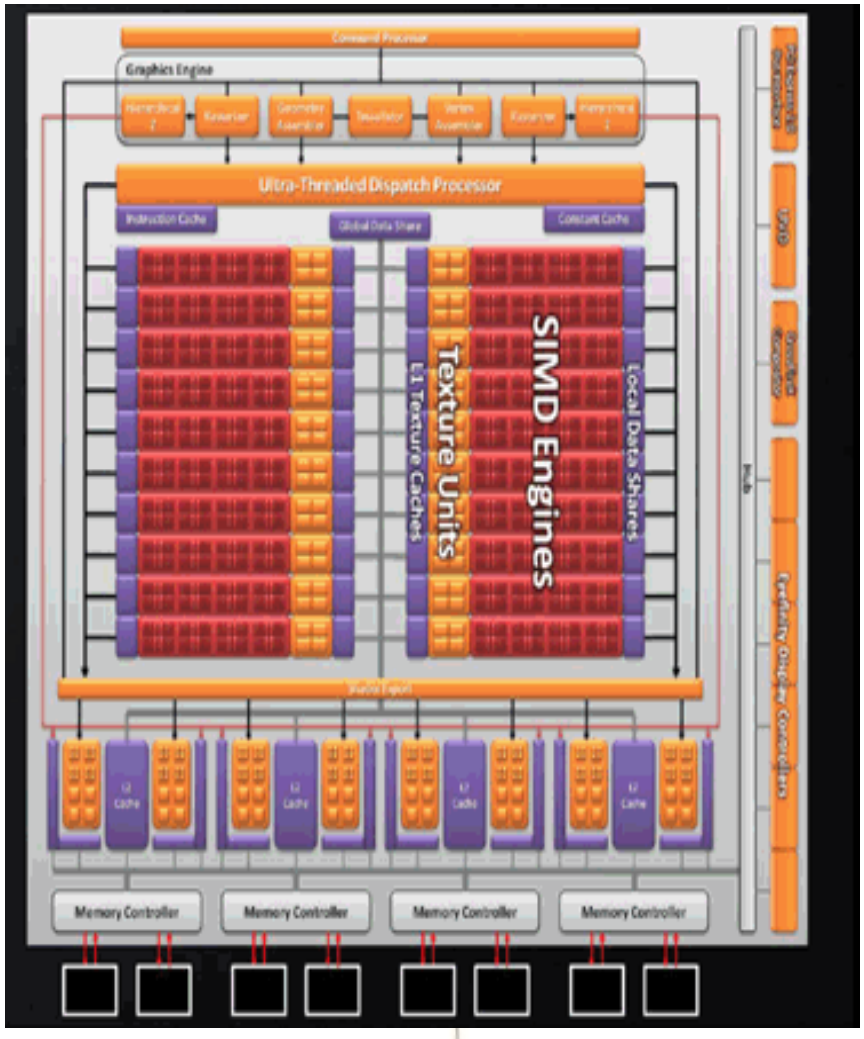


Radeon HD 5800 series

- TeraScale 2 architecture
- The first DirectX 11 support GPU
  - 2.7 TeralFlops for a single precision
  - 544 Gflops for a double precision
- Evolved from TeraScale architecture (HD 4800)
  - No revolution

TeraScale 2 Architecture

- 2X the processing power of previous Gen
  - Over 2 TeraFLOPS
  - Over 20 Gigapixels/Sec
- Major Feature and Design Enhancements:
  - Instruction set
  - Stream processing units
  - SIMD layout
  - Graphics engine
  - Texture units
  - Render back-ends
  - Display controllers



http://www.hardocp.com/image.html?image=MTI1MzU4OTM1NVlDbXBla3ZKZmsINV8xX2wuZ2lm

ATI Radeon HD 5870

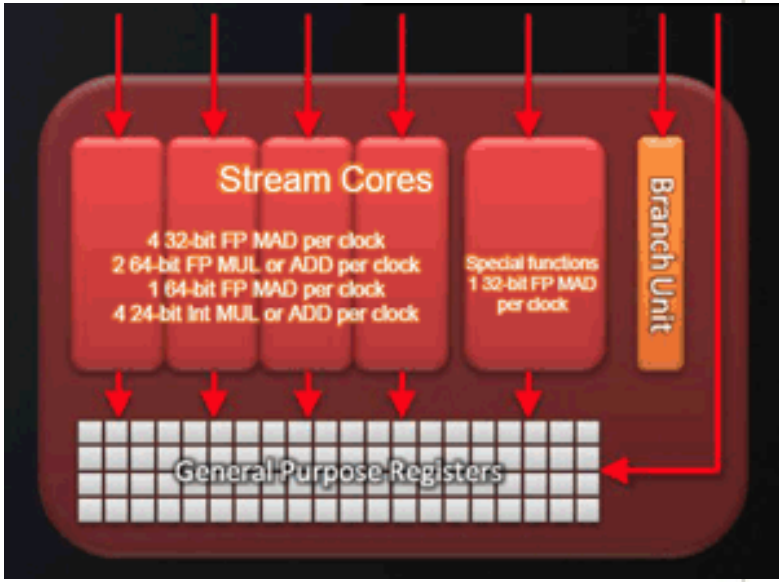
- 20 SIMD engines
  - Each with 16 thread processors
    - Each with 5 stream cores (1600 total)
- 80 Texture units
  - 4 per SIMD engine
- 150+ GB/sec GDDR5 memory interface

	ATI Radeon™ HD 4870	ATI Radeon™ HD 5870	Difference
Die Size	263 mm^2	334 mm^2	1.27x
Transistors	956 million	2.15 billion	2.25x
Memory Bandwidth	115 GB/sec	153 GB/sec	1.33x
AA Resolve	64	128	2x
Z/Stencil	64	128	2x
Texture	40	80	2x
Shader	800	1600	2x
Board Power*			
Idle	90 W	27 W	0.3x
Max	160 W	188 W	1.17x

http://www.hardocp.com/image.html?image=MTI1MzU4OTM1NVlDbXBla3ZKZmsfNV8xX2wuZ2lm

Thread Processors

- 2.7 TeraFLOPs single precision
- 544 GigaFLOPs double precision
- Increased IPC
  - More flexible dot products
  - Co-issue MUL, dependent ADD in single clock
  - Sum of absolute differences (SAD)
    - 12x speed-up with native instruction
    - Used for video encoding, computer vision
    - Exposed via OpenCL extension
  - DirectX 11 bit-level ops
    - Bit count, insert, extract, etc.
  - Fused Multiply-Add

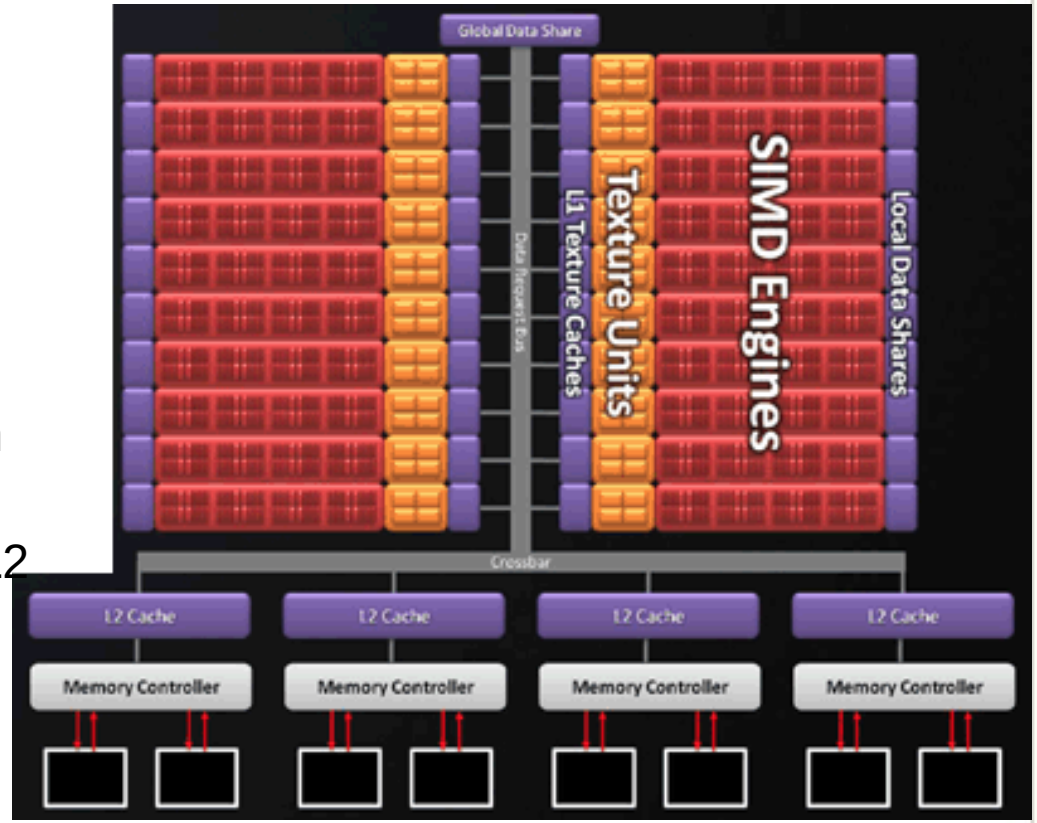


•Each thread processor includes
  - •4 stream cores + SFU
  - •Branch unit
  - •General purpose registers

http://www.hardocp.com/image.html?image=MTI1MzU4OTM1NVlDbXBla3ZKZmsfNV8xX2wuZ2lm

Texture Units and Caches

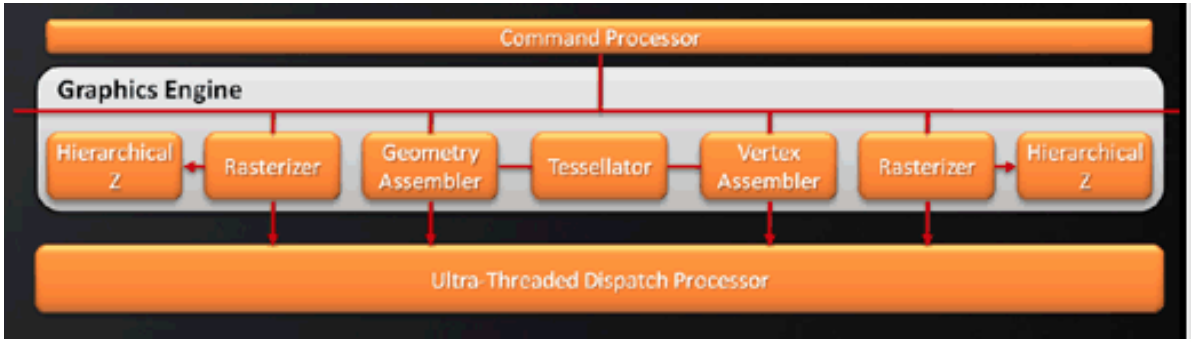
- 80 texture units
- Increased texture bandwidth
  - Up to 68 billion bilinear filtered texel/sec
  - Up to 272 billion 32-bit fetches/sec
- Increased cache bandwidth
  - Up to 1TB/sec L1 texture fetch bandwidth
  - Up to 435 GB/s between L1&L2
- Doubled L2 cache
  - 128KB per memory controller
- New DirectX 11 texture features
  - 16k x 16k max resolution
  - New 32-bit and 64-bit HDR block compression modes



http://www.hardocp.com/image.html?image=MTI1MzU4OTM1NVlDbXBla3ZKZmsfNV8xX2wuZ2lm

Graphics Engine

- Dual rasterizers
- New tessellation unit
  - 6th generation technology
  - Programmable via Direct X11 Hull & Domain shaders



- Pull model interpolation
  - New DirectX 11 feature
  - Uses stream processors for interpolation with new instructions
  - Improved flexibility, negligible performance cost
- Improved performance for constant buffer updates
- Faster geometry shading
- OpenGL enhancements
  - Improved line rendering performance and clipped speed
  - 12-bit subpixel precision

http://www.hardocp.com/image.html?image=MTI1MzU4OTM1NVlDbXBla3ZKZmsfNV8xX2wuZ2lm

Render Back-ends

- New readback path
  - Texture units can now read compressed AA color buffers
  - Improved CFAA performance
- Faster sample rate shading
- Enhanced MRT performance
- Faster color clears
- Supersample AA
  - Anti-aliases shaders & texture as well as polygon engines

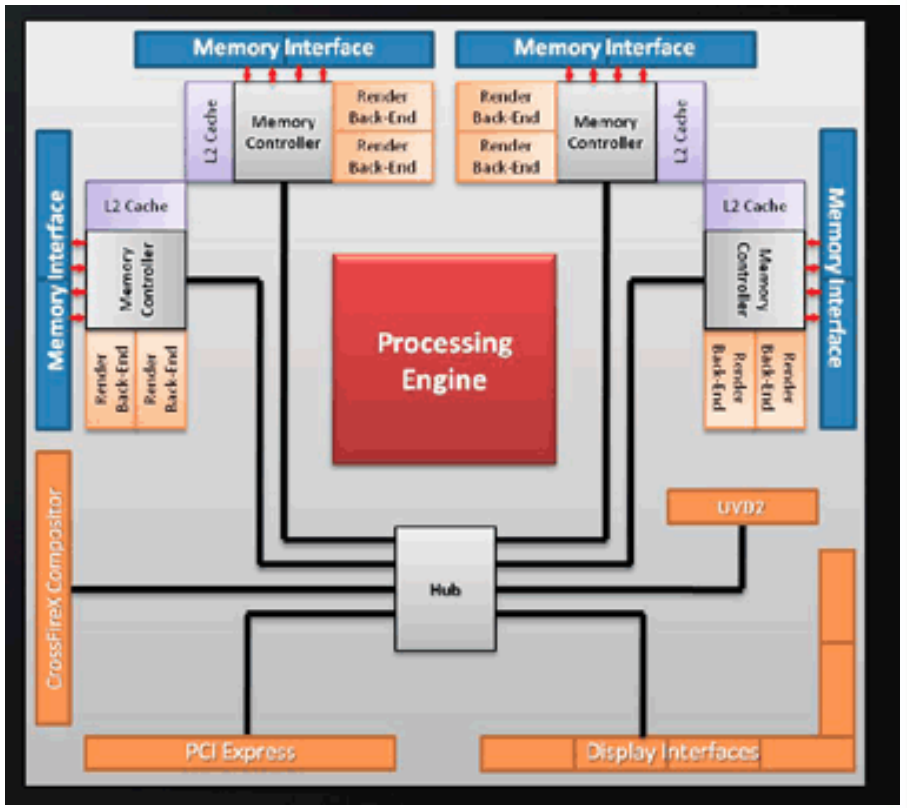
	Color	ATI Radeon HD 4800 series	ATI Radeon HD 5800 series	Difference
No MSAA	32-bit	16 pix/clk	32 pix/clk	2x
2x/4x MSAA		16 pix/clk	32 pix/clk	2x
8x MSAA		8 pix/clk	16 pix/clk	2x
No MSAA	64-bit	16 pix/clk	32 pix/clk	2x
2x/4x MSAA		16 pix/clk	32 pix/clk	2x
8x MSAA		8 pix/clk	16 pix/clk	2x
Depth/stencil only		64 pix/clk	128 pix/clk	2x

- Efficient implementation based on adaptive AA technology
- Works seamlessly with CFAA

http://www.hardocp.com/image.html?image=MTI1MzU4OTM1NVlDbXBla3ZKZmsINV8xX2wuZ2Im

GDDR5 Memory Interface

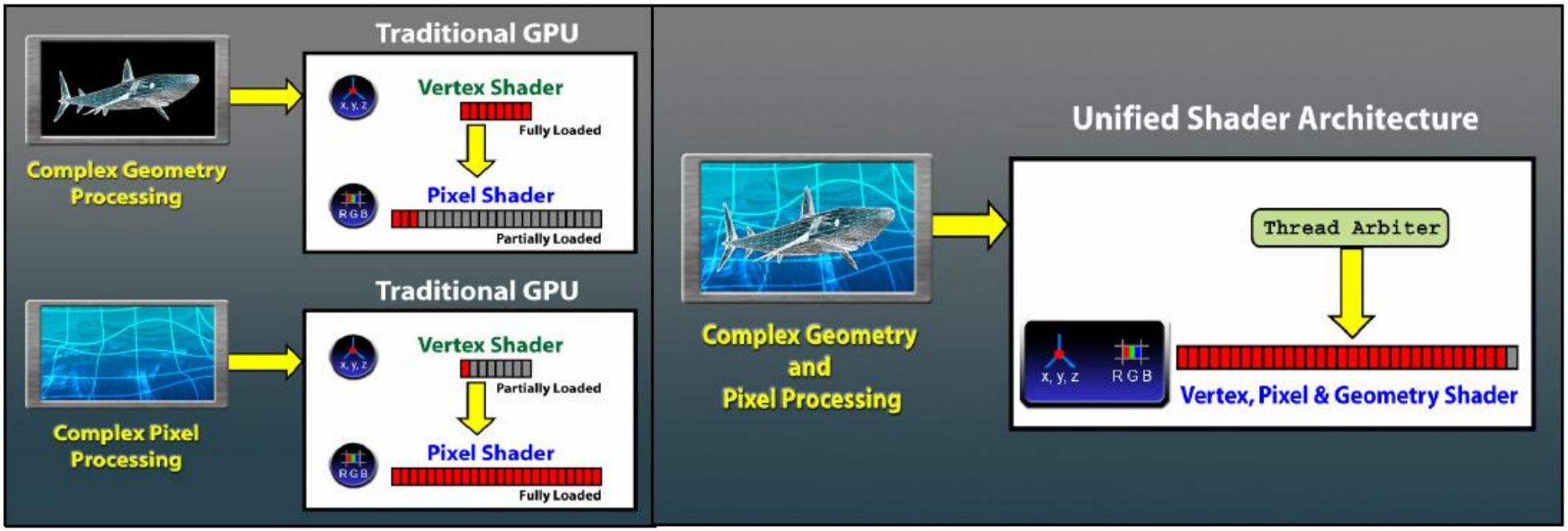
- Optimized memory controller area
- EDC (Error detection code)
  - CRC checks on data transfers for improved reliability at high clock speeds
- GDDR5 memory clock temperature compensation
  - Enables speeds approaching 5 Gbs
- Fast GDDR5 link retraining
  - Allows voltage & clock switching on the FLY without glitches



http://www.hardocp.com/image.html?image=MTI1MzU4OTM1NVlDbXBla3ZKZmsfNV8xX2wuZ2lm

Feature	Shader model 4.0	Shader model 5.0	Benefits
Thread dispatch	2D	3D	Replace multiple 2D thread arrays with a single 3D array
Thread limit	768	1024	More threads
Thread group shared memory	16KB	32KB	Increase inter-thread communication
Shared memory access	256 B write only	Full 32KB read/write	Efficient shared memory I/O
Atomic operations	Not supported	Supported	Each thread operaets on protected memory locations Easy programming CPU based algorithms
Double precision	Not supported	Supported	
Append/consume buffers	Not supported	Supported	Useful for building and accessin g data in list or stack form
Unordered access views bound to compute shader	1	8	
Unordered access views bound to pixel shader	Not supported	8	
Gather 4	Not supported	Supported	

The Benefit of Unified Shader



AMD presentations from Richard Huddy and Michael Doggett